American Society of Human Genetics 68th Annual Meeting

PLENARY AND PLATFORM ABSTRACTS

		Abstract #'s
Tuesday, October 16, 5:30-6:50 pm:		
4. Featured Plenary Abstract Session I	Hall C	#1-#4
Wednesday, October 17, 9:00-10:00 am, Concurrent Platform Session A:		
6. Variant Insights from Large Population Datasets	Ballroom 20A	#5-#8
7. GWAS in Combined Cancer Phenotypes	Ballroom 20BC	#9-#12
8. Genome-wide Epigenomics and Non-coding Variants	Ballroom 20D	#13-#16
9. Clonal Mosaicism in Cancer, Alzheimer's Disease, and Healthy Tissue	Room 6A	#17-#20
10. Genetics of Behavioral Traits and Diseases	Room 6B	#21-#24
11. New Frontiers in Computational Genomics	Room 6C	#25-#28
12. Bone and Muscle: Identifying Causal Genes	Room 6D	#29-#32
13. Precision Medicine Initiatives: Outcomes and Lessons Learned	Room 6E	#33-#36
14. Environmental Exposures in Human Traits	Room 6F	#37-#40
Wednesday, October 17, 4:15-5:45 pm, Concurrent Platform Session B:		
24. Variant Interpretation Practices and Resources	Ballroom 20A	#41-#46
25. Integrated Variant Analysis in Cancer Genomics	Ballroom 20BC	#47-#52
26. Gene Discovery and Functional Models of Neurological Disorders	Ballroom 20D	#53-#58
27. Whole Exome and Whole Genome Associations	Room 6A	#59-#64
28. Sequencing-based Diagnostics for Newborns and Infants	Room 6B	#65-#70
29. Omics Studies in Alzheimer's Disease	Room 6C	#71-#76
30. Cardiac, Valvular, and Vascular Disorders	Room 6D	#77-#82
31. Natural Selection and Human Phenotypes	Room 6E	#83-#88
32. Genetics of Cardiometabolic Traits	Room 6F	#89-#94
Wednesday, October 17, 6:00-7:00 pm, Concurrent Platform Session C:		
33. Characterization of Structural Variation in Population Controls and Disease	Ballroom 20A	#95-#98
34. Reanalysis of Sequencing Data to Increase Diagnostic Yield	Ballroom 20BC	#99-#102
35. Subclonal Somatic Mutations	Ballroom 20D	#103-#106
36. Diverse Approaches for Identifying Target Genes from GWAS Results	Room 6A	#107-#110
37. From GWAS to Function in Cancer	Room 6B	#111-#114

Copyright © 2018 The American Society of Human Genetics. All rights reserved.

38. Insights from and into Mendelian Disorders of the Liver and Pancreas	Room 6C	#115-#118
39. Scalable Tools to Enable Collaboration and Reproducible Analyses	Room 6D	#119-#122
40. Genomic Testing: Strategies and Patient Perspectives	Room 6E	#123-#126
41. Therapeutic Development: Adding or Subtracting	Room 6F	#127-#130

Thursday, October 18, 9:00-10:30 am, Concurrent Platform Session D:

42. Advances in GWAS Analytical Methods	Ballroom 20A	#131-#136
43. Genomic Insights from Diverse Populations	Ballroom 20BC	#137-#142
44. Comprehensive Descriptions of Genetic Architecture of Rare Diseases	Ballroom 20D	#143-#148
45. Learning from Genome Mutation Patterns	Room 6A	#149-#154
46. Neurodegenerative Disease: Revealing Disease Mechanisms	Room 6B	#155-#160
47. DNA Methylation in Human Disease	Room 6C	#161-#166
48. Genetic Disorders of Vision and Hearing	Room 6D	#167-#172
49. Emerging Omics Technologies	Room 6E	#173-#178
50. Discovery Studies of Cardiovascular and Metabolic Phenotypes	Room 6F	#179-#184

Thursday, October 18, 11:00 am-12:30 pm, Concurrent Platform Session E:

51. What are We Missing? Identification of Previously Underappreciated Mendelian Variants	Ballroom 20A	#185-#190
52. Single Cell Approaches to Reveal Disease Biology	Ballroom 20BC	#191-#196
53. Expanding the Scope of Mendelian Randomization	Ballroom 20D	#197-#202
54. Genetics and Functional Insights into Schizophrenia	Room 6A	#203-#208
55. The Contribution of Structural Variation to Neurological Disorders	Room 6B	#209-#214
56. Leveraging GWAS Data with Other Data Types	Room 6C	#215-#220
57. New Genes, Old Genes, and Novel Functions in the Central Nervous System	Room 6D	#221-#226
58. Skin and Bones: Mendelian Skeletal, Connective Tissue, and Ectodermal Phenotypes	Room 6E	#227-#232
59. Genetic Variant Effects on Blood Lipids	Room 6F	#233-#238

Friday, October 19, 9:00-10:00 am, Concurrent Platform Session F:

65. Mutational Processes in Cancer	Ballroom 20A	#239-#242
66. Genetics of Cancer Therapy Response and Resistance	Ballroom 20BC	#243-#246
67. Genetic Architecture of Hematopoiesis	Ballroom 20D	#247-#250
68. Enhancers and Noncoding Regions	Room 6A	#251-#254
69. Using RNA-seq to Improve DNA Sequence Interpretation	Room 6B	#255-#258
70. Cell Types, States, and Networks	Room 6C	#259-#262
71. Genetic Effects on Pregnancy and Fetal Health	Room 6D	#263-#266
72. Translating Cardiovascular Genetics to Patient Care	Room 6E	#267-#270
73. Therapeutic Development: Modulating Cellular Effectors	Room 6F	#271-#274

Friday, October 16, 5:00-6:20 pm:

87. Featured Plenary Abstract Session I	Hall C	#275-#278

Friday, October 16, 6:20-7:20 pm:

88. Late-breaking Abstract Session I	Hall C	#3568-#3570

Copyright © 2018 The American Society of Human Genetics. All rights reserved.

Saturday, October 20, 8:30-9:30 am, Concurrent Platform Session G:

89. Transcriptome Alterations in Cancer	Ballroom 20A	#279-#282
90. The Non-coding Genome and Disease	Ballroom 20BC	#283-#286
91. Biases of Polygenic Risk Scores	Ballroom 20D	#287-#290
92. Potential Genetic and Epigenetic Therapies of Disease	Room 6A	#291-#294
93. Investigating 3D Genome Structure	Room 6B	#295-#298
94. Importance of Isoform Expression in Variant Interpretation	Room 6C	#299-#302
95. Cancer Genomic Testing: Uncertainty and Decision Making	Room 6D	#303-#306
96. Leveraging The Cancer Genome Atlas	Room 6E	#307-#310
97. The Genetics and Genomics of Epilepsy	Room 6F	#311-#314

Saturday, October 20, 9:45-11:15 am, Concurrent Platform Session H:

98. Fine-mapping Using Statistical and Functional Tools	Ballroom 20A	#315-#320
99. Genetic and Epigenetic Effects on the Transcriptome	Ballroom 20BC	#321-#326
100. Understanding the Brain with Computational Genomics	Ballroom 20D	#327-#332
101. Novel Approaches for Conducting Genome-wide Association Studies	Room 6A	#333-#338
102. Increasing Functional Resolution Through Single Cell Analysis	Room 6B	#339-#344
103. Evaluating the Yield of Genetic Testing in Diverse Settings	Room 6C	#345-#350
104. Discoveries in Syndromic and Non-syndromic Congenital Heart Disease	Room 6D	#351-#356
105. Human Reproduction and Fertility	Room 6E	#357-#362
106. Genetics of Growth and Lifespan	Room 6F	#363-#368

Saturday, October 20, 11:40 am-1:00 pm:

108. Featured Plenary Abstract Session I	Hall C	#369-#372

Copyright © 2018 The American Society of Human Genetics. All rights reserved.

1

From GWAS to function: Comprehensive integrated genomic perturbation to reveal molecular mechanisms of trait associations. *M.A. Cole[1,2,3], A. Mousas[5,6], Y. Wu[1,2,3], J. Zeng[1,2,3], Q. Yao[1,2,3,4], D. Vinjamur[1,2,3], R. Kurita[7], Y. Nakamura[7,8], L. Pinello[3,4], G. Lettre[3,4], D.E. Bauer[1,2,3].* 1) Boston Children's Hospital, Boston, MA; 2) Dana-Farber Cancer Institute, Boston, MA; 3) Harvard Medical School, Boston, MA; 4) Massachusetts General Hospital, Boston, MA; 5) Universite de Montreal, Montreal, Quebec, Canada; 6) Montreal Heart Institute, Montreal, Quebec, Canada; 7) RIKEN BioResource Center, Tsukuba, Ibaraki, Japan; 8) . 7 University of Tsukuba, Tsukuba, Ibaraki, Japan.

Discovery of molecular mechanisms responsible for trait associations is hampered by difficulty in identifying causal variants. Typical assays of genetic function are low throughput or evaluate sequences in heterologous ectopic settings. Genome editing enables perturbation of trait-associated sequences within relevant genomic, chromatin and cellular context. Here we perform comprehensive analysis of genetic variants associated with red blood cell traits by pooled CRISPR screening. We performed a genome-wide knockout screen in immortalized erythroid precursors (HUDEP-2 cells) to define functional erythroid genes required for cell growth or differentiation. We evaluated 952 loci associated with nine red blood cell traits (Astle *et al*, Cell 2016) comprising 24,843 SNPs. We linked 28.9% of these SNPs to genes by at least one of four routes: sharing topological associated domain, physical proximity (<20 kb), long-range chromatin interaction, or eQTL with a functional erythroid gene. We designed ~5 sgRNAs per SNP requiring cleavage within 50 bp and exceeding an off-target score threshold, resulting in 32,710 sgRNAs testing 5,592 SNPs at 481 loci. We utilized four editors: Cas9 nuclease to produce indels, dCas9-VP64 for gene activation, dCas9-KRAB for gene repression, and dCas9 as a DNA targeting control. By pooled transduction, erythroid maturation culture, and sgRNA library deep sequencing, we found reproducible results across biological replicates, with guide count clustered by Cas9 type. Surprisingly, we found similar results for dCas9-KRAB and dCas9, suggesting the KRAB domain was largely dispensable. We performed fine-mapping of association results by Bayesian inference to calculate posterior probability of inclusion (PPI). We found a strong correlation between PPI and CRISPR significance scores indicating overall agreement between CRISPR screening and genetic fine-mapping, despite examples of validated CRISPR signals with low PPI scores. We identified numerous examples of CRISPR hits at regulatory elements including promoters and enhancers but also at noncoding sequences lacking chromatin marks, including at *BCL2L1, GFI1B,* and *KLF1*. These results represent to our knowledge the largest dataset editing human trait-associated sequences in situ and demonstrate the potential and challenges of comprehensive integrated genomic perturbation to complement genetic, biochemical, and statistical approaches to uncover molecular underpinnings of complex human traits.

2

Incidence of uniparental disomy in 2 million individuals from the 23andMe database. *P. Nakka[1,2], K. McManus[3], A. O'Donnell-Luria[4,5], U. Francke[6], J. Mountain[3], S. Ramachandran[1,2], F. Sathirapongsasuti[3], 23andMe Research Team.* 1) Center for Computational Molecular Biology, Brown University, Providence, RI; 2) Ecology and Evolutionary Biology, Brown University, Providence, RI; 3) 23andMe, Inc., Mountain View, CA; 4) Boston Children's Hospital, Boston, MA; 5) Broad Institute of MIT and Harvard, Cambridge, MA; 6) Stanford University, Palo Alto, CA.

Uniparental Disomy (UPD) is the inheritance of both homologs of a chromosome from one parent with no representative copy from the other. UPD can cause clinical phenotypes by disrupting parent-specific genomic imprinting, or by unmasking recessive alleles in blocks of homozygosity on the affected chromosome. Though over 3,300 clinical cases of UPD have been collected to date, our understanding of UPD is limited to chromosomes on which UPD causes clinical presentation. In particular, incidence of UPD, its subtypes, and their phenotypic consequences are very poorly characterized in the general population. To address this gap, we analyze instances of UPD in consented research participants from the personal genetics company 23andMe, Inc., whose database consisted of SNP data from over 2 million individuals. We estimate identity-by-descent (IBD) between 290,927 parent-child duos and identify 49 instances of UPD, corresponding to an incidence rate of 1 in 4,000 births for UPD in the general population. In the 23andMe database, UPD cases occur most frequently on chromosomes 16, 21 and 22. In contrast, existing clinical cases cluster on chromosomes 6, 7, 11, 14, and 15, which each contain genes that cause imprinting disorders. We also find that per-chromosomes rates of UPD in 23andMe are significantly correlated (r=0.67; p=7.0e-4) with rates of aneuploidy in preimplantation embryos, reflecting a potential common etiology for UPD and aneuploidy. Similarly, we find the risk of UPD significantly increases with parental age ($OR_{\text{parental age>35}}$=3.1; p=6.6e-4). We extend our analysis to individuals without genotyped relatives by detecting UPD using runs of homozygosity (ROH), which are present in certain subtypes of UPD (isodisomy and some partial isodisomies). Using a simulation-based classification framework, we identify 244 UPD cases in 974,511 Northern Europeans in the 23andMe database. Lastly, we test for associations between five categories of phenotypes (cognitive, personality, morphology, obesity, and metabolic traits) and UPD status. Besides phenotypes known to be associated with Prader-Willi Syndrome (matUPD15), we find novel associations between UPD16 and lower birth weight (p=0.007), and UPD21 and lower overall life satisfaction (p=5.0e-5). Our study presents the first estimates of UPD rates in the general population, a framework to identify UPD from ROH, as well as previously unrecognized phenotypes associated with UPD of chromosomes 16 and 21.

Copyright © 2018 American Society of Human Genetics. All rights reserved

3

Discovery and characterization of 102 genes associated with autism from exome sequencing of 37,269 individuals. *J.A. Kosmicki[1,2,3], F.K. Satterstrom[1,2,3], J. Wang[4], R.L. Collins[1,2,3], S. de Rubeis[5,6], M. Breen[5,6], S. Gerges[1,2,3], M. Peng[4], X. Xu[5,6], C. Stevens[2], JJ. Grove[7,8,9,10], A.D. Børglum[7,8,9], J.D. Buxbaum[5,6,11,12,13,14], D.J. Cutler[15], B. Devlin[16], K. Roeder[4], S.J. Sanders[17], M.E. Talkowski[1,2,3], M.J. Daly[1,2,3], Autism Sequencing Consortium.* 1) Center for Genomic Medicine, Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 2) Program in Medical & Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of Harvard and M.I.T., Cambridge, MA, USA; 3) Program in Bioinformatics and Integrative Genomics; Program in Biology and Biomedical Sciences, Harvard Medical School, Boston, MA, USA; 4) Department of Statistics and Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA; 5) Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 6) Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 7) iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research,Denmark; 8) iSEQ, Centre for Integrative Sequencing, Aarhus University, Aarhus, Denmark; 9) Department of Biomedicine–Human Genetics, Aarhus University, Aarhus, Denmark; 10) Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark; 11) Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 12) Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 13) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 14) Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 15) Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA; 16) Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA; 17) Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA.

The genetic architecture of autism spectrum disorders (ASDs) involves the interplay of common and rare variation, with hundreds of genes conferring risk to the disorder. Both identifying the genes associated with ASD and understanding their effects in the context of ASD's heterogeneous phenotypic presentation have been longstanding goals. Previous studies focused almost exclusively on de novo (newly arising) variation for gene discovery, ignoring all other sources of variation.Here, we present the largest exome sequencing study in ASD to date, with 37,269 samples from 31 sampling sources across large sequencing studies supported by SFARI, NHGRI, and NIMH. Using a Bayesian framework that incorporates de novo and case-control variation and leverages gene and regional constraint, we discover 26 genome-wide significant genes and 102 genes (FDR<0.1) associated with ASD. Thirteen of the 102 genes overlap 12 large recurrent copy number variants, indicating these genes may contribute to the autism associations in 1p36.3, 2p15, 2q37.3, 11q13, 15q11.2, 15q24, and 16p11.2 (see related abstract by Collins et al.). The 102 genes are also enriched in both gene expression regulators expressed primarily during prenatal/fetal period and neuronal communication genes expressed during later postnatal development.Meta-analyzing our results with published de novo variants from 5264 intellectual disability / developmental delay (ID/DD) trios indicates that 47 of the 102 ASD-associated genes are more strongly associated with ID/DD than ASD, as evidenced by a higher rate of de novo variants in ascertained ID/DD individuals than ASD individuals. Of the remaining 55 genes discovered, 52 are preferential towards ASD, and 3 are equally associated with both. We demonstrate that comorbid ASD-ID/DD genes are markedly different from ASD-preferential genes in terms of the degree of negative selection, phenotypic presentation, and expression from single-cell RNA sequencing in 531 cell-types. ASD patients with variants in comorbid ASD-ID/DD genes on average walk 2.6 months later (P=6e-4) and have an IQ 11.7 point lower (P=5e-04) than those patients with a variant in ASD-preferential genes.Our data constitute the most comprehensive description of the genetic and phenotypic architecture of ASD to date and differentiating ASD from ASD-ID/DD genes lends evidence that different biological processes underlie ASD-associated genes and the phenotypic presentation of ASD probands.

4

Inhibition of oxytocin signaling prevents pregnancy-associated aortic dissection in a novel mouse model of vascular Ehlers-Danlos Syndrome. *C.J. Bowen[1,2], G. Rykiel[1,2], J.C. Giadrosic[1,2], J. Habashi[1,2,3,4], M. Helmers[1,2], H.C. Dietz[1,2,3,4].* 1) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, MD; 3) Division of Pediatric Cardiology, Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD; 4) Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD.

Patients with vascular Ehlers Danlos Syndrome (vEDS) experience dissection of medium-to-large arteries. Many features of vEDS are distinctly different from other heritable vasculopathies, such as Marfan Syndrome (MFS) and Loeys-Dietz Syndrome (LDS) which have been associated with excessive TGFβ activity. These include no particular predisposition for involvement of the aortic root and dissection without prior vessel dilation. Vascular rupture in patients with vEDS is difficult to anticipate or prevent. Pregnancy specifically increases the risk of dissection, with complications occurring in over 50% of pregnancies. The predominant postpartum occurrence of vascular dissection is not consistent with a mechanism that singularly invokes hemodynamic stress. Instead, we hypothesized that oxytocin, a hormone which initiates uterine contraction and is sustained postpartum during lactation, may contribute to pregnancy-associated risk. Expression of the oxytocin receptor is induced in the aorta during pregnancy and the hormone stimulates peripheral tissues through the activation of ERK, a signaling cascade previously implicated in the pathogenesis of MFS and LDS. We showed that oxytocin-mediated ERK signaling contributes to dissection risk in mouse models of MFS and posited that this mechanism would extend to other vasculopathies, including vEDS, despite differences in phenotype and underlying disease mechanism. To explore this hypothesis, we manipulated oxytocin and its downstream signaling events in a novel knock-in mouse model of vEDS. The *Col3a1* G209S/+ mouse shows 60% lethality due to aortic dissection in the first 30 days following delivery. Prevention of lactation through pup removal after birth was able to prevent dissection and death in vEDS mice (100% survival). Further, near-complete survival (95%) was achieved upon treatment with hydralazine, which blocks the PLC/IP3/PKC/ERK axis that is activated by oxytocin; this is an important finding because hydralazine is FDA-approved and can be used safely during pregnancy. Similar protection (~90% survival) was observed upon treatment with trametinib, an FDA-approved inhibitor of MEK, the kinase that activates ERK. These therapeutic strategies have the strong potential to modify vascular risk in women with vEDS and other heritable connective tissue disorders and these data provide the first evidence that therapeutically targetable abnormalities in signaling may drive vEDS pathogenesis.

Copyright © 2018 American Society of Human Genetics. All rights reserved

5

Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance in human genes. *K.J. Karczewski[1,2], L.C. Francioli[1,2], G. Tiao[2], R.L. Collins[1,2], B.B. Cummings[1,2], J.A. Kosmicki[1,2], Q. Wang[1,2], A. Ganna[1,2], L.D. Gauthier[2], E. Banks[2], K.M. Laricchia[2], J. Alfoldi[2], T. Poterba[2], A. Wang[2], C. Seed[2], M.E. Talkowski[1,2], B.M. Neale[1,2], M.J. Daly[1,2], D.G. MacArthur[1,2], The Genome Aggregation Database Consortium.* 1) Massachusetts General Hospital, Boston, MA; 2) Broad Institute, Cambridge, MA.

Naturally occurring genetic variants predicted to cause complete loss-of-function of protein-coding genes - predicted loss-of-function (pLoF variants) - are a powerful source of information about the phenotypic consequences of gene disruption. The number of pLoF variants in a gene in a large population correlates with that gene's essentiality: genes critical for an organism's function will be severely depleted for loss-of-function variants, while non-essential genes will be tolerant to the accumulation of these variants. However, pLoF variants are typically present at very low population frequencies, and thus very large sequenced populations are required to obtain robust estimates of gene tolerance to inactivation. Here, we describe the detection of pLoF variants in 125,748 and 15,708 humans with whole exome and genome sequence data, respectively, aggregated and jointly variant-called as part of the latest version of the Genome Aggregation Database (gnomAD). We perform stringent filtering for sequencing and annotation artifacts to identify a set of high-confidence pLoF variants in this cohort. Using an updated model based on random mutational expectations and the allele frequency spectrum, we classify all human protein-coding genes based on their observed intolerance to both monoallelic and biallelic inactivation, which we correlate with incidence of predicted gene-disrupting structural variants. We identify 15% more pLoF-intolerant genes than were found from previous efforts due to our increased sample size and improved model, and we investigate the extent of population-specific constraint across major continental groups. We show that this expanded catalogue can be used to improve gene discovery power for both common and rare diseases, and that highly constrained genes are enriched for rare pLoFs in individuals with psychiatric and developmental traits as well as overall health-related traits. Additionally, highly constrained genes are 6.6-fold enriched for genes with de novo pLoF variants in cases with developmental delay compared to controls, compared to a slight depletion (0.93-fold) for unconstrained genes. The resulting map of human genic constraint will be valuable for both the interpretation of disease variants and the discovery of novel therapeutic targets.

6

A comprehensive study of essential genes, Mendelian disease genes, and loss-of-function-tolerant genes. *P.D. Evans[1], E.R. Gamazon[1,2].* 1) Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN; 2) Clare Hall, University of Cambridge, Cambridge, United Kingdom.

Characterizing the function of genes in the human genome continues to be a critical challenge. The explosion of genomic data has facilitated the development of new methodologies for identifying disease-associated genetic predisposition, but large-scale functional interpretation remains elusive. Modeling non-linear and topology-preserving representations of gene properties, including gene evolutionary history, protein evolutionary rate, protein structure, gene expression level and breadth in a broad collection of cell types or tissues from GTEx, the regulatory genetic architecture, within-biological-network characteristics, and bioenergetics fitness costs may yield important biological insights and provide a distinctly powerful approach to functional inference. Using a neural network (self-organizing map) trained on the set of known essential genes, Mendelian disease genes, and loss-of-function-tolerant genes, we show that we can significantly predict a number of molecular features for the remaining set of genes, including within-network characteristics, summary expression profile, and tissue expression breadth. Notably, we identify many novel potential essential genes and Mendelian disease genes and a number of novel potential loss-of-function-tolerant genes, using a dispensability score generated from the computational inference. We demonstrate strong predictive performance of the dispensability score, using a recent CRISPR-based screen of essential genes and the probability of loss-of-function intolerance metric derived from large-scale exome-sequencing data (ExAC). Finally, we provide a comprehensive assessment of the phenotypic impact of the novel essential genes and Mendelian disease genes using gene-based analyses of BioVU, a database of electronic medical records linked to genetic data at Vanderbilt University Medical Center. This work provides an important new tool in the prioritization of genes in a medical setting in the context of personalized medicine. Reports from patients often contain many possible disease-causing mutations of unknown significance, and it is often difficult to elucidate which mutations may truly lead to disease. This new gene-based dispensability score will be complementary to existing functional scores on SNPs such as GERP, PolyPhen, and SIFT.

Copyright © 2018 American Society of Human Genetics. All rights reserved

7

Using Exomiser for rare disease variant interpretation at scale in the 100,000 Genomes Project. *D. Smedley[1,5], J.O. Jacobsen[5], A.R. Martin[1], C. Johnson[1], C. Boustred[1], K. Smith[1], E. Thomas[1], H. Brittain[1], J. McMurry[2], M. Haendel[2], C. Mungall[3], P.N. Robinson[4], M. Cauflield[1,5], A. Rendon[1].* 1) Genomics England, Queen Mary University London, London, United Kingdom; 2) Linus Pauling Institute, Oregon State University, Corvallis; 3) Lawrence Berkeley National Laboratory, Berkeley, CA, USA; 4) Jackson Laboratory for Genomic Medicine, Farmington, CT, USA; 5) William Harvey Research Institute, Queen Mary University of London, London EC1M 6BQ.

Although genomics has revolutionized rare disease diagnostics, many cases remain unsolved, in part because of the problem of prioritizing the 10-100's of candidate variants that remain after removing those identified as common or non-pathogenic. Exomiser is an automated approach developed by the Monarch Initiative to address this very problem; it takes patient sequencing data and coded patient phenotypes and analyzes both, informed by a large corpus of gene-phenotype associations from humans and model organisms. Here we describe the use of Exomiser in the rare disease component of the UK 100,000 Genomes Project, which is transforming the National Healthcare System (NHS) through the creation of a national research database of linked genomic and clinical data. Exomiser interpretation is enabled through the systematic collection of Human Phenotype Ontology terms, with 128,000 positive and 400,000 negative terms collected to date for ~20,000 participants from 1000+ contributing clinicians. We have created a new, scalable Exomiser service architecture for the project and future NHS Genomic Medicine Service, capable of interpreting 300 cases/day for every processor node. In a blinded study of 263 patients from the project, Exomiser was benchmarked for how close it came to the diagnosis provided by NHS clinical geneticists. These cases represent 62 disease categories ranging from those with a high diagnostic rate, such as intellectual disability, to more challenging disorders such as familial colon cancer. The family structures of the cases also varied in complexity: 89 singletons, 90 duos, 111 trios, and 17 larger pedigrees. Exomiser successfully reproduced 68%, 78% or 81% of the diagnoses in the top 1, 3 and 5 matches respectively. 5% of the diagnoses were filtered out due to non-penetrance: we plan to add this functionality, along with copy number and cryptic splice variant analysis to identify additional diagnoses. Exomiser performs comparably to our virtual gene panel based pipeline (72% of diagnoses identified with a mean of 2.3 variants presented per case) but without the extra overhead of expert panel curation and clinical review to select appropriate panels. The two approaches are synergistic: combining Exomiser top 5 and panel matches yields better recall (87%) than either approach alone. In summary, Exomiser expedites phenotype-based variant prioritization at scale for improved diagnostic rates in the 100,000 Genomes Project.

8

Assessment of penetrance of 10 Mendelian disease states in 46,980 exomes. *J.K. Goodrich[1,2], J.C. Wood[2], S. Baxter[2], A. O'Donnell-Luria[1,2], B. Weisburd[1,2], Z. Zappala[1,2], J. Mercader[2,3], J. Flannick[2,4], J.C. Florez[2,3], D. MacArthur[1,2], M.S. Udler[1,2,3], AMP-T2D consortium.* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; 3) Department of Medicine, Massachusetts General Hospital, Boston, MA; 4) Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA.

Availability of large-scale sequencing data provides an opportunity to improve the evaluation of the penetrance of predicted disease-causing genetic variants. Leveraging a dataset of 46,980 exomes and linked phenotypes, ascertained from a large international consortium of type 2 diabetes (T2D) case-control studies, we assessed predicted pathogenic variants across 77 genes associated with 10 Mendelian disease states: four lipid disorders, severe obesity, neonatal and adult-onset of monogenic diabetes, partial lipodystrophy, renal disorders, and short stature syndromes.Across these genes, we identified 369 high-confidence loss-of-function variants annotated using automated filters and manual inspection of sequence reads, as well as 23 variants previously reported to be pathogenic or likely pathogenic in ClinVar and manually curated using the ACMG variant interpretation criteria. Relevant phenotypes in 1,093 variant carriers (C) and non-carriers (NC) were compared using Wilcoxon rank sum and logistic regression tests, adjusting for age, sex, T2D status, and ethnicity.Carriers of predicted pathogenic variants in genes causing the four lipid disorders had significantly different levels of relevant lipids (median [range] in mg/dl): disorders of low LDL cholesterol (*APOB*, *PCSK9*; C: 59[20-152], NC: 121[0.4-426]; P=1x10^{-27}); high LDL (*LDLR*; C: 236[92-383], NC: 121[0.4-426]; P=3x10^{-4}); high HDL cholesterol (*CETP*; C: 57[39-117], NC: 46[4-293]; P=9x10^{-6}); and high triglycerides (*APOA5*, *LPL*; C: 237[75-642], NC: 129[6-4772]; P=5x10^{-5}). We found a nominally significant association of higher BMI in carriers of predicted obesity-causing variants (median BMI in kg/m^2 C: 29.6[18.4-40.5], NC: 27.5[10.5-96]; P=0.02). There were no significant differences in T2D prevalence, creatinine levels, and height in carriers compared to non-carriers for monogenic diabetes or lipodystrophy, renal disorders, and short stature syndromes, respectively.Analysis of large-scale sequence data linked to phenotypes demonstrated incomplete penetrance of likely disease-causing variants across multiple disease states despite careful manual variant curation. Trait differences between carriers and non-carriers were most pronounced for lipid disorders, however for all disease states (with the exception of diabetes due to study design), <50% of carriers displayed expected phenotypes. Future research will investigate potential causes of incomplete penetrance including genetic modifiers.

Copyright © 2018 American Society of Human Genetics. All rights reserved

9

Joint genome-wide association study of endometrial cancer and ovarian cancer identifies novel genetic risk regions at 7p22.2 and 14q23.3. *T.A. O'Mara[1], D.M. Glubb[1], D.J. Thompson[2], A.B. Spurdle[1], OCAC, ECAC.* 1) QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia; 2) University of Cambridge, Cambridge, UK.

Ovarian cancer is the most lethal gynaecological malignancy and the sixth most common cause of death from all cancer in women, with an estimated 140,000 deaths per year worldwide. Endometrial cancer (cancer of the uterine lining) is the most commonly diagnosed gynaecological cancer, ranking sixth in incident cancer in women. Ovarian and endometrial cancer share many epidemiological, histopathologic, and tumour genetic characteristics. Meta-analyses of genome-wide association study (GWAS) datasets across aetiologically related diseases have successfully been used to increase statistical power and identify novel genetic risk regions. We hypothesised that joint meta-analysis of ovarian and endometrial cancer GWAS datasets would identify novel genetic loci for both cancers. Following quality control, summary statistics for >11 million genetic variants were available from the largest genome-wide association studies performed by the Ovarian Cancer Association Consortium (OCAC, Phelan et al Nature Genetics 2017) and the Endometrial Cancer Association Consortium (ECAC, O'Mara et al Nature Communications, accepted). A total of 35,312 cancer cases from all histological subtypes (22,406 ovarian and 12,906 endometrial cancer cases) and 149,920 controls were included in the analysis. Summary statistics were combined using an inverse-variance, fixed effects model in METAL and identified nine loci at genome-wide significance ($P < 5 \times 10^{-8}$), of which two (14q23.3 and 7p22.2) had not been previously associated with the risk of either cancer. Analyses stratifying by histological subtypes identified an additional five loci associated with endometrioid cancers and two with serous cancers. Integration of cross-cancer GWAS variants with chromatin conformation data (H3K27Ac HiChIP) from an endometrial cancer (Ishikawa) and normal (E6/E7 hTERT immortalised endometrial glandular) cell line identified 88 candidate target genes, including the oncogenes *AKT1*, *TERT* and *MYC*. Ingenuity pathway analysis found enrichment for relevant functional annotation such as tumorigenesis of the female genital tract, cell proliferation and colony formation. Generation of similar chromatin confirmation data for ovarian cell lines is currently underway and will be used identify candidate target genes common to both cancers for future study.

10

Combining genetic and exposure data significantly improve risk prediction for skin cancer. *P. Fontanillas, B. Alipanahi, M. Johnson, C. Wilson, A. 23andMe Research Team, S. Pitts, R. Gentleman, A. Auton.* 23andMe, Mountain View, CA.

Epidemiological studies have successfully identified many important risk factors associated with the development of skin cancer. However, accurate prediction of personal skin cancer risk remains challenging and unreliable. While basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) are the more common forms of skin cancer, melanoma is generally better studied due to its more aggressive nature. Over the last two decades, many predictive models have been proposed for melanoma, but consensus regarding which models provide the most medically relevant predictions remains elusive. In part, this lack of consensus is due to an incomplete knowledge of potential risk factors (including genetic risks), small sample sizes, and absence of out-of-sample validation. In this study, we evaluate and extend prediction risk models for all three skin cancer types (BCC, SCC, melanoma) in a large-scale deeply phenotyped and genotyped cohort. Over 210,000 consented 23andMe research participants responded to an online survey containing 70 questions covering personal history of skin cancer, factors relating to susceptibility and UV exposure, and details of skin cancer incidence in family history. We trained risk prediction models using data from 103,008 research participants, of whom 20,765 reported having or having had skin cancer, and held out data from the remaining 88,985 individuals as a validation set. Using this data, we tested and validated published epidemiological models, identified additional skin cancer risk factors, and investigated the value of incorporating knowledge of genetic risk into the best performing models. The best predictive models incorporate family history, pigmentation and skin sensitivity variables, mole counts, sun exposure estimates, and also a few new variables, like BMI, which is negatively correlated with skin cancer. The genetic risk scores (GRS), computed from GWAS significant loci, account for 8 to 14% of the total variance explained by the best models. The out-of-sample AUC for the best models are 0.79, 0.80, and 0.78, for BCC, SCC, and melanoma, respectively. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

11

Pan-cancer analysis detects novel genetic risk variants and shared genetic basis in the UK biobank cohort. *S.R. Rashkin[1], R.E. Graff[1], L. Kachuri[1], T.J. Meyers[1], N.C. Emami[1,2], K.K. Thai[3], S.E. Alexeeff[3], D.A. Corley[3], L.H. Kushi[3], S.K. Van Den Eeden[3,4], E. Jorgenson[3], L.A. Habel[3,5], L.C. Sakoda[3], T.J. Hoffmann[1,6], E. Ziv[6,7], J.S. Witte[1,6].* 1) Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA; 2) Program in Biological and Medical Informatics, University of California, San Francisco, San Francisco, CA; 3) Division of Research, Kaiser Permanente Northern California, Oakland, CA; 4) Department of Urology, University of California, San Francisco, San Francisco, CA; 5) Department of Health Research and Policy, School of Medicine, Stanford University, Stanford, CA; 6) Institute for Human Genetics, University of California, San Francisco, San Francisco, CA; 7) Department of Medicine, University of California, San Francisco, San Francisco, CA.

While GWAS have detected a large number of cancer risk loci, the potential shared genetic basis across cancers remains unclear. Investigating multiple phenotypes within a single, large cohort allows for studying a broad range of cancers. We analyzed genome-wide SNP data across 18 cancers in European ancestry samples from the UK Biobank cohort, comparing cases (N=471 to 13,903 across cancers) to 359,825 cancer-free controls. We detected novel GWAS hits for several cancers, including melanoma, cervical, thyroid, and pancreatic cancers. For melanoma, we identified rs183783391 (OR=0.59; P=2.68E-12) in *MITF*, a known high penetrance melanoma gene regulating melanocytes. For cervical cancer, we identified rs6755040 (OR=1.16; P=9.02E-14), a SNP in *PAX8*; rs27069 (OR=1.12; P=2.67E-9), a SNP in *CLPTM1L*; and rs117905563 (OR=0.70; P=1.44E-8), an intergenic variant. Previous GWAS have not identified any of these SNPs, but a study examining the effect of a single SNP in *CLPTM1L* (rs401681) on multiple cancers reported an association with cervical cancer. For thyroid cancer, we identified 2:173859846_TA_T (OR=1.45; P=1.5E-8), an indel in *RAPGEF4*. For pancreatic cancer, we identified rs911554 (OR=1.5; P =1.69E-8), a SNP in *TRMT61A*. We examined shared genetic factors among cancers by identifying independent regions associated with more than one cancer. Linkage disequilibrium regions were formed around index SNPs with the smallest p-values in any of the cancer GWAS and not already assigned to another region. In each region, SNPs were associated with any cancer at P<1E-6, were within 250kb of the index SNP, and had r^2>0.5 with the index SNP, and regions had to include SNPs associated with at least two cancers. One of the thirty regions identified was in *CLPTM1L*, comprised of 67 SNPs associated with melanoma, cervical, pancreatic, and lung cancers. SNPs in this region had different directions of effect for melanoma and pancreatic cancers compared with cervical and lung cancers. Other regions were in the *HLA* gene complex and on chromosome 8q24, which have previously been associated with multiple cancers. Preliminary findings of co-heritability analyses indicate a genetic correlation between breast cancer and cervical cancer (r_g=0.30; P=0.006), colon cancer (r_g=0.46; P=0.005), and endometrial cancer (r_g=0.40; P=0.006). Cervical and colon cancers were also genetically correlated (r_g=0.70; P=0.019). We plan to undertake replication analyses in the Kaiser Permanente RPGEH cohort.

12

Large scale genome-wide association meta-analysis of melanoma identifies 69 independent variants in 56 loci including immune related genes. *M.H. Law[1], J. Shi[2], D.T. Bishop[3], K.M. Brown[3], A.J. Stratigos[4], M.C. Fargnoli[5], P. Ghiorzo[6], K. Peris[7], A.E. Cust[8], J. Han[9,10], D.C. Whiteman[11], D. Schadendorf[12], G.J. Mann[13], G.L. Radford-Smith[14,15,16], N.G. Martin[17], C. Hayward[18], N.K. Hayward[19], G.W. Montgomery[20], S.V. Ward[21,22], P.D.P. Pharoah[23], C.I. Amos[24], M. Zawistowski[25], S. Puig[26], D.L. Duffy[17], F. Demenais[27,28], E. Nagore[29], S. MacGregor[1], M.M. Iles[3], M.T. Landi[2], UK Biobank, 23andMe Research Team.* 1) Statistical Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Herston, Australia; 2) Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA; 3) Section of Epidemiology and Biostatistics, Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK; 4) 1st Department of Dermatology – Venereology, National and Kapodistrian University of Athens School of Medicine, Andreas Sygros Hospital, Athens, Greece; 5) Department of Dermatology, University of L'Aquila, L'Aquila, Italy; 6) Department of Internal Medicine and Medical Specialties, University of Genoa and Genetics of Rare Cancers, Ospedale Policlinico San Martino, Genoa, Italy; 7) Institute of Dermatology, Catholic University, Rome, Italy; 8) Cancer Epidemiology and Services Research, Sydney School of Public Health, University of Sydney, Sydney, New South Wales, Australia; 9) Department of Epidemiology, Fairbanks School of Public Health, Indiana University, Indianapolis, Indiana, USA; 10) Melvin and Bren Simon Cancer Center, Indiana University, Indianapolis, Indiana, USA; 11) Population Health, QIMR Berghofer Medical Research Institute, Brisbane, Herston, Australia; 12) Department of Dermatology, University Hospital Essen, Hufelandstrasse 55, 45122 Essen, Germany; 13) Centre for Cancer Research, University of Sydney at Westmead, Millennium Institute for Medical Research and Melanoma Institute Australia, Sydney, Australia; 14) Inflammatory Bowel Diseases, QIMR Berghofer Medical Research Institute, Brisbane, Australia; 15) Department of Gastroenterology and Hepatology, Royal Brisbane & Women's Hospital, Brisbane, Australia; 16) University of Queensland School of Medicine, Herston Campus, Brisbane, Australia; 17) Genetic Epidemiology, The QIMR Berghofer Medical Research Institute, Brisbane, Herston, Australia; 18) MRC Human Genetics Unit, University of Edinburgh, Institute of Genetics and Molecular Medicine, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK; 19) Oncogenomics, QIMR Berghofer Medical Research Institute, Brisbane, Herston, Australia; 20) Molecular Biology, the University of Queensland, Brisbane, Australia; 21) Centre for Genetic Origins of Health and Disease (GOHaD), University of Western Australia; 22) Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center; 23) Department of Oncology, University of Cambridge, Cambridge, United Kingdom; 24) Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, USA; 25) Center for Statistical Genetics, University of Michigan, Ann Arbor MI, USA; 26) Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain. & Centro de Investigación Biomédica en Red en Enfermedades Raras (CIBERER), Valencia, Spain; 27) Institut National de la Santé et de la Recherche Médicale (INSERM), UMR-946, Genetic Variation and Human Diseases Unit, Paris, France; 28) Institut Universitaire d'Hématologie, Université Paris Diderot, Sorbonne Paris Cité, Paris, France; 29) Department of Dermatology, Instituto Valenciano de Oncología, València, Spain.

We report on a large (over 36,000 cases and 370,000 controls) international GWAS meta-analysis of cutaneous melanoma risk, a threefold increase over our previous meta-analysis. In addition to new GWAS data from the UK, the US, Australia and Northern/Western Europe, we also now include over 6,400 individuals with melanoma from Mediterranean populations, which are often under-represented in studies of melanoma. This new meta-analysis increases the number of melanoma risk loci from 20 to 56 (69 independent SNPs). The genome-wide SNP based heritability estimate for melanoma risk is 0.11 (SE 0.02). Annotation data for reported loci indicates the majority of the lead SNPs are in putative enhancers in melanocytes or their paracrine regulators, keratinocytes. We will also report on the overlap with melanoma and GWAS of known risk factors. In addition to better annotating melanoma loci, combining these traits allows further discoveries e.g. the combination with nevus count identifies an additional 9 novel loci associated with both traits. Discovery was not limited to new loci; the combination of increased power and greater resolution afforded by the HRC imputation panel enabled imputation of many of the rare red hair/fair skin associated *MC1R* alleles, and independently associated (and novel) SNPs at critical loci such as *CDKN2A*. As expected given the importance of sun exposure in melanoma, a number of the new risk SNPs are associated with pigmentation or tanning response (rs12078075/*RIPK5*;

Copyright © 2018 American Society of Human Genetics. All rights reserved

rs32578/*PPARGC1B*, rs10809803/*TYRP1*, rs1278763/*ATP11A*, rs5766565/*KIAA0930*). We also identify SNPs outside these pathways. These include common SNPs near (or that are eQTLs for) the familial melanoma gene *POT1*; *RAPGEF1* (apoptosis) and *MPHOSPH6* (cell division). Melanoma is deadly in part due to immune evasion, and we identify SNPs that are in (or that are eQTLs for) a *HLA* gene, *BACH2*, and *MSC* (immune response/inflammation). A number of the novel and known CMM risk SNPs are associated across multiple cancers suggesting common pathways to carcinogenesis. In addition to the established overlap with keratinocytic cancers at pigmentation/tanning genes, we identify two independent SNPs in *TP53*, one of which (rs78378222) has previously been associated with glioma and basal cell carcinoma. Novel SNPs at *CDH1* are also associated with colorectal cancer, and SNPs near *TERC* have also been associated with a range of cancers including B-cell malignancies, and telomere length.

13

High-resolution epigenomic atlas of human embryonic craniofacial development. *J. Cotney[1], A. Wilderman[1,2], J. VanOudenhove[2], J.P. Noonan[3].* 1) Genetics and Genome Sciences, UConn Health, Farmington, CT; 2) Graduate Program in Genetics and Developmental Biology, UConn Health, Farmington, CT 06030, USA; 3) Department of Genetics, Yale University School of Medicine, New Haven, CT 06510, USA.

Defects in patterning during human embryonic development frequently result in craniofacial abnormalities. The gene regulatory programs that build the craniofacial complex are likely controlled by information located between genes and within intronic sequences. However, systematic identification of regulatory sequences important for forming the human face has not been performed. Here, we describe comprehensive epigenomic annotations from human embryonic craniofacial tissues and systematic comparisons with multiple tissues and cell types. We identified thousands of tissue-specific craniofacial regulatory sequences, find likely causal variants for multiple craniofacial abnormalities, and identify a putative locus control region for the *HOXA* gene cluster activated during specifically human craniofacial development. We demonstrate significant enrichment of common variants associated with orofacial clefting in enhancers active early in embryonic development, while those associated with normal facial variation are enriched near the end of the embryonic period. This data resource represents the most detailed epigenomic annotation of any developing human tissue and greatly expand our knowledge of how craniofacial development is regulated. These data are provided in easily accessible formats for both basic craniofacial researchers and clinicians to aid future experimental design and interpretation of noncoding variation in those affected by craniofacial abnormalities. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

14

Rheumatoid arthritis heritability is concentrated in regulatory elements with CD4+ T cell-state-specific transcription factor binding chromatin signatures. *T. Amariuta[1,2,3,4], Y. Luo[1,2,3,4], E. Davenport[1,2,3,4], S. Gazal[4,5,6], B. van de Geijn[4,5,6], H.J. Westra[1,2,3,4], N. Teslovich[1,2,3,4], A. Price[4,5,6], S. Raychaudhuri[1,2,3,4].* 1) Department of Biomedical Informatics, Harvard Medical School, Boston, MA; 2) Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 3) Partners Center for Personalized Genetic Medicine, Boston, MA; 4) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; 5) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA; 6) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA.

Active regulatory regions within CD4+ T cells are known to harbor disproportionate heritability (h2) for rheumatoid arthritis (RA). We hypothesized that regulatory regions specific to pathogenic CD4+ T cell-states are enriched for RA h2. However, defining cell-state-specific regulatory elements is challenging. To this end, we introduce IMPACT (Inference and Modeling of Phenotype-related ACtive Transcription), a genome annotation strategy to identify cell-state-specific regulatory regions defined by the epigenomic profiles of transcription factor (TF) binding sites, using 398 chromatin and sequence annotations in a logistic regression model. We predicted regulatory elements in four CD4+ T cell-states (T helper 1 (Th1), 2 (Th2), 17 (Th17), and T regulatory (Treg)) using binding data of their core regulators and found that IMPACT predicts TF motif binding with high accuracy: mean AUC 0.94 (s.e. 0.03). We then identified TF-cell-state-specific chromatin profiles genome-wide and integrated these CD4+ T annotations with RA summary statistics of 38,242 Europeans and 22,515 East Asians using stratified LD score regression. In a meta-analysis of both populations, we observe that the top 5% of genome-wide predicted CD4+ T cell-state-specific regulatory regions explain 77.5% (s.e. 19.0%) of RA h2 and that these annotations capture more RA h2 than other T cell annotations (all p<0.05), including T cell super enhancers, CD4+ T cell specifically expressed genes, and CD4+ T cell specific histone marks. Finally, integration with RA fine-mapping data (N=27,345) revealed a significant enrichment (2.87, p<8.6e-03) of putatively causal variants across 20 RA associated loci in the top 1% of predicted CD4+ Treg regulatory regions. Treg-specific enrichment in the BACH2 and IRF5 loci suggests that experimental variant follow-up be performed in Tregs. We extended our analyses to 41 other complex traits and find that CD4+ T IMPACT annotations have significantly positive standardized effect sizes, defined as the proportionate change in per-SNP h2 associated with increasing the annotation value by one standard deviation, in other autoimmune traits, such as Crohn's disease and ulcerative colitis (all p<0.04). Conversely, we observe that CD4+ T IMPACT annotations do not have significantly positive standardized effect sizes for non-immune-mediated traits. As more diverse cell-state-specific data is generated, IMPACT may be used to identify other trait associated regulatory regions.

15

Uniformly collected and processed functional genomics assays provide a way to interpret non-coding variants in terms of tissue of action. *F.C.P. Navarro[1], G. Gursoy[1], J. Rozowsky[1], A. Dobin[2], T. Galeev[1], X. Kong[1], A. Vlasova[3], R. Guigo[3], M. Schatz[4], T. Gingeras[1], M. Gerstein[1].* 1) Yale University, Program in Computational Biology & Bioinformatics, New Haven, CT, USA; 2) Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; 3) The Barcelona Institute of Science and Technology, Centre for Genomic Regulation (CRG), Barcelona, Spain; 4) Department of Computer Science and Biology, Johns Hopkins University, Baltimore, MD.

One of the most important ways of interpreting non-coding variants, expression quantitative trait loci (eQTL), and genome-wide association study variants (GWAS) is through examining which tissue they are significantly active and investigating uniformly profiled chromatin derived from the same tissue. To get an accurate sense of chromatin activity one needs to integrate a wide variety of assays (e.g. ATAC, DNAse-seq, Histone Modification, Hi-C). The EN-TEX working group has been uniformly and comprehensively characterizing the genome, chromatin, and transcriptome of over 20 tissues from four adult individuals. We provide a unified chromatin and also a measure of intraindividual variation that one can use to calibrate against these measures. EN-TEX has generated more than 1,300 functional genome experiments. These experiments include RNA-seq (long and short RNA), chromatin characterization assays such as ATAC-seq, DNAse-seq, and ChIP-Seq to identify histone modification (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3) and other marks (RNAPOL-II, EP300, CTCF). Tissues also have been characterized in regard to their DNA methylation (WGBS and ChIP-array) and three-dimensional structure (Hi-C). Here, we first characterized epigenome variation by comparing functional genomic assays across tissues. We found that gene expression, H3K27ac, and H3K4me3 are highly structured and distinct between tissues. Moreover, we identified and characterized thousands of heterozygous variants and hundreds of genes and chromatin marks associated with allele-specific expression, binding, and chromatin state. Integrative analyses of personal variation, known eQTLs, and the relationships of allelic imbalances in different assays revealed variants and mechanisms that may be responsible for the allelic skew of some of the detected allele-specific genes. We compiled tissue-specific epigenome maps integrating multiple assays. Our annotation revealed a high enrichment of GWAS and eQTLs associated with promoters and enhancers. Furthermore, tissue-specific maps further enhanced our capacity to annotate potentially causal variants. In particular, pathological GWAS SNPs are frequently in linkage disequilibrium with variants disrupting tissue-specific regulatory regions. Our findings support that uniformly collection and processing of human tissue yield better tissue-specific epigenetic maps that are fundamental to a better annotation of new and existing GWAS and eQTLs SNPs.

Copyright © 2018 American Society of Human Genetics. All rights reserved

16

Development of a comprehensive, curated database of open chromatin regions of 192 cell types. *P. Shooshtari[1], C. Cotsapas[2,3].* 1) The Hospital for Sick Children, Toronto, Ontario, Canada; 2) Yale University, New Haven, CT; 3) Broad Institute, Cambridge, MA.

In the past years multiple international consortia, such as ENCODE, Roadmap Epigenome Project, GGR and Blueprint, have made several large scale epigenomics datasets publically available. These datasets are extremely useful for the study of mechanisms of gene regulation in different concepts such as disease and cell development. However, the computational pipelines used to identify regulatory regions in these datasets process each sample individually, and therefore a uniform comparison of the same regulatory regions over several samples is not directly possible. This is a bottleneck in many applications, where it is required to identify the same regulatory regions over multiple samples, an example of which is for assessing correlation between activity of specific regulatory elements and expression of genes that they control. Additionally, although most of these epigenomic datasets contain multiple biological and technical replicates of the same cell types, a comprehensive replication-based quality checking of the activity of individual regulatory elements is still lacking. In order to address the above problems, we integrated 828 DNase-I Hypersensitive Sites (DHS) data samples from ENCODE, Blueprint, Roadmap and GGR, and built a well-curated database of genome-wide DHS activity for these samples. We used the DHS peaks obtained by MACS2 peak calling algorithm and clustered the peaks across samples. This resulted in a total of 4,020,940 DHS clusters. We then checked quality of each cluster using our replication test. Only 1,449,102 (36%) of clusters pass, but explain the disease heritability attributable to all peaks, suggesting the non-replicating peaks are spurious. Since multiple centers generated these data, we observed batch effects in DHS intensities and removed them using Combat method. Then for each cell type with multiple replicates, we measured intensities of DHS clusters as the median intensity of the cluster over replicate samples. This resulted in obtaining intensities for 1,449,102 replicable DHS clusters over 192 unique cell types. We are designing a web-interface that makes it possible to query and visualize DHS activities over any genomic region of interest. Up to our knowledge, this is the most comprehensive DHS database built by integrating and curating of over 800 samples from multiple international projects and can serve as a reference for studies concerned with DHS activities and their role in gene regulation.

17

Mosaic chromosomal alterations increase proliferative loads from rare coding variants and common polygenic risk. *P. Loh[1,2], G. Genovese[2,3,4], S.A. McCarroll[2,3,4].* 1) Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; 3) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA; 4) Department of Genetics, Harvard Medical School, Boston, MA.

Clonal expansion of blood cells harboring somatic mutations (“clonal hematopoiesis”) is a common age-acquired condition that greatly increases risk of future blood cancer. The mechanisms that shape most clonal expansions in healthy individuals are not understood. We recently identified four loci at which rare inherited variants associate strongly with the development of clones in which nearby genomic segments have been deleted or made homozygous by somatic mutation (Loh*, Genovese* et al. 2018 *Nature*). We discovered these loci by analyzing array-based allele-specific intensity data from the UK Biobank interim release (*N*≈150K) and harnessing Eagle2 long-range haplotype phasing. Here we describe deeper insights from ascertaining far more (19,635) autosomal mosaic chromosomal alterations (mCAs) in the full UK Biobank cohort (*N*≈500K). We found six new loci at which somatic copy-number neutral loss of heterozygosity (CNN-LOH) events tended to duplicate or delete specific inherited risk alleles. Four such loci (*FH*, *SH2B3*, *NBN*, and *MRE11A*) involved rare variants with large effect sizes (OR=10–129). At *FH*, *SH2B3*, and *MRE11A*, we identified likely causal inherited coding variants that were made homozygous by somatic CNN-LOH. Two other loci (*TCL1A* and the imprinted *DLK1* locus) involved common variants previously implicated in mosaic Y chromosome loss. We also identified 12 independent likely causal coding variants (deleted by CNN-LOH) in *MPL* and 8 independent likely causal coding variants (duplicated by CNN-LOH) in *ATM*. Carriers of the inherited *ATM* variants exhibited increased cancer risk (OR=1.4, *p*=2e-6). CNN-LOH events tended to cause chromosomal segments with more-proliferative alleles to replace their homologous (allelic) counterparts. Averaging across all CNN-LOH events, the allelic substitutions produced by CNN-LOH significantly increased polygenic risk scores for white blood cell, neutrophil, and eosinophil counts as well as Y chromosome loss. We found that, for 14 chromosome arms, polygenic risk could predict which homologous chromosome was duplicated by CNN-LOH events, with the directionality of CNN-LOH mutations correctly predicted for 59% of 5,607 CNN-LOH events on these arms (range 52–68%). Our results demonstrate that inherited variants play a key role in clonal hematopoiesis by creating latent opportunities for clonal selection that can be unleashed by somatic chromosomal alterations.

Copyright © 2018 American Society of Human Genetics. All rights reserved

18

Clonal mosaicism in normal TCGA tissues. *Y.A. Jakubek[1], J. Fowler[1], F.A. San Lucas[1], H. Kadara[2], P. Scheet[1].* 1) MD Anderson Cancer Center, Houston, TX; 2) The American University of Beirut, Beirut, Lebanon.

Errors in DNA replication during cell division can create daughter cells with chromosomal aberrations that propagate to establish genetically distinct cell populations within an individual or tissue. This phenomenon, known as clonal mosaicism, is an important factor in the development of cancer. Recent surveys, mostly of blood, have uncovered extensive mosaicism in healthy tissues. In this work, we surveyed mosaicism in the normal tissues of TCGA patient with non-hematological malignancies. We excluded samples with possible contamination in two ways, using annotated sample files and a statistical method. We analyzed samples with hapLOH, a method that can detect changes indicative of chromosomal aberrations (gain, loss and copy-neutral LOH) even when present at low mutant cell fractions (< 5%). Our study comprised 1,714 samples derived from normal tissue adjacent to the tumor (NAT) and 7,194 samples derived from blood. We detected 338 (mega-base scale) chromosomal alterations in NAT and 178 alterations in blood. The rate of mosaicism in NAT tissues was more than twice the rate in blood (4.6% vs. 1.8%, p = 5e-11). As previously reported, blood mosaicism was positively correlated with age (p = 3e-15); by contrast, interestingly, NAT mosaicism did not show such a relationship. NAT mosaic tissues had an average of 4.3 mosaic chromosomal aberrations which was significantly higher than blood with an average of 1.4 events (p = 0.001). Both, blood and NAT mosaic chromosomal aberrations had a median size of 32 Mb and recurrent gains of 8q. The most common recurrent alteration in NAT was located on 9q, a common hot-spot for copy number changes in cancer. In blood, 13q alterations were the most common, as has been previously reported. Both blood and NAT tissues exhibited high rates of alterations on chromosome 20; however, these were 9 gains in NAT (0 losses) and 12 losses in blood (0 gains). Our results point to higher rates of clonal mosaicism in NAT tissues compared to blood, as well as important differences in the distribution of mosaic chromosomal aberrations across the genome of these tissues.

19

Identification, characterization, and modeling of genomic mosaicism among multiple tissues of healthy individuals. *Y. Huang[1,2,3], Y. Ye[2], Y. Dou[2,3], X. Yang[2], S. Wang[4], X. Zheng[4], L. Wei[1].* 1) Boston Children's Hospital, Boston, MA; 2) Peking University, Beijing, China; 3) Harvard Medical School, Boston, MA; 4) National Institute of Biological Sciences, Beijing, China.

Postzygotic single-nucleotide mosaicisms (pSNMs) are known to play critical roles in tumors as well as an increasing number of non-cancer diseases. However, the origin and patterns of pSNMs in normal tissues of healthy human individuals remain largely unknown. Using MosaicHunter, we identified and validated 164 pSNMs from the 90X whole-genome sequencing data of 27 postmortem tissue samples (including brain, liver, colon, skin, and prostate) obtained from five healthy donors. Our inter- and intra-tissue comparisons revealed two distinctive types of pSNMs, with about half originating during early embryogenesis (embryonic pSNMs) and the remaining likely to result from tissue-specific clonal expansion events that had occurred more recently (clonal expansion pSNMs). Interestingly, embryonic and clonal expansion pSNMs behaved significantly different genomic characteristics in regard to mutation spectrum, replication timing, and chromatin status, suggesting varied mutational mechanisms between these two types of pSNMs. Compared with germline mutations, embryonic pSNMs showed signatures of relaxed negative selection. We further developed a new mathematical model to describe the accumulation and allele fraction drift of pSNMs for the development of multi-cellular organisms. Applying this model, we discovered elevated postzygotic mutation rate during early embryogenesis, especially during the first cell division. We estimated that parental pSNMs occurring during the first three cell divisions contributed to approximately 10% *de novo* mutations in their children. Last but not the least, we demonstrated the possibility to reconstruct developmental lineage tree across multiple tissue types using pSNM profiles. Our analyses provide new insights into the origin and distribution of postzygotic mutations during normal human development, and shed light on understanding their pathogenic role in human diseases and aging process.

Copyright © 2018 American Society of Human Genetics. All rights reserved

20

Leveraging single-cell RNA-seq to infer cell type-specific somatic mutations and mosaicism in Alzheimer's disease. *C. Boix[1,2], M. Kousi[1,2], H. Mathys[3,4], L. Tsai[2,3,4], M. Kellis[1,2].* 1) EECS, Massachussetts Institute of Technology, Cambridge, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA; 4) Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA.

Recent studies of the aging brain demonstrated a significant accumulation of mosaic hits in post-mitotic neurons, making the case for a link between DNA damage over time and age-related increased transcriptional noise, resulting in local disruption of neuronal pathways. However, previous work focused on DNA sequencing data, enabling studies in only one cell type, neurons. This is problematic, as previous work suggests that 1-4 mutations accumulate per division in mitotic cells, indicating that clonal mosaic mutations in glial cells may play pivotal roles in driving focal neurological disorders. Here, we develop a new method for using single-cell RNA-seq technologies to infer exonic mutations in each cell, while simultaneously characterizing the identity of each cell harboring these mutations. We applied our method to Smart-seq2 data from post-mortem brain samples in a cohort of 24 Alzheimer's and 24 control patients from the ROS/MAP project. We jointly called mutations in cells from full length coding transcripts, and determined their identity across six major cell types (excitatory and inhibitory neurons, astrocytes, microglia, oligodendrocytes, and OPCs). We aligned single-cell RNA-seq reads using STAR, called genotypes using GATK's HaplotypeCaller, filtered by QD > 2.0 and FS < 30.0, and further filtered potential somatic hits using known true positive ROS genotypes to set a 10% false negative rate threshold. We established cell type-specific signatures of somatic variation by estimating the ratio of clonal versus cell specific mutations in each cell type. We called clonal events by merging single-cell alignments for each patient and re-calling genotypes to determine and filter-out germline mutations. We detected hundreds of potential clonal events per patient, the majority of which strongly co-localize in astrocytes and oligodendrocytes, emphasizing the need to study clonality beyond neuronal cells. We also estimated the mutational burden for each biological and regulatory pathway in each cell type, correcting for transcriptional strength and CDS length. This resulted in numerous noteworthy pathways that are altered in each cell type. In this work, we demonstrate that single cell RNA-seq enables us to carry out a systematic cell type-specific survey of mosaicism in the aging brain, shedding light on the role of clonality and relative pathway burden in different cell types and in disease and prioritizing mutations for functional validation.

21

Genetic studies of accelerometer-based sleep measures in 85,670 individuals yield new insights into the biology of human sleep behaviour. *S.E. Jones[1], V.T. van Hees[2], D.R. Mazzotti[3,4], P. Marques-Vidal[5], S. Sabia[6,7], A. van der Spek[8], H.S. Dashti[9,10], J. Engmann[11], D. Kocevska[8,12], J. Tyrrell[1], R.N. Beaumont[1], M. Hillsdon[13], K.S. Ruth[1], M.A. Tuke[1], H. Yaghootkar[1], J.W. Harrison[1], R.M. Freathy[1], A. Murray[1], A.I. Luik[8], N. Amin[8], J.M. Lane[9,10], R. Saxena[9,14,15], M.K. Rutter[16,17], H. Tiemeier[8,18], Z. Kutalik[19,20], M. Kumari[21], T.M. Frayling[1], M.N. Weedon[1], P. Gehrman[3,4], A.R. Wood[1].* 1) University of Exeter Medical School, Exeter, UK; 2) Netherlands eScience Center, Amsterdam, Netherlands; 3) Center for Sleep and Circadian Neurobiology, University of Pennsylvania, Philadelphia, PA, USA; 4) Perelman School of Medicine of the University of Pennsylvania, Philadelphia, Pennsylvania, USA; 5) Department of General Internal Medicine, University Hospital of Lausanne, Switzerland; 6) Research Department of Epidemiology and Public Health, University College London, London, UK; 7) INSERM, U1018, Centre for Research in Epidemiology and Population Health, Villejuif, France; 8) Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, Netherlands; 9) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; 10) Broad Institute of MIT and Harvard, Cambridge, MA, USA; 11) UCL Institute of Cardiovascular Science, Research department of Population Science and Experimental Medicine, Centre for Translational Genomics, London, UK; 12) Department of Child and Adolescent Psychiatry, Erasmus Medical Center, Rotterdam, Netherlands; 13) Sport and Health Sciences, College of Life and Environmental Sciences, University of Exeter, Exeter, UK; 14) Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA; 15) Departments of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA; 16) Division of Diabetes, Endocrinology and Gastroenterology, Faculty of Medicine, Biology and Health, University of Manchester, Manchester, UK; 17) Manchester Diabetes Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK; 18) Department of Social and Behavioral Science, Harvard TH Chan School of Public Health, Boston, MA, USA; 19) Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne, Switzerland; 20) Swiss Institute of Bioinformatics, Lausanne, Switzerland; 21) ISER, University of Essex, Colchester, Essex, UK.

Sleep is an essential human function, but many aspects of its regulation remain poorly understood. Understanding the genetic aetiology of sleep is key to untangling the complex relationships between sleep patterns and their associated diseases, including metabolic and psychiatric disorders. Large-scale genetic studies of sleep have relied on self-report data, which is subject to bias, limited in scope and does not account for inter-daily variability. We used the UK Biobank to derive and perform GWAS of objective sleep measures. We aimed to identify genetic variants influencing estimates of sleep, as recorded by activity monitors, to provide insights into the molecular regulation of sleep. We extracted sleep periods in 103,711 UK Biobank activity monitor recordings and derived 8 measures of sleep quality (sleep efficiency and number of nocturnal sleep episodes), quantity (nocturnal sleep duration and variability, and diurnal inactivity duration) and timing (sleep midpoint, timing of the least-active 5 hours (L5) and timing of the most active 10 hours (M10)). We performed a GWAS of each measure in up to 85,670 European participants with high-quality activity monitor and genetic data, and conducted replication analyses using activity monitor data collected as part of 3 independent studies (N=5,819). Using BOLT-LMM, we identified 47 genetic associations at 40 distinct loci across 7 of the 8 sleep traits ($P<5x10^{-8}$). These included 10 associations for sleep duration and 26 for sleep quality, including the *APOE* ε4 allele, well-known as a high-risk haplotype for Alzheimer's disease. Most novel variants were associated with a single sleep trait, but those previously associated with restless legs syndrome were found to be associated with multiple sleep measures. Heritability estimates ranged from 2.8% for sleep duration variability to 22.3% for number of nocturnal sleep periods. Sleep quality loci were enriched for serotonin processing genes and all sleep trait loci were enriched for genes expressed in the cerebellum. We demonstrated good directional-consistency between effect sizes in self-reported and activity monitor derived sleep duration (67 of 78 variants; $P=6x10^{-11}$) and between self-reported morningness chronotype and earlier L5 timing (316 of 341; $P=3x10^{-65}$). By performing the first large-scale GWAS we found 40 novel loci associated with objective sleep measures. Our findings provide important new insights into the biology underlying sleep behaviours.

Copyright © 2018 American Society of Human Genetics. All rights reserved

22

Moderate alcohol consumption is causally related to higher risk of coronary heart disease and stroke: A Mendelian randomization analysis in the UK Biobank. *J. Lankester[1,2], D. Zanetti[1,2], T.L. Assimes[1,2,3], E. Ingelsson[1,2,3].* 1) Stanford University, Department of Medicine, Division of Cardiovascular Medicine, Stanford, CA; 2) Stanford University, Cardiovascular Institute, Stanford, CA; 3) Stanford University, Health Research and Policy, Stanford, CA.

Observational analyses have repeatedly demonstrated a correlation between a pattern of moderate alcohol use and decreased risk of cardiovascular disease. However, a variety of other factors may influence the pattern of alcohol use and confound this relationship. Mendelian randomization (MR) mimics a randomized trial via the law of independent assortment of alleles at conception, allowing causal inference. Carriers of the Arg47His (rs1229984) in *ADH1B* variant tend to drink less alcohol due to side effects such as flushing and nausea, allowing for MR studies of alcohol intake and health consequences. We performed both multivariable-adjusted linear and logistic regression observational (n=502,617) and MR (n=337,542, non-related individuals of European ancestry) analyses in the UK Biobank to determine the causal effect of alcohol consumption on body mass index (BMI), systolic blood pressure, coronary heart disease (CHD), all stroke, atrial fibrillation, and heart failure. Our observational analyses demonstrated a U-shape association of alcohol consumption with stroke risk, with increased risk among non-drinkers (odds ratio [OR], 1.34; 95% confidence interval, 1.25-1.43) and individuals with high consumption (>3 drinks/day; OR, 1.05; 0.96-1.16) compared to individuals with moderate alcohol consumption (1 drink/day). We observed a linearly decreasing risk of CHD with increasing alcohol intake (OR 1.30; 1.23-1.37 for non-drinkers and OR 0.80; 0.74-0.86 for individuals drinking >3 drinks/day), when compared with moderate drinkers. In contrast, MR showed opposite trends. The 4.5% of individuals who were carriers (heterozygotes or homozygotes) of the *ADH1B* variant drank on average 1,852 grams/year less of alcohol (e.g., 2.1 fewer glasses of wine/week). Carriers of this variant had a reduced risk of cardiovascular risk factors and disease, including BMI (beta, -0.26 kg/m^2; -0.34 to -0.19), systolic blood pressure (beta, -0.91 mmHg; -1.24 to -0.57), CHD (OR, 0.85; 0.77-0.94), and stroke (OR, 0.83; 0.73-0.94), compared to non-carriers (who drink more alcohol). We conclude that alcohol use is causally related to an increased risk of obesity, hypertension, CHD, and stroke. Our results suggest that the association between moderate alcohol use and cardiovascular disease is strongly confounded and that moderate alcohol use should not be encouraged as part of a heart healthy diet.

23

GWAS for anorexia nervosa identifies eight loci and suggests it is both a psychiatric and metabolic disorder. *N. Martin[1], C. Bulik[2], Anorexia Nervosa Genetics Initiative (ANGI).* 1) Queensland Institute of Medical Research, Brisbane, Queensland, Australia; 2) Psychiatry, University of North Carolina.

Background: Characterized primarily by extremely low BMI, anorexia nervosa (AN) is a complex, serious, and commonly misunderstood illness. AN predominantly affects women, carries high risks of mortality from wasting and suicide, and treatment outcomes remain poor. No medications exist that are effective and improved treatments will likely depend on deeper understanding of the core biology of this illness. Genetic factors clearly play an etiological role with twin-based heritability estimates of 50-60%. **Procedure:** Combining samples from the Anorexia Nervosa Genetics Initiative (ANGI) and the Eating Disorders Working Group of the Psychiatric Genomics Consortium (PGC-ED), we conducted a genome-wide association study (GWAS) meta-analysis of clinical, population, and volunteer cohorts and identified 8 independent genome-wide significant loci in 16,991 AN cases and 56,059 controls. **Results:** Our analyses reveal that the genetic architecture of AN mirrors its clinical features and comorbidities, with high genetic correlations with obsessive-compulsive disorder (OCD), major depressive disorder (MDD), and anxiety disorders. Notably, the etiology of AN had the strongest metabolic and anthropometric genetic components of any psychiatric disorder yet examined. Specifically, we observed strong negative genetic correlations between AN and fasting insulin levels, body fat percentage, and BMI as well as a strong positive genetic correlation between AN and high-density lipoprotein cholesterol. **Conclusions:** These results support a broadened reconceptualization of AN etiology, underscoring that AN should be considered as both a psychiatric and metabolic disorder. Deeper understanding of the metabolic component is a critical next step and attention to both components may be necessary to improve treatment efficacy.

Copyright © 2018 American Society of Human Genetics. All rights reserved

24

Genome-wide association study of tobacco smoking among European Americans in Million Veteran Program (MVP). *B. Li[1], A.C. Justice[1,2,3], K.A. McGinnis[3], N. Sun[1], R.V. Smith[3,4,5], C. Dao[1,3], J.P. Tate[3], W.C. Becker[2,3], J. Concato[2,6], J. Gelernter[2,3], H.R. Kranzler[4,7], H. Zhao[1,2,3], K. Xu[2,3], VA Million Veteran Program.* 1) Yale School of Public Health, New Haven, CT; 2) Yale School of Medicine, New Haven, CT; 3) Veterans Affairs Connecticut Healthcare System, West Haven, CT; 4) Crescenz Veterans Affairs Medical Center, Philadelphia, PA; 5) University of Louisville School of Nursing, Louisville, KY; 6) VA Clinical Epidemiology Research Center (CERC), VA Connecticut Healthcare System, West Haven, CT; 7) University of Pennsylvania Perelman School of Medicine, Philadelphia, PA.

To better understand genetic architecture of complex diseases, the Million Veteran Program (MVP) is recruiting veterans who volunteer to complete questionnaires, allow access to their electronic health record (EHR), and provide a blood sample for genomic analysis. Here, we report a genome-wide association study (GWAS) on smoking among European-American veterans (N=209,000) from the MVP. We first replicated the association of *CHRNA3-A5-B4* on chromosome 15 with quantitative cigarette consumption (**rs55781567**, p=1.3E-16). We then identified 20 genomic regions significantly associated with smoking status (current, past, or never). Specifically, we replicated 5 previous findings of smoking-associated genomic loci mapped to *CHRNA2*, *CHRNA4*, *DRD2*, *DBH*, and *FAM163B,* and we identified 15 novel genomic regions associated with smoking status that were mapped to *PHACTR4*, *NEGR1*, *LRRIQ3*, *CAMKMT*, *SIX3-AS1*, *TEX41*, *PABPC1P2*, *ZBTB20*, *FAM160A1*, *MAD1L1*, *EPHX2*, *CLU*, *TMPRSS5*, *SPATS2*, and *CYB5B*. One of the lead SNPs was **rs11780471** (p=2.1E-14), reportedly a significant marker of lung cancer that maps to the intergenic region of *CHRNA2* and *EPHX2*. Existing literature suggested that a cis-eQTL of **rs11780471** affects *CHRNA2* expression in brain, and *EPHX2* and its related pathways are involved in xenobiotic metabolism. To understand potential functions of the identified non-coding SNPs, we performed LD score regression analysis, together with GenoSkyline-Plus tissue and cell type-specific functional annotations. The heritability of smoking was significantly enriched in the brain and gastrointestinal tract. The highlighted tissue and cell types included brain anterior caudate, normal human astrocytes primary cells, gastric, and fetal stomach. Finally, we used GNOVA to estimate the genetic correlations between smoking in the MVP and published GWASs of multiple complex phenotypes from LD Hub. After Bonferroni correction, 38 of 207 phenotypes showed significant genetic associations with smoking, including lung cancer (corr=0.44, p=2.9E-31), coronary artery disease (corr=0.31, p=2.3E-20), and depressive symptoms (corr=0.36, p=7.1E-19). Results from the largest GWAS of smoking to date highlight the genetic risks of smoking and smoking-related traits. Extracting phenotypes from an EHR in a large sample enables identification of multiple novel genetic loci contributing to the risk of complex diseases.

25

Discovering genotype specific effects of inducing pluripotency in a mix pool of many donor cells without the use of exogenous DNA barcodes or single-cell sequencing. *Y. Chan[1,2], E.T. Lim[1,2], G.M. Church[1,2].* 1) Harvard Medical School, Boston, MA; 2) Wyss Institute for Biological Inspired Engineering, Boston, MA.

Performing genotype associations on cellular phenotypes for many different donors is a laborious process. One approach to multiplex this process is to screen a mixed pool of cells from many different donors. However, tracking donor identity for each cell currently requires single-cell sequencing or inserting into each donor cell a unique DNA barcode prior to mixing. Current single-cell sequencing technologies is expensive and tedious and can only assay a limited number of cells. While insertion of unique DNA barcodes can be achieved by lentiviral delivery, it is still time-consuming and costly. Primary cells, non-dividing cells and cells with limited number of passages cannot be effectively barcoded. Here, we describe a method that accurately predicts each donor proportion from the pool without requiring exogenous DNA barcodes. Instead, we use single nucleotide polymorphisms (SNPs) as a natural barcode for donor identity. It is difficult to use SNPs because they are usually biallelic and are distributed sparsely throughout the genome. Current commonly used sequencing technologies have read-lengths <1,000 base pairs, making it nearly impossible to ascribe any read to any particular donor. To solve this, our method utilizes genome wide SNP profiles for every donor and whole-genome sequencing from DNA extracted from the mixed pool. Our method employs an Expectation Maximization algorithm that uses an iterative process to discover the proportions that best fit the observed sequencing data given whole genome SNP profiles of every donor. From simulations, we showed that sequencing 500,000 SNPs at 1X coverage is sufficient to accurately predict the individual proportion of a mixed pool of 100 donors but higher coverage (30X) is required when the number of donors is 1000. We further showed that our method can accurately predict the proportion of actual samples from mixed pool of 102 different donors. Finally, we applied our method to test for genotype specific effects of inducing pluripotency by nucleofection of Yamanaka factors into a mixed pool of donor cells from the Harvard Personal Genome Project as well as 1000 Genomes Project. We discovered that some donor cells are significantly more susceptible to reprogramming than others and the difference can be explained by specific genotypes. Our method can be thus be broadly applied to test genotype specific effects for other cellular phenotypes and enables the multiplexed testing of diverse donor cells *en masse*.

Copyright © 2018 American Society of Human Genetics. All rights reserved

26

Targeted full-length transcriptome sequencing to confirm gene models identifies novel isoforms. *E. Tseng[1], G. Sheynkman[2], D. Hill[2], M. Vidal[2].* 1) Pacific Biosciences, Menlo Park, CA; 2) Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School Boston, MA.

Long read, third generation sequencing technologies enable sequencing of full-length transcripts for all expressed isoforms. However, the lower throughput compared to traditional RNA-seq methods makes it difficult to capture low abundance genes and rare isoforms in a cost-effective manner. We developed 'ORF Capture-Seq' (OCS) to gain the highest sensitivity and precision to sequence novel, full-length isoforms. This method capitalizes on the CCSB human ORFeome collection, a resource of 'ready-to-express' open reading frame (ORF) clones for nearly every protein-coding gene in humans. A set of desired gene targets are selected (1 to ~2,000 genes) and the genes' ORFs, in the form of clones in GATEWAY entry plasmids, are used to generate biotinylated PCR fragements across the length of the ORF that are then used in hybridization-based capture to enrich for full-length isoforms. The enriched fraction can then be sequenced on the PacBio sequencing platform. We present a pilot dataset targetting 736 transcription factors (TFs) for isoforms from 8 different human tissues, followed by sequencing on a single Sequel SMRT cell. We recovered a total of 4,372 distinct transcripts, 1,907 of which were on-target, corresponding to 386 of the 736 TFs of interest. As transcription factors are generally low abundance, this allowed an unprecedented insight into alternative splice forms in the TF genes. We also present bioinformatics advances in the Iso-Seq bioinformatics software developed by PacBio. The Iso-Seq pipeline outputs full-length, high-quality, consensus sequences that can be used directly for ORF prediction. The latest version of the Iso-Seq software, Iso-Seq3, is capable of processing 10 Sequel SMRT cells of the UHRR sample, which contains 2 million full-length reads, in under 6 hours. The software obtained ~70k unique isoforms covering 14,000 known genes and > 2,000 novel genes. Saturation studies show that using only half or less of the sequencing depth, 80% of the transcripts could already be recovered. Combined with the OCS probe enrichment method, it is now possible to sequence hundreds of genes at high resolution using only 1 SMRT cell while having bioinformatics answers within matters of hours.

27

Characterization of de novo mutations in an unascertained large clinical cohort. *A. Blumenfeld[1], J. Staples[1], O. Gottesman[1], J.D. Overton[1], J.G. Reid[1], D. Carey[2], M. Murray[2], C. Gonzaga-Jauregui[1], Geisinger-Regeneron DiscovEHR.* 1) Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY; 2) Geisinger Health System, Danville, PA.

De Novo mutations (DNMs) are changes in DNA that occur naturally and spontaneously in the genomes of living organisms during DNA replication. These changes can occur during early embryonic development or in the parental germline, where they can be passed to the offspring. Since these mutational events happen mostly at random and across the genome, the majority of *de novo* events are benign or unlikely to show a phenotype. However, *de novo* mutations in coding regions can alter the normal function of genes and lead to disease.METHODS: We performed whole-exome sequencing in 92,455 participants from the MyCode Precision Medicine Initiative through our Geisinger-Regeneron DiscovEHR collaboration. We calculated identity-by-descent (IBD) estimates between pairs of samples and derived familial 1st degree relationships to identify parent-child trios for analysis of *de novo* mutations. We performed trio analysis and *de novo* variant identification in 2,602 reconstructed parent-child trios through an automated bioinformatics pipeline. We identified all potential coding *de novo* single nucleotide variants (SNVs) and insertion/deletion (InDel) variants and further characterized their sequence context, and potential functional consequences.RESULTS: We identified 5201 *de novo* SNVs and InDels, of which 3,409 were moderate and high-quality DNM calls after visual confirmation of a subset across all quality categories to estimate false-positive rates. Initially, the number of DNMs varied from 0 to 48 per individual, but patients with >10 had a high error rate and were excluded (mean = 1.31; IQR=2.0). The ratio of transitions to transversions was 2:1 (n=2038 to n=1007, respectively), with most being C->T (n=781) and G->A (n=739). Maternal and paternal age were independently correlated with DNMs/individual (paternal: median age=26.6, 0.012 DNMs/yr, $p=6.8*10^{-4}$; maternal: median age=24.8, 0.013 DNMs/yr, $p=7.8*10^{-4}$). Most DNMs were nonsynonymous SNVs (n=1992), followed by synonymous SNVs (n=830), frameshift InDels (n=187), splice site variants (n=152) and stop-gain (n=112). 625 (18.3%) DNMs were predicted to be pathogenic by SIFT, Provean and PolyPhen-2-HumDiv algorithms, and 413 (12.1%) by 2 of the 3 algorithms. CONCLUSIONS: We analyzed *de novo* mutations (DNMs) in coding regions in a clinically unascertained precision medicine cohort. This study broadens our understanding of naturally occurring *de novo* mutations in the general population unbiased by specific ascertained phenotypes.

Copyright © 2018 American Society of Human Genetics. All rights reserved

28

Private information leakage from raw functional genomics data: Theoretical quantifications & practical privacy-aware file formats. *G. Gursoy[1], A. Harmanci[2], M. Green[1], F. Navarro[1], M. Gerstein[1].* 1) Yale University, New Haven, CT; 2) University of Texas Health, Houston, TX.

Functional genomics experiments on human subjects present a privacy conundrum. On one hand, many of the conclusions we infer from these experiments are not tied to the identity of individuals but represent universal statements about biology and disease. On the other hand, by virtue of the experimental procedure, the sequencing reads are tagged with small bits of patients' variant information, which presents privacy challenges in terms of data sharing. There is great desire to share data as broadly as possible. Therefore, measuring the amount of variant information leaked in a variety of experiments, particularly in relation to the amount of sequencing, is a key first step in reducing information leakage and determining an appropriate "set point" for sharing with minimal leakage. To this end, we derived information-theoretic measures for the private information leaked in experiments and developed various file formats to reduce this during sharing. We show that high-depth experiments such as Hi-C provide accurate genotyping that can lead to large privacy leaks. Counterintuitively, low-depth experiments such as ChIP and single-cell RNA sequencing, although not useful for genotyping, can create strong quasi-identifiers for re-identification through linking attacks. We show that partial and incomplete genotypes from many of these experiments can further be combined to construct an individual's complete variant set and identify phenotypes. We provide a proof-of-concept analytic framework, in which the amount of leaked information can be estimated from the depth and breadth of the coverage as well as sequencing biases of a given functional genomics experiment. Finally, as a practical instantiation of our framework, we propose file formats that maximize the potential sharing of data while protecting individuals' sensitive information. Depending on the desired sharing set point, our proposed format can achieve differential trade-offs in the privacy-utility balance. At the highest level of privacy, we mask all the variants leaked from reads, but still can create useable signal profiles that give complete recovery of the original gene expression levels.

29

An atlas of human and murine genetic influences on osteoporosis. *J.A. Morris[1,2], J.P. Kemp[3,4], S.E. Youlten[5], L. Laurent[2], J.G. Logan[6], R. Chai[6], N.A. Vulpescu[7], V. Forgetta[2], A. Kleinman[8], S. Mohanty[5], C.M. Sergio[5], J. Quinn[5], L. Nguyen-Yamamoto[9], A.L. Luco[9], J. Vijay[10], C.L. Gregson[11], N.C. Harvey[12,13], E. Grundberg[10,14], D. Goltzman[9], D.J. Adams[15], C.J. Lelliott[15], D.A. Hinds[8], C.L. Ackert-Bicknell[16], Y-H. Hsu[17], M.T. Maurano[7], P.I. Croucher[5], G.R. Williams[6], J.H.D. Bassett[6], D.M. Evans[3,4], J.B. Richards[1,2,18,19], GEFOS Consortium.* 1) Department of Human Genetics, McGill University, Montréal, Québec, Canada; 2) Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada; 3) University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland, Australia; 4) MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK; 5) Garvan Institute of Medical Research, Sydney, New South Wales, Australia; 6) Molecular Endocrinology Laboratory, Department of Medicine, Imperial College London, London, UK; 7) Institute for Systems Genetics, New York University Langone Medical Center, New York, New York, USA; 8) Department of Research, 23andMe, Mountain View, California, USA; 9) Research Institute of the McGill University Health Centre, Montréal, Québec, Canada; 10) McGill University and Genome Quebec Innovation Centre, Montréal, Québec, Canada; 11) Musculoskeletal Research Unit, Department of Translational Health Sciences, University of Bristol, Bristol, UK; 12) MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, UK; 13) NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust, Tremona Road, Southampton, UK; 14) Children's Mercy Hospitals and Clinics, Kansas City, Missouri, USA; 15) Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK; 16) Center for Musculoskeletal Research, Department of Orthopaedics, University of Rochester, Rochester, New York, USA; 17) Institute for Aging Research, Hebrew SeniorLife, Boston, Massachusetts, USA; 18) Departments of Medicine and Epidemiology, Biostatistics & Occupational Health, McGill University, Montréal, Québec, Canada; 19) Department of Twin Research and Genetic Epidemiology, King's College London, London, UK.

Identifying genetic determinants of osteoporosis will help to highlight drug targets and understand its biological causes. To do so, methods to map associated SNPs to causal genes are needed. We therefore performed GWAS' of estimated bone mineral density measured by quantitative heel ultrasound (eBMD, N=426,824) and of bone fracture (N_{cases}=53,184) in White-British UK Biobank participants, replicating fracture findings in 367,900 cases and 363,919 controls from 23andMe. We empirically tested methods to map SNP associations to causal genes with in-depth functional genomics tools and murine models. We found 1,106 lead SNPs associated with eBMD, mapping to 518 loci (301 novel) and explaining 20% of eBMD variance. For fracture, we found 14 lead SNPs, all associated with eBMD. Statistical fine-mapping and human functional genomics data, featuring osteoblast Hi-C contact domains and open chromatin sites, enabled us to map 2,530 plausibly causal SNPs (97% non-coding), to 551 target genes. The target genes were enriched for osteoporosis drug targets or genes causing Mendelian forms of osteoporosis (odds ratios up to 58, $P=10^{-75}$). We next undertook skeletal phenotyping of 126 target gene knockout mice and found an increased frequency of abnormal skeletal phenotypes (P<0.0001), compared to unselected genes. *DAAM2, CBX1, RGCC, WAC* and *YWHAE* were highlighted as novel genes for osteoporosis. To study *DAAM2* further, we generated additional *Daam2* hypomorphic mice and observed increased cortical porosity and impaired bone quality, leading to decreased bone strength (2.1 standard deviations below control strength). We then introduced CRISPR/Cas9-mediated *DAAM2* edits in osteoblast cell lines and observed a 90% decrease in this crucial bone forming cell's ability to mineralize. In summary, this atlas of human and murine genetic determinants of osteoporosis has increased the number of associated BMD loci from previous studies 2.5-fold to 518, increasing the variance explained by genetic factors 1.5-fold to 20%. We identified methods that mapped associated SNPs to genes strongly enriched for known causal genes. Moreover, genome editing of the *DAAM2* target gene in human bone cells led to decreased mineralization and in mice to reductions in bone strength and increased cortical bone porosity. This set of identified genes greatly improves our understanding of the genetic determinants of osteoporosis, providing methods to identify causal genes underlying SNP associations.

Copyright © 2018 American Society of Human Genetics. All rights reserved

30

Qsox1 is a novel genetic determinant of bone size in mice. *B.M. Al-Barghouthi[1,2], G.M. Calabrese[1], L.D. Mesner[1], K. Nguyen[1], M.L. Bouxsein[3], D. Brooks[3], M.C. Horowitz[4], C.J. Rosen[5], S.M. Tommasini[4], P. Simecek[6], G.A. Churchill[6], C.L. Ackert-Bicknell[7], D. Pomp[8], C.R. Farber[1,2,8].* 1) Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22911; 2) Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22911; 3) Center for Advanced Orthopedic Studies, Beth Israel Deaconess Medical Center, Department of Orthopedic Surgery, Harvard Medical School, Boston, MA 02215; 4) Department of Orthopaedics and Rehabilitation, Yale School of Medicine, New Haven, CT 06520; 5) Maine Medical Center Research Institute, 81 Research Drive, Scarborough, ME 04074; 6) Center for Musculoskeletal Research and Department of Orthopaedics & Rehabilitation, University of Rochester Medical Center, Rochester, NY, 14627; 7) Department of Genetics, University of North Carolina, Chapel Hill, NC 27599; 8) Department of Public Health Sciences, University of Virginia, VA 22911.

Osteoporosis is a complex disease of reduced bone strength and increased risk of fracture. Many characteristics of bone contribute to its strength; however, other than bone mineral density (BMD), which has been interrogated using genome-wide association studies (GWAS) in humans, we know little of the genetics of other strength-related traits such as bone size. To identify novel genetic factors affecting bone size, we used a powerful new murine resource, the Diversity Outbred (DO). The DO is an outbred population derived from eight genetically diverse inbred founder strains. We measured femoral size (femoral length (FL) and medial-lateral (ML)/anterior-posterior (AP) femoral widths) in a cohort of 12-16 week-old DO mice of both sexes (N=602). A significant (LOD=9.5; permutation P<0.05) quantitative trait locus (QTL) affecting ML was identified on Chr1@155Mbp. The QTL mapped to a confidence interval of ~2.8 Mbp and explained 6.6% of the variance in ML. We replicated the ML QTL in an independent cohort of 12 week-old DO mice of both sexes (N=312). To identify the gene(s) responsible, we queried the locus for non-synonymous mutations and expression QTLs (eQTLs). We imputed a complete list of variants from whole genome sequences of DO founder strains within the DO cohort and performed single variant association tests. None of the most significant variants were non-synonymous. We next identified eQTL for all genes in proximity of the ML QTL using tibial RNA-seq profiles from 192 DO mice. Of the 14 genes with eQTL, Quiescin Sulfhydryl Oxidase 1 (*Qsox1*) was the only one in which the ML QTL and eQTL founder haplotype effects were concordant, with the WSB/EiJ founder haplotype being associated with increased ML and decreased *Qsox1* levels. QSOX1 is a secreted catalyst of disulfide bond formation that has not been previously linked to the regulation of bone size. Using CRISPR/Cas9, we generated five *Qsox1* mutant lines with mutations ranging from a 1 bp frameshift to a ~1300 bp deletion encompassing the first exon of *Qsox1*. QSOX1 activity in serum was abolished in mutant mice from all lines. Across the five lines, we observed significantly ($P<4.1 \times 10^{-6}$) increased ML as a function of the number of *Qsox1* mutant alleles. These data identify *Qsox1* as a genetic determinant of bone size and highlight the power of the DO for the genetic analysis of complex traits.

31

***COPB2* loss of function leads to disrupted collagen trafficking and juvenile osteoporosis.** *R. Marom[1], L.C. Burrage[1], M. Jain[1,2], I. Grafe[1], D.A. Scott[1], M. Shinawi[3], J.A. Rosenfeld[1], J.D. Heaney[1], D. Lanza[1], Y.C. Lee[1], I.W. Song[1], J.M. Sliepka[1], D. Batkovskyte[1], Z. Jin[1], B. Dawson[1], S. Chen[1], Y. Chen[1], M.M. Jiang[1], V.R. Sutton[1], C. Kuzawa[4], R. Venditti[5], M.A. Weis[6], A. Clément[7], B. Tremp[7], B. Blanco-Sánchez[7], M. Westerfield[7], D. Eyre[6], C.G. Ambrose[4], M.A. De Matteis[5], B.H. Lee[1], Members of the Undiagnosed Diseases Network.* 1) Department of Molecular and Human Genetics, Baylor College of Medicine , Houston, TX; 2) Department of Bone and OI, Kennedy Krieger Institute, Baltimore, MD; 3) Division of Genetics and Genomic Medicine, Department of Pediatrics, Washington University School of Medicine, St. Louis, MO; 4) Department of Orthopaedic Surgery, UT Health McGovern Medical School, Houston, TX; 5) Telethon Institute of Genetics and Medicine, Pozzuoli (Naples), Italy; 6) Department of Orthopaedics and Sports Medicine, University of Washington, Seattle, WA; 7) Institute of Neuroscience, University of Oregon, Eugene, OR.

Abnormal collagen trafficking has been implicated in a number of skeletal dysplasias. In a 7-year-old female patient with juvenile osteoporosis, we identified a *de novo*, heterozygous loss-of-function (LOF) variant in *COPB2*, a component of the coatomer complex COPI that is involved with membrane trafficking in the Golgi complex and between the ER and the Golgi complex. The patient presented with recurrent fractures and low bone mineral density. RNA studies in the patient's lymphoblast cells showed decreased expression of *COPB2*, consistent with haploinsufficiency. A second patient presenting with a heterozygous, LOF variant in *COPB2* and low bone mineral density was identified via GeneMatcher. To better understand the molecular consequences of *COPB2* haploinsufficiency, we assessed collagen trafficking by immunofluorescence microscopy. In *COPB2* siRNA-treated fibroblasts, we observed delayed collagen trafficking with retention of type I collagen in the ER and Golgi, and altered subcellular distribution of Golgi markers. To characterize the effect of *COPB2* loss of function on skeletal development, we generated a zebrafish model carrying a frameshift variant (p.Lys10Thrfs*12) resulting in an early stop codon in the *copb2* gene. *copb2*-null embryos showed early lethality. Alizarin-red staining of heterozygous larvae at 7 days post-fertilization was reduced, suggesting delayed mineralization. We then generated a mouse model carrying a *Copb2* deletion by CRISPR technology. $Copb2^{-/-}$ mice were embryonic lethal. In the heterozygous female mice, µCT analysis showed a 15-20% reduction in spine BV/TV, trabecular number and trabecular thickness. Biomechanical testing of femurs by 3-point bending demonstrated decreased bone strength, with reduced maximal load, stiffness and rigidity in femurs of $Copb2^{+/-}$ female mice, that was associated with decreased cortical thickness and cross-sectional bone area. Interestingly, in spite of the delay in collagen trafficking observed *in vitro*, collagen post-translational modifications were not altered in bones of $Copb2^{+/-}$ mice. In summary, we identified variants in *COPB2* in 2 unrelated patients with juvenile osteoporosis. Cell studies, zebrafish and mouse models support *COPB2* haploinsufficiency as the mechanism of low bone mass in these patients, via disruption of intracellular collagen trafficking. This study suggests that pathogenic variants in *COPB2* should be considered in the genetic differential diagnosis of early-onset osteoporosis.

Copyright © 2018 American Society of Human Genetics. All rights reserved

32

Variant to gene mapping strategies to maximize gene identification from a GWAS of grip strength in UK Biobank. *D.M. Waterworth[1], D. Rajpal[2], K. Guo[1], J. Freudenberg[2], K.B. Sieber[1], A. Pandey[2], I. Tachmazidou[3], R.A. Scott[3].* 1) Genetics, GlaxoSmithKline, Collegeville, PA; 2) Computational Biology, GlaxoSmithKline, Collegeville, PA; 3) Genetics, GlaxoSmithKline, Stevenage, UK.

Frailty is a significant component of morbidity in old age and a predictor of mortality, although aetiology remains unclear. While many components of frailty are difficult to measure on large scale, grip strength is easily measured, is a proxy for overall strength and strong predictor of morbidity and mortality. Also, it is measured on all UK Biobank participants, allowing large-scale genetic analyses to be performed. We performed a GWAS of maximal grip strength (adjusted for age, sex, BMI and height) in UK Biobank (n=448,861) and identified 208 genome-wide significant loci. Identification of effector genes at each of these loci is a key step in understanding of biological mechanisms and vital to any aspirations of therapeutic translation. We used a series of established and novel variant to gene (V2G) mapping methods along with manual curation. To identify potentially causal sets of SNPs, we performed approximate conditional analyses using GCTA, followed by Bayesian fine mapping. In addition to identifying putative coding variants (n=11), we used the Coloc package to identify GTEx eQTLs that colocalized with the with grip strength signals in relevant tissues. We supplemented these results with those generated using PICCOLO, which is an implementation of Coloc which allowed allows us to use data from a larger number of eQTL and pQTL studies for which we did not have access to genome-wide summary statistics. We identified 96 high confidence (H4>0.8) colocalizations across 77 loci. In 10 loci, the GWAS signal implicated multiple genes. Lastly, we annotated each locus with chromatin states (IDEAS), FANTOM5, and DHS correlations to elucidate the potential gene-regulatory mechanisms. All loci were reviewed for variants which were present in these tissue-specific enhancers and promoters and/or correlated with gene expression. This genomic information along with single gene loci, literature review and lower confidence eQTLs facilitated the identification of a further 72 potential effector genes. We also aggregated multiple (n=23) gene expression studies of muscle atrophy and response to exercise and overlaid those results with the GWAS genes in MetaCore to identify convergent pathways. Overall, we implicated 179 potential effector genes at 164 loci, which are now being taken to high-throughput functional validation, and promise to be a valuable source of putative targets for muscular function.

33

Proactive genetic testing in a primary care setting reveals unexpected results. *J. Gu[1], A. Hazell[1], M. Zarb[1], J. Furnival[1], L. Velsher[1,2].* 1) Medcan, Toronto, Ontario, Canada; 2) North York General Hospital, Toronto, Ontario, Canada.

Advances in genomic research and DNA sequencing technologies have revolutionized our ability to offer large gene panels at reasonable costs, and, as a result, our understanding of the genetic variation associated with human disease. Limited data exist on the true prevalence or penetrance of “rare” actionable genetic conditions in the unselected (“healthy”) population. Initial studies, largely based on the 2013 ACMG recommendations for the return of incidental findings, suggest that 1 to 9% of individuals harbor pathogenic or likely pathogenic mutations in medically actionable genes. Additionally, evidence continues to support the clinical utility of genetic testing in determining appropriate medical management for these “healthy” individuals. Identifying medically actionable mutations can result in early identification of disease and reduced mortality.Medcan, a preventive health and wellness clinic in Toronto, Ontario, serves over 60,000 patients annually. In September 2017, we began offering a proactive 139-gene panel for actionable Mendelian disorders (based on ACMG59 with additional relevant genes) through Invitae. The laboratory reports only pathogenic and likely pathogenic variants; variants of uncertain significance are reported as negative. Clients that elect to undergo our “Proactive Genetic Screening (PGS)” service meet with a genetic counselor for family history review and informed consent, then again to review the results and their implications for the client’s health risks and medical management.Our mission is to integrate genetics into mainstream medicine to proactively prevent and/or minimize the burden of disease. Here we report on our experience with proactive genetic testing and results to date. As of June 1st, 2018, 761 clients have undergone PGS, and we anticipate that this number will surpass 1000 by October 2018. Our overall positive rate is higher than anticipated at 14.9%. Only 30.6% of our positive results would have been identified using the ACMG59 recommendations. In addition, 85.7% of positive clients received management recommendations that differ from their current screening protocol. These data have significant implications for our understanding of the prevalence and penetrance of genetic conditions. Our experience with elective genetic testing in a high volume clinical setting also provides insight into potential challenges, implications, and recommendations for the future of genomic screening in the “healthy” population.

Copyright © 2018 American Society of Human Genetics. All rights reserved

34

The Alabama Genomic Health Initiative. *B.R. Korf[1], G.S. Barsh[2], K.M. Bowling[2], A. Cannon[1], J.J. Cimino[3], G.M. Cooper[2], W.A. Curry[3], K. East[2], J. Edberg[3], M. Fouad[3], A.C. Hurst[1], M. Might[3], S.J. Knight[3], T. May[2], I.P. Moss[1], M. Nakano[4], J.H. Schach[3], B.M. Shaw[1], S.O. Sodeke[5].* 1) Department of Genetics, University of Alabama at Birmingham, Birmingham, AL; 2) HudsonAlpha Institute for Biotechnology, Huntsville, AL; 3) Department of Medicine, University of Alabama at Birmingham, Birmingham, AL; 4) Department of Medical Education, University of Alabama at Birmingham, Birmingham, AL; 5) Bioethics Center, Tuskegee University, Tuskegee, AL.

The Alabama Genomic Health Initiative (AGHI) is funded by the State of Alabama and is intended to engage a diverse group of citizens from all 67 counties of the state. The overall goals are to establish the efficacy of a population-wide program to return results of medically-actionable genomic variants to adult volunteers; to utilize whole genome sequencing to establish diagnoses in children and adults with undiagnosed chronic disorders; and to develop a “genomics ready” community, including both citizens and health providers. Two cohorts are being recruited. The “population cohort” includes adult volunteers not selected for any specific medical condition, with an ultimate goal of 10,000 individuals. DNA from these volunteers is genotyped using the Illumina Global Screening Array. Pathogenic or likely pathogenic variants in any of the 59 genes on the ACMG list of secondary findings are verified in a CLIA-certified laboratory and returned by a genetic counselor free-of-charge. Individuals who receive a positive result are referred to appropriate health providers for care related to their area of risk. The second cohort, referred to as the “affected cohort,”are children and adults with undiagnosed conditions thought to have a genetic basis. These individuals are offered whole genome sequencing, including testing of both parents if available. Funds are available to sequence up to 1,000 individuals over five years (including parents). Enrollment for AGHI commenced in May 2017 following IRB approval. To date, 1,891 individuals have been enrolled in the population cohort, of which 1,644 have been fully analyzed. Pathogenic or likely pathogenic variants were found in 25 participants, for a total of 1.5%. For the affected cohort, 44 families have been enrolled; 37 of these have been analyzed, and 11 harbored a pathogenic, likely pathogenic, or variant of unknown significance that was returned. One received a secondary finding. Participants in both cohorts can opt into allowing their genomic data to be maintained in a database with linkage to questionnaires and electronic health record information, and a sample of their DNA and plasma stored for future research. More than 90% have agreed to participate in the biobank. AGHI is accompanied by a community engagement effort that includes a facilitated deliberative process group approach to assess community expectations and needs, and also a medical education program aimed at health providers in the state.

35

The Integrating Pharmacogenetics in Clinical Care (I-PICC) study: Baseline characteristics of participants in a point-of-care randomized trial. *S. Advani[1], C.A. Brunette[1], S.J. Miller[1], N. Majahalme[1], L. MacMullen[1], C. Hau[1], A.J. Zimolzak[1,2], J.L. Vassy[1,3,4].* 1) VA Boston Healthcare System, Boston, MA; 2) Boston University School of Medicine, Boston, MA; 3) Harvard Medical School, Boston, MA; 4) Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA.

Background: The association between the *SLCO1B1* rs4149056 C allele and simvastatin myopathy is well validated, but the impact of its use in patient care is unknown. The I-PICC Study is a randomized controlled trial (RCT) of preemptive point-of-care *SLCO1B1* genotyping in primary care and women’s health clinics of the VA Boston Healthcare System. **Methods:** Eligible patients for the I-PICC Study are statin-naïve but have elevated cardiovascular disease (CVD) risk by American College of Cardiology/American Heart Association (ACC/AHA) guidelines. Study staff identify eligible patients of enrolled providers through medical record data and obtain telephone informed consent. Consented patients are then enrolled during usual clinical care if and when their primary care provider (PCP) signs an order for *SLCO1B1* genotyping of an extant clinical blood sample. Enrolled patients are randomized to have their PCPs receive the results at baseline (PGx+) or after 12 months (PGx-). The primary outcome is change in LDL cholesterol. Secondary outcomes include concordance with ACC/AHA guidelines for statin therapy (CVD prevention) and Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for simvastatin therapy (drug safety). Exploratory outcomes include patient-reported medication side effects, beliefs about medications, and recall of genetic test results. **Results:** As of April 18, 2018, 772 patients have consented to study participation, and 386 have declined. Of consented patients, 32 later became ineligible, 55 had eligible blood samples for which their PCP did not sign the PGx order, and 8 patients had insufficient samples for enrollment. Of the first 300 of 408 planned enrolled patients (cared for by 39 PCPs), mean age at enrollment was 64 years, 7% were women, and 13% with available race/ethnicity data were non-white. Enrollees were eligible for inclusion because of preexisting CVD (26%), diabetes (28%), and LDL cholesterol ≥190mg/dL (3.3%); the remaining 47% had 10-year CVD risk ≥7.5%. Genotypes have balanced distribution between the arms: 75% vs 67% TT, 22% vs 31% TC, and 2.8% vs 1.9% CC in the PGx+ vs PGx- arms, respectively. **Discussion:** The baseline characteristics of the first 300 I-PICC Study enrollees demonstrate the feasibility and patient and PCP acceptance of a point-of-care RCT of PGx testing in a healthcare system. The I-PICC Study outcomes will inform the clinical utility of preemptive *SLCO1B1* testing in the routine practice of medicine.

Copyright © 2018 American Society of Human Genetics. All rights reserved

36

Prevalence of pathogenic variants in actionable genes for neurological disorders in the Geisinger MyCode Initiative. *V. Abedi[2,3], R. Zand[1,3], T.N. Person[2], A. Sadeghi[1], S.J. Thakur[1], Y. Zhang[4], M.T.M. Lee[2,4], N.E. Andary[1], C. Griessenauer[1], A.M. Michael[5], M.C. Sandulescu[1], N. Holland[1], A. Sarkar[1], C.M. Schirmer[1], N. Martin[1], M.S. Williams[4], D.J. Carey[6].* 1) Department of Neuroscience, Geisinger Health System, Danville, PA, USA; 2) Biomedical and Translational Informatics Institute, Geisinger Health System, Danville, PA, USA; 3) Biocomplexity Institute, Virginia Tech, Blacksburg, VA, USA; 4) Genomic Medicine Institute, Geisinger Health System, Danville, PA, USA; 5) Autism & Developmental Medicine Institute, Geisinger Health System, Lewisburg, PA, USA; 6) Weis Center for Research, Geisinger Health System, Danville, PA, USA.

Background: Precision Health, a patient-centric idea, describes an innovative prevention and treatment approach that is targeted at individual level based on unique characteristics, including genetics and clinical biomarkers. Neurological disorders are a well-suited target for precision health given their prevalence and impact. The goal of this study was to define the prevalence of pathogenic variants in actionable genes for nervous system disorders among patients enrolled in the Geisinger MyCode Community Health Initiative. We also described the variants' characteristics and implications in precision health.Method: We utilized whole exome sequence data from more than 92,000 MyCode participants from Geisinger Health System. All variants were called through the GATK (genome analysis tool kit) best practices pipeline, and filtered with Variant Quality Score Recalibration (VQSR) and a Genotyping Quality (GQ) of 20. The resulting variant call files were then annotated with Ensembl Variant Effect Predictor. The "everything" flag was used to get additional frequency annotations from publicly available databases. Variants that fell into one of two categories, possible loss of function (pLOF) variants or annotated pathogenic/likely pathogenic in the ClinVar database were then selected.Results: Using a set of 76 actionable genes, we identified 22 genes that were related to 13 neurological conditions. In our cohort, we have found 321 variants, corresponding to 225 individuals (prevalence 25/10,000) with a total of 80 different pathogenic/likely pathogenic variants. Our results indicate that the frequency of several pathogenic gene variants in the MyCode population are significantly different from the reported prevalence of the associated diseases in the general population. Hereditary Paraganglioma-Pheochromocytoma syndrome pathological variants were found at a 100-fold higher rate in the MyCode population as compared to the prevalence of the disease in the general population. We also found a rate of 1/31,000 for variants associated with Loeys-Dietz Syndrome. The prevalence of Loeys-Dietz Syndrome in the general population is unknown.Conclusion: Prevalence of pathogenic variants in actionable genes for nervous system disorders in the Geisinger MyCode population is approximately 0.24% or 1/400. We believe that Geisinger is a pioneer and a role-model at leveraging and integrating genomic discoveries into precision health.

37

Genetic and environmental effects disrupt molecular co-regulation. *A.J. Lea[1], M. Subramaniam[2], A. Ko[3], T. Lehtimäki[4], E. Raitoharju[4], M. Kähönen[4], I. Seppälä[4], N. Mononen[4], O. Raitakari[5], M. Ala-Korpela[6,7,8,9,10], P. Pajukanta[3], N.A. Zaitlen[2], J.F. Ayroles[1].* 1) Princeton University, Princeton, NJ; 2) UCSF, San Francisco, CA; 3) UCLA, Los Angeles, CA; 4) University of Tampere, Tampere, Finland; 5) University of Turku, Turku, Finland; 6) Baker Heart and Diabetes Institute, Melbourne, Australia; 7) University of Oulu and Biocenter Oulu, Oulu, Finland; 8) University of Eastern Finland, Kuopio, Finland; 9) University of Bristol, Bristol, UK; 10) Monash University, Melbourne, Australia.

Functionally related genes work together in networks, and often show strong patterns of correlation at the mRNA or protein level that reflect shared, homeostatic regulation. These strong correlations are thought to break down when individuals experience disease, environmental perturbations, or genetic mutations, pointing toward dysregulation of homeostasis. We tested this prediction using genome-wide gene expression and NMR metabolite data from three human cohorts (n=214, 1672, and 2477), combined with a newly developed, flexible approach, to ask whether the degree of correlation between two molecular traits is affected by a predictor variable of interest. First, using mRNA-seq data from monocytes exposed or unexposed to a bacterial compound, we show that simulated bacterial infection causes up to 73% of transcript pairs that are tightly correlated in uninfected cells to become uncorrelated. Second, we use whole blood-derived NMR metabolite data from healthy individuals and those with metabolic syndrome to demonstrate a similar pattern of correlation breakdown: 74% of metabolite pairs lose correlation following the onset of metabolic syndrome. Using longitudinal data, we also show that the degree of dysregulation of specific metabolite pairs predicts whether an individual will develop metabolic syndrome in the future. Finally, leveraging our novel approach, we demonstrate that the degree of correlation between mRNA transcripts is under genetic control. Specifically, we map and replicate 484 SNPs that affect the magnitude of correlation between pairs of mRNA transcripts in whole blood. We show that these 'correlation QTLs' often modulate correlations involving transcription factors or their known target genes. Together, our results support the idea that stressful environmental exposures (here, infection and metabolic syndrome) do not simply lead to mean changes in mRNA transcript or protein levels, but also to changes in correlation structure and a loss of homeostasis. In other words, under stress, genes or metabolites that are typically co-regulated no longer exhibit such tight co-regulation. Further, we show that correlation structure is under genetic control. While we have focused here on mapping genetic and environmental effects on molecular co-regulation, our approach could be paired with many additional data types to address questions of how and why correlations at the molecular or organism-level vary across individuals.

Copyright © 2018 American Society of Human Genetics. All rights reserved

38

Exposome-wide association study (EWAS) identifies link between pregnancy anxiety and various autism spectrum traits. *A. Verma[1], A. Lucas[1], I. Hertz-Picciotto[2], Y. Ludena-Rodriguez[2], RJ. Schmidt[2], MD. Ritchie[1].* 1) Department of Genetics and Institute for Biomedical Informatics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA; 2) Department of Public Health Sciences, University of California, Davis, CA.

Autism spectrum disorder (ASD) is a complex trait that affects social, intellectual, and behavioral development in young children. There are a number of studies showing links between environmental exposures and increased risk in developmental disorders such as autism. We used environmental exposure data from Childhood Autism Risks from Genetics and the Environment (CHARGE), a population-based study of children aged 2-5 years, in three categories: ASD, Developmental Delay (DD), and General Population/ Typical Development (TD). We designed a multi-step study, in which we first performed an exposome-wide association study (EWAS) to identify predictive environmental factors with ASD. Then, we tested the top exposure variable associated with ASD for associations with 56 clinical assessment variables and 11 child measurement variables recorded on CHARGE participants to explain its link with the exposure further. Using an EWAS approach, we evaluated 874 demographic, maternal health, maternal lifestyle, chemical exposure, residential, and other pregnancy variables from 1286 participants (778 ASD, 508 TD) using logistic regression, adjusting for sex, age, and race. We found six exposure variables that met the Bonferroni significance threshold (α=0.05). The association with maternal nervousness or anxiety during pregnancy (maternal health) was the most significant (p-value = 4.05×10^{-6}, OR= 2.21 [1.56, 3.14]). A variety of previous studies have demonstrated that prenatal anxiety or stress is linked to developmental disorders and autism spectrum traits. Subsequently, we found 18 clinical assessment variables significantly associated with maternal nervousness or anxiety using a multi-phenotype approach similar to what is used in phenome-wide association studies, with “Irritability” in children being the most significant association with p-value = 1.82×10^{-8} and beta = 4.20. Other clinical assessment variables significantly associated with pregnancy nervousness or anxiety were hyperactivity, inattention, stereotypy, lethargy, and novel associations with several developmental quotient variables. These findings highlight the potential links between maternal health during pregnancy and developmental delay in children. In the future, we will investigate the combined effect of these exposure variables on autism spectrum traits.

39

Do gene-environment interactions matter in multifactorial diseases? *V. Laville[1], Y.J. Sung[2], T.W. Winkler[3], M. Province[4], K. Rice[5], S. Kardia[6], J. Gauderman[7], D.C. Rao[2], H. Aschard[1], CHARGE Gene-Lifestyle Interactions Working Group.* 1) Institut Pasteur, Paris, France; 2) Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA; 3) Department of Genetic Epidemiology, University of Regensburg, Regensburg, D-93051, Germany; 4) Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO, USA; 5) Department of Biostatistics, University of Washington, Seattle, WA, USA; 6) School of Public Health, University of Michigan, Ann Arbor, MI, USA; 7) Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA.

Compared to marginal genetic and environmental effects, the importance of gene-by-environment (GxE) interactions in human multifactorial phenotypes is often considered limited. For risk prediction in the general population, this view may be reasonable. Indeed, in many realistic scenarios, marginal additive effects can capture a substantial proportion of the phenotypic variance due to interaction effects. However, this does not rule out the potential importance of interaction effects from a biological perspective, and experimental data and model organism studies suggest that GxE interactions may be critical in the etiology of complex phenotypes. A central question to move the field forward is to determine whether the genetic mechanisms underlying multifactorial phenotypes in humans differ depending on the individual’s environment. Here we argue that this question can be partly answered by combining some recent methodologies. We first implement a method we recently developed to estimate the relative contribution of interaction effects to phenotypic variance. Second, for the best candidates identified at the previous step, we assess whether the cell-types involved in the genetic effect tend to differ across the exposure range. For the latter, we used the recently published stratified LD score regression (Finucane et al. 2018), which allows testing for disease heritability enrichment in genes with high specific-tissue expression based on the GTEX data.We applied our approach in gene-by-smoking interactions of diastolic (DBP) and systolic (SBP) blood pressure performed in 70,000 individuals from the CHARGE consortium. We found differences in heritability of SBP between current smokers and non-current smokers (h2 = 0.09 and h2 = 0.14 respectively, Phet = 0.19). Partitioning SBP heritability overall we first found enrichment for variants expressed in cardiovascular cell types (median P-value for enrichment = 0.24 across 14 cardiovascular tissues). When stratified by current smoking status, we observed differential enrichment for several cell-types, in particular in blood-immune tissues in smokers (median P = 0.22 across 40 blood-immune tissues) but not in non-smokers (median P = 0.52). This is in agreement with recent gene expression studies showing that smoking□related genes are enriched for immune system and blood coagulation, thus confirming the ability of our approach to unravel potential exposure-specific biological processes.

Copyright © 2018 American Society of Human Genetics. All rights reserved

40

Dynamic human environmental exposome revealed by longitudinal personal monitoring. *C. Jiang, X. Wang, X. Li, J. Inlora, T. Wang, Q. Liu, M. Snyder.* Department of Genetics, Stanford University, Palo Alto, CA.

Background Human health is greatly impacted by genetics, environmental exposure, and lifestyle. In recent years, significant effort has been dedicated to understanding how genetics and lifestyle can influence our health. However, our understanding of the human environmental exposures, especially at the personal level, is quite limited. Information about environmental exposures, both biotics (e.g. pollens, microbes, and microbes) and abiotics (e.g. chemicals) can be important for understanding and monitoring numerous diseases such as respiratory diseases, allergy and asthma, chronic inflammatory diseases, and even cancer. **Methods** We have developed a novel highly sensitive method to monitor personal airborne biological and chemical exposures (collectively referred to as the environmental exposome) longitudinally by integrating a wearable device and multiple-omics measurements. We applied this method to track 15 different individuals spatial-temporally, among which three people were tracked up to 890 days and 201 time points, to provide an extensive personal profiling of the environmental exposome. **Results** We demonstrated that individuals are potentially exposed to thousands of pan-domain species and thousands of chemical compounds, including insecticides and carcinogens. In aggregate, over 2500 species were identified with great intraspecies diversity. We found that personal biological and chemical exposomes are highly dynamic and vary spatial-temporally, even for individuals located in the same general geographical region. We were able to construct a season-predictive model based on the pan-domain genera profile. Integrated analysis of biological and chemical exposomes revealed strong location-dependent relationships. Finally, we built an exposome interaction network and demonstrated the presence of distinct yet interconnected human- and environment-centric clouds, depicting extensive inter-species relationships derived from various interacting ecosystems such as human, flora, pets and arthropods. **Conclusions** Overall, we describe a method to capture and analyze personal environmental exposures, and demonstrate that human exposomes are diverse, dynamic, spatiotemporally-driven interaction networks that have the potential to impact human health. Both the data and approach are expected to be of general value to a broad spectrum of scientific fields, such as public health, biotechnology, microbiome, environmental science, evolution, and ecology.

41

Yeast surrogate genetic approaches for functional predictions of genetic variants related to inborn errors of metabolism. *A. Sirr[1], A. Scott[1], G. Cromie[1], M. Heyesus[1], A. El-Hattab[2], F. Alkuraya[3], A. Dudley[1].* 1) Pacific Northwest Research Institute, Seattle, WA; 2) Division of Clinical Genetics and Metabolic Disorders, Pediatrics Department, Tawam Hospital, Al-Ain, United Arab Emirates; 3) Department of Genetics, King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia.

Many genes in the human genome are currently “actionable”, i.e. detection of a pathogenic variant can inform medical management. However, for the vast majority of polymorphisms in human disease genes, there is not enough evidence to determine whether the sequence change is pathogenic or benign. Such variants of uncertain significance (VUS) cannot be used to diagnose disease or inform treatment. We have developed a high throughput assay for measuring the functional impact of genetic variation in human protein coding sequences using the model organism *Saccharomyces cerevisiae*. While the approach can be adapted for any human gene that is able to functionally replace its yeast ortholog, our initial applications are focused on inborn errors of metabolism (IEMs). These diseases are often actionable via dietary restriction or supplementation. Further, IEMs are generally monogenic loss of function mutations in genes encoding core metabolic enzymes that are often highly conserved between yeast and humans. Here we focus on a rare, but highly actionable IEM, Serine Deficiency Syndrome/ Neu-Laxova Syndrome (NLS). Loss of function mutations in any of the genes in the phosphorylated serine biosynthesis pathway (*PHGDH, PSAT1*, and *PSPH*) cause severe neurological symptoms (e.g. congenital microcephaly, intractable seizures, and severe psychomotor retardation), which may be reduced and even eliminated with genotype-guided diagnosis and prenatal serine supplementation. Using our yeast complementation assay, we have phenotyped all known disease-causing alleles of the second enzyme in the pathway, *PSAT1*. We have assessed the functionality of each allele in isolation (i.e. as haploids) and in combination with other alleles (i.e. as heterozygotes), recreating both the disease carrier parental genotypes as well as the patient genotype to establish a model of serine deficiency in yeast. Our results show a range of levels of enzymatic dysfunction across the panel of human disease alleles tested which generally mirror the severity of symptoms ascribed to the individual patients. With a ‘disease-causing’ threshold and range established, we are currently expanding our phenotyping to all known alleles of *PSAT1*.

Copyright © 2018 American Society of Human Genetics. All rights reserved

42

Variant interpretation practice amongst clinical genetic counselors: Assessment of training and resource needs to support clinical practice. *K.E. Wain[1], D.R. Azzariti[2], J. Goldstein[3], A. Knight Johnson[4], P. Krautscheid[5], J.M. O'Daniel[3], D. Ritter[6], J.M. Savatt[1], C.L. Martin[1], E.R. Riggs[1] on behalf of the ClinGen Education Working Group.* 1) Geisinger, Lewisburg, PA; 2) Laboratory for Molecular Medicine, Cambridge, MA; 3) University of North Carolina at Chapel Hill, Chapel Hill, NC; 4) University of Chicago, Chicago, IL; 5) ARUP Laboratories, Salt Lake City, UT; 6) Baylor College of Medicine, Houston, TX.

Broad-scale genomic testing is now incorporated into clinical care across specialties. Genetic counselors (GCs) are increasingly performing post-test variant interpretation activities, such as independently assessing evidence for the interpretation of reported variants to inform patient counseling; however, training and resource needs to support such activities are not well defined. The ClinGen Education Working Group administered an online survey to National Society of Genetic Counselors (NSGC) members providing part- or full-time clinical counseling to assess current variant interpretation practices and determine training and resource needs. Respondents (n=239) represented all major clinical specialties and demographics were generally consistent with the 2018 NSGC Professional Status Survey. The majority (68.3%) performed interpretation activities for most reported variants. The most commonly used resources were ClinVar (used “often” by 73%), population databases to obtain allele frequency (24%), and published sequence interpretation guidelines (20%). Other resources, such as genome browsers, curated databases, and structural variant resources, were associated with lower comfort and less common use. Application of variant interpretation knowledge was assessed via three case vignettes, in which respondents indicated if specific variant details or evidence examples would support a benign, neutral, or pathogenic interpretation. Results indicated a general understanding of how to apply key evidence types (population data/allele frequency, *de novo*/segregation, published cases) based on overall concordance, low self-reported uncertainty, and author expert review. Self-reported uncertainty was greatest for assessing gene-level data from population databases (e.g., constraint metrics, etc.) and the implications of a specific variant type (e.g., missense vs. splice variant). Lack of time was the most commonly reported barrier to completing interpretation activities (73.9%). GCs reported that improved laboratory reports, increased access to literature and licensed data/tools, and more laboratory submissions to a single public database, such as ClinVar, would help support their practice. Clinical GCs are actively assessing evidence relevant to variant interpretation for their patients. These results highlight areas for additional training and resources to support variant interpretation practice by clinical GCs.

43

Variants unmasked by recurrent deletions modify disease presentations of genomic disorders. *B. Yuan[1,2], W. Bi[1,2], N.A. Batzir[2], S. Gu[1,2], W. Zhu[1,2], F. Bocanegra[3], C. Fong[4], J. Holder[5], J. Nguyen[6], J. Zhang[1,2], C. Shaw[1,2], C. Schaaf[2,7], C. Eng[1,2], Y. Yang[1,2], P. Liu[1,2].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Baylor Genetics, Houston, TX; 3) Instituto de Referencia Andino, Bogotá, Colombia; 4) Department of Pediatrics, University of Rochester Medical Center, Rochester, NY; 5) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 6) Department of Pediatrics, University of Texas Health Science Center, Houston, TX; 7) Institute of Human Genetics, University Hospital Cologne, Cologne, Germany.

Background: Incomplete penetrance and variable expressivity of genomic disorders have been revealed by extensive phenotype-genotype correlation studies. The disease presentation of a genomic disorder resulting from a copy number deletion largely depends on the genomic span of the deletion. On the other hand, a deletion may unmask a deleterious rare variant at a recessive locus, which could modify the disease severity or add new phenotypic features to the spectrum. Mediated by nonallelic homologous recombination between low copy repeats, some genomic disorder associated deletions can recur in unrelated patients with the identical span and genomic content. Selection of patients sharing common deletions can constitute unique patient cohorts sharing homogenized genomic profiles for comparison. We hypothesize that the phenotypic spectrum of individuals carrying recurrent deletions may be modified by variants on the second allele unmasked by the deletion *in trans*. **Methods:** We used an in-house developed recurrent deletion map to identify recurrent deletions from the Baylor Genetics clinical exome sequencing patient cohort (N>10,000), and characterized detailed phenotypes of the patients with these deletions. We then investigated the sequence variants that may contribute to the atypical clinical features of these patients. **Results:** We identified 116 patients harboring 15 recurrent deletions characterized with incomplete penetrance/variable expressivity. Among those, exome sequencing identified a molecular diagnosis outside of the deleted region in 35 patients. Focusing on the rare variants uncovered by the recurrent deletions in the 116 patients, we identified hemizygous variants in autosomal recessive disease genes (*COX10*, *ERCC6* and *NDE1*) that potentially contributed to the atypical clinical presentation of the deletion. We also identified hemizygous variants in genes (*PRRT2* and *OTUD7A*) as candidate disease modifiers, which were proposed to cause full penetrance or a more severe disease phenotype with the presence of deletion. **Conclusion:** Our study demonstrated the contribution of the unmasked rare deleterious variants to the penetrance or increased disease severity of recurrent deletions. Elucidating the role of genetic variants within such recurrent genomic disorder regions could provide a precise molecular answer with regards to the specific disease phenotypes, which will be useful in guiding clinical decisions.

Copyright © 2018 American Society of Human Genetics. All rights reserved

44

GeneMatcher: Analysis of five years of experience. *N. Sobreira[1], E. Wohler[1], F. Schiettecatte[2], R. Martin[1], Z. Akdemir[3], S. Jhangiani[3], J. Posey[3], J. Lupski[3], D. Valle[1], A. Hamosh[1].* 1) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) FS Consulting LLC, Seattle, WA; 3) Department of Molecular and Human Genetics, Baylor College of Medicine Houston, TX.

We created GeneMatcher 5 years ago to facilitate communication and improve the search for patients with pathogenic variants in novel candidate genes. Here we analyze the data in GeneMatcher over the last 5 years aiming to understand better how to identify characteristics of the genes associated to disease phenotype. In GeneMatcher, matches can be made based on gene name, genomic location, OMIM® number, and on phenotypic features. GeneMatcher is a founding member of the Matchmaker Exchange and is linked to PhenomeCentral, DECIPHER, MyGene2, *matchbox* and Australian Genomics Health Alliance. As of 1 June 2018, there have been 22,455 submissions to GeneMatcher of 9,138 genes by 5,372 submitters (4,055 researchers, 2,805 health care providers,162 patients) from 76 countries. For 1,094 submitters (20.4%) no matches have occurred. 749 (14%) submitted only one submission. Most of the users come from US (41%), followed by France, Netherlands, UK, Germany and Canada. GeneMatcher includes 93 submissions (156 genes) with data related to model organisms, 103 genes related to mouse models. Of the total 9,138 submitted genes, 5,738 are genes not associated with an OMIM phenotype. Currently, 4,683 (51%) have no matches while 4,455 (49%) genes have matched at least one time. Of the 4,683 genes that have not matched, 2,767 have no OMIM phenotype. Of the matched genes, 289 genes have had more than 10 matches. We will investigate features of both matched and unmatched genes including gnomAD pLI and missense z scores, RVIS scores and gene ontology terms in these 2 different groups to better understand the characteristics of genes that match and those that do not and their relationship to disease phenotype. Finally, the matches in GeneMatcher have supported more than 91 publications describing at least 64 novel disease genes. Thirty-seven of them causing autosomal dominant phenotypes, 22 causing autosomal recessive phenotypes, 1 causing a X-linked recessive phenotype, 1 causing an imprinted phenotype and 3 causing phenotypes that could be autosomal dominant or autosomal recessive. An analysis of these 64 genes showed that they usually matched four times, while genes with more than 10 matches were rarely published. We expect that the analysis of these data will enable better prioritization of candidate genes to increase the specificity and speed of the gene discovery rate.

45

ClinGen allele and evidence registries catalyze the emergence of an open ecosystem of variant data and knowledge. *R.Y. Patel[1], P. Pawliczek[1], L. Babb[8], A.R. Jackson[1], S. Paithankar[1], L.R. Ashmore[1], C. Bizon[2], T. Nelson[3], B. Powell[4], R. Freimuth[5], N. Shah[1], M.W. Wright[6], S. Dwight[6], J. Zhen[6], P. McGarvey[7], H.L. Rehm[9], C. Bustamante[6], S.E. Plon[1,10], A. Milosavljevic[1].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine Houston, TX; 2) Renaissance Computing Institute, University of North Carolina, Chapel Hill, NC; 3) Geisinger Autism and Developmental Medicine, Lewisburg, PA 17837; 4) University of North Carolina, Department of Genetics, Chapel Hill, North Carolin; 5) Department of Health Sciences Research, Mayo Clinic, Rochester, MN; 6) Stanford University School Of Medicine, Palo Alto, CA; 7) Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC; 8) Sunquest Information Systems Company, Boston, MA; 9) Laboratory for Molecular Medicine at the Partners HealthCare Center for Personalized Medicine; 10) Department of Pediatrics, Baylor College of Medicine, Houston, TX.

Recent past has witnessed a rapid increase in the number of known human variants and associated annotations. However, it is increasingly difficult to aggregate all the information required for research and clinical applications in a single database. This problem calls for an easily accessible distributed infrastructure and web-interfaces for aggregation and exchange of variant information that comes from many sources and participants. To seed the emergence of an open linked data ecosystem that will help integrate and harness variant information, we developed the ClinGen Allele Registry (CAR) and Evidence Registry (ER). These resources are accessible via user-friendly web interface and well-documented APIs. The CAR provides readily available globally unique variant identifiers that enable aggregation of information from different sources. A core element of the CAR is a canonicalization service that groups variants denoting the same nucleotide or amino acid changes and assigns a unique/dereferenceable canonical identifier (CAid). The canonicalization service is implemented using a highly optimized in-memory sequence alignment-based index that hosts over 500,000 known reference sequences. More than 650 million distinct variants are currently registered, including those from key resources such as gnomAD, dbSNP, ClinVar and a smaller number registered by users. The ER complements the CAR by providing a means for sharing evidence and interpretations about a variant's pathogenicity using ACMG/AMP or similar guidelines. The ER currently hosts over 5,000 ACMG guideline based variant interpretations imported from multiple sources. For interoperability with curation interfaces and other components (for import from or export to the ER), the interpretations are represented using the ClinGen SEPIO (semantically interoperable variant interpretation model). We demonstrate the utility of CAR and ER by 1) showing how the variant information linked by the CAR provides raw material for reasoning about pathogenicity; 2) demonstrating API-based integration of the ER through the ClinGen Pathogenicity Calculator, where ACMG/AMP-style variant interpretations are communicated from ER by a few clicks; and 3) analyzing over 5,000 ACMG guideline based variant interpretations currently present in the ER.Availability:Allele Registry: https://reg.clinicalgenome.orgEvidence Registry: https://ereg.clinicalgenome.orgPathogenicity Calculator: https://calculator.clinicalgenome.org.

Copyright © 2018 American Society of Human Genetics. All rights reserved

46

A rigorous interlaboratory examination of the need to confirm NGS-detected variants in clinical genetic testing. *S. Lincoln[1], R. Truty[1], C. Lin[2,3,7,8], J. Zook[4], J. Paul[1], V. Ramey[1], M. Salit[4,5], H. Rehm[2,3,6,7,8], R. Nussbaum[1,9], M. Lebo[2,3,7,8].* 1) Invitae, San Francisco, CA; 2) Partners HealthCare Laboratory for Molecular Medicine, Cambridge, MA; 3) Brigham and Women's Hospital, Boston, MA; 4) National Institute of Standards and Technology, Gaithersburg, MD; 5) Joint Initiative for Metrology in Biology, Stanford, CA; 6) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 7) Harvard Medical School, Boston, MA; 8) The Broad Institute of MIT and Harvard, Cambridge, MA; 9) Department of Medicine, University of California San Francisco.

Background: The use of orthogonal assays (e.g., Sanger sequencing) to confirm variants identified by next-generation sequencing (NGS) is standard practice in many laboratories to reduce the risk of delivering false positive results. In clinical genetic testing, confirmation may be particularly important for variants that can suggest irreversible interventions or substantial treatment changes for the patient. Moreover, because clinical NGS methods often emphasize sensitivity (to avoid missing clinically important variants), FP rates can be elevated compared with those in research NGS. Published clinical studies have examined this issue, concluding that confirmation of the highest quality NGS calls may not be necessary. However, these studies are generally small, omit statistical justification, and explore limited aspects of the underlying data. The rigorous definition of criteria that separate high-accuracy NGS calls that do not benefit from confirmation from those of intermediate quality that do remains a critical and pressing issue.**Methods**: Five reference samples and over 80,000 patient specimens from two clinical genetics laboratories were analyzed. Quality metrics for almost 200,000 variant calls with orthogonal data, including 1684 false-positives, were examined. **Results**: A novel classification algorithm used these data to identify a suite of criteria that flag 100% of false positives as requiring confirmation (CI lower bound: 98.5–99.8%, depending on variant type) while minimizing the number of flagged true positives. These criteria identify false positives that the currently published criteria miss. Sampling analysis demonstrated that substantially smaller datasets would result in less effective criteria. **Discussion**: Although we found limitations with the currently published approaches, our large, multi-laboratory study reaffirms prior findings that high accuracy variant calls can be separated from those of intermediate confidence. Our methodology for rigorously determining test and laboratory-specific criteria can be generalized into a practical approach that many laboratories could use to reduce the cost and time burden of confirmation without impacting clinical accuracy.

47

Integrated germline and somatic analysis identifies clinically actionable cancer predisposing pathogenic germline variants in patients with lung cancers. *S. Mukherjee[1], P. Srinivasan[2], C. Bandlamudi[3], Y. Kemel[1], A. Zehir[2], D.L. Mandelker[2], M. Walsh[1], M. Zauderer[1], M.D. Hellmann[1], M.E. Selvan[4], Z.H. Gümüs[4], S.M. Lipkin[5], M. Ladanyi[2], D.B. Solit[3,2], M. Robson[1], L. Zhang[2], J. Vijai[1], D. Jones[1], C. Rudin[1], B.S Taylor[3], Z.K. Stadler[1], M.F. Berger[2], K. Offit[1].* 1) Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY; 2) Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY; 3) Marie-Josée and Henry R. Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, New York; 4) Icahn School of Medicine at Mount Sinai, New York, NY; 5) Department of Medicine, Weill Cornell Medicine, New York, New York.

Most of the heritability for lung cancer, estimated at 18% in population studies, remains unexplained. We determined the prevalence of clinically actionable germline cancer-susceptibility variants in patients with lung cancer using paired tumor-normal next-generation sequencing. Susceptibility variants were correlated with clinical data, somatic mutation profile, and mutational signature characteristics to understand the interplay between germline and somatic mutations. Patients (pts) with lung cancer who underwent sequencing analysis for 468 genes by MSK IMPACT between January 2014 and May 2016 were analyzed. Fully anonymized germline and somatic mutation data were generated using a secure hash algorithm. The pathogenicity of germline variants was determined according to American College of Medical Genetics (ACMG) criteria. Frequencies of pathogenic/likely pathogenic (P/LP) germline variants were compared to population frequencies in genome aggregation database (gnomAD) as control. The purity and ploidy of each tumor specimen were inferred using FACETS algorithm. Somatic mutational signature decomposition was performed in samples with >=15 mutations. Among 2687 lung cancer pts (median age 68), 210 (8%) pts harbored P/LP germline variants in 27 cancer predisposition genes. The frequency of P/LP variants was similar in squamous cell carcinoma (8.5%), lung adenocarcinoma (8%), and small cell carcinoma (7.5%). 93 (3.5%) pts carried pathogenic germline variants in high or moderate penetrant genes known with cancer susceptibility (5 *TP53*, 6 *ATM*, 1 *EGFR*,19 *BRCA2*, 16 *BRCA1*, 9 *FH,*25 *CHEK2*, 1 *PALB2*, 2 *RAD51C*, 2 *RAD51D,* 2 *PMS2*, 1 *BRIP1*, 6 *NBN*). 50% of the P/LP variants in *CHEK2, APC, MUTYH, BRCA1* were population “founder” mutations. Pathogenic variants in *ATM, BRCA2, BRCA1, NBN, TP53* and*FANCC* were significantly enriched in pts with lung adenocarcinoma compared to gnomAD controls. 15% pts with germline variants had loss of heterozygosity (LOH) of the wild-type allele in tumor, with highest frequency of LOH in high and moderate penetrant genes (e.g. 4/5 (80%) TP53 variants, 9 /19 (47%) BRCA2 variants). Mutation signature attributed to BRCA1/2 dysfuction was observed in 4.4% pts and was more common in squamous cell compared to lung adenocarcinoma (OR=2.8, CI=1.3-5.7, P=0.006). In conclusion, we identified 8% prevalence of clinically actionable germline mutations in patients with lung cancer and provide evidence for the biological role of BRCA2 in lung cancer etiology.

Copyright © 2018 American Society of Human Genetics. All rights reserved

48

Assessing causality of pathogenic and likely pathogenic germline variants by integrating somatic and germline sequencing in children with cancer enrolled on the "Genomes for Kids" (G4K) sequencing study at St. Jude Children's Research Hospital. *C. Kesserwan[1], K. Hamilton[1], S. Newman[2], E. Quinn[1], R. McGee[1], R. Nuccio[1], S. Hines-Dowell[1], L. Harrison[1], S. Brady[2], M. Rusch[2], J. Nakitandwe[3], JM. Valdez[1], A. Ouma[1], E. Gerhardt[1], L. Taylor[1], S. Foy[2], A. Silkov[2], A. Patel[2], M. Edmonson[2], D. Hedges[3], S. Shurtleff[3], E. Azzato[3], DW. Ellison[3], J. Downing[3], J. Zhang[2], K. Nichols[1].* 1) Department of Oncology, Division of Cancer Predisposition, St. Jude Children's Research Hospital, Memphis, TN; 2) Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN; 3) Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN.

Introduction Germline (GL) panel testing for mutations in cancer predisposition genes (CPG) is increasingly offered for children with cancer. Given this multi-gene testing approach, when pathogenic (P) or likely pathogenic (LP) variants are found in genes not typically linked with the child's cancer, it is often challenging to determine whether the variant is actually causal for the child's cancer or simply an incidental finding. **Methods** WGS, WES and RNA-seq were done on paired tumor-germline samples from 300 children with cancer. GL variants were classified per ACMG guidelines. To determine whether P/LP variants were likely incidental (LI) or likely causal (LC), we examined tumor data for: 1) loss of heterozygosity (LOH), 2) aberrant RNA splicing, 3) variant affecting a known hotspot, 4) high mutation burden and 5) mutational signature consistent with homologous recombination defects (HRD). P/LP variants in known driver genes of a tumor type were automatically considered LC. P/LP variants in genes not linked to tumor type were sub-classified as: 1) LC if tumor data supported causality, 2) LI in the absence of supportive evidence, 3) Uncertain causality (UC) if tumor data is lacking or inconclusive. **Results** We identified **51 P/LP** variants among 156 CPG in 47 patients (16%). **20 P/LP** variants from 18 patients were in known driver genes for tumor type. Tumor data were available on 10/18 and all showed evidence of likely causality including: 1) LOH in *RB1* (n=4), *APC, NF1* and *PTCH1*, 2) a hypermutator phenotype in biallelic *PMS2* mutations (n=2), and 3) a hotspot in *TP53.* **31 P/LP** variants in 29 patients were in genes not linked to tumor type. Tumor data were available for 24 patients who harbored 26 P/LP variants. 11 P/LP variants from 10 patients were classified as LC and include *NF1* in ALL, *APC* in craniopharyngioma, *SLX4* in glioma, *BAP1* in Ewing sarcoma, *MSH2* & *ATM* in glioblastoma, *SMARCA4* in neuroblastoma, *FANCM* in germinoma and *RECQL4* in spindle cell sarcoma, osteosarcoma and ALL. Tumor data from the 3 patients with *RECQL4* variants, showed an HRD signature (n=2) and LOH (n=1). The remaining P/LP variants were classified as LI (n=10) and of UC (n=5). **Conclusion** By integrating tumor data, we are able to determine causality for 42% of P/LP variants in genes not linked to tumor type. Our study highlights the importance of integrating tumor data to elucidate the role of GL variants in tumorigenesis and define the spectrum of cancers associated with specific syndromes. .

49

Beyond FISH and karyotype: Complex genomic rearrangements uncovered by clinical mate-pair sequencing in B-lymphoblastic leukemia/lymphoma. *C.J. Zepeda Mendoza[1], S.A. Smoley[1], S.H. Johnson[2], J.B. Smadbeck[2], L.B. Baughn[1], P.T. Greipp[1], G. Vasmatzis[3], N.L. Hoppman[1], R.P. Ketterling[1].* 1) Department of Laboratory Medicine and Pathology. Mayo Clinic, Rochester, MN. 55905, USA; 2) Center for Individualized Medicine. Mayo Clinic, Rochester, MN. 55905, USA; 3) Department of Molecular Medicine. Mayo Clinic, Rochester, MN. 55905, USA.

B-lymphoblastic leukemia/lymphoma (B-ALL) is a genetically heterogeneous hematologic malignancy characterized by recurrent acquired chromosome abnormalities. Accurate identification of such chromosome abnormalities is essential to the diagnosis, prognosis, and predictive therapeutic targets of patients with B-ALL, and is typically carried out by standard cytogenetic approaches such as fluorescence *in situ* hybridization (FISH) and chromosome karyotyping. With the introduction of clinical next-generation sequencing strategies such as mate-pair sequencing (MPseq), it is now possible to detect chromosome rearrangements with a high sensitivity and uncover additional genomic complexity in human congenital disorders and cancer. In this study, we characterized with MPseq the genomic profiles of 16 B-ALL individuals with normal, complex, and/or inconclusive FISH and karyotype results. Clinical MPseq refined 61% of karyotype and 37% of FISH calls, further defining translocation partners in abnormalities such as t(9;14)[*JAK2/C14orf93*], t(5;12)[*IL3/ETV6*], t(8;22)[*FGFR1/BCR*], among others. An additional layer of complexity was revealed, which included complex translocations like t(1;3;4), t(1;8;1;6), t(4;14;9;8), and a 9p chromothripsis event associated with a *JAK2* deletion that was interpreted as an insertion/inversion by FISH. In total, 136 predicted gene fusions were identified from the 16 individuals, including 25 previously reported fusions, 76 instances with at least one partner known to form fusions, and 38 fusions which had not been previously reported. Importantly, at least eight of these fusions are amenable to targeted therapeutic treatment, including *TPR/FGFR1*, *BCR/FGFR1*, *EBF1/JAK2*, among others. As expected, while MPseq provided a comprehensive molecular characterization of gene and rearrangement identities, it was out-performed by karyotype in the detection of iso, iso dicentric, and pseudo dicentric chromosomes, and by FISH in the detection of clonal abnormalities with a frequency of less than 10%. Overall, MPseq results from these 16 B-ALL individuals highlight the utility of using a highly-sensitive clinical molecular test to aid standard karyotype and FISH approaches in the characterization of B-ALL chromosome abnormalities. These molecular studies will not only improve our understanding of the manifold genetic causes of hematological malignancies, but will further refine individual prognosis and disease management of these complex leukemic patients.

Copyright © 2018 American Society of Human Genetics. All rights reserved

50

Pan-cancer analyses of germline and somatic mitochondrial DNA mutations in pediatric malignancies. *X. Gai[1,2], P. Triska[1,2], K. Kaneva[1,3], D. Merkurjev[1,2], M.J. Falk[4,5], J.A. Biegel[1,2].* 1) Center for Personalized Medicine, Department of Pathology & Laboratory Medicine, Children's Hospital Los Angeles, Los Angeles, California, USA; 2) Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, California, USA; 3) Division of Hematology, Oncology, and Blood and Marrow Transplant Program, Children's Center for Cancer and Blood Diseases, Department of Pediatrics, Children's Hospital Los Angeles, Los Angeles, California, USA; 4) University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA;; 5) Division of Human Genetics, Department of Pediatrics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.

Little is known regarding the spectrum of mitochondrial DNA (mtDNA) mutations across pediatric malignancies. We analyzed matched tumor and normal whole genome sequencing (WGS) data from 616 pediatric patients with hematopoietic malignancies, solid tumors, and brain tumors of 23 subtypes. The WGS data provided very high coverage across the mtDNA genome in both the tumors (~7,130±5,800X) and the matched normal tissues (~4,100±3,000X). As controls, we sequenced the mtDNA genomes of blood samples from 249 children with no history of cancer to a median depth of 6,930X. We identified 391 somatic mtDNA mutations in 284 of 616 tumors (45.7%) that varied by cancer type and mtDNA haplogroup. Among tumor subtypes, the low-grade gliomas (LGG) had the fewest and no Loss-of-Function (LoF) somatic mtDNA mutations. The adrenocortical carcinomas (ACC) had the highest number of somatic mtDNA mutations per tumor and a significant percentage (25%) of the tumors had LoF mtDNA mutations. The number of somatic mtDNA mutations per tumor also differed significantly between the two largest macro-haplogroups: 0.66 for H, which is mostly Eurasian-specific, and 0.47 for L, which is mostly African-specific (p = 0.03, Kruskal-Wallis test). Interestingly, four pediatric cancer patients harbored homoplasmic variants in their matched tumor and normal samples that are causal of Leber's Hereditary Optic Neuropathy (LHON) and hearing loss. Three cancer patients' matched tumor and normal samples had heteroplasmic mtDNA variants known to cause classic mitochondrial diseases. Most strikingly, the 45 LoF mutations from all tumors clustered at 4 statistically significant hotspots ($p=5.9\times10^{-3}$ to 1.8×10^{-28}) in *MT-CO3*, *MT-ND4*, and *MT-ND5*. We also found a mutation hotspot in *MT-TM* ($p=2.361\times10^{-5}$). Furthermore, a skewed ratio (4.83) of non-synonymous versus synonymous (dN/dS) mtDNA mutations with high statistical significance ($p = 5.19\times10^{-4}$) was identified based on Monte Carlo simulations in the tumors. By comparison, opposite ratios of 0.44 and 0.93 were observed in 616 matched normal tissues and in 249 blood samples from children without cancer, respectively. These results suggest that, while deleterious mtDNA mutations in general contribute to tumorigenesis, specific mitochondrial genes, functions, and related pathways may play a more important role for tumorigenesis of pediatric cancers. Together, our study establishes the landscape of germline and somatic mtDNA mutations across childhood cancers.

51

Experience with germline confirmation for TP53 variants identified by tumor-only sequencing in pediatric cohorts. *M. Luo, F. Lin, G. Akgumus, D. Gallo, X. Zhao, H. Jung, J. Tang, E. Romasko, L. Conlin, G. Wertheim, L. Surrey, M. Li.* Department of Pathology and Laboratory Medicine, The Children S Hospital of Philadelphia, Philadelphia, PA.

Both somatic and germline changes in *TP53* are very common in cancers, making identification of potential germline variants from tumor genomic profiles very important for patient care. Challenges include highly variable Li-Fraumeni syndrome (LFS)-related neoplasms and complex tumor genomic alterations. In this study, we investigated *TP53* changes in 1500 retrospective pediatric tumors tested with targeted NGS panels interrogating both SNVs and CNVs in genes associated with solid tumors or hematological malignancies. Somatic and germline variant classification was based on the practice guidelines. Germline confirmation was performed after genetic counseling. A total of 83 Tiers 1/2 SNVs were detected in 74 tumors (41 with LFS core tumors) from 70 individuals. About 61% of 74 tumors harbored two *TP53* alterations indicating biallelic loss of TP53. Besides *TP53* alterations, additional Tiers 1/2 variants were detected in 76% of all tumors. Germline testing was only performed on 23 patients, eleven of whom were found to carry a germline *TP53* variant and 3 of whom had a second germline change in other cancer genes. For the remaining 12 patients, four had only one germline change in different cancer genes. No significant difference among patients with or without *TP53* germline variants was found regarding variant allele fraction (VAF), tumor type, or presence of other somatic changes. Reviewing all 83 *TP53* SNVs, 72% can be classified as pathogenic or likely pathogenic if germline, and 82% of them are located in somatic mutation hotspots. The remaining variants have uncertain significance, but 60% of them are also located in hotspots. The overall ratio of P/LP vs VOUS was similar in the 11 confirmed *TP53* germline variants. Our study demonstrated that germline confirmation for *TP53* should be routinely assessed in patients with pediatric cancers, regardless of VAF, tumor type, and/or identification of other germline variants. In addition, our data also showed that variant classification should be performed using both somatic and germline criteria as somatic Tiers 1/2 variants may include both VOUS and pathogenic germline changes. While some somatic data (e.g. second hit and hotspot) may be used in support of pathogenicity for germline variant classification, this application is not fully addressed in the current standards. Additional studies may be needed to prove the pathogenicity of *TP53* variants beyond the qualification for Tiers 1/2.

Copyright © 2018 American Society of Human Genetics. All rights reserved

52

Sequencing of whole genome, exome and transcriptome for pediatric precision oncology: Somatic variants and actionable findings from 253 patients enrolled in the Genomes for Kids study. *S. Newman[1], E.M. Azzato[2], J. Nakitandwe[2], S. Shurtleff[2], C. Kesserwan[3], D. Hedges[2], S. Foy[1], A. Silkov[1], Y. Liu[1], Y. Liu[1], S. Brady[1], J. Gu[2], M.N. Edmonson[1], A. Patel[1], M. Wilkinson[1], K. Hamilton[3], R. McGee[3], E. Quinn[3], R. Nuccio[3], L. Harrison[3], A. Bahrami[2], J.M. Klco[2], B. Orr[2], A. Pappo[4], G. Robinson[4], M. Rusch[1], D.W. Ellison[2], J.R. Downing[2], K.E. Nichols[3], J. Zhang[1].* 1) Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN; 2) Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN; 3) Division of Cancer Predisposition, St. Jude Children's Research Hospital, Memphis, TN; 4) Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN.

The Genomes for Kids study (G4K) assessed the feasibility and utility of combined whole genome (WGS), exome and transcriptome sequencing in a real-time pediatric oncology service. Unlike previous studies focusing on specific diagnoses or high risk/relapsed cases, testing by our CAP/CLIA-certified laboratory was done prospectively for all patients. We sequenced tumor and matched normal tissue from 253 patients (123 hematological, 63 brain, 67 other solid tumors) at >45X for WGS, >100X for exome and >100 million reads for transcriptome. Our analysis pipeline cross-validated variants discovered in each sequence type, yielding high-accuracy calls at a low limit of detection (15% of SNV/indels detected at <0.1 VAF). This removed the need for time-consuming confirmatory and reflex testing in most cases, enabling efficient assessment by a multidisciplinary tumor board and reporting in less than 40 days by the end of the study. We reported a mean of four pathogenic/likely pathogenic variants per case (range 0-18); only two cases (<1%) had no reportable findings. These abnormalities consisted of biologically relevant SNVs (22.6%); indels (11%); gross chromosomal losses, gains and LOH (31.8%); sub-arm copy number changes (20.3%) and structural variants, gene fusions, and ITDs (14.3%). Of the reported events, 15.1% were diagnostic, 21.8% were prognostic, and 6.8% were directly or indirectly targetable; actionable mutations included common events such as *ETV6-RUNX1*, *EWSR1-FLI1* and *BRAF* V600E, as well as rarer events such as *MN1-CXXC5* fusion (diagnostic for HGNET-MN1), *CDK6@-MECOM* (likely prognostic in AML), and *IGH-EPOR* (indirectly targetable with a JAK inhibitor in B-ALL). Of note, 26% of reported and 8% of actionable findings were in genes not part of the current COSMIC cancer gene census. On a per patient basis, 79% had at least one somatic finding that could be used to guide care. G4K highlights the benefit of incorporating a prospective NGS test including WGS into pediatric oncology diagnostics. All data are available as a free community resource hosted in the cloud (https://stjude.cloud). We demonstrate how users may replicate our analysis using their own bioinformatics workflows, visually explore thousands of coding and non-coding variants, or drill down to individual pathogenicity classifications. Finally, we challenge the clinical genomics community to find actionable mutations in the 21% of cases in which we found none despite this comprehensive testing.

53

Rare biallelic variants in Deoxyhypusine synthase (DHPS), an enzyme involved in the hypusination of eukaryotic translation initiation factor 5A, are associated with neurodevelopmental delay and seizures. *M. Ganapathi[1], L.R. Padgett[2], K. Yamada[3], O. Devinsky[4], R. Willaert[5], R. Person[5], P-Y. Billie Au[6], J. Tagoe[6], M. McDonald[7], D. Karlowicz[7], Y. Shen[8], V. Okur[9], L. Deng[9], C.A. LeDuc[9], R.G. Mirmira[3], M.H. Park[10], T.L. Mastracci[2,11], W.K. Chung[9,12].* 1) Columbia University Medical Center, New York, NY; 2) Regenerative Medicine & Metabolic Biology, Indiana Biosciences Research Institute, Indianapolis, IN, 46202, USA; 3) Department of Pediatrics and the Center for Diabetes and Metabolic Diseases, Indiana University School of Medicine, Indianapolis, IN, 46202, USA; 4) NYU Langone Medical Center Department of Medicine, New York, NY, 10016, USA; 5) GeneDx, Gaithersburg, MD, 20877, USA; 6) Department of Medical Genetics, Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, Alberta, T2N 4N1, Canada; 7) Division of Medical Genetics, Department of Pediatrics, Duke University Medical Center, Durham, NC, 27710, USA; 8) Department of Systems Biology, Columbia University Medical Center, New York, NY, 10032, USA; 9) Department of Pediatrics, Columbia University Medical Center, New York, NY, 10032, USA; 10) National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD, 20892-4340, USA; 11) Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, 46202, USA; 12) Department of Medicine, Columbia University Medical Center, New York, NY, 10032, USA.

Hypusine is formed post-translationally from lysine and is found only in a single cellular protein, eukaryotic translation initiation factor-5A (eIF5A). Biosynthesis of hypusine is a two-step reaction involving the enzymes deoxyhypusine synthase (DHPS) and deoxyhypusine hydroxylase (DOHH). eIF5A is highly conserved throughout eukaryotic evolution and plays a role in mRNA translation, cell proliferation, differentiation, and inflammation. DHPS is also highly conserved in eukaryotes and essential as *Dhps* homozygous deletion knockout mice are embryonic lethal. Using exome sequencing, we identified rare biallelic, recurrent, predicted pathogenic variants in *DHPS* segregating in four affected individuals from three unrelated families who all share a phenotype of neurodevelopmental disability that includes global developmental delay, seizures, and a prenatal history of maternal HELLP syndrome or hypertension. Two of the three affected females also have short stature. All four affected children share a recurrent missense variant (c.518A>G;p.N173S) in trans with a likely gene disrupting variant, either c.1014+1G>A or c.912_917delTTACAT;p.Y305_I306del. Expression studies demonstrated that the c.1014+1G>A variant causes aberrant splicing. We generated recombinant DHPS enzymes with either the p.N173S or p.Y305_I306del mutation and demonstrated reduced or absent *in vitro* activity. The $DHPS^{N173S}$ enzyme had approximately 20% of wild-type enzyme activity, whereas the $DHPS^{Y305\_I306del}$ enzyme had no activity. Moreover, we co-transfected constructs expressing HA-DHPS (wildtype or mutant) and GFP-eIF5A into HEK293T cells to determine the effect of these mutations on DHPS function, and observed that both the p.N173S and p.Y305_I306del alleles result in reduced hypusination of eIF5A. Additionally, an *in vitro* stability assay determined that the $DHPS^{Y305\_I306del}$ enzyme was degraded. Our data suggest that rare biallelic variants in *DHPS* result in reduced or absent enzyme activity which limit the hypusination of eIF5A and are associated with a severe neurodevelopmental disorder.

Copyright © 2018 American Society of Human Genetics. All rights reserved

54

Mutations in *ZFP92*, a novel KRAB Zinc-finger Protein, results in an X-linked intellectual disability and mitochondrial dysfunction disorder. *C.E. Schwartz[1], J.W. Norris[1], M. Harr[2], A. Orrico[3], E. Zackai[2].* 1) JC Self Research Institute, Greenwood Genetic Center, Greenwood, SC; 2) Center for Applied Genomics, Children's Hospital of Philadelphia, PA; 3) Molecular Medicine and Genetics. Az. Osp. Univ.Senese - Siena and Clinical Genetics. USLsudest - Grosseto.

X-linked intellectual disability (XLID), a chronic congenital disability, is one of the most common causes of intellectual disability, especially among males. To date, 141 XLID genes have been identified. Nonetheless, not all XLID genes have been found. The *ZFP92* gene, located in Xq28, codes for a protein containing a KRAB domain and eight C2H2 zinc finger domains. It is predicted to act as a transcription factor based on the presence of these domains. Three males in an XLID family and an unrelated young boy all of who had ID, developmental delay, hypotonia and behavioral issues were found to have mutations in *ZFP92*, p.R171H and p.L188_P211 deletion respectively. Utilizing the CRISPR/Cas9 technology, these mutations were created in a male fetal mesencephalon cell line, VM. Studies of these altered cell lines indicated decreased neurite outgrowth and increased expression of *BSN* and *SLC8A2* as compared to normal VM cells. Both of these genes are co-expressed with *ZFP92*. BSN (Bassoon) is a large synaptic protein and a core component of the cytomatrix at the active zones of both excitatory and inhibitory synapses. SLC8A2 mediates the exchange of Ca^{2+} and Na^{+} ions across the plasma membrane, thereby contributing to the Ca^{2+} homeostasis in excitable cells. *SLC8A2* is located in 19q13.32 and is deleted in the 19q13.32 microdeletion syndrome. Metabolic studies of both patient lymphoblasts and the altered VM cells, utilizing the Seahorse platform, indicated the two *ZFP92* mutations resulted in abnormal mitochondrial function. Specifically, both cell types exhibited lower glycolytic activity and mitochondrial respiration. So for lymphoblastoid cells and edited VM cells, those with the patient mutations were not able to match the energy production exhibited by normal cells. The patient with the deletion mutation showed a positive response to treatment with a mitochondrial “cocktail” despite blood and urine metabolites not being suggestive of mitochondrial deficiency. Thus, patients with *ZFP92* mutations can be regarded as having a novel XLID syndrome associated with hypotonia, developmental delay and features potentially consistent with a mitochondrial disorder.

55

Biallelic variants in *DYNC1I2* cause syndromic microcephaly with intellectual disability, global developmental delay and dysmorphic facial features. *M. Ansar[1], F. Ullah[2], A. Darius[3], A. Lai[4], S.A. Paracha[5], M.T. Sarwar[5], J. Iwaszkiewicz[6], Z. Agha[7], L. Pais[8], E. Falconnet[1], E. Ranza[1,8], F.A. Santoni[1,10], V. Zoete[6], J. Ahmed[5], P. Makrythanasis[1,11], N. Katsanis[2], C. Walsh[4], E.E. Davis[2], S.E. Antonarakis[1,9,12].* 1) Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland; 2) Center for Human Disease Modeling, Duke University, Durham, North Carolina, USA; 3) Atlantic Health System, Morristown, NJ, USA; 4) Howard Hughes Medical Institute and Division of Genetics and Genomics, Children's Hospital Boston, and Neurology and Pediatrics, Harvard Medical School Center for Life Sciences, Blackfan Circle, Boston, MA, USA; 5) Institute of Basic Medical Sciences, Khyber Medical University, Peshawar, Pakistan; 6) Swiss Institute of Bioinformatics, Molecular Modeling Group, Batiment Genopode, Unil Sorge, Lausanne, Switzerland; 7) Department of Biosciences, COMSATS University, Islamabad, Pakistan; 8) Medical and Population Genetics Program and Center for Mendelian Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 9) Service of Genetic Medicine, University Hospitals of Geneva, Geneva, Switzerland; 10) Department of Endocrinology Diabetes and Metabolism, University Hospital of Lausanne, Switzerland; 11) Biomedical Research Foundation of the Academy of Athens, Athens, Greece; 12) iGE3 Institute of Genetics and Genomics of Geneva, Geneva, Switzerland.

DYNC1I2 (Dynein Cytoplasmic 1 Intermediate Chain 2) (OMIM#603331) encodes a protein that is part of the cytoplasmic intermediate dynein 1 complex. It is involved in motor activity of microtubules and participates in cargo transport. We characterized a consanguineous Pakistani family of three affected individuals with microcephaly, severe intellectual disability (ID), developmental delay and facial dysmorphism. Exome sequencing and genotyping identified a homozygous variant in a splice donor site of *DYNC1I2* (NM_001378:c.607+1G>A), predicted to produce an early termination. Through GeneMatcher, we identified two additional cases from the USA who harbored biallelic deleterious variants in *DYNC1I2*: One was compound heterozygous with a p.(Tyr247Cys) change and a complete gene deletion. The second unrelated case had compound heterozygous variants, p.(Gln290Ter) and p.(Tyr247Cys). These two individuals displayed overlapping neuroanatomical and facial phenotypes with the index pedigree. A brain MRI in one affected individual revealed mega-cisterna magna, absence of the rostrum and genu of the corpus callosum and the septum pellucidum, and partial absence of the splenium. Structural analysis suggested that p.(Tyr247Cys) disrupts the interactions between light and intermediate chain of the dynein 1 complex; furthermore, it might influence protein folding by creating nonnative disulfide bonds. To gain insight into the pathomechanism of this novel syndrome, we developed DYNC1I2 loss of function models in zebrafish. Larvae with either high-efficiency mosaic targeting (F0 CRISPR/Cas9), or transient morpholino-induced suppression of *dync1i2a* displayed significantly altered craniofacial patterning with concomitant reduction in head size. We monitored cell cycle progression and apoptosis in *dync1i2a* morphants and found both increased phospho-Histone H3 and TUNEL positive cells, respectively, offering a possible explanation for the cellular basis of microcephaly. Finally, *in vivo* complementation assays in zebrafish using quantitative readouts of head size and mandible structures demonstrate that p.Tyr247Cys confers a loss of function. In sum, we provide strong evidence that *DYNC1I2* is a novel cause of an autosomal recessive microcephaly syndrome with ID and facial dysmorphism. Other members of the dynein complex cause similar features (DYNC1H1; OMIM#600112), and our data highlight further the importance of the dynein complex in neurodevelopment.

Copyright © 2018 American Society of Human Genetics. All rights reserved

56

***De novo* mutations in *CNOT1,* a master regulator of gene expression on DNA, RNA, and protein level, cause neurodevelopmental delay.** *L.E.L.M. Vissers[1], S. Geuer[1], S. Kalvakuri[2], M. Oud[1], I. van Outersterp[1], M. Kwint[1], D.L. Polla[1], A. Begtrup[3], A. Ruiz[4], J.E.V. Morton[5], C. Griffith[6], K. Weiss[7], C. Gamble[8], J. Bartley[9], M. Mori[10], H. Vernon[11], K. Brunet[12], C. Ruivenkamp[13], P. Kruska[7], A. Afenjar[14], K. Nugent[15], F.L. Raymond[16], M. Cho[3], H. van Bokhoven[1], T. Kleefstra[1], R. Bodmer[2], M. Muenke[7], A.P.M. de Brouwer[1], the DDD study.* 1) 1.Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, P.O. Box 9101, 6500 HB, Nijmegen, The Netherlands; 2) 2.SBP Medical Discovery Institute, 10901 North Torrey Pines Rd, La Jolla, California, USA; 3) GeneDx, Gaithersburg, MD 20877, USA; 4) 4.Laboratori de Genètica, UDIAT-Centre de Diagnòstic, Corporació Sanitària Parc Taulí, Sabadell, Spain; 5) West Midlands Regional Clinical Genetics Service and Birmingham Health Partners; 6) Department of Pediatrics, University of South Florida, Tampa, Florida, USA; 7) National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA; 8) Cook Children's, Fort Worth, Texas, USA; 9) Pediatric Specialty Clinics, Loma Linda University, Loma Linda, California, USA; 10) Division of Human Genetics, Hasbro Children's Hospital, The Warren Alpert Medical School of Brown University, Providence, Rhode Island, USA; 11) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, USA; 12) Porcupine Health Unit, Timmins, Ontario, Canada; 13) Department of Clinical Genetics, Leiden University Medical Centre, Leiden, The Netherlands; 14) APHP, Centre de Référence déficiences intellectuelles de causes rares Département de génétique médicale, Sorbonne Université, GRC n°19, pathologies Congénitales du Cervelet-LeucoDystrophies, Hôpital Armand Trousseau, F-75012 Paris, France; 15) Children's Hospital of San Antonio, San Antonio, Texas, USA; 16) Department of Medical Genetics, University of Cambridge, Cambridge, United Kingdom.

Over the last decade, exome and genome sequencing studies have shown that a vast majority of neurodevelopmental disorders such as developmental delay (DD) can be explained by a *de novo* mutation in one of thousands of genes involved in neurodevelopment. Yet, many genes underlying DD still await discovery. Here, we report 17 patients with DD and a heterozygous *de novo* mutations in *CNOT1*. Deep-phenotyping of these patients revealed a spectrum of features centered around DD including intellectual disability, motor delay, speech delay, seizures, hypotonia, and behavioral problems, such as autism. CNOT1 is a member of the CCR4-NOT complex, which is a master regulator that orchestrates different levels of gene expression. It has been implicated in several aspects of mRNA and protein expression, including transcription initiation, elongation, mRNA degradation, ubiquitination, and protein modification. Population genetic signatures for *CNOT1* indicate that it belongs to the 2% human protein-coding genes most intolerant to variation (RVIS -1.9), and that *CNOT1* is depleted from loss-of-function mutations in large databases such as ExAC (pLi=1, z-score=7.44), suggestive for such mutations being under strong purifying selection. CNOT1 functions as a scaffolding protein binding other subunits of the CCR4-NOT complex, such as CNOT2, CNOT4, and CNOT7 through CNOT11. To elucidate the pathophysiological effects of the *de novo* variants observed in our patients on CNOT1 scaffolding capacity, we generated 8 different *CNOT1* mutation constructs and transfected these in COS1 cells. Using a PalmMyr-CFP-tagged construct, targeting CNOT1 to the cell membrane, co-localization studies have confirmed the interaction of wildtype (wt) CNOT1 with its partners CNOT2, CNOT4 and CNOT8. Analyses of the mutation-specific constructs are currently ongoing. To further examine the effect of *CNOT1 de novo* mutations on neurodevelopment, we generated mutation-specific Drosophila models, which showed learning and memory defects. Introduction of human wt *CNOT1* was able to rescue this phenotype. Moreover, introduction of the mutation-specific constructs were not, supporting our hypothesis that *CNOT1* impairment results in DD. In summary, we show that *de novo CNOT1* mutations cause neurodevelopmental delay. In addition, our data confirm the role of CNOT1 as a master regulator in CCR4-NOT complex formation, demonstrating the essential central role of the CCR4-NOT complex in normal human brain development.

57

Mutations in *FAM50A* cause Armfield XLID Syndrome: A spliceosomopathy impacting a global repertoire of transcripts involved in neurodevelopment. *K. Khan[1], Y.R. Lee[2], K. Armfield-Uhas[3], J.W. Norris[4], K. Gripp[5], K.A. Aleck[6], C. Li[7], J. Edward Spence[8], T. Moreland[4], C. Skinner[4], R.E. Stevenson[4], C.H. Kim[2], E.E. Davis[1], C.E. Schwartz[4].* 1) Center for Human Disease Modeling, Duke University, NC USA, Durham, NC; 2) Department of Biology, Chungnam National University, Daejeon, Republic of Korea; 3) Children's Healthcare of Atlanta at Scottish Rite, Atlanta, GA; 4) Greenwood Genetic Center, Greenwood, SC; 5) Division of Medical Genetics, A. I. duPont Hospital for Children, Wilmington, DE; 6) Genetics and Metabolism, Phoenix Children's Medical Group, Phoenix, AZ; 7) Clinical Genetics Program, McMaster University Medical Center, Hamilton, Ontario, Canada; 8) Division of Pediatric Genetics and Metabolism, University of North Carolina School of Medicine, Chapel Hill, NC.

Intellectual disability (ID) is a clinically and genetically heterogeneous group of disorders characterized by deficits in both intellectual function and adaptive behavior. Twenty-five to thirty percent of ID is associated with genomic/genetic variants on the X-chromosome (X-linked ID; XLID). Armfield XLID syndrome is a condition we previously localized to Xq28 in a large family. The syndrome is characterized by postnatal growth retardation, craniofacial anomalies, seizures and glaucoma. We identified a missense mutation (p.D255G) in *FAM50A* that segregated with the phenotype in the males. Using GeneMatcher, we identified an additional five sporadic male cases with overlapping clinical features who harbor mutations in *FAM50A* that were identified by whole exome sequencing. To study the functional relevance of FAM50A to the human phenotype, we conducted studies in zebrafish. First, we generated a *fam50a* knockout (KO) zebrafish model. Mutants recapitulated the human phenotype including shortened body length, and abnormal development of anterior structures. RNAseq analysis of the *fam50a* KO transcriptome from mutant heads showed dysregulation of ~12% of the genes (FDR corrected $p<0.05$). Further, HPO and OMIM disease gene annotation highlighted significant enrichment of genes implicated in syndromic and non-syndromic ID. These data were supported by gene set enrichment analysis which identified downregulation of relevant pathways such as collagen trimer, neuronal synaptic transmission, glutamate receptor activity, and cell adhesion. Additionally, we observed a significant upregulation of genes involved in the spliceosome complex. We validated 22 genes with either augmented or depleted expression by RNA *in situ* hybridization in *fam50a* KOs. Finally, using a transgenic reporter of cartilage patterning, we showed craniofacial defects in *fam50a* KO or *fam50a* zebrafish morphants. *In vivo* complementation assays indicated that human mRNAs containing the missense variants identified in cases are hypomorphic. Together our study shows that mutations in *FAM50A* result in compromised neuronal development or function with concomitant upregulation of genes associated with the spliceosome complex, highlighting the need for an intricate balance of mRNA processing during neurodevelopment. Therefore, we propose Armfield XLID syndrome as a novel spliceosomopathy.

Copyright © 2018 American Society of Human Genetics. All rights reserved

58

Exome sequencing identifies a novel gene *RNF170* for autosomal recessive hereditary spastic paraplegia. *Y. Jamshidi[1], M. Wagner[2,9], I. Gehweiler[3], S. Bakhtiari[4,5], E. Ozkan[1], R. Maroofian[1], R. Boostani[6], E. Ghayoor Karimiani[7], S. Padilla-Lopez[4,5], K. Vill[10], H. Darvish[8], D.P.S. Osborn[1], M.C. Kruer[4,5], J. Winkelmann[2,9], R. Schüle[3].* 1) St George's University of London, London, United Kingdom; 2) Institut für Humangenetik, Munich, Germany; 3) Center for Neurology and Hertie Institute for Clinical Brain Research, Tübingen, Germany; 4) Cerebral Palsy & Pediatric Movement Disorders Program, Barrow Neurological Institute, Phoenix Children's Hospital, Arizona, USA Children's Hospital, Arizona, USA; 5) Departments of Child Health, Neurology, and Cellular & Molecular Medicine, University of Arizona College of Medicine, Phoenix, Arizona, USA; 6) Department of Neurology, Mashhad, Iran; 7) Next Generation Genetic Clinic, Mashhad, Iran; 8) Department of Medical Genetics, Semnan University of Medical Sciences, Semnan, Iran; 9) Institute for Neurogenomics, Helmholtz-Zentrum Munich, Munich; 10) Division of Pediatric Neurology, Developmental Medicine and Social Pediatrics, Center for Neuromuscular Disorders in Childhood. Dr. von Hauner Children's Hospital, University Hospital, LMU Munich, Germany.

Introduction: Hereditary Spastic Paraplegias (HSP) are a rare group of genetically heterogeneous neurological disorders characterized by progressive spasticity and weakness of the lower limbs in 'pure' HSP patients, and additional neurological and non-neurological symptoms in 'complicated' HSP. **Methods and Results:** Using exome sequencing and co-segregation analysis in a consanguineous Iranian family with four children affected by early onset complex HSP, we identified a homozygous missense mutation in *RNF170*. RNF170 encodes an E3 ubiquitin ligase, that mediates the ubiquitination of the inositol-1,4,5-triphosphate receptor 1 (IP3). The mutant residue lies in the RING domain of RNF170 which is required for ligase activity. Three additional unrelated families with biallelic variants in *RNF170* were also identified using GeneMatcher - a homozygous deletion of exons 4-7 and two frameshift variants. The patients aged four to 53 all present with infantile onset complicated HSP with optic atrophy, and axonal sensorimotor peripheral neuropathy. To further investigate the consequence of *RNF170* loss of function we designed antisense morpholino oligonucleotides targeting the zebrafish *RNF170* orthologous gene. Using acetylated alpha tubulin immunostaining of zebrafish larvae at 72 hours post fertilization we show disorganization of motor neurons, and also observe microphthalmia, and defects in motility. Finally, we show pronounced upregulation of IP3 receptor in patient fibroblasts compared to controls, likely indicating failure of RNF170 to degrade the receptor. **Conclusions:** We conclude that *RNF170* is a novel causative gene for autosomal recessive complex HSP. RNF170 ubiquitination of IP3 is dependent on the ERLIN1/2 protein complex, and biallelic loss of function variants in ERLIN1/2 have recently been associated with HSP. Our data therefore reinforces the role of the endoplasmic reticulum associated degradation pathway in the pathogenesis of HSP.

59

Genome wide discoveries in 13,000 whole genome sequenced rare disease cases and controls. *K.E. Stirrups[1,2] on behalf of the NIHR BioResource - Rare Diseases and the 100,000 Genomes Project.* 1) NIHR BioResource, Cambridge University Hospitals NHS Foundation, Cambridge Biomedical Campus, Cambridge, UK, CB2 0QQ; 2) Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK, CB2 0PT.

To study genetic sequence variants underlying unresolved Mendelian disorders and improve interpretation of already identified high penetrance variants, a collection of 13,000 individuals with a rare disease and their relatives has been whole genome sequenced with an average 30x coverage. Participants were mainly recruited at NHS hospitals in the UK using approved eligibility criteria for 15 different rare disease domains. We describe the population structure including ethnicity and relatedness estimation, high level phenotypes collected using Human Phenotype Ontology (HPO) terms and quality control and summary metrics for samples and variants. The resource contains over 165 million unique variants (including 90, 3 and 6% SNVs, small insertions and deletions respectively) in the 10,258 genetically independent samples with 47% of variants previously unobserved in other large scale publically available genome datasets (e.g. gnomAD, HGMD, UK10K). We summarise the curation of gene lists and pertinent findings in 2,000 unique diagnostic-grade genes for the 15 domains. Over 1200 reports assigning pathogenic or likely pathogenic causal variants have been issued following review by Multi-Disciplinary Teams. The diagnostic yield varied across the different domains from 0.5 to 55%, while the proportion of novel (compared to HGMD) causal variants ranged between 25 to 73%; causal variants in 10 genes have been reported that involve cross-domain findings, where the same gene is linked to different clinical phenotypes. We show the power of a recently developed rapid Bayesian association test, BeviMed, to identify novel genes (n>30) and potentially causal variants in the non-coding space of the genome and to provide independent validation of recent rare disease gene discoveries by others. Some of these novel genes have been reported and led to changes in patient management. Finally we review the infrastructure required to host and analyse such a large dataset and how this has been used to undertake a realignment to genome build 38 and the increase in sensitivity of variant detection generated by this upgrade. The rare disease pilot of the 100,000 Genomes Project has shown the feasibility of using WGS across a national health system to deliver a molecular diagnosis for patients with inherited rare diseases and how a national resource of genotype accompanied by HPO-coded phenotypes provides a powerful platform for the identification of novel diagnostic-grade genes.

Copyright © 2018 American Society of Human Genetics. All rights reserved

60

Genetic associations of replication timing reveal chromatin and sequence determinants of DNA replication origin activity in humans.

Q. Ding[1], C. Charvet[1], J. Hsiao[2], X. Zhu[2], F.T. Merkle[3], R.E. Handsaker[4,5], S. Ghosh[4,5,6], K. Eggan[4,6], S.A. McCarroll[4,5], M. Stephens[2], Y. Gilad[2], A.G. Clark[1], A. Koren[1]. 1) Department of Molecular Biology and Genetics, Cornell University, Ithaca 14853, NY, USA; 2) Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA; 3) Wellcome Trust - Medical Research Council Institute of Metabolic Science, University of Cambridge, Cambridge, UK; 4) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; 5) Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; 6) Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

In eukaryotic cells, DNA is replicated according to a strict spatiotemporal program. Replication origins are activated at different times within the S phase; as a consequence, different genomic loci replicate at different times. However, the mechanisms that influence replication origin specification and DNA replication timing in human cells are poorly understood. Here, we exploited inter-individual differences in replication timing to identify sequence determinants of replication timing at near base-pair resolution, and used this information to reveal novel insights into the mechanistic basis of human DNA replication. We inferred replication timing profiles for 108 human embryonic stem cell lines (hESCs) based on deep whole-genome sequencing data. We identified 613 *cis* replication timing quantitative trait loci (rtQTLs), at 10% false discovery rate. rtQTLs were cumulatively associated with the replication timing of 13.54% of the genome, and were enriched at replication origins, suggesting that replication origin activity is extensively influenced by sequence determinants. rtQTLs associated with 396 unique genomic regions, among which 142 were associated with at least two (and up to six) independent rtQTLs, revealing that human replication timing is jointly and additively determined by multiple *cis*-acting sequence elements. rtQTLs were enriched for active chromatin marks and the central pluripotency transcription factors POU5F1 (Oct4) and NANOG. Among individuals, loss of Oct4 or NANOG motifs were associated with later replication timing, indicating that transcription factor binding may be integral to replication timing control. In summary, we suggest that human DNA replication timing is subject to a complex regulatory code in which multiple sequence and chromatin determinants cooperate to influence the activity of individual DNA replication origins.

61

Characterizing the impact of genetic loss of function in inflammation and pain response.

C. Lam[1], C.R. Bauer[2], S.A. Pendergrass[2]. 1) Geisinger Health System, Danville, PA, USA; 2) Geisigner Health System, Rockville, MD, USA.

The NIH estimates that nearly 50 million Americans suffer from chronic or severe pain. Treatment of pain symptoms is notoriously difficult when proximate causes cannot be identified, and many of the most effective medications, most notably opioids, carry substantial risk of addiction, drug interactions, or even death. Genomic testing may help to better identify variability in pain response, optimize treatment plans to individuals, or identify new potential drug targets. Thus, there is an urgent need to characterize genetic factors that may lead to pain-related disorders or impact patient responses to specific interventions. To close these knowledge gaps, we have evaluated the effects of predicted loss of function (LoF) mutations within a large patient population with comprehensive and longitudinal clinical data. The DiscovEHR collaboration integrates whole exome sequencing data with the electronic health records of patients at Geisinger who have enrolled in the MyCode Community Health Initiative. Within a subset of 62,971 unrelated individuals of European descent, we conducted an exome wide screen across 7,746 genes for low frequency variants (MAF < .05) that are predicted to alter gene function and identified associations with diagnoses for 49 groups of disorders related to pain symptoms, neural degeneration, or inflammation of nerves. Our results include highly significant associations with loss of function (LoF) mutations in the genes *POC5* and *ACAD9* for alcoholic polyneuropathy ($p = 7x10^{-36}$ and $p = 4x10^{-14}$, respectively). We also found that mutations in *KL*, *DRD2*, and *SGO1* were significantly associated with complications of herpes infection or postherpetic polyneuropathy ($p = 2x10^{-28}$, $p = 3x10^{-23}$, and $p = 3x10^{-18}$, respectively). *ACADS*, *DNAJC10*, and *TMEM175* were associated with complex regional pain syndrome ($p = 2x10^{-14}$, $p = 4x10^{-12}$, and $p = 4x10^{-12}$, respectively). Finally, we found that mutations in the gene *PSEN1* are associated with sciatica ($p = 8x10^{-32}$). A secondary PheWAS analysis of these genes across 2,363 diagnoses revealed that *PSEN1* LoF variation is associated with a wide variety of conditions ranging from disorders of vision and hearing to arthritis, degeneration of intervertebral discs, and kidney disease. Future directions include efforts to replicate these findings in an independent dataset as well as deeper investigations into the specific nature of the implicated variants.

Copyright © 2018 American Society of Human Genetics. All rights reserved

62

Polygenic localization of disease heritability using functional annotations. *O. Weissbrod[1], F. Hormozdiari[1], B. van de Geijn[1], A. Schoech[1], S. Gazal[1], Y. Reshef[2,3,4], C. Márquez-Luna[1], L. O'Connor[1], H. Finucane[4], A.L. Price[1,4,5].* 1) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA; 2) Department of Computer Science, Harvard University, Cambridge, MA; 3) Harvard/MIT MD/PhD Program, Boston, MA; 4) Broad Institute of MIT and Harvard, Cambridge, MA; 5) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA.

Because diseases and complex traits are extremely polygenic, existing fine-mapping methods focused on genome-wide significant loci can localize only a very small fraction of disease heritability. Here, we propose a method for polygenic localization of genome-wide causal variants: we aim to identify the smallest set of SNPs causally explaining a given proportion of disease heritability, by leveraging functional annotations that are enriched for causal disease heritability. Our method makes use of stratified LD score regression (S-LDSC; Finucane et al. 2015 Nat Genet). Although S-LDSC produces robust estimates of functional enrichment of causal heritability, the approach of ranking SNPs based on (noisy) estimates of per-SNP heritability from S-LDSC, and then estimating the proportion of heritability explained by the top SNPs using the same data, is susceptible to overfitting. Instead, we (1) run a regularized extension of S-LDSC on even (resp. odd) chromosomes; (2) use the resulting functional enrichment estimates to predict per-SNP heritability for each SNP on the remaining chromosomes, and rank these SNPs accordingly; and (3) run S-LDSC on the remaining chromosomes to estimate the heritability explained by the top-ranked SNPs, producing an estimate that is not susceptible to overfitting. We confirmed via simulations using real genotypes that our method is well-calibrated, producing estimates of heritability explained by top-ranked SNPs that are not inflated by overfitting. We applied our method to 130 diseases and complex traits (average N=240K) using the baseline-LD model (Gazal et al. 2017 Nat Genet). The average fraction of common SNPs required to explain 50% of causal disease heritability was 11.3% (s.d. 7.5%). Polygenic localization was most successful for autoimmune diseases (e.g. psoriasis: 2.7%; celiac disease: 2.8%; lupus: 2.8%; Crohn's disease: 3.9%) and blood cell traits (e.g. eosinophil count: 3.6%), which have the largest functional enrichments. It was least successful for brain-related traits (e.g. subjective well-being: 33.6%; ever smoked: 32.2%; autism: 25.0%; bipolar disorder: 23.1%; schizophrenia: 18.2%), which have smaller functional enrichments. Our results help bridge the gap between fine-mapping and functional partitioning of heritability and can help prioritize variants and genes for functional follow-up.

63

Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank. *I. Tachmazidou[1], K. Hatzikotoulas[2], L. Southam[2,3], J.E. Gordillo[1], V. Haberland[4], J. Zheng[4], T. Johnson[1], M. Koprulu[2,5], E. Zengini[6,7], J. Steinberg[2,8], J.M. Wilkinson[6], S. Bhatnagar[9], J. Hoffman[10], L. Yerges Armstrong[10], G. Davey Smith[4], T. Gaunt[4], R.A. Scott[1], L. McCarthy[1], E. Zeggini[2], arcOGEN Consortium.* 1) Target Sciences - R&D, GSK Medicines Research Centre, Stevenage, UK; 2) Human Genetics, Wellcome Sanger Institute, Wellcome Genome Campus, UK; 3) Wellcome Centre for Human Genetics, University of Oxford, UK; 4) MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, UK; 5) Department of Medical Genetics, University of Cambridge, UK; 6) Department of Oncology and Metabolism, University of Sheffield, UK; 7) 5th Psychiatric Department, Dromokaiteio Psychiatric Hospital, Greece; 8) Cancer Research Division, Cancer Council NSW, Australia; 9) Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Canada; 10) Target Sciences - R&D, GSK Medicines Research Centre, King of Prussia, US.

Osteoarthritis (OA) is the most common musculoskeletal disease and the leading cause of disability globally with no effective treatment. Thirty-four distinct OA association signals have been discovered in European and Asian populations, accounting for a small fraction of OA genetic heritability. Here, we performed the largest genome-wide association study for OA to date (70,588 cases and 263,702 controls), using hospital-diagnosed and self-reported data from the UK Biobank resource and identify 43 genome-wide significant signals following meta-analysis with OA data from the arcOGEN consortium. Thirty-two of the 43 signals are novel, thus doubling the number of established disease loci. Three novel signals fine map to a single causal variant (missense variant rs13107325 in *SLC39A8*, OR 1.10, P=5.7x10-17; intergenic rs75621460 near *TGFB1*, OR 1.15, P=3.0x10-14; and predicted deleterious missense variant rs4252548 in *IL11*, OR 1.32, P=2.9x10-12), and ten additional signals fine map to less than 10 variants. Proteomics analysis on intact and degraded cartilage samples (N=24) showed increased expression for both *IL11* and *SLC39A8* (FDR=5.65x10-3 and 6.36x10-5, respectively). We observe strong evidence of eQTL colocalization in at least one GTEx tissue for 28 out of our 43 loci, 21 of which are at newly associated signals. Transcriptome-wide association analyses between predicted expression levels and OA were completed using MetaXcan and identify 106 unique genes in at least one GTEx tissue with this number reducing to 39 genes after filtering results by eQTL colocalization probabilities. We find signal enrichment for genes underlying monogenic forms of bone development diseases (P=3.6x10-5), and for the collagen formation (FDR=0.03) and extracellular matrix organization biological pathways (FDR=0.04). We enhance our understanding of the relationship of complex physiological and molecular traits with OA through causal inference analysis.

Copyright © 2018 American Society of Human Genetics. All rights reserved

64

Automated annotation for re-analysis of whole exome sequencing data: A pilot study. *A. Ferrer[1], P.H. Duffy[2], J.A. Johnson[2], C. Kaiwar[3], M.A. Meiners[2], R.M. Moore[1], M.B. Mundy[2], G.R. Oliver[1], D.N. Rider[2], M.E. Williams[2], F. Vairo[1], M.A. Cousin[1], A. Gupta[1], J.P.A. Kocher[1], E.W. Klee[1].* 1) Center for individualized medicine, Mayo Clinic, Rochester, MN; 2) Information Technology, Mayo Clinic, Rochester, MN; 3) Invitae Corporation, San Francisco, CA.

Genomic sequencing for rare and undiagnosed disease has a reported diagnostic rate of 25-50%, depending on the level of prior testing. While this has had a transformative impact on genetic disease diagnosis, a majority of patient cases still remain unsolved following this testing for a variety of reasons, including the lack of information about the variants identified, causing many to be classified as “Uncertain Significance” (VUS) by ACMG guidelines. New variant evidence is reported in the literature and collected in databases every day and consequently any VUS can be reclassified as a result of emerging findings in public disease-variant databases, turning previously unsolved cases into diagnosed patients. Consequently reanalyzing cases regularly can be highly informative. Due to the high number of unsolved cases, reanalysis using a manual approach is unrealistic and to the best of our knowledge no clinical lab is routinely performing this task. At the Center for Individualized Medicine (Mayo Clinic, MN) we are addressing this issue using an automated bioinformatics pipeline that quarterly identifies changes in the classification and/or newly reported variants in three of the major genetic databases (ClinVar, HGMD and OMIM). This information is then used to automatically reannotate genomic data from unsolved cases and identify variants warranting further review. Using additional filters that take into account inheritance pattern of the disease and population frequency, we are able to focus on relevant information and reanalyze undiagnosed cases in minutes. To test this protocol, we introduced results from 50 unsolved cases tested with exome sequencing into the pipeline and analyzed changes in the databases from August, 2015 to October, 2017. An average of 1.5 (ClinVar), 7.5 (HGMD) and 55.5 (OMIM) variants were identified per case, and further review found this information relevant for 9 cases, including identifying a suspected causative variant in 3 of previously unsolved cases. Out of these 3 newly solved cases, 2 of them used new information from OMIM while the third one was solved using updated information from HGMD. With the increasing use of clinical genomic sequencing and rising number of undiagnosed cases, we believe reanalysis approaches will become very relevant in the near future and our automated bioinformatics pipeline offers a fast and practical approach to this issue and are working towards a future state where this process is fully automated.

65

NGS gene panel for newborn screening, a case-control study. *G.L. Yamamoto[1,2], F.P. Vairo[3], K.M. Rocha[2], M.L. Magalhães[2], M.S. Naslavsky[2], R.S. Honjo[1], C.A. Kim[1], R. Giugliani[3], M.R.S. Passos-Bueno[2], D.R. Bertola[1,2].* 1) Instituto da Criança - FM - University of São Paulo, Sao Paulo, Sao Paulo, Brazil; 2) Centro de Pesquisa Sobre o Genoma Humano - IB - University of São Paulo, Sao Paulo, Sao Paulo, Brazil; 3) Hospital de Clínicas de Porto Alegre - Universidade Federal do Rio Grande do Sul.

Criteria for newborn screening tests have been quite well stablished since 1968 by Wilson e Jungner. However, with novel techniques such as Tandem Mass Spectrometry some disorders that would not fit the original criteria but would be screened “free of additional costs” were included. A similar reasoning could be applied to support newborn screening through Next Generation Sequencing (NGS). Two major issues must be taken into account: the cost of NGS and the risk of false positives when genotype is not predictive of phenotype. The cost of NGS has been dropping steadily since 2007 and sequencing panels instead of complete exomes or genomes can further lower the costs. The risk of false postives can be addressed following strict guidelines such as ACMG 2015 for variant report. We proposed to do, in a research setting, a case-control blind analysis with 45 “negative patients” from CEGH-CEL-IB-USP and 45 “positive cases” from HCPA-UFRGS that were sequenced with a panel of about 500 genes for mendelian disorders. The samples were de-identified and randomized prior to analysis. Positive cases had patients from 3 major groups: inborn errors of metabolism (IEM), muscular disorders and neurological disorders; some of them confirmed by Sanger sequencing. Results: 43/45 negative patients would have been given true negative reports, 1 negative patient was given a true positive result (a Marfan hypothesis negative for *FBN1*, *TGFBR1* and *TGFBR2* that turned out to be positive for *CBS*-homocystinuria), and 1 negative patient had a false positive for a hemizygous missense variant in *FMR1* that was classified in the literature as pathogenic but ACMG criteria would be between VUS and likely pathogenic. In the positive patients 22/45 were true positives, with a diagnostic yield of 49%. Among the IEM, 19/26 (73%) were true positives, whereas in the muscular and neurologic groups the rates were lower (2/14 and 1/4 respectively). In two IEM patients the condition was not identified due to the fact that the corresponding genes were not in the panel list. In conclusion we have observed a high sensibility for IEM (73%) in spite of a blind analysis. The sensibility could be even higher with a broader gene panel. But, most importantly, applying only ACMG criteria and a blind analysis no false positive in 90 cases would be clinically reported. We believe that this study strength the evidence for the introduction of NGS tests in newborn screening.

Copyright © 2018 American Society of Human Genetics. All rights reserved

66

Utilization of rapid whole genomic sequencing (rWGS) in hospital intensive care units demonstrates a significant improvement in clinical care, healthcare costs and diagnostic rate across a diverse range of patient phenotypes. *S.A. Nahas[1], S. Chowdhury[2], J. Friedman[2], J. Gleeson[1], Y. Ding[1], M. Wright[1], M. Tokita[1], K. Ellsworth[1], N. Sweeney[1], D. Dimmock[1], S. Kingsmore[1].* 1) Genomics, Rady Children's Institute for Genomic Medicine, San Diego, CA; 2) UCSD/Rady Children's Hospital San Diego and Rady Children's Institute for Genomic Medicine, San Diego, CA.

Introduction: Implementation of rapid Whole Genome Sequencing (rWGS) in Hospital NICUs has resulted in reduced time to diagnosis and improved diagnostic rates compared to the current standard of care. The current diagnostic odyssey is time consuming and costly for ill infants and healthcare providers. We sought to investigate the clinical and economic utility of rWGS in infants from hospital NICUs. **Methods:** After consent, blood was collected, gDNA isolated, library preparation performed, and Illumina rWGS was undertaken to ~45X coverage. Phenotypic features were translated into HPO terms. **Results:** To date, rWGS was interpreted in 487 patients, yielding diagnostic information in 163 (~33%). On average, diagnosis occurred within 96 hours (fastest 37 hours). Changes in management in those with a diagnosis were identified in 117 families (~72%). The greatest impact was in infants initially presenting with seizures, congenital heart disease (CHD) and inborn errors of metabolism (IEM). In 18 patients with seizures who received diagnosis, 8 (44%) were in epileptic encephalopathy genes and 7/8 (88%) resulted in a change in management. In those with a suspected CHD, 46% (11/24) received positive or negative genomic results that resulted in a change in management that ranged from cardiac transplantation, avoidance of intraoperative cholangiogram, and transfer to a pediatric lung transplant center among others. 82% (9/11) obtained a definitive diagnosis that explained their cardiac associated phenotype. Nine patients were diagnosed with IEM and medically actionable results were identified in six, resulting in medication/dietary changes, avoidance in further morbidity and invasive diagnostic testing. Specific examples include a 7 d/o with status epilepticus diagnosed with ALDH7A1-related pyridoxine-dependent epilepsy within 48 hours of admission, whose seizures promptly ceased following treatment with a modified diet and pyridoxine, a CHD patient with LVNC and a pathogenic TPM1 missense variant resulting in heart transplant course adjustments, and a 2m/o infant admitted with cholestasis and hepatosplenomegaly who was diagnosed with Niemann Pick Type C and as a result, experimental therapy was instituted prior to the onset of neurologic symptoms. In total, rWGS resulted in ~$3M in net cost savings over projected standard care. **Conclusion:** rWGS improves clinical care, preventing unnecessary procedures, while reducing acute care costs among hospital inpatient infants.

67

Next generation children project: Whole genome sequencing for rapid diagnosis of severely ill children in intensive care. *F.L. Raymond[1,2,3], C. French[1], H. Dolling[1,2], A. Sanchis-Juan[1,2], I. Delon[3], E. Dewhurst[1,2], T. Austin[3], R. Armstrong[3], G. Belteki[3], M. Bohatschek[3], S. Bowdin[3], R.G. Branco[3], S. Broster[3], H. Firth[3], S. Park[3], A. Parker[3], C.G. Woods[1,3], S. Abbs[3], D. Rowitch[1,3], NIHR BioResource - Rare Diseases.* 1) University of Cambridge, Cambridge, UK; 2) NIHR BioResource - Rare Diseases, Cambridge, UK; 3) Cambridge University Hospitals, Cambridge, UK.

Children in neonatal or paediatric intensive care units (N/PICU) may have an underlying genetic condition. The Next Generation Children Project (NGC) aims to achieve a rapid diagnosis via trio whole genome sequencing that would augment clinical decision-making and to establish appropriate criteria for such testing. Careful consideration was given to approaching families and a two-stage consenting process was used successfully. 49% of eligible families consented to the study and of these 88% provided a two-parent and child trio sample. Recruitment was specifically declined by 15% of those families approached for diverse reasons. The remainder of families approached (36%) were either undecided by the time of discharge, lost to follow-up or repatriated to a local hospital. To date, 145 probands (age range neonate - 15 years) have been sequenced and analysed (420 samples at ~38 X coverage). The time from sample receipt to a research finding is < 2 weeks and with confirmed diagnostic reports through the NHS, a further 7 days. Variants of uncertain significance and incidental findings were not reported. A likely diagnosis was reported in 17% of cases, including encephalopathies, myopathies, skeletal dysplasias and rare syndromes. Each proband has on average 11 HPO terms. The cohort was clustered by phenotypic similarity and the group with the highest diagnostic rate was enriched for neurodevelopmental delay. Some of the lowest diagnostic rates were for congenital heart defects or hypoxic ischaemic encephalopathy. When comparing the phenotype of the diagnosed proband to the associated phenotypes of genes in OMIM, the diagnosed gene was in the top 10 most similar gene/phenotype in only 13% of cases. The reduced correlation between genotype and phenotype in intensively ill children suggests that balancing early detection of rare conditions with an unusual neonatal presentation against predictive testing for childhood or early adult onset conditions irrelevant to the child's current clinical episode is sometimes challenging. These results demonstrate a need to understand parental acceptability of genomic testing in the N/PICU setting, the value of genotype-driven whole genome trio analysis and the need for long-term research and clinical follow-up of cases to evaluate outcomes based on early diagnosis.

Copyright © 2018 American Society of Human Genetics. All rights reserved

68

Rapid exome testing of 500 acutely ill newborns and infants provides equivalent diagnostic sensitivity to whole-genome sequencing. *J. Juusola, D. Copenheaver, S. Yang, J. Hare, E. Butler, S. Bale, J. Scuffins, B.D. Solomon, K. Retterer.* GeneDx, Gaithersburg, MD.

Several recent studies emphasized the utility of rapid whole-genome sequencing (WGS) for NICU patients. Farnaes et al. (PMID29644095) tested 42 patients <1yo with 18 positive findings; Petrikin et al. (PMID29449963) diagnosed 10/32 patients <4mo; Willig et al. (PMID25937001) tested a similar cohort with 20/35 positive findings. The combined diagnostic rate of these studies was 44% (48/109). The median time to WGS analysis was 26hrs-5 days and to final report 14-23 days. We performed a retrospective study to compare sensitivity and other parameters between WGS and rapid exome sequencing (ES). Since 2015, our laboratory has performed trio-based rapid ES in >700 individuals, including 500 acutely ill newborns/infants. Of those, 48% (n=242) had a diagnostic result (pathogenic/likely pathogenic variants in known disease genes), demonstrating a slightly higher yield than WGS (p=0.46, Fisher's exact test). In the newborn (<1mo) cohort, 47% (76/161) were positive, while 49% (166/339) of infants (1mo-2yo) tested positive. In addition, 27% (n=136) of patients had VUSs in phenotypically-relevant genes, and 21% (n=105) had relevant variants in candidate genes. The median time to ES analysis and verbal result was 5 days with a final written report issued at 9.5 days, after variant confirmations. Historically, WGS offered better breadth of coverage of coding regions compared to exome, but improved exome capture methods, such as Agilent Clinical Research Exome or IDT xGen Exome v1.0, now offer nearly equivalent coverage and sensitivity and have other advantages. On average, 40x WGS yields 98.8% coverage of coding regions at 10x while 60-120x ES yields 98.5%, representing only a marginal loss of sensitivity. Higher sequencing depth also improves our ability to detect mosaic variants, which represent 2.9% of our positive findings (7/242). Of the three WGS studies, structural variants (SV) were only systematically assessed by Farnaes, and accounted for 17% (3/18) of positive findings. In our rapid ES cohort, CNVs comprised 12% (30/242) of diagnostic findings, demonstrating that first-tier ES can routinely detect clinically-relevant CNVs at a rate comparable to WGS (p=0.71, Fisher's exact test) as well as clinically-relevant mobile element insertions (MEIs). While ES has certain limitations, such as lack of coverage in deep non-coding regions, it is considerably cheaper than WGS, can also produce rapid results, and has diagnostic yield that is equal or better to WGS.

69

The impact of newborn genomic sequencing on families: Findings from the BabySeq Project. *S. Pereira[1], D. Petersen[1], J.O. Robinson[1], L. Frankel[2], K.D. Christensen[3,4], S.E. Waisbren[3,5], I.A. Holm[3,5], A.H. Beggs[3,5], R.C. Green[3,4,6,7], A.L. McGuire[1], The BabySeq Project Team.* 1) Center for Medical Ethics & Health Policy, Baylor College of Medicine, Houston, TX; 2) Department of Psychological, Health and Learning Sciences, University of Houston, Houston, TX; 3) Harvard Medical School, Boston, MA; 4) Division of Genetics, Brigham and Women's Hospital, Boston, MA; 5) Division of Genetics and Genomics, The Manton Center for Orphan Disease Research, Boston Children's Hospital & Harvard Medical School, Boston, MA; 6) Partners Healthcare Personalized Medicine, Cambridge, MA; 7) Broad Institute of MIT and Harvard, Cambridge, MA.

The familial impact of newborn genomic sequencing (nGS) is unknown. Family systems theory suggests that events affecting one family member will affect the entire family system. We therefore assessed the impact of nGS on families as part of the BabySeq Project, a randomized, controlled trial assessing the impact of integrating nGS into the clinical care of newborns. Parents completed surveys with validated and novel measures at enrollment in the neonatal period, after a disclosure session when they reviewed their child's NBS results and family history alone (control arm) or with nGS results (nGS arm), and at 3 and 10 months after results disclosure. Within the nGS arm, parents whose infants were found to have monogenic disease risk findings (n=7 parents of 4 infants) perceived their children to be more vulnerable than parents of infants who did not (n=105 parents of 74 infants) 10 months after disclosure ($p<0.001$), but no differences were observed in parent-child bonding scores ($p>0.05$). We observed no differences in parental depression and anxiety scores by randomization arm. We also found no evidence of harmful impact of nGS or disclosure of a monogenic disease risk on self- or partner-blame. Counterintuitively, results demonstrated that more parents in the control arm blamed themselves for passing on potentially harmful genes to their child compared to the nGS arm (25% control arm vs. 12% nGS arm, $p<0.05$). Similarly, more parents in the control arm blamed their partner for passing on potentially harmful genes to their child compared to the nGS arm (17% control arm vs. 7% nGS arm, $p<0.05$). Overall, preliminary findings suggest that providing nGS and disclosing monogenic risk findings to parents causes no significant short-term harm to families. In fact, findings related to self and partner blame suggest that providing nGS to families may provide peace of mind to parents, though this may reflect concerns among the subset of families that volunteered for this particular research, misunderstanding, or overinterpretation of results. Continued data collection is necessary to explore these findings in greater depth.

Copyright © 2018 American Society of Human Genetics. All rights reserved

70

Genomic sequencing (GS) in the neonatal intensive care unit (NICU) demonstrates a significant disease burden in acutely ill infants: Interim results of the NSIGHT2 study. *D.P. Dimmock[1], M.N. Bainbridge[1], S. Batalov[1], W. Benson[1], J.A. Cakici[1], J. Carroll[1,2], S. Caylor[1], S. Chowdhury[1], M.M. Clark[1], C. Clarke[1], Y. Ding[1], K. Ellsworth[1], M. Gaughan[1], L. Farnaes[1,2], A. Hildreth[1,2], S. Nahas[1], L. Salz[1], E. Sanford[1,2], N.M. Sweeney[1,2], M. Tokita[1], N. Veeraraghavan[1], K. Watkins[1], K. Wigby[1,2], T. Wong[1], M. Wright[1], S.F. Kingsmore[1], RCIGM Investigators.* 1) Rady Children's Institute for Genomic Medicine, San Diego, CA; 2) University of California, San Diego, San Diego, CA.

Background Genomic sequencing (GS) has demonstrated significant improvement in the early identification of genetic disorders in the NICU in cases where genetics experts have a suspicion of monogenic disorders. More recent data has demonstrated that the early identification of such children improves clinical outcomes and may reduce the cost of care. However, the majority of NICUs do not have access to genetic services. Thus the questions of the burden of genetic disease in the NICU, and how this potentially life saving approach can apply to regional NICU's are important to address in clinical trials. We have embarked upon an NIH funded multi-year single site double blind randomized controlled study to evaluate WGS vs WES (NCT03211039). Acutely ill inpatients < 4 months within 96 h of admission without isolated prematurity, unconjugated hyperbilirubinemia, HIE, confirmed genetic etiology, transient neonatal tachypnea or sepsis with a normal response to therapy were eligible for enrolment in this study. These inclusion criteria reflect ~3% of all liveborn children in the US. To date the study has not performed interim analysis with regard to the primary endpoint. However, as part of this study we are evaluating the incidence of undiagnosed genetic disorders in the NICU population. Methods After informed consent, neonates underwent a phenotype driven analysis of proband WGS or WES with reflex to trio analysis if a diagnosis was not made. A case was considered positive if the genetic etiology explained why the child was in the ICU whereas a partial diagnosis was assigned if it explained part of the child's phenotype. Medically actionable incidental findings were returned on cases but secondary findings were not sought. Results To date, 753 infants <4 m of age were admitted to our ICUs, 422(56%) had conditions ineligible for enrollment. Parents of 331 infants were approached and 93(28%) declined enrolment or timed out of the 96 hour window. 137 families had completed analysis. 36(26%) cases had a monogenic diagnosis. If our results are generalizable ~ 1/4 of all Level 4 NICU admissions not due to prematurity are as a result of undiagnosed genetic disorders that are currently clinically reportable. Summary This is the first systematic study to evaluate the burden of genetic disease in the NICU. If these cases are representative of the national birth cohort, 18,000 infants < 4 months old are admitted yearly to high acuity ICU settings with an undiagnosed genetic disorder.

71

Fine-mapping of Alzheimer's disease risk loci to identify potential novel drug targets. *J.Z. Liu[1], J. Schwartzentruber[2], K.D. Nguyen[1], S. Cooper[3], E. Bello[3], B.C. Grabiner[1], J.C. Barrett[2,4], T. Johnson[5], A.R. Bassett[3], K. Estrada[1].* 1) Biogen, Cambridge, MA; 2) Open Targets, European Bioinformatics Institute, Hinxton, UK; 3) Open Targets, Wellcome Sanger Institute, Hinxton, UK; 4) Current affiliation: Genomics plc, Oxford, UK; 5) GlaxoSmithKline plc, Stevenage, UK.

Genome-wide association studies (GWAS) have identified over 30 loci associated with Alzheimer's disease (AD) risk. Identifying and experimentally validating the causal variants and genes at these risk loci will shed light on biological mechanisms of AD, and may lead to novel drug targets for disease treatment and prevention. We first performed a GWAS meta-analysis combining summary statistics from the largest AD GWAS (Lambert et al., Nat Genet. 2013) with a GWAS-by-proxy of AD family history in the UK Biobank (total sample size of 25,550 cases, 56,772 proxy cases and 427,836 controls). We identified 35 genome-wide significant ($P < 5x10^{-8}$) loci associated with AD risk, including two novel risk loci near *TMEM163* and *NDUFAF6*. Common AD risk variants were also genome-wide significant for the first time near *APP* the gene encoding the amyloid precursor protein which is causal for early onset AD. At each genome-wide significant locus, we performed conditional analysis to identify the number of independent signals, and then used Bayesian fine-mapping to construct credible sets of causal variants. At 18 of these loci we were able to narrow the 90% credible set down to 10 or fewer causal candidates; at four loci, we could narrow the signal down to a single causal variant. To identify putatively causal genes, we assessed statistical colocalization between each risk locus and nearby quantitative trait loci (QTLs) for gene expression and other molecular traits from GTEx, AMP-AD, and Blueprint. Twenty-five loci showed strong colocalization evidence (>80% posterior probability) between AD risk and at least one QTL. These include an eQTL for *CCDC6* in the brain cortex, methylation QTLs near an enhancer at *CD2AP*, and an sQTL at *INPP5D*. To select variants for experimental follow-up, we annotated candidate variants using data from Roadmap, FANTOM, and ATAC-seq peaks from primary microglia and iPSC-derived cell lines (macrophages, neural progenitors, and neurons). As many functional variants overlapped candidate AD genes, we are applying CRISPR-Cas9 editing to assay the effects of 30 selected variants on transcription of 8 genes which they overlap. We anticipate that the readouts from these assays will allow us to estimate the variants' causal effects on gene regulation to generate novel (or validate existing) hypotheses on the functional mechanisms that drive the genetic risk for AD, a crucial step to narrow down a list of potential targets for novel therapeutic approaches.

Copyright © 2018 American Society of Human Genetics. All rights reserved

72

Using RNA-sequencing to compare the transcriptional profile of primary human microglia and *in-vitro* microglial models. *F. Calvert[1], A. Young[2,3], N. Kumasaka[1], A. Knights[1], N. Murphy[3], C. McMurran[3], M. Segel[3], P. Hutchinson[2], R. Franklin[3], D. Gaffney[1,4].* 1) Wellcome Sanger Institute, Hinxton, Cambridgeshire , United Kingdom; 2) Division of Academic Neurosurgery, Department of Clinical Neurosciences, Cambridge University Hospital, Cambridge, UK; 3) Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK; 4) Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK.

Dementia is estimated to effect nearly 50 million people worldwide, with that number expected to triple by 2050. Alzheimer's disease (AD), the most common form of dementia, is complex with many factors appearing to affect disease development and progression. Multiple lines of evidence, including large-effect AD risk variants in myeloid-specific genes such as *TREM2* and a striking enrichment of AD loci risk variants near genes with immune function, have suggested that microglia, the tissue resident macrophages of the CNS, as key players in AD. Here, we analyse the largest collection to date of single cell and bulk RNA-seq data from human primary microglia and macrophages, collected from 30 neurosurgical patients from a range of clinical presentations. We use our single cell data to illustrate how the population composition of microglia changes between different patient clinical phenotypes and across patient age from early adulthood to old age. In parallel, we compare how closely primary microglia are represented by a range of *in-vitro* models, including human IPS-derived macrophages and microglia, and widely-used leukemic macrophage cell models, THP-1 and U937. We show that primary microglia still retain a distinct transcriptional profile, maintaining the expression of key transcriptional regulators such as *SALL1* and marker genes such as *P2RY12* and *CX3CR1*, that are weakly expressed in all available alternative models. Using gene co-expression analysis, we identify the regulatory networks that are active in primary microglia but silent in most *in-vitro* models. Finally, we use our expression data identify candidate causal genes at known AD risk loci and examine how expression is conserved or diverges between primary microglia and their existing *in-vitro* models. Our work provides a roadmap for the derivation of improved *in-vitro* microglia cell models and identifies AD risk loci that are likely to be well captured by existing model systems, and those that remain poorly represented.

73

Transcriptomic meta-analysis identifies sex-specific and APOE-specific gene signatures in Alzheimer's disease using single gene and network approaches. *M. Paranjpe[1], K. Zalocusky[2], A. Taubes[2], M. Sirota[1].* 1) University of California San Francisco, San Francisco, CA; 2) Gladstone Institute of Neurological Disease, Gladstone Institutes, San Francisco, CA.

Alzheimer's disease (AD) is a heterogeneous cognitive disorder with multiple etiologies and phenotypes. In spite of evidence of females having a greater lifetime risk of developing AD and greater apolipoprotein E-related (ApoE4) AD risk compared to males, molecular signatures underlying these findings remain elusive. We compiled expression data from brains of 862 AD patients and age-matched controls to identify sex and ApoE genotype-specific gene signatures in AD. After adjusting for ApoE status, region and age, we observed 692 differentially expressed genes in AD vs controls unique to females and 65 unique to males. 384 genes were commonly differentially expressed in both males and females. Pathway analysis revealed genes uniquely upregulated in females with AD compared to controls were enriched for Il-1 signaling, WNT signaling, cytokine-cytokine receptor interaction and proinflammatory processes while genes uniquely downregulated in females compared to controls were enriched for monoamine GPCRs. Genes commonly upregulated in both males and females were enriched for complement activation, glutathione metabolism and carbon metabolism while downregulated genes were enriched for BDNF signaling, serotonin signaling and circadian regulation related pathways. Male specific genes lacked pathway enrichment. Weighted Gene Correlation Network Analysis (WGCNA) identified networks of correlated genes related to AD and ApoE4 carrier status. After controlling for region, age, ApoE4 status and sex, WGCNA revealed 3 gene networks whose eigengene was positively correlated with AD and 2 negatively correlated with AD. Stratifying by sex revealed 2 more modules enriched for inflammation and fatty acid oxidation and apoptosis to be positively correlated with AD in females but not males. Interestingly, eigengenes corresponding to four modules (three positively associated with disease; one negatively), enriched for complement activation, notch signaling, fatty acid oxidation and BDNF signaling, exhibited a significant disease:ApoE carrier interaction effect in females but not males. In summary, this meta-analysis revealed a female-specific AD disease gene signature involving immune-related processes at the single gene and network level. Network analysis revealed four gene modules that exhibit a significant disease:ApoE carrier interaction effect in females only, offering a novel transcriptomic explanation for differential ApoE4-associated AD risk in males and females. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

74

Brain somatic mutations in Alzheimer's disease are associated with dysregulated phosphorylation of Tau and aging. *J.S. Park[1], J.H. Lee[3], E.S. Jung[4], J.H. Lee[1,2].* 1) Biomedical Science and Engineering Interdisciplinary Program, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea; 2) Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea; 3) Center for Computational Science Platform, National Institute of Supercomputing and Networking, Korea Institute of Scinece and Technology Information, Daejeon, Republic of Korea; 4) Department of Biochemistry & Biomedical Sciences, College of Medicine, Seoul National University, Seoul, Republic of Korea.

Somatic mutations can occur in any tissue throughout its lifetime. In spite of the prominent role in human cancer, the presence of brain somatic mutations in Alzheimer's disease(AD) has been elusive. We examined whether brain somatic mutations arise in the hippocampal formation of AD patients and have significant roles in AD pathogenesis. To identify somatic single nucleotide variations(SNV) in AD, we performed deep whole-exome sequencing(mean x585.2) on matched brain and blood tissues of 48 AD patients and 11 non-demented controls. We found average 11.1 and 11.55 somatic SNVs in brain and 57.12 and 57.17 SNVs in blood of AD patients and controls, respectively. There were no significant differences between the AD and control groups in terms of the average number of SNVs, mean variant allele frequency, and proportion of mutation types. Notably, the number of somatic SNVs in both blood and brain linearly increased with age but the speed of its accumulation in brain was five times slower than that in blood. Interestingly, analysis of mutational signature showed that brain somatic mutations are significantly associated with defects in DNA mismatch repair, which is distinct from those in blood. We further processed enrichment test and quantified pathogenicity scores using genes with non-synonymous SNVs. Surprisingly, the mutated genes in the AD brain significantly enriched in PI3K-Akt and AMPK pathway, associated with phosphorylation of Tau. In addition, the mutations in AD brain showed higher mean pathogenic scores than that of control brain. To examine biological function of non-synonymous SNVs in the AD brain, we tested a missense mutation in PIN1 regarding its protein expression and subsequent phosphorylation state of Tau. We found that the mutated PIN1 showed almost complete depletion of its expression *in vitro* mutagenesis experiments(p<0.0001), suggesting haploinsufficient expression state of Pin1. We found that knockdown of Pin1 in HT22 cell line, mimicking the expression state, increased phosphorylation of Thr231 and aggregated form of Tau. Overall, our findings showed that brain somatic mutations in AD patients increase with age and have strong potential to induce Tau pathology by adversely affecting expression level of mutated proteins associated with protein phosphorylation. This study demonstrates more elaborated ways of analyzing deep whole-exome data, which can give new insights to AD as well as other neurological diseases with somatic mutations.

75

Decoding the genomic architecture of LOAD using single cells analyses. *J. Barrera[1,2], L. Song[2,3], A. Safi[2,3], G.E. Crawford[2,3], O. Chiba-Falek[1,2].* 1) Department of Neurology, Duke University Medical Center, Durham, NC; 2) Center for Genomic and Computational Biology, Duke University, Durham, NC; 3) Department of Pediatrics, Division of Medical Genetics, Duke University Medical Center, Durham, NC.

Large multi-center genome-wide association studies (GWAS) found associations between >20 genomic loci and late-onset Alzheimer's disease (LOAD). The majority of GWAS associated SNPs are in noncoding, intergenic and intragenic, regions of the genome, suggesting regulatory function. Moreover, expression quantitative trait (eQTL) analyses using brain tissues vulnerable to disease neuropathology revealed several eSNPs that overlap with LOAD-GWAS regions. However, the eQTL studies were conducted using homogenized brain tissues in which the heterogeneity of brain tissue makes it difficult to molecularly characterize specific cell-types. In addition, brain homogenates encounter the limitation of sample-to-sample variability in the proportion of each brain cell-type. Nuclei sorting allows us to detect cell-type specific signals and to study disease relevant cell-types. We performed single cells analyses using brain regions affected in LOAD. We applied the NeuN Fluorescence Activated Nuclei Sorting (FANS) method to sort neuronal vs. non-neuronal nuclei from frozen human *postmortem* brains. We then analyzed the NeuN$^+$ and NeuN- sorted nuclei using ATAC-seq and the NanoString nCounter gene expression assay to identify neuronal specific vs. non-neuronal quantitative differences in chromatin accessibility and gene expression, respectively, followed by data integration with known LOAD-risk loci. We identified ca. 190,000 significant cell-type-specific differential ATAC-seq signals, with ca. 110,000 neuronal specific sites and 70,000 non-neuronal specific sites. Comparisons between 22 LOAD brain samples to matched normal controls discovered 5,344 and 1,145 differential chromatin accessibility sites in neuronal and non-neuronal cells, respectively. Noteworthy, 11 neuronal specific open chromatin sites coincide with LOAD GWAS regions extended to +/-100 kb from the associated SNP position, and 4 non-neuronal specific open chromatin regions overlay genomic regions implicated in LOAD, indicative of candidate regulatory elements that are likely causative in LOAD. In the post-GWAS era the key challenge is moving forward from association to causation towards uncovering the genetic etiologies of LOAD. Our research focuses on mechanistic understanding of noncoding regions in loci associated with LOAD. We present a strategy to decode GWAS-discoveries that combines *in vivo* and *in silico* approaches, and applies genomic technologies, single cell techniques and bioinformatics tools.

Copyright © 2018 American Society of Human Genetics. All rights reserved

76

Identification of distinct immune cell-types associated with restricting development of Alzheimer's disease from fresh human brains. *Y. Kim, J. Fullard, Y. Wang, K. Beaumont, R. Sebra, P. Roussos.* Genetics and Genomic Science Department, Icahn School of Medicine at Mt. Sinai, New York, NY.

Better understanding of regulatory architectures and underlying disease etiology substantially enhance targeting effective risk variants or biological entities in complex neurodegenerative diseases including Alzheimer's disease (AD). Although there existed many studies showing the contribution of systemic immunity and tissue-resident microglia to AD onset and its progression, the results based upon a small set of markers remained highly controversial mainly due to the limitation of resolving cellular heterogeneity and niche-specific complexity of immune cell types. Recently, one study deployed scRNAseq analyses into matched healthy and AD-transgenic mice, identified a novel microglia cell type (DAM), and proposed associated pathways and regulatory mechanism controlling a transition from homeostatic microglia to DAM. Transferring the insightful model into applicable therapeutics in human neurodegeneration is more challenging because single-cell experiments in human individuals are not controlled for covariates (i.e. age and gender) and their heterogeneity is presumably more complex. In this study, we sorted all immune cells (CD45+) from biopsies and postmortem human brain tissues in twelve healthy and AD patients and generated around 100,000 cells via drop-seq single cell protocols. We develop scalable and robust computational frameworks by incorporating parametric and nonparametric modeling approaches for efficiently handling the unwanted effects of inevitable stochastic noise, large sparcity, and various batch effects. With the rigorous statistical analysis, we found a comparable number of immune cell populations also in human brain and identified both known and novel associated markers from the distinct cell types. We further investigate the enrichments of known loci and genes for neurodegenerative diseases including AD in the cell sub-populations and identify novel cell types likely responsible for their states at different disease stages. Combining data from the transcriptional analysis of various immune cells at single-cell levels and from the human genome-wide association studies (GWASs) would provide potential evidence on the function and mechanism of distinct cell types in the diseased human brains. We hope that our approach assists substantial improvement of an immunological therapy targeting the universal and intrinsic mechanism of fighting against neuronal damages possibly shared across multiple neurodegenerative diseases.

77

Deep learning of cardiac morphology from UK Biobank MRI data reveals genome-wide associations for bicuspid aortic valve. *A. Cordova-Palomera[1,2], J. Fries[3,4], P. Varma[5], V.S. Chen[3], M. Fiterau[3], K. Xiao[1], H. Tejeda[1], B. Keavney[6], H.J. Cordell[7], E. Ashley[8], J.R. Priest[1,2].* 1) Department of Pediatrics, Stanford University, Stanford, CA; 2) Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA; 3) Department of Computer Science, Stanford University, Stanford, CA; 4) Center for Biomedical Informatics Research, Stanford University, Stanford, CA; 5) Department of Electrical Engineering, Stanford University, Stanford, CA; 6) Division of Cardiovascular Sciences, University of Manchester, Manchester, United Kingdom; 7) Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom; 8) Department of Medicine, Stanford University, Stanford, CA.

With a prevalence of 1-2% in the general population, bicuspid aortic valve (BAV) is the most common congenital heart disease (CHD) and accounts for more morbidity and mortality than all other CHDs combined. Although reported heritability estimates are as high as 89%, specific molecular genetic markers of BAV risk remain to be discovered.Here, 14,000 magnetic resonance imaging (MRI) sequences from the UK Biobank resource were used to classify aortic valves as either BAV or normal (tricuspid) using a novel deep learning algorithm (https://www.biorxiv.org/content/early/2018/06/05/339630). A standard genome-wide association study was conducted on the subset of unrelated European-ancestry participants (595 BAV, 9207 tricuspid aortic valve) using PLINK on both common and rare imputed variants (minor allele count (MAC) >= 30, imputation score >= 0.8). External validation of the genetic findings was performed on imputed data from a case-control study of up to 2594 cases representing eight CHD types and 5159 healthy subjects from the Wellcome Trust Case Control Consortium 2 (WTCCC2) [Cordell et al. 2013, Nat Genet 45, 822–824].Markers at three rare-variant loci (minor allele frequency (MAF) < 5%) displayed statistically significant associations with BAV, including a variant on chromosome 12 near *IGF1* and *LINC00485* (rs146357447, 12:103025165, MAF = 1.3%, odds ratio (OR): 3.2, p = 6.1e-9), an intronic locus on *MIR28* (rs550423221, 3:188508236, MAF = 0.2%, OR = 9.6) and a marker on chromosome 2 (rs192377594, 2:140363901, MAF = 0.6%, OR = 4.1). In the external dataset rs146357447 was associated with risk for atrial septal defect (MAC in cases = 680, MAC in controls = 10318, OR = 1.9, p = 0.033), and the *MIR28* marker displayed an association with non-specific/mixed CHD (MAC in cases = 774, MAC in controls = 10318, OR = 1.9, p = 0.013).The results suggest novel candidate loci as determinants of genetic risk for BAV in the general population, and indicate a shared genetic architecture with different types of CHD. Imputed rare variants require confirmation by direct genotyping, and selection/survivorship biases limit the use of adult populations to study severe CHD. Leveraging imaging and diagnostic information from large-scale public records offer an unprecedented opportunity to study the genetic architecture of congenital diseases, such as BAV.

Copyright © 2018 American Society of Human Genetics. All rights reserved

78

RERE deficiency leads to downregulation of *Gata4* and the development of ventricular septal defects. *B. Kim[1], H. Zaveri[1], V. Jordan[2], A. Hernández-García[1], D. Scott[1,2].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, TX.

Deletions of chromosome 1p36 affect approximately 1 in 5000 newborns and are associated with a high incidence of congenital heart defects (CHD). The arginine-glutamic acid dipeptide repeats gene (RERE) is located in a critical region for CHD on chromosome 1p36 and encodes a cardiac-expressed nuclear receptor co-regulator. Mutations in *RERE* cause atrial and ventricular septal defects in humans and RERE-deficient mice also develop VSDs. Although RERE clearly plays a critical role in septal development, the morphogenetic and molecular mechanisms by which RERE deficiency causes septal defects are not known. During cardiac development, mesenchymal cells are generated when the endocardial cells in the atrioventricular (AV) canal undergoes epithelial-to-mesenchymal transition (EMT). These mesenchymal cells proliferate to fill the AV cushions and then form part of the atrioventricular septum. In this study, we demonstrated that RERE-deficient mouse embryos have reduced numbers of mesenchymal cells in the AV cushions due to decreased levels of EMT at E10.5. In addition, proliferation of mesenchymal cells was decreased in these embryos at E12.5. Using *Rere* conditional knockout mice, we also showed that tissue-specific ablation of *Rere* in the endocardium leads to hypocellularity of the AV cushions, defective EMT and VSDs. In the AV cushions, RERE colocalizes with GATA4, a transcription factor required for EMT and mesenchymal cell proliferation and expression of *Gata4* was downregulated in *Rere* deficient embryo hearts at E10.5. Luciferase activity driven by the *Gata4* promoter was significantly increased with overexpression of *Rere* and significantly decreased by *Rere* siRNA *in vitro*. We also demonstrated that *Rere* and *Gata4* interact genetically in the development of congenital heart defects. Taken together, these results suggest that RERE functions to positively regulate the expression of GATA4 in the developing AV canal and that a deficiency of RERE leads to the development of VSDs through its effects on EMT and mesenchymal cell proliferation.

79

***PALMD* calcific aortic valve stenosis risk variant disrupts a NFATC2 binding site at distant-acting enhancer and activates a fibrogenic program.** *M. Rosa, M.C. Boulanger, A. Chignon, M. Lamontagne, R. Devillers, G. Mkannez, D. Argaud, G. Rhéaume, N. Gaudreault, S. Thériault, Y. Bossé, P. Mathieu.* Institut universitaire de cardiologie et de pneumologie de Québec-Université Laval, Quebec City, QC, G1V 4G5, Canada.

Background: Calcific aortic valve stenosis (CAVS) is a prevalent heart valve disorder. Recently, genome-wide association and Mendelian randomization studies have highlighted that lower expression of PALMD is causally associated with CAVS. **Methods and Results:** In this work, we characterized the regulation of *PALMD* risk locus and we investigated its impact on the biology of human valve interstitial cells (VICs), the main cellular component of the aortic valve. Integrative weighted-scoring analysis identified rs6702619, a highly conserved noncoding variant located at 65 kb proximal to *PALMD*, as being functionally relevant. Further prioritization using promoter capture Hi-C showed that a region spanning ~10 kb, including rs6702619, has significant enrichment of contact signal with the promoter region of *PALMD*. CRISPR-mediated gene activation at rs6702619 increased the expression of PALMD by 30%. The reference allele T at rs6702619 has strong H3K27ac and H3K4me1 marks, revealing an enhancer signal. The risk allele G is, however, associated with a drastic reduction in H3K4me1 level. Analysis of the locus by position weight matrix score including variants in strong linkage disequilibrium showed that only the risk allele G at rs6702619 disrupts a transcription factor binding site (TFBS) for NFATC2, a transcription factor involved in heart valve morphogenesis. Atomic resolution structure data showed that the risk allele affects base readout in the major groove, whereas in the minor groove indirect readout is impacted by risk allele-induced modification in DNA shape and electronegativity. In reporter assay, the risk allele G decreased substantially the response to NFATC2. DNA-binding ELISA assay performed with 30-mer double stranded oligonucleotides centered on rs6702619 showed that the risk allele impedes significantly the binding of NFATC2. Weighted gene co-expression network analysis in 233 valves showed that PALMD module is enriched in Gene Ontology terms for actin filament based-process. Functional assays showed that variant-induced lower expression of PALMD in VICs promoted the polymerization of actin and the activation of SRF-MRTF signalling, driving a fibrogenic program. **Conclusion:** Risk allele at *PALMD* locus disrupts a TFBS for NFATC2 located in a distant-acting enhancer, resulting in lower expression of PALMD in the aortic valve, and promotes a fibrogenic program, a key underpinning process involved in the development of CAVS.

Copyright © 2018 American Society of Human Genetics. All rights reserved

80

Down-regulation of *Sox7* impairs epithelial-to-mesenchymal transition and endocardial cushion morphogenesis. *A. Hernandez-Garcia[1], M. Wat[1], R. Udan[2], A. Renwick[1], Z. Yu[1], C.A Shaw[1], M.E. Dickinson[2], Y. Li[3], D.A Scott[1,2].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX., USA; 2) Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, TX., USA; 3) Single Cell Genomics Core, Baylor College of Medicine, Houston, TX., USA.

SOX7 is located in a critical region for congenital heart defects (CHD) on chromosome 8p23.1 that is recurrently deleted in individuals with septal defects. *SOX7* encodes a DNA binding transcription factor that is highly expressed in vascular endothelium and the endocardium during early embryonic development. To explore the role of SOX7 in cardiovascular development, we developed standard and conditional *Sox7* knockout mice. *Sox7* $^{-/-}$ embryos die around E11.5 with signs of heart failure, pericardial edema, and failure of vasculature remodeling. The same phenotype was observed when *Sox7* was ablated in endothelial cells using a Tie2-Cre. This suggests that SOX7 plays a critical role in vasculogenesis. We also observed that *Sox7*$^{-/-}$ embryos had hypocellular endocardial cushions with severely reduced numbers of mesenchymal cells and that one out of two *Sox7* $^{flox/flox}$;Tie2-Cre embryos that escaped early lethality had a ventricular septal defect. This led us to hypothesize that SOX7 might also be playing a critical role in the endocardium of the AV canal. To test this hypothesis, we performed explant studies in which we harvested the AV canals of E9.5 *Sox7*$^{-/-}$ embryos and their wild-type littermates, and cultured them on a collagen gel for 48 hours. By counting the number of migrating mesenchymal cells associated with each culture, we demonstrated that SOX7 deficiency leads to a severe reduction in endothelial-to-mesenchymal transition. Since SOX7 is a transcription factor, we assumed that it must function in the endocardium by regulating the expression of one or more genes that play a critical role in EMT. To identify these EMT-related genes in an unbiased manner, we performed RNA-seq on E9.5 hearts harvested from *Sox7*$^{-/-}$ embryos and their wild-type littermates. In these studies, we found that *Wnt4* transcript levels were severely downregulated. Previous studies have shown that WNT4 is expressed in the endocardium, and promotes EMT by acting in a paracrine manner to increase the expression of BMP2 in the myocardium. Consistent with these findings, we found that *Bmp2* transcript levels were also decreased in *Sox7*$^{-/-}$ embryonic hearts. Based on these findings, we conclude that SOX7 promotes EMT in the AV canal by modulating the expression of *Wnt4*, and that decreased expression of SOX7 contributes to the congenital heart defects seen in individuals with recurrent 8p23.1 microdeletions.

81

Rational therapeutic epigenetic modulation in the treatment of syndromic thoracic aortic aneurysm. *B.E. Kang[1], R. Bagirzadeh[1], D. Bedja[2], H.C. Dietz[1,3,4].* 1) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) Department of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD; 3) Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD; 4) Howard Hughes Medical Institute, Bethesda, MD.

Many syndromic presentations of thoracic aortic aneurysm including Marfan syndrome (MFS), Shprintzen-Goldberg syndrome (SGS), and Loeys-Dietz syndrome (LDS) show substantial clinical overlap in the craniofacial, skeletal, cutaneous and cardiovascular systems including a strong predisposition for aortic root aneurysm (AoRA) and tear. Phenotypic similarities reflect commonality of mechanism; MFS and LDS manifest dysregulated TGFβ signaling due to failure of an extracellular regulator of cytokine bioavailability or altered function of TGFβ ligands, receptor subunits or intracellular signaling effectors, respectively. We previously showed that SGS is caused by heterozygous mutations in the Sloan-Kettering Institute proto-oncogene (*SKI*) that encodes a prototypical suppressor of transforming growth factor beta (TGFβ) target gene expression through inhibition of the histone acetyl-transferase (HAT) activity of CBP/P300. We also reported that mouse lines harboring germline or vascular-smooth muscle cell (VSMC)-specific orthologous amino acid substitutions (G34D) in *Ski* recapitulate highly penetrant aortic root aneurysm in association with a strong signature for enhanced H3K27 acetylation and an increased TGFβ transcriptional response (TTR) in the aortic wall. In consideration of mechanism, we hypothesized therapeutic potential for pharmacologic HAT inhibitors (HATi) in the management of AoRA. A proof of concept study in SGS mice was extended to include a knock-in mouse model of MFS harboring a heterozygous cysteine substitution in an epidermal growth factor-like domain, the most common class of mutation causing human MFS. In response to C646, a selective P300 inhibitor, both SGS and MFS mice showed full normalization of aortic root growth rate in association with normalization of the TTR (e.g. *Col3a1, Fn1, Smad7, Mmp2* and *Mmp9* expression) when compared to their wild-type vehicle-treated mutant littermates. Histological and morphometric analyses of the aortic wall also showed complete preservation of aortic wall architecture including collagen and elastin content and elastic fiber integrity in C646-treated mouse models of SGS and MFS. Systemic delivery of a potent and selective P300 inhibitor was not associated with any apparent toxicity. These data document excessive H3K27 acetylation and hence TGFβ target gene expression in the pathogenesis of inherited presentations of aortic root aneurysm and the therapeutic potential of pharmacologic epigenetic modulation.

Copyright © 2018 American Society of Human Genetics. All rights reserved

82

Whole genome sequencing of Kyrgyz highlanders reveals novel insight into the genetic basis of high altitude pulmonary hypertension. *T. Stobdan[1], A. Iranmehr[2], D. Zhou[1], O. Poulsen[1], K. Strohl[3], A. Aldashev[4], A. Telenti[5], H.M. Wong[6], E. Kirkness[6], J.C. Venter[6,7], V. Bafna[8], G.G. Haddad[1,9,10].* 1) Pediatrics, University of California San Diego, La Jolla, CA; 2) Department of Electrical & Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA; 3) Department of Medicine, University Hospitals Cleveland Medical Center, Cleveland, OH, USA; 4) National Academy of Sciences, Bishkek 720071, Kyrgyz Republic; 5) Department of Integrative Structural and Computational Biology, Scripps Research Institute, La Jolla, CA 92037, USA; 6) Human Longevity Inc., San Diego, CA 92121, USA; 7) 7 J. Craig Venter Institute, La Jolla, CA 92037, USA; 8) Department of Computer Science & Engineering, University of California, San Diego, La Jolla, CA 92093, USA; 9) Department of Pediatrics, Department of Neurosciences, University of California, San Diego, La Jolla, CA 92093, USA; 10) Rady Children's Hospital, San Diego, CA 92123, USA.

Background: The Central Asian Kyrgyz highland population provides a unique opportunity to address genetic diversity and understand the genetic mechanisms underlying hypoxia-induced pulmonary hypertension (HAPH). While a significant fraction of the population is unaffected, there are susceptible individuals who display HAPH in the absence of any lung, cardiac or hematologic disease. **Results:** We report herein the analysis of the whole genome sequencing of healthy individuals compared with HAPH patients and other controls. Genome scans reveal selection signals in various regions encompassing multiple genes. We present here cumulative evidence of three candidate genes *MTMR4, TMOD3* and *VCAM1* that are functionally associated with well-known molecular and pathophysiological processes and which likely lead to HAPH in this population. These processes are a) dysfunctional BMP-signaling, b) disrupted tissue repair processes and c) abnormal endothelial cell function. **Conclusions:** We sequenced and analyzed whole genomes and uncovered novel candidate genes that belong to several pathways central to the pathogenesis of HAPH. These studies on high altitude human populations are pertinent to the understanding of sea level diseases involving hypoxia as a main element of their pathophysiology.

83

Building genealogies for tens of thousands of individuals genome-wide identifies evidence of directional selection driving many complex human traits. *S.R. Myers[1,2], L. Speidel[1].* 1) Department of Statistics, University of Oxford, Oxford, United Kingdom; 2) Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom.

For a variety of species, large-scale genetic variation datasets are now available. All observed genetic variation can be traced back to a genealogy, which records historical recombination and coalescence events and in principle captures all available information about evolutionary processes. However, the reconstruction of these genealogies has been impossible for modern-scale data, due to huge inherent computational challenges. As a consequence, existing methods usually scale to no more than tens of samples. We have developed a new, computationally efficient method for inferring genome-wide genealogies accounting for varying population sizes and recombination hotspots, robust to data errors, and applicable to thousands of samples genome-wide in many species. This method is >10,000 times faster than existing approaches, and more accurate than leading algorithms for a range of tasks including estimating mutational ages and inferring historical population sizes. Application to 2,478 present-day humans in the 1000 Genomes Project, and wild mice, provides dates for population size changes, merges, splits and introgressions, and identifies changes in underlying evolutionary mutation rates, from 1000 years, to more than 1 million years, ago. Using our mutational age estimates, we developed an approach quantifying evidence of natural selection at each SNP. We compared resulting p-values to existing GWAS study results, finding widespread enrichment (>2.5-fold in Europeans and East Asians) of GWAS hits among individual SNPs with low selection p-values (Z>6), stronger than the 1.5-fold increase observed at nonsynonymous mutations, and with enrichment increasing with statistical significance. We found evidence that directional selection, impacting many SNPs jointly, has shaped the evolution of >50 human traits over the past 1,000-50,000 years, sometimes in different directions among different groups. These include many blood-related traits including blood pressure, platelet volume, both red and white blood cell count and e.g. monocyte counts; educational attainment; age at menarche; and physical traits including skin colour, body mass index and (particularly in South Asian populations) height. Our approach enables simultaneous testing of recent selection, ancient natural selection, and changes in the strength of selection on a trait through time, and is applicable across a wide range of organisms.

Copyright © 2018 American Society of Human Genetics. All rights reserved

84

FADS1 and the timing of human adaptation to agriculture. *I. Mathieson[1], S. Mathieson[2].* 1) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 2) Department of Computer Science, Swarthmore College, Swarthmore, PA.

Human history has seen a number of major transitions in diet. The most recent were the transition to a modern “industrialized” diet and, before that, the transition in many parts of the world to a diet heavily based on the products of agriculture. Variation at the *FADS1/FADS2* gene cluster is functionally associated with differences in lipid metabolism and is often hypothesized to reflect adaptation to an agricultural diet. Here, we test the evidence for this relationship using both modern genomes, and ancient DNA data from over 1030 ancient humans. We also investigate the evidence for the relationship between agriculture and other hypothesized genetic adaptations. We first infer pre-out-of-Africa selection for both the derived and ancestral *FADS1* alleles and show that almost all the inhabitants of Europe carried the ancestral allele until the derived allele was introduced approximately 8,500 years ago by Early Neolithic farming populations. However, we also show that it was not under strong selection in the Early Farmers. Further, we find that this allele was not strongly selected until the Bronze Age, 2,000-4,000 years ago. Therefore, selection at the *FADS* locus was not tightly linked to the initial development of agriculture many thousands of years earlier. Although the *FADS1* locus is strongly associated with lipid levels in present-day populations, patterns of population differentiation at lipid-associated alleles are not consistent with those at *FADS1*, so selection at *FADS1* was not driven by more general selection on lipid levels. We then investigate other adaptive alleles at *LCT/MCM6*, *SLC22A4* and *NAT2*, and show that these alleles were similarly not strongly selected until the Bronze Age. Similarly, increased copy number variation at the salivary amylase gene *AMY1* is not linked to the development of agriculture although in this case, the putative adaptation precedes the agricultural transition. Our analysis shows that selection at *FADS1* and other proposed agricultural adaptations were not tightly linked to the development of agriculture. Further, it suggests that the strongest signals of recent human adaptation may not have been driven by the agricultural transition but by more recent changes in environment, for example a Bronze Age shift towards more intensive cereal agriculture, or by increased efficiency of selection due to increases in effective population size.

85

The impact of modern human specific sites on human phenotypes. *C.R. Robles[1], S. Sankararaman[1,2].* 1) Human Genetics, University of California Los Angeles, Los Angeles, CA; 2) Computer Science, University of California Los Angeles, Los Angeles, CA.

A major question in human evolution centers around understanding genetic changes that make anatomically modern humans “unique”, i.e. genetic mutations that have enabled cultural and technological breakthroughs over the past 200,000 years. The comparison of modern human genomes with genome sequences from our closest evolutionary relatives, Neanderthals and Denisovans, can be used to study the function of these genetic changes. We identified a list of 745 SNPs in the coding sequences of genes or regulatory elements in which the derived mutation is nearly fixed in African individuals in the 1000 Genomes Project but not present in at least some of the sequenced high-coverage archaic genomes. These Fixed Derived Mutations (FDMs) are mutations that rose to high frequency in modern humans since the split from the archaic humans and might provide clues to the biology that causes modern humans to differ from our closest relatives. One approach to study the functional consequences of the FDMs uses the idea that even though these mutations are nearly fixed for the derived allele in African populations, a small fraction of European individuals (~2%) likely carry the ancestral variant at several of these sites due to Neanderthal introgression around 50,000 years ago. As a result, several of the FDMs are polymorphic in Europeans and their impact can be studied by genotyping these mutations in large numbers of individuals with phenotypic information. We added the FDMs to the UK Biobank Axiom array to study the impact of these mutations, and analyzed about 120 FDMs across 107 distinct phenotypes measured in up to 495,000 people of European ancestry in the UK Biobank dataset. We discovered 102 independent associations of FDMs with 37 phenotypes with p-values that pass a threshold accounting for both number of FDMs and phenotypes tested. These hits are in phenotypes including standing height (13 hits), bone heel density (5), and thyroid-related diseases (2). After appropriately controlling for frequencies and the linkage disequilibrium patterns at these variants, we find that the contribution of FDMs to phenotypic variation at FDMs is significantly depleted in traits including whole body fat mass (Z-score of -4.21), trunk fat mass (Z=-4.91), heel bone mineral density (Z=-6.64), and heel quantitative ultrasound index (Z=-7.18). Our results provide insights into genetic and phenotypic changes that were important for modern human biology since their split from archaic humans.

Copyright © 2018 American Society of Human Genetics. All rights reserved

86

Using high-throughput functional assays of protein variants to advance evolutionary inference of deleterious variants. *A.G. Clark, R. Fragoza, H. Yu.* Cornell University, Ithaca, NY.

Protein-coding variants that have a known deleterious impact on fitness produce a set of population-level features that are well understood by population geneticists, including: (1) they have a frequency spectrum that is skewed toward rarer classes, (2) they have a mean coalescence time that is shorter than neutral alleles, and (3) they are more likely to be population-specific than neutral variants. Despite this, efforts to use population attributes to infer functional impact of variants remain error prone. In addition, there is a gap between inference of altered protein function and the impact on reproductive fitness. Here we explore the interface of three large datasets: (1) the very large variant database of gNOMAD (123,136 exome sequences), (2) the coding gene alignments of >100 mammals, and (3) a collection of high-throughput functional assays of protein variants that include disruption of protein-protein interactions and protein stability. The latter includes our experimental assessment of 2,008 nonsynonymous variants that impact 2,181 protein-protein interactions, including a selected set from the Human Gene Mutation Database. While these latter variants clearly disrupt protein function, it appears that they do not universally have strongly deleterious fitness effects. Furthermore, many of the variants, even those that ablate protein interactions, have attained substantial frequency in the population. Possible reasons for the mismatch between a variant's impact on protein function and reproductive fitness include, (1) the rarity and recessiveness of most of the variants, reflecting the recent arrival time of most variants in the population, and (2) a buffering effect of variants in gene regulatory networks that can have parallel and compensatory pathways. One bit of evidence that selection detects protein disruptive variants is that although >50% of the HGMD disease-associated mutations were found to be disruptive, only 17.8% of all coding variants were. Furthermore, there is an excess of disruptive mutations in genomic regions of inferred positive selection, consistent with these regions being augmented for functional relevance. Ultimately, this approach will help us construct a better classifier for inference of *in vivo* protein dysfunction based on genome sequence and *in vitro* functional data. This is strongly motivated by the fact that over half the missense variants in ClinVar are variants of uncertain significance.

87

Whole genome sequencing reveals ancient African substructure and local adaptation. *S. Fan[1], M.E.B. Hansen[1], M.H. Beltrame[1], A. Ranciaro[1], E. Mbunwe[1], J. Chan[2], A.K. Njamnshi[3,4], C. Fokunang[5], S.W. Mpoloka[6], G.G. Mokone[7], T. Nyambo[8], D.W. Meskel[9], Y.S. Song[2,10,11], S.A. Tishkoff[1,12].* 1) Department of Genetics, University of Pennsylvania, Philadelphia, PA; 2) Department of Electrical Engineering and Computer Sciences, University of California–Berkeley, Berkeley, CA 94704, USA; 3) Neurology and Clinical Neurophysiology, Neuroscience Laboratory, Faculty of Medicine and Biomedical Sciences, University of Yaoundé, Cameroon; 4) Neurology Department Central Hospital Yaoundé Brain Research Africa Initiative, Yaoundé PO Box 25625, Yaoundé, Cameroon; 5) Medical and Pharmacotherapeutics Research Group Immunology, Biochemistry & Biotechnology Laboratory Department of Toxico-Pharmacology & Pharmacokinetics Faculty of Medicine and Biomedical Sciences University of Yaoundé, Cameroon; 6) Department of Biological Sciences, University of Botswana, Gaborone, Botswana; 7) Department. of Biomedical Sciences, University of Botswana School of Medicine, Gaborone, Botswana; 8) Department of Biochemistry, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania; 9) Department of Biology, Addis Ababa University, Addis Ababa, Ethiopia; 10) Department of Statistics, University of California–Berkeley, Berkeley, CA 94704, USA; 11) 11.Chan Zuckerberg Biohub, San Francisco, CA 94158, USA; 12) 12.Department of Biology, University of Pennsylvania, Philadelphia, PA 19104.

Africa is the origin of modern humans and is the source of human migrations across the globe within the last 100,000 years. However, Africans are still underrepresented in modern human genomic studies. We conducted high coverage (>30X) whole genome sequencing of 180 individuals from 12 indigenous African populations from Eastern, Western, and Southern Africa, which cover a wide range of subsistence patterns, languages, and phenotypic variation. We found a total of 33.6 million SNPs of which ~5 million SNPs were not reported in the dbSNP database (version 150) and 34% were predicted to be of functional significance, indicating the importance of including ethnically diverse Africans in genomics studies. Phylogenetic reconstruction found that the San populations have lineages that are basal to all modern human lineages. The location of other African populations on the phylogeny largely correlates with their current geographic locations, with the exception of the Pygmies, whose lineages cluster near the San. Admixture analysis identified a shared ancestry among African hunter-gatherer populations from southern, eastern and central Africa, suggesting an ancient connection among these populations. Based on coalescence analysis, we find evidence of population structure in Africa emerging as early as 250 kya, with the earliest divergence between the ancestors of the San and other African populations. We also see evidence for an ancient divergence of all hunter-gatherer populations (San, Pygmy, Hadza, Sandawe) >80 kya. PSMC analysis indicates that ancient effective population sizes of populations begins to differentiate at ~250 kya, close to the time of emergence of modern humans. Despite a small current census size, the ancestors of the San and Pygmy populations maintained the largest (Ne) from ~50–250 kya. We also observe evidence for a recent population bottleneck in the Hadza that resulted in a current size of only ~1,000 individuals. We also identified signatures of positive selection in each population, which may contribute to their local adaptation and phenotypic variation. For example, positively selected SNPs in the Pygmy population are statistically enriched for pathways involving bone growth and cartilage development, which may relate to their short stature. In the San, the SNPs that are under positive selection are significantly enriched in pathways that are associated with skin pigmentation.

Copyright © 2018 American Society of Human Genetics. All rights reserved

88

The genetic prehistory of the Andean highlands 7,000 years BP though European contact. *J. Lindo[1,2], R. Haas[3], C. Hofman[4], M. Apata[5], M. Moraga[5], R. Verdugo[5], J. Watson[6], C. Llave[11], D. Witonsky[2], E. Pacheco[8], M. Vilena[8], R. Soria[8], C. Beall[9], C. Warinner[4,7], J. Novembre[2], M. Aldenderfer[10], A. Di Rienzo[2].* 1) Emory University, Atlanta , GA; 2) University of Chicago, Chicago, IL; 3) University of California, Davis, CA; 4) University of Oklahoma, Norman, OK; 5) Universidad de Chile, Santiago, Chile; 6) University of Arizona, Tucson, AZ; 7) Max Planck Institute for the Science of Human History, Jena, Germany; 8) Instituto Boliviano de Biología de Altura, La Paz, Bolivia; 9) Case Western University, Cleveland, Ohio; 10) University of California Merced, Merced, California; 11) Peruvian Register of Professional Archaeologists, Peru.

The peopling of the Andean highlands above 2500m in elevation was a complex process that included cultural, biological and genetic adaptations. Here we present a time series of ancient whole genomes from the Andes of Peru, dating back to 7,000 calendar years before present (BP), and compare them to 64 newly sequenced genome-wide datasets from both high and lowland populations. We infer three significant features: a split between low and high elevation populations that occurred between 9200-8200 BP; a highland effective population size reduction of 27% after European contact that is significantly lower than an estimated 96% reduction experienced by South American lowlanders; and evidence for positive selection at genetic loci related to starch digestion and plausibly smallpox resistance after European contact. Importantly, we do not find selective sweep signals related to known components of the human hypoxia response, which may suggest more complex modes of genetic and cultural adaptation to high altitude.

89

Meta-analysis in 283,579 East Asians identifies 28 new loci associated with type 2 diabetes. *C.N. Spracklen[1], X. Sim[2], Y.J. Kim[3], M. Horikoshi[4] on behalf of the AGEN and DIAMANTE consortia.* 1) Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) Saw Swee Hock School of Public Health, National University of Singapore, Singapore; 3) Division of Structural and Functional Genomics, Center for Genome Science, Korean National Institute of Health, Song, Chungchungbuk-do, South Korea; 4) RIKEN, Centre for Integrative Medical Sciences, Japan.

Large-scale meta-analyses of genome-wide association studies (GWAS) have identified >250 loci associated with type 2 diabetes (T2D), primarily in analyses of individuals of European ancestry. With differences in linkage disequilibrium (LD) structure and allele frequencies among ancestry groups, large genetic studies in non-European samples may reveal additional loci, identify loci that are shared across ancestry groups, and detect loci that are ancestry-specific. We conducted the largest East Asian GWA meta-analyses to date in up to 283,579 individuals from six East Asian countries using genotypes imputed up to 1000G Phase 3. Models were adjusted for age, sex, and other study-specific covariates, assuming an additive genetic model. Study-specific association summary statistics were combined using a fixed-effects inverse variance weighted meta-analysis approach (~28.2 million variants meta-analyzed). We identified 101 loci associated with T2D (P<5x10^{-8}), including 73 known and 28 novel loci. Among loci with the greatest magnitude of effect, most are low frequency (MAF: 0.01-0.04) variants at novel loci, including *ATG16L1* (OR=1.27)*, ZNF257* within a *ZNF*-cluster on chr19 (OR=1.25), *NKX6-1* (OR=1.24), and *JPH1* (OR=1.19), all of which are rare (MAF < 0.01) in European populations. The lead variant at *ZNF257* is located within a known East Asian-specific 415 kb inversion (MAF=0.04) that disrupts *ZNF257* and creates a fusion transcript. Of the 28 novel loci, the strongest associated variant maps to an intron of *GRM8* (rs117737118, OR=1.19, P=8.0 x 10^{-23}). *GRM8* encodes a G-protein coupled receptor for glutamate, which has been shown to influence human eating behaviors and obesity, and *Grm8* knockout mice demonstrate increased adiposity. Three additional novel loci are located near genes that encode key pancreatic transcription factors, *SIX3-SIX2, NKX6-1*, and *FOXA2*. Effect sizes for significant loci are consistent with effect sizes from European meta-analyses (r^2=0.88; data from DIAGRAM 2017). Expression quantitative trait loci (eQTL) data suggest 4/28 novel GWAS loci are colocalized with eQTLs in adipose tissue (*DCK*, *ARHGAP19*), pancreatic islets (*SIX3, SIX2, SIX3-AS1*), skeletal muscle (*MED23*), and/or whole blood (*MED23*), indicating possible target genes. Taken together, we identified novel T2D loci that inform new biological understanding of T2D, suggesting some etiological differences in T2D risk between the European and East Asian populations.

Copyright © 2018 American Society of Human Genetics. All rights reserved

90

MultiMuTHER – A longitudinal multi-omic study of whole blood gene expression and metabolite levels enabling integrative analysis of multi-omics with genetics and phenotypic trajectories. *J.S. El-Sayed Moustafa[1], M. Abdalla[2,3,4], Y. Jiao[2,3,4], G. Leday[5], M. Stevens[6,7], C. Menni[1], G. Nicholson[4], C. Holmes[4], T.D. Spector[1], M.I. McCarthy[2,3,8], S. Richardson[5], E.T. Dermitzakis[6,7], K.S. Small[1].* 1) Department of Twin Research and Genetic Epidemiology, King's College London, London, UK; 2) Wellcome Centre Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7BN, UK; 3) Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, OX3 7LE, UK; 4) Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK; 5) Medical Research Council Biostatistics Unit, Cambridge, United Kingdom; 6) Institute of Genetics and Genomics of Geneva, Geneva, Switzerland; 7) Department of Genetic Medicine and Development, Faculty of Medicine, University of Geneva, Geneva, Switzerland; 8) Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, OX3 7LE, UK.

There are now a plethora of multi-omics studies, but the vast majority are cross-sectional. TwinsUK is a deeply-phenotyped cohort of mono- and dizygotic twins with extensive 'omics data and repeat phenotypic measures. Using this cohort, we sought to investigate how gene expression and metabolite levels in blood, and their associated heritabilities, genetic and phenotypic associations, change over time within an individual and at the population level, and what genetic or environmental factors drive this. We also sought to determine whether longitudinal gene expression and metabolomics, separately or combined, can be employed to identify multi-omic signatures whose longitudinal trends associate with phenotypic measures and disease outcomes. To address this, we have established the MultiMuTHER study. We have generated gene expression (RNASeq) and metabolomics (Metabolon–1,197 metabolites) data in whole blood at three or more timepoints per individual from 332 TwinsUK subjects. Participants range in age from 32 to 80 years old (median 61yrs), and time between first and last visit ranges from 2 to 8 years (median 6yrs). We observe moderate correlation in metabolite levels across timepoints (median rho=0.46), with significantly lower correlations between timepoints 1-3 (median rho TP1-3=0.43; $P=3.35x10^{-8}$). Metabolite heritabilities calculated using the median of all timepoints (median metabolite heritability=0.40) were higher than single-timepoint heritability estimates (median single-timepoint heritability=0.29), supporting the utility of repeat 'omics measurements. We also show temporal experimental and technical covariates including time of blood collection, month of visit and year of sample collection to be significantly associated (Bonferroni threshold $P<5x10^{-5}$) with metabolite levels (20%, 9% and 18% of metabolites, respectively), emphasising the importance of their inclusion in downstream analyses. Metabolite associations were replicated in a sample of 2,000 twins with three timepoints per subject. Finally, approximately 12% of metabolites showed significant longitudinal associations ($P<5x10^{-5}$), with highly significant enrichment for metabolites within the peptide class ($P=6.48x10^{-15}$). Pre-processing of MultiMuTHER RNASeq data is underway, as well as definition of uni- and multivariate analysis pipelines for integration of these multi-omic data with genetic and phenotypic data, in the largest such longitudinal multi-omic variation study to date.

91

Large-scale GWAS of human plasma metabolome. *P. Surendran[1], I.D. Stewart[2] on behalf of the mGAP (metabolome Genetic Architecture Programme) Investigators.* 1) Cardiovascular Epidemiology Unit (CEU), Department of Public Health and Primary Care, University of Cambridge, UK; 2) MRC Epidemiology Unit, University of Cambridge, Cambridge, UK.

Genetically influenced metabotypes (GIMs) constitute intermediate phenotypes that link GWAS risk loci to disease endpoints. They reveal new functions of uncharacterized genes, the identity of unknown metabolites, validate new therapeutic targets, and are instrumental in a Mendelian randomization approach to infer causality. To date, ~150 genetic loci associated with GIMs have been recognized, many of which overlap with loci of inborn errors of metabolism (IEM). To enhance our understanding of the role of GIMs in complex diseases, we performed the largest GWAS of human plasma metabolome, with twice as many participants (>14,000) and metabolites (>900) as previously reported. In a subset of INTERVAL and EPIC-Norfolk studies, metabolites were measured using the Metabolon untargeted Discovery HD4™ platform and genetic variants were genotyped (UK Biobank Axiom™ array) or imputed (UK10K+1000G/HRC reference). Genetic analyses were performed within study, meta-analysed, and further validated using an independent set of up to ~5,700 EPIC-Norfolk participants. We identified 1,847 sentinel variant associations with 646 metabolites in 330 genetic regions after stringent Bonferroni correction ($P<5.48x10^{-11}$). Conditional analyses uncovered a total of 2,599 independent genotype-metabolite associations corresponding to 488 GIMs. 314 variants associated with 270 metabolites were rare or low frequency (MAF:0.1%–5%); 160 of these were predicted to be protein altering. Majority of the regions (n=206) had more than one metabolite association with the *FADS1/FADS2* cluster being the most pleiotropic with >100 metabolites associated. GIMs overlapped with complex disease GWAS loci, such as *DMGDH* locus co-associated with Type 2 diabetes. Furthermore, we identified 5 unique GIMs containing specific sphingolipids that enable the characterisation of the observed *PCSK9/APOE* association with CHD at a metabolic level. Proximity to genes involved in IEM was observed, for instance, several rare variant GIMs in solute carrier (SLC) genes, *ALDH4A1*, *CTH*, *ACADM* and *DPYD*. Integration of transcriptomics (GTEx/S-PrediXcan) and clinical annotations (ClinVar) and phenome scans (PhenoScanner, SNiPA) showed an association of GIMs with gene expression, clinical outcomes and complex diseases. Our results represent the most comprehensive characterisation of genetic influence on human plasma metabolome to date. Dissemination and interpretation of results will be facilitated by an interactive web server.

Copyright © 2018 American Society of Human Genetics. All rights reserved

92

Obesity represses mitochondrial transcriptional activity in human adipose tissue. *Z. Miao[1,2], M. Alverez[1], Y. Bhagat[1], K.L. Mohlke[3], T. Tusie-Luna[4,5], C. Aguilar-Salinas[5], M. Laakso[6], P. Pajukanta[1,2].* 1) Dept. of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA; 2) Bioinformatics Interdepartmental Program, University of California, Los Angeles, USA; 3) Department of Genetics, University of North Carolina, Chapel Hill, NC, USA; 4) Instituto Nacional de Ciencias Médicas y Nutrición, Salvador Zubiran, Mexico City, Mexico; 5) Instituto de Investigaciones Biomédicas de la UNAM, Mexico City, Mexico; 6) Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.

Mitochondria produce cellular energy, and their dysfunction may contribute to type 2 diabetes and obesity. However, how mitochondrial transcriptional activity relates to obesity in different human tissues is poorly understood. We investigated the association between mitochondrial gene expression and body mass index (BMI) across 1,409 human transcriptomes in multiple tissues to examine whether the mitochondrial expression is associated with BMI, specifically in adipose tissue. Type 2 diabetes subjects were included in our study. First, using 335 adipose RNA-seq samples from the Finnish METSIM cohort, we calculated the percent of expression by all genes encoded by mitochondria in overall gene expression, and corrected that for several technical factors. The corrected mitochondrial expression percent (CMTEP) is significantly associated with BMI in METSIM ($p=2.35x10^{-8}$), the individuals with higher BMI exhibiting a lower CMTEP. Furthermore, we identified that 13 of 14 protein-coding mitochondrial genes showed significantly more expression (adjusted $p<0.05$) in the lean (BMI<=25) individuals compared to the obese (BMI>30). The 13 differentially expressed (DE) mitochondrial genes contribute to 86.9% of the overall expression by mitochondrially encoded genes. We replicated this finding in the adipose RNA-seq data from GTEx (n=308), where the CMTEP is also associated ($p=1.62x10^{-4}$) with BMI in the same direction as in METSIM. In GTEx, 7 of the 13 mitochondrial protein-coding genes contribute to 59.8% of the overall mitochondrial expression and showed the same differential expression as in METSIM. Moreover, we also investigated 107 Mexican adipose RNA-seq samples from obese (BMI>30) bariatric surgery subjects and lean (BMI<=25) Mexicans, and observed the same CMTEP difference ($p<1.55x10^{-8}$). Next, we tested the association between the CMTEP and BMI in other tissues of GTEx, such as muscle, liver, and whole blood (total n=691); however, none of these showed an association between mitochondrial expression and BMI. Our gene-based analysis suggests that the DE of CMTEP is caused by a change in the overall transcriptional activity of mitochondria rather than by a small number of specific genes. Taken together, the obese individuals have significantly less mitochondrial transcriptional activity in subcutaneous adipose tissue compared to the lean individuals in 3 cohorts, supporting a robust repressive effect of obesity on mitochondrial activity across diverse populations.

93

Using genetics to test the phenotypic consequences of higher adiposity uncoupled from its adverse metabolic effects. *T.M. Frayling, Y. Ji, S.E. Jones, R. Beaumont, M.A. Tuke, J. Harrison, K.S. Ruth, A. Murray, R. Freathy, M.N. Weedon, A.R. Wood, H. Yaghootkar, J. Tyrrell.* University of Exeter Medical School, University of Exeter, Exeter, UK, Devon, United Kingdom.

Recent studies have identified common alleles associated with higher adiposity but lower risk of type 2 diabetes, hypertension and coronary artery disease. We tested the hypothesis that such “favourable adiposity” alleles could be used to test the causal role of higher adiposity uncoupled from its adverse metabolic consequences in obesity-related conditions. We used “favourable adiposity” alleles identified from a genome wide association study of body fat percentage in 451,000 people and a multivariate genome wide association study of body fat percentage and six biomarkers of metabolic disease. To test the role of higher adiposity uncoupled from its adverse metabolic consequences we constructed two high-adiposity genetic scores – one with, and one without, its adverse metabolic effects. We then used Mendelian randomisation approaches to test the role of the two types of adiposity in 91 traits in the UK Biobank. We identified 620 variants associated with higher adiposity at p<5x10-8, of which 14 had a “favourable adiposity” phenotype. After excluding known adiposity and metabolic traits such as diabetes and hypertension, we observed 17 associations between a “favourable adiposity” genetic score and conditions at $p<0.05$ when we would only expect 5 by chance. A 1 SD higher fat percentage without its adverse metabolic effects was associated with longer accelerometer-derived sleep duration (0.23 SDs 0.11-0.36), higher grip strength (0.36 SDs 95%CIs: 0.1-0.62) a higher risk of major depressive disorder (odds ratio 1.23 95%CIs: 1.03-1.47) and lower risk of endometriosis (0.41 95%CIs: 0.19-0.86). In comparison, a 1SD higher fat percentage with its adverse metabolic effects was associated with smaller effects on sleep duration (-0.04 SDs 95%CIs: -0.22-0.13), grip strength (0.043 95%CIs: -0.17-0.26) major depressive disorder (odds ratio 1.05 95%CIs: 0.91-1.21) and endometriosis (odds ratio 0.53 95%CIs: 0.33-0.85). In conclusion, higher adiposity is likely to alter the risk of several diseases and traits regardless of its adverse metabolic effects. These results imply that that the psychological and mechanical effects of higher adiposity may have stronger effects on certain conditions than the adverse metabolic effects. Careful characterisation of genetic variants associated with higher adiposity provides a genetic tool for testing the consequences of higher adiposity with and without its adverse metabolic effects. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

94

Tissue-wise sub-typing of complex trait based on genetics. *A. Majumdar[1], N. Cai[2,3], C. Giambartolomei[1], H. Shi[4], J. Flint[5], B. Pasaniuc[1,4].* 1) Department of Pathology & Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, 10833 Le Conte Ave, Los Angeles, CA, USA; 2) Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK; 3) European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SD, UK; 4) Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, USA; 5) Brain Research Institute, University of California, Los Angeles, CA, USA.

Analyzing gene-expression and GWAS data together can prioritize tissues or cell types relevant for a complex trait which is often unknown. If multiple tissue/cell-type specific causal pathways underlie an overall phenotype, the phenotype can be classified into sub-types stratified according to different causal mechanisms. For example, BMI can be regulated by genes expressed only in brain tissue, or adipose tissue, or both but with differential expression levels and an individual's BMI can be regulated more by genes specifically expressed in brain compared to adipose. We aim to learn about such hidden sub-phenotype structure of a complex trait for a group of individuals based on their marginal phenotype data and genotype data for sets of eQTLs, each corresponding to a set of genes specifically expressed in a tissue. We implement an expectation-maximization (EM) algorithm to estimate the posterior probability of an individual being assigned a sub-type (corresponding to a tissue/cell-type). Our simulation study shows that the accuracy of correctly inferring the sub-type depends on the heritability of each sub-type explained by the corresponding set of SNPs. For example, we consider 30,000 (30K) people with half of them having one sub-type and the other half having another. Suppose, two non-overlapping sets of 100 SNPs regulate each sub-type and explain 10% heritability of each. For this synthetic data, the AUC quantifying the classification accuracy is estimated as 60%. However, the individuals falling in the tail region of the posterior probability spectrum are of most interest. For 5.2K individuals whose posterior probability of being assigned either sub-type are > 65%, the AUC was estimated as 73% offering more reliable classification for these selected individuals. We applied our method in the UK Biobank cohort for BMI. We obtained two set of genes specifically expressed in cerebellum brain and subcutaneous adipose tissues from Finucane et al. (Nature Genetics, 2018). We took the eQTLs for the two set of genes from GTEx and filtered each set for LD. We ran the EM algorithm to infer sub-type for a set of 150K people with their BMI (adjusted for relevant covariates) and genotype data for the tissue-specific sets of eQTLs. Based on a 65% posterior probability cut-off, 11K individuals were assigned to cerebellum brain and 9K individuals to subcutaneous adipose. In summary, we present a novel approach to identify genetically defined sub-type of complex trait.

95

Refining the map of genomic disorder loci and associated driver genes by integrating microarray data from 102,257 genomes and exome sequencing of 37,269 individuals. *R.L. Collins[1,2,3], K. Mohajeri[1,2,3], J. Kosmicki[1,2,3], F.K. Satterstrom[1,2], J. Wang[4], J.Y. An[5], J. Buxbaum[6], D. Cutler[7], B. Devlin[8], S. Sanders[5], K. Roeder[4], H. Brand[1,2,3], M. Daly[1,2,3], M.E. Talkowski[1,2,3], The Autism Sequencing Consortium.* 1) Center for Genomic Medicine, Analytical and Translational Genetics Unit, and Department of Neurology, Massachusetts General Hospital, Boston, MA; 2) Program in Medical & Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of Harvard and M.I.T., Cambridge, MA; 3) Program in Bioinformatics and Integrative Genomics, Program in Biological and Biomedical Sciences, Department of Neurology, and Department of Medicine, Harvard Medical School, Boston, MA; 4) Department of Statistics and Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA; 5) Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA; 6) Seaver Autism Center for Research and Treatment, Department of Psychiatry, Friedman Brain Institute, and Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY; 7) Department of Human Genetics, Emory University School of Medicine, Atlanta, GA; 8) Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA.

Genomic disorders (GDs) are large, recurrent copy number variants (CNVs) that are often mediated by long homologous DNA segments. While individually rare, GDs collectively represent a penetrant source of risk for neurodevelopmental disorders (NDDs). Here, we aimed to expand the existing catalog of GD loci and identify plausible dominant genes contributing to their phenotypes by integrating microarray-based CNV data from 102,257 genomes (63,629 NDD cases & 38,628 controls) with coding variants from whole-exome sequencing (WES) of 37,269 individuals, including 6,429 NDD trios. We identified 51 GD segments associated with NDDs, including 18 “reciprocal” GDs (RGDs; significant for deletion & duplication). All RGD loci encompassed >11 genes, underscoring the challenge of pinpointing dominant gene(s) within these regions. In parallel, we identified 102 NDD-associated genes with burdens of damaging coding variants in the WES cohort (“WES-significant genes”). Intersection of the GD loci and WES-significant genes revealed at least one WES-significant gene in 21.6% (12/51) of GD loci – ~2.5-fold more than expected by chance (P<0.001) – and nominated three genes not previously associated with NDDs within GD segments (*SKI*, *BCL11A*, *SIN3A*) as well as four WES-significant genes that did not match previously proposed GD driver genes (*HDLBP*, *SUV420H1*, *GABRB3*, *CORO1A*). Genes within GD loci were ~1.5-fold enriched for damaging *de novo* mutations (DNMs) *en masse* in NDD probands (P<0.004), which persisted (at ~1.2-fold) after excluding WES-significant genes. We noted several differences between RGDs and non-homologous GDs: RGDs were less likely to harbor a WES-significant gene (P=0.072 vs. P=0.008), whereas the DNM burden enrichment remained in RGDs, but not GDs, after exclusion of WES-significant genes (1.29-fold vs. 1.05-fold). Finally, we conducted CNV-based association tests for all genes, identifying 47 and 33 genes enriched for rare deletions and duplications, respectively. CNV-associated genes were overrepresented among WES-significant genes (deletions: 11.5-fold, P<0.001; duplications: 8.0-fold, P=0.007. Collectively, deletion-associated genes also exhibited a significant burden of loss-of-function DNMs in NDD cases (3.1-fold, $P=6.62\times10^{-30}$), whereas duplication-associated genes did not (1.0-fold). These results demonstrate that rare CNVs and point mutations converge on a shared set of genes and pathways in NDDs, with intriguing distinctions among RGDs and duplications.

Copyright © 2018 American Society of Human Genetics. All rights reserved

96

Structural variation across human populations and families in more than 37,000 whole-genomes. *W. Salerno[1], A. Carroll[2], F.J. Sedlazeck[1], O. Krasheninina[1], G. Jun[3], A. Mansfield[1], J. Farek[1], Z. Khan[1], V. Menon[1], G. Metcalf[1], E. Boerwinkle[3,1], R.A. Gibbs[1].* 1) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 2) DNAnexus, Mountain View, CA; 3) The University of Texas Health Science Center at Houston, Houston, TX.

While the impact of small variation in well-characterized genomic regions is still being realized, it is clear that clinical-quality understanding of the full spectrum of genetic disease requires accurate assessment of large, complex variants across the entire genome for populations that span phenotypic space, including gender and ethnicity. Such structural variants (SVs) pose specific challenges with respect to detection accuracy, validation, allele reconciliation and the cost of these methods. Here we address these challenges and present the aggregation of multiple SV methods applied to whole-genome sequencing across a large human population and families. This data set comprises quality control metrics and six variant calls sets across more than 37,000 individuals that were short-read sequenced with multiple experimental protocols resulting in heterogeneous average coverage, insert sizes and sequencing platforms. Mapped with an NIH-compliant GRCh38 WGS protocol, all samples were processed with the latest cloud-deployments of Parliament and SURVIVOR, structural variant tools that scalably merge calls from five SV detection methods and provide project-level genotyping across reconciled alleles. Of the 6.2 million putative deletions, duplications and inversions identified, approximately 1,000,000 exist at an allele frequency greater than 0.03 and were of sufficient quality to genotype via an orthogonal computational method. The observed SV allele frequency spectrum is similar to those of SNVs and indels: most SVs in an individual are shared but rare across a large population. We will describe the computational and logistical challenges of executing this analysis at scale, quality control measures to account for data artifacts, how high-confidence structural variants are recapitulated in 111 family structures, and the aggregate degree of structural variation across multiple whole-genome contexts in large human populations.

97

Mapping and characterization of structural variation in 23,559 deeply sequenced human genomes. *I.M. Hall[1,2,3], H.J. Abel[1,3], D. Larson[1,3], C. Chiang[1], R. Layer[4,5], A. Regier[1,2], K. Kanchi[1], I. Das[1], N. Stitziel[1,2,3].* 1) McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO; 2) Department of Medicine, Washington University School of Medicine, St. Louis, MO; 3) Department of Genetics, Washington University School of Medicine, St. Louis, MO; 4) Department of Human Genetics, University of Utah, Salt Lake City, UT; 5) USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT.

A core goal of whole genome sequencing (WGS) based human genetics studies is to conduct comprehensive trait mapping analyses that include all forms of genome variation. The key challenges are variant detection and interpretation, and in both cases our knowledge of structural variation (SV) has lagged behind that of smaller-scale variants. There is a need to develop improved SV analysis approaches that scale to the massive sample sizes of ongoing studies, and to generate and disseminate high quality SV catalogs from large populations to improve variant interpretation in personal genomes. Here, we developed a fast and scalable toolkit (svtools) and cloud-based pipeline for assembling high quality SV maps – including deletions, duplications, inversions, mobile element insertions and complex rearrangements – via joint analysis of tens to hundreds of thousands of deeply sequenced human genomes. We used these tools to map and characterize SV in 23,559 ancestrally diverse genomes derived primarily from the NHGRI Center for Common Disease Genomics program. We identified 356,948 high-confidence SVs and show that our map is extremely high quality and high resolution, with low Mendelian error rates and 73% of SVs mapped to single base resolution. We describe the public release of site-frequency information for the 17,589 individuals permitted for aggregate-level sharing – a novel and valuable community resource. We next exploit this dense SV map to explore the contribution of SV to the burden of rare deleterious variation in the human population. On average, each individual harbors 2.3 rare high-impact gene-altering SVs, and SVs account for 6.3% of deleterious coding variants when compared to loss-of-function (LoF) mutations caused by single nucleotide (SNV) and insertion/deletion (indel) variants in a common set of samples. An independent genome-wide analysis using variant impact prediction tools suggests that SVs comprise 17.3% of rare deleterious mutations, a surprisingly large fraction of which are non-coding CNVs. Analysis of 184,113 ultra-rare SVs (mean 10.9 / person) reveals many dramatic examples of recent structural mutation including megabase-scale CNVs (0.01 / person), reciprocal translocations (0.002) and complex rearrangements involving three or more breakpoints (0.12). Finally, we will present a genome-wide dosage sensitivity analysis that reveals intriguing patterns of mutational intolerance across regulatory elements and cell-types.

Copyright © 2018 American Society of Human Genetics. All rights reserved

98

CNVs cause autosomal recessive genetic diseases with or without involvement of SNVs. *W. Bi[1,2], L. Wang[2], P. Liu[1,2], C. Shaw[1,2], H. Dai[1,2], L. Cooper[2], F. Xia[1,2], R. Xiao[1,2], X. Wang[1,2], L. Meng[1,2], A. Braxton[1,2], P. Ward[1,2], S. Peacock[1,2], F. Vetrini[2], W. He[2], T. Chiang[3], D. Muzny[3], R.A. Gibbs[1,3], A.L. Beaudet[1], A. Breman[1,2], J. Smith[1,2], S.W. Cheung[1], C. Bacino[1,4], C.M. Eng[1,2], Y. Yang[1,2], J.R. Lupski[1,3,4,5], B. Yuan[1,2].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Baylor Genetics Laboratory, Houston, TX; 3) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 4) Texas Children's Hospital, Houston, TX; 5) Department of Pediatrics, Baylor College of Medicine, Houston, TX.

Background: Autosomal recessive genetic disease traits are caused by defects on both alleles of a gene (i.e. biallelic variants). Involvement of copy number variants (CNV) in recessive disorders has been increasingly recognized with the advancing diagnostic technologies. Next generation sequencing can readily detect single nucleotide variants (SNV), insertions/deletions (indel) as well as homozygous/hemizygous gross deletions. Chromosomal microarray analysis (CMA) is a powerful genome-wide assay in CNV detection of genomic disorders, and has been also used for detection of intragenic CNVs through arrays with exonic coverage for disease genes. In this study, we investigated the different ways by which CNVs contributed to the molecular diagnosis of recessive-disorders. **Method:** We retrospectively investigated the CNVs in the patients who were subjected to whole exome sequencing (WES) and/or CMA testing at Baylor Genetics; most had neurodevelopmental problems. WES was performed between 2012 and 2018 (N=~12,000) with CNVs being detected from WES read depth data and by the concurrent Illumina SNP array serving as a quality measurement. CMA has been performed since 2004 (N=~80,000) mainly using customized Agilent arrays. **Results:** CNVs were identified to be the cause of recessive disorders in 87 patients. Of those patients, 75 had CNVs affecting the same gene on both alleles, and 12 had compound heterozygous CNV and SNV/indel alleles on the opposite chromosomes. The most frequently affected gene was *TANGO2* (8), followed by *VPS13B* (6), *HBA1/HBA2* (5), *WWOX* (4), *NPHP1* (4), *TBCK* (4), *CLDN1* (3), *SLCO1B3/SLCO1B1* (3), and 45 other genes (<3). The vast majority of CNVs were deletions except for one duplication. Among the 75 patients with bi-allelic CNVs, 69 had homozygous CNVs, 5 had overlapping but distinct CNVs, and 1 had non-overlapping CNVs *in trans*. Large deletions encompassing multiple genes were identified in four patients and considered as the second molecular diagnosis in addition to the recessive condition. **Conclusion:** Our study demonstrates the importance of CNVs in molecular diagnosis of recessive disorders. CNVs detected by WES and CMA contributed to recessive conditions with or without involvement of SNVs. Combined CNV and SNV/indel analyses are warranted to provide a precise genetic diagnosis. In addition, our finding of high frequency of homozygous CNVs (79%=69/87) highlights the role of recurrent CNVs in recessive diseases.

99

Assessing variants in genes of unknown significance: The quest for novel gene discoveries at the NIH Undiagnosed Diseases Program. *C. Lau[1], E. Macnamara[1], B. Pusey[1], N. Balanda[1], P. Kendrick[1], M. Malicdan[1], C. Toro[1], C. Tifft[1], W. Gahl[1], UDN. Undiagnosed Diseases Network[2], D. Adams[1].* 1) NIH Unidiagnosed Diseases Program, Bethesda, MD; 2) Undiagnosed Diseases Network (UDN).

The NIH Undiagnosed Diseases Program (UDP), which is now part of the Undiagnosed Diseases Network (UDN), enrolls patients with diseases that remain undiagnosed despite extensive diagnostic evaluation and clinical testing. Clinical whole exome or genome sequencing (WES or WGS) are frequently utilized for UDP cases, but the majority produce no clinical diagnosis. As a result, the UDP invests extensive resources in generating and evaluating variants in genes of unknown significance (GUS). Our protocol for generating and evauating GUSs includes five major components: 1.) Re-evaluation of clinical data collected during evaluation of the study participants; 2.) re-analysis of WES or WGS data using updated bioinformatics pipeline and annotation sources, and leveraging SNP microarray; 3.) re-interpretation of variants based on updated knowledge bases and new publications; 4.) follow-up laboratory testing for sequence variants or biochemical analytes; and 5.) functional studies to validate biological function of variants. We convene weekly variant assessment conferences where a panel of multi-disciplinary experts gather to assess and prioritize variants from WES or WGS. We have worked out ranking criteria that is facilitating the triage of the variants into these categories: (i) strong candidates: those likely to be the diagnosis pending confirmation of additional patients or validation through functional studies; (ii) moderate candidates: those that might prove crucial when more information of the genes become available or benefit from matching of additional cases using tools such as PhenomeCentral and the Matchmaker Exchange; and (iii) weak candidates: those on final review are thought to be unlikely to contribute to the presenting phenotype. To date, we have assessed over 212 cases of previously negative clinical WES or WGS cases using this workflow. Of these, we have reached a diagnosis and characterized disease-causing variants in 21 cases and further identified strong candidates for patient matching and research follow-up in 61 cases. In summary, the variant re-assessment strategies adopted by the NIH UDP currently are yielding diagnoses (10%), strong candidates (29%), and moderate candidates (50%), and importantly, tangible action plans, for a substantial portion of previously negative clinical exome cases. This represents an important step forward towards our goal of bringing a diagnosis to each patient who comes through the NIH UDP.

Copyright © 2018 American Society of Human Genetics. All rights reserved

100

Iterative reanalysis provides diagnostic avenue for previously unsolved rare and complex disease cases. *M. Velinder[1,2], J. Carey[3], L. Botto[3], R. Layer[1,2], B. Pedersen[1,2], A. Farrell[1,2], A. Andrews[3], P. Bayrak-Toydemir[4], R. Mao[4], A. Quinlan[1,2], G. Marth[1,2].* 1) USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT; 2) Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT; 3) Division of Medical Genetics, Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, UT; 4) Molecular Genetics and Genomics, ARUP Laboratories, Salt Lake City, UT.

Precision medicine critically relies on the ability to rapidly identify clinically actionable genetic variants in affected patients. However, most diagnoses made through clinical sequencing relate to well defined monogenic disorders for which causative genes have been established and known pathogenic variants have been identified. A significant number of patients who receive clinical sequencing remain undiagnosed, unable to benefit from targeted molecular genetic therapies that could alter their clinical management and potentially improve their quality of life. To address this disparity we have contributed our lab's bioinformatics and computational expertise to the Penelope Undiagnosed and Rare Disease Program. While the program has achieved an impressive diagnostic rate, several cases have remained undiagnosed despite exhaustive analysis by a team of trained bioinformaticians, medical geneticists and clinicians. We subsequently enrolled these cases into a research protocol where raw sequencing data was released to our lab and proceeded through our in-house alignment and variant calling pipeline. We then applied novel genomic analysis tools developed at UCGD including gene.iobio (gene.iobio.io), RUFUS, GEMINI and Peddy, among others. Using this iterative reanalysis workflow we were able to diagnose an additional number of cases. One of the cases that benefited from this workflow was a 5-year-old boy with subependymal gray matter heterotopia, pre- and post-natal growth delay, multiple congenital anomalies, bifid uvula and an unusual episode of altered behavior and consciousness. Upon applying our workflow we readily identified a de novo frameshift variant in the *SON* gene. Based on the predicted functional impact, mode of inheritance and clinical phenotypes consistent with those reported in the literature, we concluded that this *SON* variant was causative of the disorder. This particular case and the others we have solved by this approach demonstrate its utility. We propose that releasing clinical exomes into an academic research setting where novel genomic algorithms and tools can be applied in an iterative manner would provide a continued diagnostic avenue for unsolved cases. More broadly, this approach ensures that patients who lack an initial diagnosis are not abandoned by our healthcare system and preserves the promise of precision medicine for these patients and their families.

101

Description of a systematic approach for the reanalysis of clinical whole exome sequencing data. *H. Alsharhan, I. Ward, H. Ayoubieh, C. Applegate, F. Schiettecatte, D. Valle, A. Hamosh, N. Sobreira.* McKusick-Nathans Institute of Genetic Medicine and Department of Pediatrics. Johns Hopkins University, Baltimore, MD.

Whole exome sequencing (WES) is the primary method and a cost-effective tool for identifying the molecular basis of many rare Mendelian disorders. But, more than 50% of the clinical WES analyses are non-diagnostic. The field of genetics is dynamic: >200 novel disease gene discoveries are made yearly, variants are reclassified and ~200 novel syndromes are added to OMIM per year. Per Nambot et al. (2017), the reanalysis of 156 unsolved WES performed 12 months after the initial analysis yielded additional diagnoses in 15.5% of cases. For these reasons, a systematic approach to the reevaluation of clinical WES is needed. We described the approach that was implemented in our Genetics clinic at Johns Hopkins Hospital for the reanalysis of clinical WES. Patients are consented to have VCF files released from the clinical laboratory and stored in our database, PhenoDB. The database automatically converts VCF files to ANNOVAR files and performs standard analyses. Monthly, the system reviews the annotations for the genes and variants identified in the standard analyses, creates a new file with the updated information, and clinicians and counselors following the patients are informed of the updates by email. At any time, clinical providers can access the stored files or perform new analyses of the original files. We currently store 262 VCF files related to 117 probands without a diagnosis. We reanalyzed the VCF files of 86 unsolved cases sequenced from 2013 to 2018. 55 cases were sequenced more than a year prior to the reanalysis. In 61 cases, the proband and at least one additional family member were sequenced, and in 25 cases, only the proband. With our reanalysis approach, we identified candidate variants in known disease genes related to the phenotype of our patients, but not reported by the clinical laboratory, in 29 of the 86 (33.72%) cases. In other 5 of the 86 cases, the classification of a variant reported by the clinical laboratory changed allowing us to establish the diagnosis in our patients. Reanalysis also allowed for the identification of candidate novel disease genes, which were submitted to GeneMatcher. In one case, this has already led to the establishment of a new diagnosis. In conclusion, our approach allows continued reanalysis of clinical WES data and increases the solve rate over time. We plan to analyze the clinical unsolved cases together with research cases from the Baylor-Hopkins Center for Mendelian Genomics to increase the solve rate.

Copyright © 2018 American Society of Human Genetics. All rights reserved

102

Molecular diagnostic and clinical genomics outcomes of post-reporting reanalysis of exome data over time. *P. Liu[1,2], L. Meng[1,2], E.A. Normand[1], F. Xia[1,2], A. Ghazi[1], J. Rosenfeld[1], P. Magoulas[1,3], A. Braxton[1,2], P. Ward[1,2], H. Dai[1], B. Yuan[1], W. Bi[1,2], R. Xiao[1,2], X. Wang[1,2], T. Chiang[5], F. Vetrini[2], W. He[2], H. Cheng[2], J. Dong[2], C. Gijavanekar[2], V.R. Sutton[1,2,3], A.L. Beaudet[1], D. Muzny[5], R.A. Gibbs[1,5], J.E. Posey[1,3], S. Lalani[1,3,4], C. Shaw[1,2], C.M. Eng[1,2], J.R. Lupski[1,3,4,5], Y. Yang[1,2], Additional coauthors will appear in the presentation.* 1) Baylor College of Medicine, Houston, TX; 2) Baylor Genetics Laboratory, Houston, TX; 3) Texas Children's Hospital, Houston, TX; 4) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 5) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX.

Importance: Reassessment of existing exome data to continuously provide additional molecular diagnostic findings is novel in Medicine. Objective: To perform reanalysis of genomic data generated from clinical exome sequencing after clinical and genomic knowledge accrual. To investigate its impact on clinical decision making and patient management and associated challenges to such clinical implementation of genomics. Design: Systematic reanalysis was performed for two previously published clinical exome cohorts (N=250, 2011-2012, manual analysis; N=2000, 2012-2013, semi-automated analysis) using information from a data freeze in Dec 2017. From the first cohort, follow-up data regarding impacts on clinical management and challenges encountered while communicating updated molecular diagnostic findings were available for 42 out of the 64 cases that received a new diagnosis. Estimated costs for reanalysis were based on published data and experience from our clinical laboratory. Main Outcomes: The increment in molecular diagnostic yield by reanalysis and the systematic breakdown into disease genes involved, variant types, and molecular diagnostic categories. Benefits of employing semi-automated analysis for clinical practice sustainability. The impact of new molecular diagnostic findings on clinical management. Results: Post sign-out reanalysis increased the molecular diagnostic rates from 24.8% and 25.2% to 47.6% and 37.3% of referred patients, respectively. The most significant contributing reason for new diagnosis is new knowledge regarding novel disease genes and specific variants. Other major factors include the receipt of additional clinical phenotype information or additional samples for family segregation analysis, and the identification of pathogenic copy number variants. Among the 42 individuals who received new diagnosis and participated in follow-up studies, 30 received genetic counseling for the updated results, and clinical management impacted 57% (17/30) as a result. Semi-automated analysis improved analysis efficiency and sustainability. Conclusions: Reanalysis increases molecular diagnostic yield benefitting patients, families and physicians. Challenges exist in communicating updated results to the physicians and patients. Continued medical genetics knowledge accrual results in new understanding of existing raw genomics data warranting periodic cost-effective reanalysis to the benefits of all stakeholders including patients, families and physicians.

103

Association of clonal hematopoiesis of indeterminate potential with adverse outcomes in a diverse hospital-based biobank. *K.T. Nead, K.N. Maxwell, B. Wubbenhorst, R.L. Kember, J. Renae, M. Levin, H. Williams, D. Birtwell, D.J. Rader, S.M. Damrauer, K.L. Nathanson.* University of Pennsylvania Perelman School of Medicine, Philadelphia, PA.

Background: Clonal hematopoiesis of indeterminate potential (CHIP) is defined as the presence of age related acquired mutations resulting in the clonal expansion of blood cells in the absence of other hematologic abnormalities. Previous studies have supported an association of CHIP with hematologic malignancy, cardiometabolic disease, and all-cause mortality. Here we characterize CHIP in a large hospital-based biobank. **Methods**: We used electronic health record and whole-exome sequencing data from peripheral blood cells in 10,996 individuals enrolled in the Penn Medicine BioBank. We evaluated 74 genes recurrently mutated in hematologic cancers for known pathogenic mutations. We examined the association of CHIP with hematologic malignancy and overall survival in multivariable regression models adjusting for age at enrollment, age at enrollment squared, genetic ancestry, and gender. **Results**: We identified 388 individuals (3.5%) with 426 CHIP mutations. Individuals with CHIP were older (mean 70.0 years, standard deviation [SD] 12.6) and less likely to be of African ancestry (11%) compared to those without CHIP (mean 61.2 years, SD 12.6; 20% African ancestry; p<0.001 for both). CHIP frequency increased with age: 70-79 (6.2%), 80-89 (8.6%), and ≥90 years (11.8%). The most frequently mutated genes were *DNMT3A* (n=135), *TET2* (n=45) and *ASXL1* (n=35). We found an association between CHIP and an increased risk of prevalent hematologic malignancy (odds ratio, 4.56; 95% confidence interval [CI], 2.90-7.16). We further found an association between CHIP and decreased overall survival (hazard ratio [HR], 1.61, 95% CI, 1.32-1.96), which was dose-dependent (1 CHIP mutation, HR, 1.47, 95% CI, 1.19-1.82; >1 CHIP mutation, HR, 3.40, 95% CI, 2.11-5.50; p-trend <0.001). The association with overall survival differed significantly by gene (p-heterogeneity=0.041). Phenome-wide association studies revealed differential phenotype associations by gene. We observed no association between CHIP and overall survival among those with African genetic ancestry (HR, 0.80, 95% CI, 0.36-1.80; n=2,133), which significantly differed from other ancestral groups (p<0.001). **Discussion**: In a large, diverse, practice-based biobank we support the association of CHIP with hematologic malignancy and all-cause mortality. We observed differences in the association of CHIP with survival and non-hematologic phenotypes by gene and genetic ancestry.

Copyright © 2018 American Society of Human Genetics. All rights reserved

104

Human glioblastoma arises from the subventricular zone harboring low-level driver mutations. *J.H. Lee[1,2], J.E. Lee[2,3], J.Y. Kahng[2,4], S.H. Kim[5], S.J. Yoon[6], J.Y. Um[7], W.K. Kim[2], J.S. Park[2], J.K. Lee[2], J. Park[6], E.H. Kim[6], J.H. Lee[6], J.H Lee[4], W.S. Chung[4], Y.S. Ju[2], S.H. Park[2,7], J.H. Jang[6], S.G. Kang[6], J.H. Lee[2,8].* 1) Department of Radiation Oncology, Seoul National University Hospital, Republic of Korea; 2) Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea; 3) Department of Internal Medicine, College of Medicine, Chungnam National University, Republic of Korea; 4) Department of Biological Sciences, KAIST, Republic of Korea; 5) Department of Pathology, Brain Korea 21 project for medical science, Yonsei University College of Medicine, Republic of Korea; 6) Department of Neurosurgery, Brain Tumor Center, Severance Hospital, Yonsei University College of Medicine, Republic of Korea; 7) Department of Bio and Brain Engineering, KAIST, Republic of Korea; 8) Center for Synaptic Brain Dysfunctions, Institute for Basic Science, Republic of Korea.

Glioblastoma (GBM) is a devastating and incurable brain tumor, with a median overall survival of 15 months from time of diagnosis. Identifying the cell of origin that harbors mutations driving GBM could provide a fundamental basis for understanding disease progression and developing novel treatments. Given that the accumulation of somatic mutations is implicated in gliomagenesis, studies have suggested that neural stem cells (NSCs), with their self-renewal and proliferative capacities, in the subventricular zone (SVZ) of the adult human brain may be the cells from which GBM originates. However, there is a lack of direct genetic evidence thereof in human GBM patients. Here, we describe direct molecular genetic evidence from patient brain tissue and genome-edited mouse models that show astrocyte-like NSCs in the SVZ to be the cell of origin that harbors the driver mutations of human GBM**.** First, we performed deep sequencing of triple-matched tissues, consisting of i) radiologically and pathologically normal SVZ tissue away from the tumor mass, ii) tumor tissue, and iii) normal cortical tissue (or blood), from 28 patients with primary GBM (isocitrate dehydrogenase-wild type) or other types of brain tumors. In doing so, we found that normal SVZ tissue away from the tumor in 56.3% of primary GBM patients contained low-level GBM driver mutations (down to ~1% of the mutational burden) that were observed at high levels in their matching tumors. Moreover, via single cell sequencing and single cell cloning, we found that the NSCs harbring driver mutations in the SVZ clonally evolve to tumor. Laser microdissection analysis of patient brain tissue showed that mutations are enriched in the astrocyte ribbon area. Furthermore, using genome editing of the postnatal and adult mouse model, we discovered that NSCs with driving mutations migrated away from the mutated SVZ site and then formed the high grade malignant glioma in the distant cortical region through aberrant growth of oligodendrocyte precursor lineage. Altogether, our results highlight NSCs in human SVZ tissue as the cell of origin that harbors the driver mutations of GBM. Furthermore, NSCs harboring cancer-driving mutations can remain in the human SVZ even after the development and resection of the tumor.

105

Somatically acquired variants contaminate public germline variant population databases. *B. Coffee, H. Cox, M. Jones, J.P. De La O, S. Manley, L. Esterling, K. Bowles, B. Roa.* Myriad Genetic Laboratories, Inc., Salt Lake City, UT.

Background: Germline *TP53* pathogenic variants (PVs) cause Li-Fraumeni syndrome (LFS), a severe, early-onset familial cancer syndrome with significantly reduced life expectancy. Previously, we demonstrated that somatically acquired *TP53* PVs are detected during hereditary pan-cancer gene panel testing, often in older adults at low NGS read frequencies, though the NGS read frequency can increase to a level similar to what is expected for a germline PV. This observation suggests that a significant proportion of these *TP53* PVs are due to age-associated clonal hematopoiesis. Population databases, such as gnomAD, contain exome and genome sequences from older adults, increasing the risk that they contain somatically acquired PVs due to age-associated clonal hematopoiesis that may inflate the observed pathogenic allele frequency of PVs in some genes. As this data is often used in germline variant classification, we evaluated the presence of somatic variants in gnomAD. **Methods:** The prevalence of PVs in gnomAD was determined for *TP53* and *ASXL1*, two genes known to demonstrate age-associated clonal hematopoiesis, and compared to the prevalence of the corresponding gene-associated diseases. NGS read frequencies for these PVs were evaluated relative to the ratio expected for a heterozygous germline PV (~50:50). **Results:** For *TP53*, 40/123,000 (~1/3000) individuals included in gnomAD carried a PV. This is ~7 times greater than the ~1/20,000 general population prevalence estimated for LFS. For *ASXL1*, 409/123,000 (~1/300) individuals carried a PV, which is significantly greater than expected for the ultra-rare Bohring-Opitz syndrome (<50 individuals described in medical literature). NGS read frequencies for PVs in *TP53* and *ASXL1* were skewed from the ratio expected for germline variants, suggesting that the increased PV prevalence in gnomAD is due to the presence of somatically acquired variants. **Conclusions:** These results demonstrate that the gnomAD population database likely harbors somatically acquired variants that are due to age-associated clonal hematopoiesis. This contamination falsely inflates the number of PVs in these genes potentially leading to an overestimation of the prevalence of disease. As allele prevalence data in gnomAD can be used as evidence in clinical variant classification, it is critical that testing laboratories include safeguards for somatic contamination to prevent incorrect variant classification and, ultimately, inappropriate patient management.

Copyright © 2018 American Society of Human Genetics. All rights reserved

106

Driving mosaicism: Presence of somatic driver variants in population databases and effect on rare Mendelian diseases. *V. Avramovic[1,2,3], M. Brkic[1,4], M. Tarailo-Graovac[2,3].* 1) Department of Neurobiology, Institute for Biological Research, University of Belgrade, Belgrade, Serbia; 2) Departments of Biochemistry, Molecular Biology and Medical Genetics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; 3) Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada; 4) VIB Center for Inflammation Research, Ghent University, Ghent, Belgium.

Exome Aggregation Consortium (ExAC) with exome data for 60,706 unrelated individuals, gnomAD dataset with 123,136 exome and 15,496 whole-genome sequences, as well as the BRAVO database with 62,784 genomes are publicly available sources of population variant information. These resources on “normal variation” play crucial role in accurate interpretation of variant pathogenicity in individuals affected with genetic diseases, rare diseases in particular. Previously, others and we reported on unexpected presence of *ASXL1* pathogenic variants in the ExAC population. The heterozygous *ASXL1* variants had been implicated in severe pediatric autosomal dominant condition, Bohring–Opitz syndrome. Closer inspection of the ExAC data revealed that those pathogenic *ASXL1* variants are somatic rather than germline, highlighting the importance of considering somatic mosaicism in variant interpretation. Thus far, the *ASXL1* and the related *DNMT3A* (Tatton-Brown-Rahman syndrome) are the two main examples of the effect of somatic mosaicism on interpretation of Mendelian conditions. To gain further insights, we compiled and manually curated the list of 1476 genes implicated in autosomal dominant diseases and assessed the prevalence of dominant likely/pathogenic variants (as per ClinVar classification + supported by at least one PMID record) in the above population datasets. From thousands of likely/pathogenic variants, only a subset displayed evidence of somatic mosaicism. The vast majority of the variants with clear evidence of somatic mosaicism affected the hematopoietic genes (n=20) suggesting clonal hematopoiesis of cells with driver mutations. Of the 160 genes that are recurrently mutated in hematologic cancers (including *ASXL1* and *DNMT3A*), 53 result in autosomal dominant conditions mostly due to *de novo* germline variants. Our data shows that beyond the reported *ASXL1* and *DNMT3A*, somatic variants are likely to affect interpretation of diseases like: Noonan and related syndromes (*BRAF*, *CBL*, *KRAS*, *NRAS, PTPN11*, *RIT1*), Weaver syndrome (*EZH2*), Leukoencephalopathy (*CSF1R*), various forms of immunodeficiencies (*CARD11*, *IKZF1*), as well as other known and yet to be discovered Mendelian conditions associated with germline variation of hematopoietic genes. Furthermore, our results show the importance of continued efforts to expand the population databases for variant classification, novel discoveries as well as uncovering pathogenic mechanisms common to rare Mendelian diseases and cancers.

107

An integrated approach to functionally characterize GWAS. *A.C. Joslin[1], D.R. Sobreira[1], G. Hansen[1], N.J. Sakabe[1], I. Aneas[1], L. Montefiori[1], D. Lehman[2], M. Nobrega[1].* 1) Human Genetics, University of Chicago, Chicago , IL; 2) Department of Medicine, University of Texas Health Science Center, San Antonio, TX.

The mechanistic dissection of Genome Wide Association Studies (GWAS) remains a challenge. In order to identify causal mechanisms behind non-coding associations, we devised a platform to screen for SNPs that disrupt enhancer activity in obesity GWAS associated loci and implicate likely obesity risk genes in both a primary human adipose cell line across differentiation, as well as human iPSC derived hypothalamic neurons across differentiation. We applied this methodology to 97 loci associated with Body Mass Index (BMI). Using a massively parallel reporter assay (MPRA), we tested 2,396 variants in high LD with the 97 lead GWAS variants for their ability to drive allele specific enhancer activity and identified 94 variants across 40 of these loci with allele-specific enhancer modulating properties in adipose and/or brain cells. The enhancers identified in these obesity-associated loci are highly enriched for transcription factor motifs critical for adipogenesis and maintenance of circadian rhythm – two homeostatic mechanisms important for body weight regulation. We tied these enhancers to their target genes using a combination of in-situ promoter capture HiC in human adipose and hypothalamic neurons across differentiation as well as GTEx eQTL data. To better understand the developmental timepoint in which these variants may impart their phenotypic effect, we overlaid this information with chromatin accessibility. In combination, these data will provide strong genetic support for, or guide researchers away from, genes playing a role in obesity risk. We demonstrate that a large number of candidate causal SNPs for these associations lie within enhancers with activity in brain and adipose tissue, highlighting an interesting pleiotropic nature of these genetic associations. We describe how our integrative platform revealed two independent GWAS loci physically and genetically converging to regulate the same target gene, *SH2B1,* a gene critical for insulin and leptin signaling. Our experimental methodology highlights the importance of integrating multiple datasets to understand mechanisms underlying GWAS associations with complex diseases.

Copyright © 2018 American Society of Human Genetics. All rights reserved

108

Simultaneous analysis of open chromatin, promoter interactions and gene expression across a 24hr period in primary T-cells implicates GWAS SNPs with causal genes. *J. Yang[1], A. McGovern[2], P. Martin[2], K. Duffus[2], M. Imran[1], P. Fraser[4], M. Rattray[1], S. Eyre[2,3].* 1) Division of Informatics, Imaging & Data Sciences, University of Manchester; 2) Division of Musculoskeletal and Dermatological Sciences, The University of Manchester, UK; 3) NIHR Manchester BRC Central Manchester NHS Foundation Trust; 4) Frank and Yolande Fowler Endowed Chair of Biological Science, Center for Genomics and Personalized.

Background. One of the consistent findings from GWAS studies in complex diseases is how the vast majority of associated variants are found within non-coding regions of the genome, often residing in cell type specific enhancers. Here we combined 3 methodologies, simultaneously in the same primary cells, across a 24hr period following cell stimulation, to determine dynamic active DNA, linked to promoters and the effect on gene regulation. We used primary T-cells, stimulated with CD3/CD28 beads and created ATAC-seq libraries, HiC libraries, Capture HiC libraries and nuclear RNA -seq libraries at 6 time points across a 24 hour period. **Methods**. Primary human CD4+ T-cells were isolated from PBMCs (EasySep) and stimulated with CD3/CD28 Dynabeads over a period of 24-hours. Hi-C libraries were generated from fixed CD4+ T-cells from three individuals, pooled at the lysis stage to give ~25 million cells. CHi-C enriched for promoters, from dynamic or GWAS implicated genes. Nuclear RNA-seq was used to quantify nascent transcription to determine changes through time. Five million CD4+ T-cells were harvested, stored in Qiagen RNAprotect solution and the nuclear RNA isolated. Libraries for RNA-seq were prepared using the NEB Next Ultra Directional RNAseq Protocol using 100ng of nuclear RNA as input. ATAC-seq libraries were generated from 50,000 CD4+ T-cells from three individual samples with Illumina Nextera DNA Sample Preparation Kit. Each ATAC-seq and RNA-seq libraries were sequenced on half a lane of an Illumina HiSeq2500, CHiC on a full lane. For the CRISPR/Cas9 analysis, 3 guides were designed within the intronic enhancer of the COG6 gene and virally transduced into a HEK-p300 stable cell line. RNA was extracted and expression of COG6 and FOXO1 was determined by qPCR **Results**. Linking dynamic ATAC-seq peaks to their target gene through physical interactions, we show how RNA expression is correlated to the number of enhancer interactions, that genes that physically interact are enriched for correlated expression and enhancers that are physically linked to promoters and correlate with gene expression are enriched for autoimmune GWAS SNPs, often over 300kb from their target gene, providing compelling, often novel, evidence for the target GWAS gene. We go on to confirm how an enhancer within the *COG6* gene, shown to interact and correlate with the expression of *FOXO1,* 800kb away, can influence *FOXO1* gene expression through CRISPR/CAS9 technology.

109

Chromatin structure guided approach to evaluate genetic risk reveals oligodendrocyte intrinsic genetic contribution to multiple sclerosis. *O. Corradin[1], D.C. Factor[2], P.A. Hall[1], S. Nisraiyya[1], A. Barbeau[1], P.C. Scacheri[2,3], P.J. Tesar[2,4].* 1) Whitehead Institute for Biomedical Research, Cambridge, MA; 2) Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH; 3) Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH; 4) National Center for Regenerative Medicine, Case Western Reserve University, Cleveland, OH.

Multiple sclerosis (MS) is an autoimmune disorder characterized by attack on myelinating oligodendrocytes within the central nervous system. >150 risk loci have been identified, and most are thought to perturb gene enhancer activity in B and T cells. Current MS therapies focused on preventing further immune attack are largely palliative. Patients continue to face progressive disability due to a failure of oligodendrocyte progenitor cells (OPCs) to regenerate new myelin suggesting other cellular pathologies. We developed an approach that utilizes 3D chromatin structure to evaluate DNA variants that physically interact with MS risk genes as part of a genes' regulatory circuitry. We identify "outside variants," SNPs in weak LD with GWAS risk SNPs that physically interact with the same target promoter and significantly modify clinical risk. We hypothesized that the chromatin state at outside variants could be used to identify gene targets of MS risk loci and predict novel cell types that contribute to MS pathology. To test the robustness of our approach, we integrated MS risk loci with Hi-C and ChIP-seq data from 15 blood cell types. Significant outside variants were found at 72% of risk loci. As expected, many loci were predicted to dysregulate gene targets in T-cells. However, a subset were predicted to modulate risk through disrupting gene enhancer activity in non-lymphoid cell types. For example outside variants at 16q24 are predicted to dysregulate *IRF8* expression in the monocytes. This prediction is supported by studies in mice, in which knock out of IRF8 in monocytes, and not T-cells, confers resistance to MS. Excitingly, when we applied the approach to brain tissues, we identified 2 risk loci predicted to dysregulate genes involved in transcriptional pause release. Outside variants at these two loci were highly enriched for H3K27ac in oligodendrocytes, potentially implicating this cell type in MS risk. High throughput screening and chemical genetic approaches functionally validated inhibition of transcriptional pausing as a dominant pathway blocking oligodendrocyte generation and myelination from OPCs. Using microarray and immunohistochemistry we demonstrate that these pause release factors are dysregulated in MS patient brain tissue. These data implicate oligodendrocyte intrinsic aberrations in MS risk and suggest that therapeutic modulation of transcriptional elongation in the brain may be an effective strategy to overcome remyelination block in MS.

Copyright © 2018 American Society of Human Genetics. All rights reserved

110

Functional annotation of IBD GWAS loci by enrichment analysis of differentially expressed genes identifies loci with shared biological effects and defines individual genetic immune landscapes. *R. Kosoy[1,2], A. Hart[3], A.F. Di Narzo[4], S. Huang[5], K. Hao[1,2], H. Irizar[6], B. Losic[1,2], A. Castillo[7], J. Rogers[7], A. Atreja[7], A. Hurley[1,2], L.A. Peters[1,2,8], J.R. Friedman[3], F. Baribaud[3], C. Monast[3], C. Brodmerkel[3], S. Plevy[3], J.F. Colombel[7], M. Dubinsky[9,10], J. Cho[9], B.E. Sands[7], E.E. Schadt[1,2,8], A. Kasarskis[1,2], C.A. Argmann[1,2], M. Suarez-Farinas[1,5].* 1) Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York City, NY; 2) Icahn Institute for Genomics and Multiscale Biology, New York City, NY; 3) Janssen R&D, Spring House, PA; 4) Antonio Di Narzo Consulting, Kraków , Poland; 5) Population Health Science and Policy, Icahn School of Medicine at Mount Sinai , New York City, NY; 6) Division of Psychiatry, University College London, London, UK; 7) Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York City, NY; 8) Sema4, a Mount Sinai venture, Stamford, Connecticut, USA; 9) Pediatric GI and Hepatology , Icahn School of Medicine at Mount Sinai , New York City, NY; 10) Susan and Leonard Feinstein IBD Clinical Center, Icahn School of Medicine at Mount Sinai , New York City, NY.

While over 200 IBD risk loci have been identified, their function in disease development is largely unknown, as just few loci have been mapped to individual genes. Alternatively, we can elucidate biological impact of the etiologic variants without having to identify the genes directly affected by these variants. We functionally annotated IBD loci using paired genotype and expression data from 1120 patients in Mount Sinai Crohn's and Colitis Registry cohort. We generated differentially expressed gene (DEG) signatures for each IBD locus using RNA-seq expression data from blood samples (n=1003), from inflamed (n=700) and non-inflamed (n=1715) biopsies from ileum, rectum, and non-rectum colon. Each locus was tested using linear mixed-effect additive models with core covariates including genetic ancestry, demographic information, and current medication use. Since most of the resulting IBD locus-based DEG signatures contained few genes at an FDR <0.1, we utilized enrichment analyses of the top DEG signatures using function-based gene-sets. Thus, we functionally annotated each locus in a tissue-specific manner according to the biological consequences of carrying IBD susceptibility alleles. Among the 226 IBD loci tested, we observe diverse patterns of functional annotation depending on the tissue, with unique pattern of upregulated and downregulated processes. Highlighted processes include TNFα signaling via NF-κB, Interferon α and Interferon γ responses, KRAS signaling, and epithelial mesenchymal transition processes in blood, inflamed, and non-inflamed biopsies, revealing GWAS loci with shared function and, likely mechanisms, of contribution to IBD. Additionally, we derived genetic predisposition scores by aggregating the functional annotations per IBD-associated allele across all IBD loci, providing a measure of inflammatory predisposition in a tissue- and pathway-specific manner. When tested for association with IBD subtypes, we observe that compared to non-IBD patients, genetic landscapes in CD and UC patients have lower functional annotation for processes involved in MTORC1 signaling and MYC targets in blood, higher estrogen response and glycolysis processes in inflamed intestinal biopsies, and lower hypoxia and p53-pathways processes in non-inflamed biopsies. Comparing to CD, UC patients' genetic landscape has lower TNFα signaling via NF-κB functional annotation in blood. This approach can be generalized for any disease with complex genetic associations.

111

Functional characterization of the 14q24 renal cancer susceptibility locus implicates SWI/SNF complex member DPF3. *L. Machado Colli, L. Jessop, T. Myers, M. Machiela, J. Choi, M. Purdue, K. Yu, K. Brown, S. Chanock.* DCEG, National Cancer Institute - NIH, Rockville, MD, MD.

Using the Massively Parallel Reporter Assay (MPRA), a powerful tool to identify promising regulatory regions under GWAS signals, one of the strongest signals identified mapped to the 14q24 Renal Cancer risk region. We now show that rs4903064 is a transcriptional enhancer with C allele preferential activity and have confirmed the differential effect for the C allele 'risk' allele by luciferase assay in three RCC cell lines. Using electromobility shift assays (EMSA), the T allele of rs4903064 showed preferential binding, whereas motif analysis revealed that the T allele binds the transcriptional repressors IRX2 and IRX5. Also, motif analysis shows that the C allele is predicted to create a HIF1A binding site. These results suggest that the risk rs4903064-C allele has higher transcriptional activity than the T allele, which is consistent with TCGA and IARC eQTL analysis showing rs4903064-C allele associated with higher expression of the double PHD Fingers 3 gene (*DPF3)*. DPF3 is member of SWI/SNF complex, in which one or more genes can be somatic mutated in 40% of RCC, demonstrating the importance of this complex for RCC. To understand the effect of DPF3 on RCC, we create stable DOX-inducible DPF3 isoforms (a or b), separately in distinct cell lines: two RCC- ACHN and UOK-121 and one normal kidney cell line- HK2. The results suggest an oncogenic effect of both DPF3 isoforms; in cell lines over expressing DPF3a or DPF3b, a higher growth rate was observed compared to controls. RNA-seq analysis of DPF3a or DPF3b over expressing cell lines showed deregulation of *CEMIP*, an apoptotic gene, and IL23R, which is involved in immune-evasion. Knockdown of CEMIP by si-RNA reduced the effect of DPF3a and DPF3b on growth rate, suggests that at least part of the DPF3 effect on cell growth can be attributed to *CEMIP* deregulation. Using AnexinV and 7-AAD flow cytometery and Caspase-3 and PARP Western Blot, we found that DPF3a/b overexpression cell lines shows reduced apoptosis levels. ELISA of intrinsic apoptosis components suggested that DPF3a/b overexpression reduces BAX from mitochondrial membrane. In summary, the RCC risk C-rs4903064 at the 14q24 locus increases expression of SWI/SNF complex gene DPF3, leading to dysregulation of CEMIP and a reduction of apoptosis.

Copyright © 2018 American Society of Human Genetics. All rights reserved

112

Prostate cancer risk SNP rs10993994 is a trans-eQTL for *SNHG11* mediated through *MSMB*. *M. Bicak, X. Wang, X. Gao, M. Middha, R.J. Klein.* Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY.

GWAS studies have successfully identified numerous prostate cancer risk SNPs. However we still lack understanding of how these SNPs function to alter an individual's risk of prostate cancer. Specifically, how they affect the expression of remote genes as trans eQTLs remains unknown, partially due to insufficient sample sizes in prostate eQTL studies. To increase sample size, we conducted a meta-analysis eQTL association analysis on existing data from prostate tumor, adjacent normal from prostate cancer patients, and normal prostate tissue from individuals without prostate cancer, resulting in a sample size of 496 tumor and 602 normal samples, where we tested if 117 SNPs associated with prostate cancer risk are correlated with gene expression changes in nearby and remote genes. Previously unreported cis-eQTLs were found, such as *NOTCH4* and rs30967026. Furthermore, 46 trans-eQTLs were found at an FDR of 5%, for which 9 SNPs seem to be trans-eQTLs for a minimum of 2 and a maximum of 12 genes, suggesting these are prostate cancer associated trans-eQTL hotspots. These include: (i) SNPs that are at loci for prostate-secreted proteins, such as rs10993994 (*MSMB*) and rs17632542 (*KLK3/PSA*), as well as (ii) SNPs that are at loci with transcription factors, such as rs12653946 which is confirmed as cis-eQTL with *IRX4*; rs1512268 which is near *NKX3-1*, a transcription factor known to play a role in prostate cancer; and rs339331 which has been shown to associate with *RFX6* activity. Mediation testing was applied to trans-eQTL hotspots; we demonstrate that *MSMB* is a cis-acting potential mediator for *SNHG11* ($p < 0.01$). Using isogenic LNCaP cell lines where one copy of the heterozygous rs10993994 was removed by CRISPR/Cas9 editing, we found that the allele correlating with an over 100-fold increase in *MSMB* expression resulted in a 5-fold increase in *SNHG11* expression. Colocalization analysis confirms that the same set of SNPs associated with *MSMB* expression are associated with *SNHG11* expression (posterior probability of shared variants is 66.6% in tumor and 91.4% in normal). Furthermore, this analysis demonstrates that the variants driving *MSMB* expression differ in tumor and normal cells, suggesting evidence of regulatory network rewiring in the transformation process.

113

Functional analysis revealed a prostate-cancer risk associated germline variant that modulates PSA activity and glycosylation. *S. Srinivasan[1,2], A. Buckle[3], The PRACTICAL Consortium[4], The APCB[1,2], J. Clements[1,2], J. Batra[1,2].* 1) Queensland University of Technology, Brisbane, Queensland, Australia; 2) Australian Prostate Cancer Research Centre-Queensland, Institute of Health & Biomedical Innovation, Translational Research Institute, Queensland University of Technology, QLD, Australia; 3) Biomedicine Discovery Institute, Monash University, Victoria, Australia; 4) Centre for Cancer Genetic Epidemiology, Cambridge, UK.

Background: Prostate cancer is the second most common cancer in men world-wide with a complex genetic etiology. Prostate-specific antigen/kallikrein-3 is the primary non-invasive diagnostic biomarker. PSA play pivotal role in prostate cancer progression. We tested 383 SNPs spanning 420 Kb at 19q13.3 kalllikrein (KLK) locus in 49,941 prostate cancer cases and 32,001 disease-free controls. We identified three independent prostate cancer risk associated germline variants in the *KLK* locus: two intronic variants, rs62113212 and rs266883; and a nonsynonymous variant rs61752561 to be independntly associated with prostate cancer risk. Interestingly, the nonsynonymous rs17632542 SNP (Ile161Thr substitution) is in high linkage disequilibrium with the most significant SNP at this locus, rs62113212 was previously shown by us to alter PSA expression and enzyme function. However the biological mechanism that underlies the association of the second non-synonymous rs61752561:G>A with prostate cancer risk is not yet known. We thus aim to understand the contribution of exon-3 *KLK3* rs61752561 SNP with prostate cancer risk at this locus. **Methods:** Structural, biochemical and gene over-expression studies were utilized to evaluate the effects of the germline variant on glycosylation, proliferation and migration of PCa cells. **Results and conclusions:** Prostate cancer risk associated rs61752561: G>A (Odds-Ratio=0.85, $P=1.7X10^{-10}$, minor-allele frequency=0.04) SNP leads to Asp84Asn substitution. Stable over-expression of wild-type PSA in PC3 cells increased proliferation and migration compared to SNP isoform. Structural analysis suggested the rs61752561 SNP is located at the start of the kallikrein loop, a characteristic loop important for controlling substrate recognition and substrate binding. Interestingly, a deglycosylation assay indicated the SNP to create an extra-glycosylating site which may hamper the complexing ability with serum inhibitors and thus influence clinically measured free/total PSA, which was confirmed in clinical samples (n=900). Our study show that the rs61752561 SNP may be a functional contributor to prostate cancer risk. Thus the clinically measured free/total PSA needs to be carefully reassessed in men carrying the SNP allele and replaced with precise diagnostic methods accounting for the allele specific affects of the rs61752561 SNP.

Copyright © 2018 American Society of Human Genetics. All rights reserved

114

Chromatin interactome mapping identifies target genes at 147 independent breast cancer risk signals. *S.L. Edwards[1], H. Sivakumaran[1], J. Beesley[1], M.M. Marjaneh[1], K.M. Hillman[1], S. Kaufmann[1], L. Fachal[2], D.F. Easton[2,3], A.M. Dunning[2], G. Chenevix-Trench[1], J.D. French[1].* 1) QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia; 2) Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK; 3) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.

Genome-wide association studies (GWAS) for breast cancer have identified 196 independent signals associated with increased risk. The majority of risk-associated variants within these signals fall in regulatory sequences, such as enhancers, that control gene expression. In this study, we perform *in situ* Capture Hi-C using a high-resolution breast cancer susceptibility Variant Capture array (VCHi-C), which includes probes to cover all fine-mapped candidate causal variants. We apply VCHi-C and Promoter Capture Hi-C (PCHi-C) to link risk variants to their target genes in six human mammary epithelial and breast cancer cell lines. We use the CHiCAGO pipeline to assign confidence scores to interactions, apply a strict threshold, and identify between 10-27,000 high-confidence interactions per cell type. Hierarchical clustering of CHiCAGO interaction scores stratifies cell lines by estrogen receptor status, suggesting cell-type specificity of the interactomes. Global analysis of promoter-interacting regions (PIRs) shows strong enrichment for cell-type specific accessible chromatin (ATAC-seq, DNAse-seq), histone marks for active enhancers (e.g. H3K27ac, H3K4me1) and transcription factor binding motifs (e.g. GATA3, FOXA1), supporting the regulatory potential of many PIRs. Similarly, analysis of variant-interacting regions (VIRs) shows enrichment of expressed genes in the relevant cell type. In total, reciprocally validated CHiCAGO-identified interactions results in 647 candidate target genes at 147 breast cancer risk signals. To further prioritise the CHi-C-derived chromatin interactions, we use a recently developed Bayesian framework, to fine-map the direct contacts. Importantly, the combined PCHi-C and VCHi-C contact fine-mapping enables us to prioritize 1832 out of 7375 highly-correlated risk variants at 118 signals, including 51 signals which are potentially reduced to less than five risk variants, and lowers the total number of candidate target genes to 393. One example which makes evident the utility of this dual approach is the 16q24 risk region, where contact fine-mapping decreases the number of statistically indistinguishable variants from 85 to 12, and the predicted protein-coding target genes from 10 to 3 (*MTHFSD*, *FOXC2* and *FOXL1)*. Our results demonstrate the power of combining genetics, computational genomics and molecular studies to streamline the identification of key variants and target genes at GWAS-identified risk regions.

115

Identification of a missense variant in the WFS1 gene that causes a mild form of Wolfram syndrome and is also associated with risk for type 2 diabetes in Ashkenazi Jewish individuals. *V. Bansal[1], B.O. Boehm[2,3,4], A. Darvasi[5].* 1) University of California San Diego, La Jolla, CA; 2) Department of Internal Medicine I, Ulm University Medical Centre, Ulm, Germany; 3) Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore; 4) Imperial College London, UK; 5) Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Givat Ram, Jerusalem, Israel.

Wolfram syndrome is a rare, autosomal recessive syndrome characterized by juvenile-onset diabetes and optic atrophy. Wolfram syndrome is caused by bi-allelic mutations in the WFS1 gene. In a recent targeted sequencing study of monogenic diabetes genes in ~6900 individuals, one individual with juvenile-onset diabetes (age at diagnosis = 14 years) was observed to be homozygous for a rare missense variant in the *WFS1* gene that had a very low allele frequency (MAF = 0.07%). The same variant was previously reported to be homozygous in another individual with an atypical presentation of Wolfram syndrome (Lieber et al., BMC Medical Genetics 2012). To investigate this rare variant further, we analyzed the allele frequency of the missense variant in multiple variant databases. Using the GnomAD database, we found that the missense variant had an allele frequency of 1.4% in individuals of Ashkenazi Jewish ancestry, 60-fold higher than other populations. Therefore, we genotyped the variant in 475 individuals with type 1 diabetes and 2130 controls of Ashkenazi Jewish ancestry. Surprisingly, we detected 8 homozygotes among the 475 individuals with type 1 diabetes compared to none in 2130 controls (genotype relative risk = 135.3, $p = 3.4 \times 10^{-18}$). The age at diagnosis of diabetes for these 8 individuals (17.8 ± 8.3 years) was several times greater than for typical Wolfram syndrome (5 ± 4 years). Further, optic atrophy was observed in only one of the 8 individuals while another individual had the Wolfram syndrome relevant phenotype of neurogenic bladder. Analysis of sequence and genotype data in two case-control cohorts of Ashkenazi Jewish ancestry demonstrated that this variant is also associated with an increased risk of type 2 diabetes in heterozygotes (OR = 1.81, p = 0.004). In conclusion, we have identified a low frequency coding variant in the *WFS1* gene that is enriched in Ashkenazi Jewish individuals and causes a mild form of Wolfram syndrome characterized by young-onset diabetes and reduced penetrance for optic atrophy. This variant should be considered for genetic testing in individuals of Ashkenazi ancestry diagnosed with young-onset non-autoimmune diabetes and should be included in carrier screening panels. Our results have important clinical implications for the diagnosis of juvenile onset diabetes and highlight the genotype-phenotype complexity of mutations in the *WFS1* gene.

Copyright © 2018 American Society of Human Genetics. All rights reserved

116

Mendelian form of nonalcoholic fatty liver disease and/or dyslipidemia due to monoallelic *ABHD5* mutations. *L. Youssefian[1,2], H. Vahidnezhad[1,3], AH. Saeidian[1], S. Pajouhanfar[1], A. Touati[1,4], S. Sotoudeh[5], P. Mansouri[6], S. Zeinali[3,7], MA. Levine[8], K. Peris[9,10], R. Colombo[10,11], J. Uitto[1,12].* 1) Department of Dermatology and Cutaneous Biology, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, Pennsylvania, USA; 2) Department of Medical Genetics, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran; 3) Molecular Medicine Department, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran; 4) Drexel University College of Medicine, Philadelphia, Pennsylvania, USA; 5) Department of Dermatology, Children's Medical Center, Center of Excellence, Tehran University of Medical Sciences, Tehran, Iran; 6) Skin and Stem Cell Research Center, Tehran University of Medical Sciences, Tehran, Iran; 7) Kawsar Human Genetics Research Center, Tehran, Iran; 8) Division of Endocrinology, Children's Hospital of Philadelphia, Philadelphia, USA; 9) Institute of Dermatology, Faculty of Medicine, Catholic University, Rome, Italy; 10) IRCCS Policlinico Gemelli - University Hospital, Rome, Italy; 11) Institute of Clinical Biochemistry, Faculty of Medicine, Catholic University, Rome, Italy; 12) Jefferson Institute of Molecular Medicine, Thomas Jefferson University, Philadelphia, Pennsylvania, USA.

Non-alcoholic fatty liver disease (NAFLD) is a common disorder, highly associated with the metabolic syndrome, which may progressively lead to simple steatosis, nonalcoholic steatohepatitis (NASH), cirrhosis, hepatic failure and hepatocellular carcinoma. The prevalence of the disease is predicted as high as 27-38% in the US. While NAFLD is clearly multifactorial, in this study, we report a rare Mendelian form of NAFLD in association with rare pathogenic variants of *ABHD5* in six distinct Italian and Iranian multi-generation families, including 24 individuals affected by NAFLD grade I to II. In a large Italian family, whole-exome sequencing identified a heterozygous stop codon mutation in *ABHD5*. Segregation analysis of 22 members of this extended family showed the presence of the *ABHD5* mutation in all 9 patients, while those 13 members without evidence of NAFLD showed wild-type sequence only. While we identified a monoallelic *ABHD5* mutation in NAFLD in adults, biallelic *ABHD5* mutations were previously associated with Chanarin-Dorfman syndrome (CDS) (OMIM #275630), an extremely rare, autosomal recessive disease with multi-systemic manifestations including severe ichthyosis and neonatal NAFLD. Therefore, to explore the potential presence of NAFLD in heterozygous carriers of the mutations in families with CDS, we subjected individuals from five Iranian families, including a total of 11 CDS patients and 16 normal appearing *ABHD5* carriers who did not consume alcohol, for clinical studies. A total 14 out of 16 carriers showed NAFLD along with dyslipidemia (hypercholesterolemia and/or hypertriglyceridemia). Two young carriers did not show any lipid abnormalities. The youngest carrier who demonstrates NAFLD was 29 years of age. For estimation of the disease-allele frequency of the *ABHD5* locus related to this late-onset disease, we analyzed the DNA sequence data for 60,706 normal unrelated individuals in the ExAC database. By a stepwise bioinformatics strategy, only 26 out of 357 variants survived as pathogenic. These 26 variants were present in 72 individuals out of 60,706, predicting the prevalence of NAFLD due to monoallelic *ABHD5* mutations in 1 out of 833 individuals, if the penetrance of mutations was hypothetically considered complete. In summary, for the first time, we report a Mendelian form of NAFLD caused by monoallelic *ABHD5* mutations in adults.

117

Bi-allelic mutations in Phe-tRNA synthetase identified from four families are associated with a multi-system disease and support ex-translational function. *Z. Xu[1,2,3], W.S. Lo[1,2], D.B. Beck[4,5], L. Schuch[6], M. Oláhová[7], R. Kopajtich[8,9], Y.E. Chong[3], C.L. Alston[7], E. Seidl[6], L. Zhai[1], D. Timchak[10,11], C.A. LeDuc[10], A.C. Borczuk[12], A.F. Teich[13], J. Juusola[14], C. Sofeso[15], C. Müller[16], G. Pierre[17], T. Hilliard[17], P.D. Turnpenny[18], M. Wagner[8,9,19], M. Kappler[6], F. Brasch[20], J.P. Bouffard[21], L.A. Nangle[3], R.W. Taylor[7], H. Prokisch[8,9], M. Griese[6], W.K. Chung[4,10], P. Schimmel[1,22,23].* 1) IAS HKUST - Scripps R&D Laboratory, Institute for Advanced Study, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China; 2) Pangu Biopharma, Edinburgh Tower, The Landmark, 15 Queen's Road Central, Hong Kong, China; 3) aTyr Pharma, 3545 John Hopkins Court, Suite 250, San Diego, CA; 4) Department Medicine, Columbia University, New York, NY; 5) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 6) Dr. von Hauner Children´s Hospital, Division of Pediatric Pneumology, University Hospital Munich, German Center for Lung Research (DZL), Lindwurmstr. 4, 80337 München, Germany; 7) Wellcome Centre for Mitochondrial Research, Institute of Neuroscience, The Medical School, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK; 8) Institute of Human Genetics, Technical University Munich, 81675 Munich, Germany; 9) Institute of Human Genetics, Helmholtz Zentrum München, Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Ingolstädter Landstr. 1, 85764 Neuherberg, Germany; 10) Department Pediatrics, Columbia University, New York, NY; 11) Goryeb Children's Hospital, Atlantic Health System, Morristown, NJ; 12) Department of Pathology, Weill Cornell Medicine, New York, NY; 13) Department Pathology and Cell Biology, Columbia University, New York, NY; 14) GeneDx, Gaithersburg, MD; 15) Zentrum für Humangenetik und Laboratoriumsdiagnostik, Lochhamer Str. 29, 82152 Martinsried, Germany; 16) Department of Pediatrics and Adolescent Medicine, University Medical Center, Medical Faculty, University of Freiburg, 79085 Freiburg, Germany; 17) Bristol Royal Hospital for Children, University Hospitals Bristol NHS Foundation Trust, Bristol, BS2 8BJ, UK; 18) Royal Devon & Exeter NHS Foundation Trust, Exeter EX2 5DW, UK; 19) Institut für Neurogenomik, Helmholtz Zentrum München, Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Ingolstädter Landstr. 1, 85764 Neuherberg, Germany; 20) Klinikum Bielefeld Mitte, Institute for Pathology, Teutoburger Straße 50, 33604 Bielefeld, Germany; 21) Department Pathology, Morristown Memorial Hospital, Morristown, NJ; 22) The Scripps Laboratories for tRNA Synthetase Research, The Scripps Research Institute, 10650 North Torrey Pines Road, La Jolla, CA; 23) The Scripps Laboratories for tRNA Synthetase Research, Scripps Florida, 130 Scripps Way, Jupiter, FL.

The aminoacyl-tRNA synthetases (AARS) are essential enzymes for protein synthesis. Recent studies associated AARS with various human diseases and highlighted ex-translational functions of these proteins. For example, human neurological disorders such as Charcot Marie Tooth have been attributed to dominant gain-of-function mutations in some AARS. Recessive loss-of-function mutations have also been increasingly identified in AARS, which can potentially elucidate ex-translational activities. Here we present data on bi-allelic mutations in *FARSB* that encodes the beta chain of the $alpha_2beta_2$ phenylalanine-tRNA synthetase (FARS). One 5'-splice junction non-coding variant (SJV) and 6 missense variants (one shared by unrelated individuals) were identified by exome sequencing in five individuals from four families. All five individuals with bi-allelic variants in *FARSB* presented a multi-system disease with common features of hypotonia, interstitial lung disease with cholesterol pneumonitis. Additional variable features included vascular, neural, hepatic, renal, intestinal, connective tissue and distinctive facial features. We confirmed exon-skipping and frame-shifted transcripts resulted from the SJV. The 6 missense mutations are highly conserved in eukaryotes, but none is known to be directly involved in the translational activity of FARS. The bi-allelic combination of the SJV with a Arg305Gln missense mutation in two individuals led to severe disease; however, cells from the compound heterozygous individual had no defect in protein synthesis. These results support a disease mechanism independent of protein translation and suggest that this FARS activity is essential for normal function in multiple organs.

Copyright © 2018 American Society of Human Genetics. All rights reserved

118

***RINT1* biallelic alterations: A novel cause of infantile onset recurrent liver failure with dysostosis multiplex.** *M.A. Cousin[1,2], E. Conboy[1,3], J.S. Wang[4,5], D. Lenz[6], M. Williams[7], T.L. Schwab[1,7], R.S. Abraham[8], S. Barnett[8], M. El-Youssef[9], R. Graham[8], L.H. Gutierrez Sanchez[9], L. Hasadsri[8], G.F. Hoffmann[6], N.C. Hull[10], R. Kopajtich[11,12], R. Kovacs-Nagy[11,12], J. Li[4], D. Marx-Berger[13], V. McLin[14], H. Prokisch[11,12], D. Rymen[15], C. Staufner[6], Y. Yang[4], K.J. Clark[1,7], B.C. Lanpher[1,3], E.W. Klee[1,2,3,8].* 1) Center for Individualized Medicine, Mayo Clinic, Rochester, MN; 2) Department of Health Sciences Research, Mayo Clinic, Rochester, MN; 3) Department of Clinical Genomics, Mayo Clinic, Rochester, MN; 4) Department of Pediatrics, Jinshan Hospital, Fudan University, Shanghai, China; 5) Center for Pediatric Liver Diseases, Children's Hospital of Fudan University, Shanghai, China; 6) Department of General Pediatrics, Division of Neuropediatrics and Pediatric Metabolic Medicine, University Hospital Heidelberg, Heidelberg, Germany; 7) Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, MN; 8) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN; 9) Department of Pediatric Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN; 10) Department of Radiology, Division of Pediatric Radiology, Mayo Clinic, Rochester, MN; 11) Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany; 12) Institute of Human Genetics, Technische Universität München, Munich, Germany; 13) Pediatric Nephrology, University Children's Hospital Zurich, Switzerland; 14) Pediatric Gastroenterology Unit, University Hospitals Geneva, Switzerland; 15) Department of Metabolic Diseases, University Children's Hospital Zurich, Switzerland.

Pediatric acute liver failure (ALF) is life threatening with genetic, immune, and environmental etiologies. Approximately 50% of pediatric ALF remain undetermined. Reports of recurrent ALF (RALF) in infants describe repeated episodes of severe liver injury with recovery of hepatic function between crises. We describe biallelic *RINT1* alterations as a novel cause of a multisystem disorder including RALF and dysostosis multiplex. Golgi-ER retrograde transport is a highly regulated process necessary for protein and vesicle sorting, membrane recycling, and organelle homeostasis. Genes involved in this process have been implicated in RALF including *NBAS* and *SCYL1* with variable skeletal findings. Skeletal phenotypes of lysosomal storage disorders including dysostosis multiplex are caused by dysregulated autophagy, with UVRAG being a key regulator. RINT1 interacts with NBAS and UVRAG, with loss of RINT1 associated with abnormal vesicle transport and autophagy. Three unrelated individuals with biallelic *RINT1* variants are reported with RALF onset ≤3 years of age. All have a splice alteration at the same position (c.1333+1G>A or G>T) in trans with a missense (p.A368T or p.L370P) or in-frame deletion (p.V618_K619del). ALF episodes are concomitant with fever/infection and not all patients have complete normalization of liver enzymes between episodes. Skeletal surveys on two patients revealed dysostosis multiplex. Liver biopsies showed only non-specific liver damage including fibrosis or steatosis. The splice variant leads to skipping of exon 9 and an out-of-frame product. The in-frame deletion showed increased aggregation and proteasome degradation through poly-ubiquitination exacerbated by cell culture in increased temperature. Consistently, the ER-Golgi intermediate compartment was dispersed in primary patient fibroblasts after temperature challenge. We describe the first cohort of patients with biallelic *RINT1* variants expressing a complex phenotype including RALF and dysostosis multiplex and showing disrupted cellular vesicle trafficking. During nutrient depletion or infection, Golgi-ER transport is suppressed and autophagy is promoted by UVRAG regulation by mTOR. RINT1 falls at the regulatory fulcrum of Golgi-ER retrograde transport and the autophagy-lysosomal pathway. This may explain both the patients' liver and skeletal findings. Clarifying the pathomechanism underlying this new gene-disease relationship may suggest therapeutic opportunities.

119

Scalable and accurate implementation of generalized mixed model for region-based association tests in large biobanks and cohorts. *W. Zhou[1,2], J.B. Nielsen[3], L.G. Fritsche[2], J. LeFaive[2], M.B. Elvestad[4], K. Hveem[5,6], G.R. Abecasis[2], C.J. Willer[1,3,7], S. Lee[2].* 1) Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, 48109, United States of America; 2) Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, 48109, United States of America; 3) Department of Internal Medicine, Division of Cardiology, University of Michigan Medical School, Ann Arbor, Michigan, 48109, United States of America; 4) K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, NTNU, Norwegian University of Science and Technology, Trondheim, Norway; 5) HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, 7600 Levanger, Norway; 6) Department of Medicine, Levanger Hospital, Nord-Trøndelag Health Trust, 7600 Levanger, Norway; 7) Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan, 48109, United States of America.

Although population-based biobanks provide large sample sizes to identify novel genetic associations for complex traits, single-variant association test is still underpowered for rare variants (MAF < 1%). The region-based test, SKAT-O, is widely used to analyze rare variants, allowing for different directions of variant effects and regions with a small proportion of causal variants. To adjust for sample relatedness, which is a major confounder for association tests, mixed model SKAT-O (mmSKAT-O) has been developed. However, large biobanks pose challenges to apply mmSKAT-O, mainly because of the high computation and memory cost to handle the large data size. Here, we have developed a method called SAIGE-SKAT-O for region-based association analysis of very large samples (> 400,000 individuals). Similar to BOLT-LMM and SAIGE, the single-variant association methods using mixed models for large sample sizes, our method utilizes state-of-art optimization strategies to reduce the computation cost for fitting null mixed models. To further improve the time and memory efficiency of the mmSKAT-O test, we approximate the variance of score test statistics with the full genetic relationship matrix (GRM) using the variance with a sparse GRM, which is constructed by thresholding small values in the full GRM, and ratios of these two variance estimates calculated from a subset of genetic markers. Our method can analyze 69,479 samples and 14,458 genes in 10 hours using 16 threads and < 5 GByte of memory and the computing time scales linearly with increasing numbers of samples and markers. In contrast, the existing mmSKAT-O software package is projected to require more than 200 CPU hours and 300 GByte of memory. We applied SAIGE-SKAT-O to analyze 14,458 genes, with rare (MAF < 0.5%) missense and stop-gain variants for four blood lipid traits and thyroid stimulating hormone (TSH) in 69,500 Norwegian samples from a population-based Nord Trøndelag Health Study (HUNT) with substantial sample relatedness. 18 genes for LDL, 12 for TC, 4 for TG, 10 for HDL, and 4 for TSH reached the exome-wide significant threshold (P-value < 2.5×10^{-6}). Most genes are located in the previously reported genome loci for the traits of interest, while potentially novel genes are also identified. For example, the gene *B4GALNT3* that has been previously suggested to play a role in thyroid cancer is found to be significantly associated with TSH, even with thyroid cancer cases excluded (p-value 2.7×10^{-15}).

Copyright © 2018 American Society of Human Genetics. All rights reserved

120

Analyzing the world's largest public human variation resources in less than a day: Massively scalable software for genomic analysis. *T. Poterba[1,2,3], L. Abbott[1,2,3], J. Bloom[1,2,3], C. Churchhouse[1,2,3], R. Cownie[1,2,3], L.C. Francioli[2,3], J. Goldstein[1,2,3], D. Howrigan[1,2,3], K.J. Karczewski[1,2,3], D. King[1,2,3], D. Palmer[1,2,3], P. Schultz[1,2,3], G. Tiao[2,3], R. Walters[1,2,3], A. Wang[1,2,3], D.G. MacArthur[2,3], B.M. Neale[1,2,3], C. Seed[1,2,3].* 1) Stanley Center for Psychiatric Genetics, Broad Institute of Harvard and MIT, Cambridge, MA; 2) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston 02114, MA, USA; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.

Recent growth in the size and complexity of genetic data to petabyte-scale datasets of hundreds of thousands of individuals has made even simple analyses cumbersome or intractable. To address this challenge, we have developed Hail, an open-source Python framework that includes infrastructure for data exploration, standard algorithms used in the analysis of genomic data, functionality for linear algebra, and a suite of statistical methods for common and rare variant association. Hail natively distributes computation, scaling from a laptop to a large compute cluster or cloud. While Hail is designed primarily for analysis of genomic data, it provides general interfaces that can be used to process other forms of data, such as imputed genotype data and single-cell RNA-Seq expression matrices. Hail has been used for small and large projects in academic and industry groups worldwide. Two examples in particular testify Hail's value to the genomics community: the analysis and QC of gnomAD (gnomad.broadinstitute.org) and the UK Biobank Rapid GWAS (www.nealelab.is/blog). The gnomAD project required iterative exploration and quality control of 125,748 whole exomes and 15,708 whole genomes (over 40 terabytes of compressed VCF files) to characterize and address sequencing artifacts and batch effects, followed by thousands of combinatorial aggregations to produce site-level summary data that accrued over 5 million pageviews since its public release two years ago. Hail provided both a means to scale analysis to immense compute clusters to enable QC computations in less than a day, and the ability to easily and concisely express complex queries over the data, including building models of gene constraint and characterizing compound heterozygotes. Without either of these requirements, the public data release would have been delayed or impossible. The UK Biobank Rapid GWAS was faced with a computational problem: simultaneously performing GWAS on 2000 traits in one of the largest genetic datasets ever assembled (337K individuals). The analysis team leveraged Hail's optimized regression algorithms and scalability to produce 2 terabytes of association results in 24 hours for public release, for about $1 per trait. As we develop infrastructure to process datasets of millions of whole genomes expected in the near term, we welcome the scientific community to leverage this toolkit to develop and share new methods at scale to enable analyses that would otherwise be impossible.

121

Kipoi: Accelerating the community exchange and reuse of predictive models for regulatory genomics. *J. Gagneur[1], Z. Avsec[1], R. Kreuzhuber[2], J. Israeli[3], N. Xu[3], J. Cheng[1], A. Shrikumar[3], L. Urban[2], D. Kim[3], W.H. Ouwehand[4], A. Kundaje[3], O. Stegle[2].* 1) Technical University of Munich, Munich, Germany; 2) European Molecular Biology Laboratory, European Bioinformatics Institute; 3) Stanford University; 4) Department of Human Genetics, The Wellcome Trust Sanger Institute.

Machine learning methods have allowed for recasting one of the most central problems of biology and genetics, namely to predict phenotypic consequences from genotype. Using large compendia of high-throughput datasets, complex machine learning systems are now trained to predict molecular consequences of DNA sequence including transcription factor binding, chromatin accessibility, and splicing efficiency. Once trained, such predictive models hold the promise to allow for probing regulatory dependencies in silico, which among other applications enables interpreting functional variation in personal genomes. However, there is no standard for sharing *trained models*, in striking contrast to well-established repositories for bioinformatics methods or archives for genomic raw data. This lack of standardization has led to models that are difficult to deploy, to apply to new data, to retrain, and to combine with each other. Here we present Kipoi, a collaborative project to support the sharing and re-use of trained models in genomics. Kipoi's repository, at https://kipoi.org, contains over 2,000 trained models covering canonical prediction tasks in transcriptional and post-transcriptional gene regulation. Kipoi's specification standard grants automated setup of software dependencies and provides a unified application programming interface to execute and interpret models and to apply them on standard file formats including the variant call format VCF. We demonstrate the power of these features through use cases implemented in a few lines of code including i) a benchmark of diverse models predicting transcription factor binding, ii) interpreting regulatory sequences with multiple models, and iii) building rapidly a performant predictor of DNA accessibility for a given cell line by tuning an existing large deep neural network pre-trained on hundreds of other cell lines. Finally, we show how Kipoi streamlines the development and objective benchmarking of variant scoring methods. To this end, we leverage Kipoi's unified variant effect prediction module and combine complementary sequence-based splicing models into a pathogenicity score for variants located near splice sites. Our score outperforms state-of-the-art models reaching 97% auROC on ClinVar variants. Altogether, Kipoi promises to catalyze the integration of models predicting the impact of genetic variation and to accelerate the translation of genomics research advances into personal genome interpretation.

Copyright © 2018 American Society of Human Genetics. All rights reserved

122

DRAMS: A tool to Detect and Re-Align Mixed-up Samples leveraging multi-omics data. *Y. Jiang[1,6], Y. Xia[1,7], L. Han[2], G. Giase[3], K. Grennan[3], L. Sloofman[4], S. Liu[5], C. Chen[1], B. Li[6], C. Liu[1,7].* 1) Center for Medical Genetics, Central South University, Changsha, Hunan, China; 2) Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA; 3) University of Illinois at Chicago, Chicago, IL, USA; 4) Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 5) Yale University, New Haven, CT, USA; 6) Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN, USA; 7) Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA.

Background: As the number of multi-omics studies increased, it is worth noting that sample mix-up happens frequently for many different reasons, which will reduce statistical power and result in false findings. Correct sample alignment is critical for integrative analysis of -omics data. Several tools, including MixupMapper and MODMatcher, have provided a few solutions in correcting mixed-up sample IDs leveraging multi-omics data. However, the tools can only be applied to a maximum of three -omics data types. In addition, these tools use eQTL as a mediator to estimate sample relatedness between DNA and RNA -seq data, which may bias the results as eQTL is often tissue-specific with variable quality depending on the sources. Therefore, it is important to develop a tool to accurately estimate sample relatedness and take full advantage of sample relationships among -omics data of increasing dimensions to detect and correct mixed-up samples. **Methods**: Here we present DRAMS, a tool to detect and realign mixed-up samples. The tool calls genotypes from -omics data and estimates pairwise sample relatedness by comparing genotypes directly. Then, we estimate switch direction for each mismatched sample pair based on sample relatedness with other -omics data as well as the match of nominal and SNP-inferred sexes. The true sample IDs can be determined by integrating all the matched and directional mismatched pairs. **Application**: We applied DRAMS to the PsychENCODE BrainGVEX project, which produced data from six platforms, including whole genome sequencing (n=285), Psych v1.1 BeadChips (n=263), Affymetrix 5.0 450K SNP-array (n=137), Assay for Transposase-Accessible Chromatin using sequencing (n=295), RNA sequencing (n=426), and ribosome profiling (n=103). In total we detected that 10.3% of samples were mixed-up. As a validation step, we calculated the Pearson correlation between genotypes and allele frequencies in the corresponding ethnicity in the 1000 Genomes Project for each sample. We found that the correlation coefficient increased or remained the same after realigning sample IDs for all the cross-ethnicity sample switches. As a result, after realigning sample IDs, the number of eQTLs (FDR<0.05) increased from 775,761 to 1,228,398 (1.58×). **Conclusion**: As being applied to the BrainGVEX datasets, the statistical power increased significantly after correcting sample IDs. It is predicted that DRAMS will perform better as the more dimension of -omics data are used.

123

Patient-data sharing of whole exome sequencing results with GenomeConnect informs variant interpretation and gene-disease relationships. *J.M Savatt[1], D.R Azzariti[2], E. Palen[1], J. Hart[3], B.L Kattman[3], M.J Landrum[3], S.M Harrison[2,6], V. Rangel Miller[4], J. Rhode[4], J.A Vidal[4], D.H Ledbetter[1], H. Rehm[2,5,6,7], W.A Faucett[1], E.R Riggs[1], C.L Martin[1].* 1) Geisinger, Lewisburg, PA, USA; 2) Laboratory for Molecular Medicine, Partners Personalized Medicine, Boston, Massachusetts, USA; 3) National Center for Biotechnology Information, Bethesda, Maryland, USA; 4) Invitae, San Francisco, California, USA; 5) The Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA; 6) Harvard Medical School, Boston, Massachusetts, USA; 7) Center for Genomic Medicine, Massachusetts General Hospital, Boston, USA.

The implementation of whole exome sequencing (WES) has improved the detection of genetic etiologies for rare conditions, but increased the number of genomic variants with an uncertain impact on health. Broad sharing of genetic and health data is needed to inform variant interpretation and improve patient care. GenomeConnect (GC), the NIH-funded Clinical Genome Resource (ClinGen) online patient registry, is open to anyone who has had genetic testing and wishes to broadly share their deidentified genetic and health data. Participants complete health surveys and upload genetic testing reports that are curated by the GC team. Here we present the results of 155 GC participants who uploaded WES results, describe how this data has been shared and utilized, and demonstrate the benefits of patient-data sharing. Participants ranged in age from 2 months to 62 years at time of testing; 94.2% (n=146/155) had at least one variant reported. In total, 423 unique variants with deidentified health information were submitted to NCBI's ClinVar database, including 420 sequence variants; 55.1% of variants were novel to ClinVar (n=238/423) showing the importance of patients as a genomic data source. As variant interpretation can change over time, GC also offers participants the option to receive variant classification updates if the reporting laboratory submits to ClinVar; 3.9% (n=4/103) of patient submitted variants had a conflict with the reporting laboratory's current classification in ClinVar. Of variants submitted by GC and at least one other clinical laboratory besides the reporting institution, 39.3% (n=44/112) had a discrepancy with another laboratory's submission. GC is working with ClinGen's Discrepancy Resolution effort to encourage clinical laboratories to address these discrepancies; to date, five discrepancies have been resolved. GC participants are informed of any updated classifications from their reporting laboratory. GC also shares results in candidate genes with GeneMatcher. Thus far, 16 variants have been submitted with all yielding potential matches. Three participants were provided contact information for a researcher or clinician interested in their gene based on the match. Finally, GC facilitates matching between participants; 30.3% of participants with WES (n=47/155) had a gene in common allowing for participant matching. Overall, we show that broad data sharing through GC empowers patients to help contribute knowledge to improve variant interpretation.

Copyright © 2018 American Society of Human Genetics. All rights reserved

124

Blood biospecimen donation and consent to share data among African American women. *C. Wang[1], L. Barber[2], L. Rosenberg[2], J.R. Palmer[2].* 1) Community Health Sciences, Boston University School of Public Health, Boston, MA; 2) Slone Epidemiology Center, Boston University, Boston, MA.

Efforts to increase the racial and ethnic diversity of biospecimens for precision medicine research necessitate a better understanding of biospecimen donation behaviors and consent to share data among minority populations. The Black Women's Health Study (BWHS), an ongoing prospective cohort study of 59,000 self-identified African American women from across the U.S. followed for 22 years, initiated a four-year effort in 2013 to collect blood biospecimens from women in the cohort. Women were also consented on their willingness to share data via a government database such as dbGaP. The present study examined blood donation and data sharing patterns in the BWHS and identified sociodemographic predictors of these patterns. Of 48,956 BWHS participants invited to participate in the blood study, 27% (n=13,037) provided a sample. Among women who provided a blood sample, 63% (n=8229) consented to share their data via dbGaP. The strongest predictors of providing a blood sample were recent history of a physical exam (multivariable odds ratio (OR) 3.90 [95% CI: 3.70-4.11]) and recent history of cancer screening (ORs 3.51 [3.33-3.70] for mammography and 2.37 [2.26-2.47] for PAP). Other associations were considerably weaker: women who were more educated, lived in the Midwest, or had a family history of cancer had a higher odds of providing a blood sample (ORs ranging from 1.16 for family history to 1.30 for >16 vs. ≤12 years of education), and women with more frequent experiences of racism in their daily lives had a lower odds (OR 0.83 [0.77-0.89] for highest vs. lowest quartile of racism score). There were no strong predictors of willingness to share data: ORs were 1.31 [1.15-1.50] for women aged ≥70 vs. <50, 1.16 [1.05-1.28] for having had a physical exam in the past two years, and 1.31 [1.20-1.43] for living in the Northeast relative to all other regions. A history of cancer was not associated with either sample provision or agreement to share. The proportion of BWHS participants who provided a blood sample, 27%, is similar to the proportion in the predominantly white Nurses' Health Study, which offered blood collection many years after enrollment. However, 37% of BWHS sample providers declined to allow data from the samples to be shared via dbGaP. Efforts to facilitate engagement in precision medicine research will need to consider important sociodemographic factors to ensure broader representation of diverse populations within these research endeavors.

125

Effectiveness of the Genomics ADvISER decision aid for the selection of incidental genome sequencing results: Randomized clinical trial. *Y. Bombard[1,2], M. Clausen[2], C. Mighton[2], S. Shickh[2], S. Casalino[2], T.H.M. Kim[9], L. Carlsson[6], E. Joshi[7], L. McCuaig[1,2], N. Baxter[1,2], A. Scheer[1,2], C. Elser[6], A. Eisen[7], M. Evans[1,2], R.H. Kim[6,8,9], S. Panchal[8], T. Graham[7], M. Aronson[1,8], C. Piccinin[8], L. Winter-Paquette[1,8], K. Semotiuk[1,8], J. Lerner-Ellis[1,8], J.C. Carroll[1,8], J. Hamilton[5], E. Glogowski[3], K. Schrader[4], K. Offit[5], M. Robson[5], K. Thorpe[1], A. Laupacis[1,2].* 1) St. Michael's Hospital, Toronto, ON,; 2) University of Toronto, Toronto, ON,; 3) GeneDx, Gaithersburg, MD,; 4) BC Cancer Agency, Vancouver, BC,; 5) Memorial Sloan Kettering Cancer Center, New York, NY; 6) University Health Network, Toronto, ON; 7) Sunnybrook Health Sciences Centre, Toronto, ON; 8) Mount Sinai Hospital, Sinai Health System, Toronto, ON; 9) Hospital for Sick Children, Toronto, ON.

Background: The volume and complexity of incidental results (IR) from genome sequencing (GS) makes engaging in pre-test shared decision-making with each patient infeasible. Novel e-health tools such as decision aids (DAs) are needed to fill this critical care gap. We created an interactive, online DA (www.genomicsadvisor.com) to guide patients' selection of IR by "binning" IR into 5 categories: 1. Medically actionable & pharmacogenetic variants 2. Common disease SNPs 3. Mendelian disease variants 5. Early-onset neurological variants 6. Carrier results **Aim**: To evaluate the effectiveness of the Genomics ADvISER DA compared to genetic counseling. **Methods**: We conducted a superiority RCT of the Genomics ADvISER DA among adult patients who received uniformitive results from past genetic testing for colorectal or breast cancer, and who could thus be considered for GS as a second tier test. Patients in the intervention arm used the Genomics ADvISER DA to select IR they would hypothetically want to learn, then briefly spoke with a genetic counselor (GC) about their selections. Patients in the control arm spoke with a GC to select IR. The primary outcome was decisional conflict (DC) and secondary outcomes were knowledge, preparation for decision making (PDM), and satisfaction with decisions (SWD). **Results**: A total 133 patients enrolled (90% female; 60% ≥50yo). DC scores did not significantly differ between intervention and control groups (P=0.60, treatment difference is -1.41 with 95% CI of -6.75 – 3.93). However, mean DC scores in both groups were <25, which is associated with implementing decisions. Knowledge of the benefits of GS was higher in the intervention group (P=0.014), whereas knowledge of limitations of GS did not differ significantly (P=0.42). PDM and SWD scores did not vary significantly between groups (P=0.95, P=0.06), however there was a trend toward higher SWD scores for DA users. **Conclusion**: The Genomics ADvISER DA was not superior to GC. However, participants in both groups felt equally prepared for and satisfied with their decisions, experienced similar levels of DC and scored below the cut-off associated with implementing decisions. Interestingly, higher knowledge of benefits scores in the intervention group indicate the DA's effectiveness in educating patients on GS and IR. Thus, the Genomics ADvISER DA could serve as an educational tool for IR, reducing in-clinic education time and potentially health care costs.

Copyright © 2018 American Society of Human Genetics. All rights reserved

126

Canadian indigenous peoples & genomics: Starting a conversation.

P.H. Birch[1], R.R. Coe[1], R. Lesueur[2], R. Kenny[2], N. Makela[1], R. Price[2], W.H. McKellin[1], A. Lehman[1], J. Morgan[2]. 1) University of British Columbia, Vancouver, B.C., Canada; 2) BC Women's and Children's Hospitals, Vancouver, B.C., Canada.

Background Compared to Europeans, Indigenous Canadians are more likely to have uninterpretable genomic sequencing (GS) diagnostic tests due to the lack of Indigenous representation in reference databases. **Aim** We aimed to start a conversation with Indigenous British Columbians to raise awareness of, and give voice to, this issue. Our goal was to co-create a video explaining genomic non-representation that included diverse Indigenous views. **Methods** The project leaders are of European and Indigenous ancestry. We used a community-based participatory research model to promote trust. We audio-recorded 4 focus groups of Indigenous adults in various settings in Vancouver. Groups were guided by a professional facilitator and an Indigenous Elder. Groups began with a 5 minute video we developed using a culturally appropriate analogy of blanket-weaving to explain GS and lack of Indigenous representation. Opinions were sought on relevance of GS, perceived value of a genomic database, and control of, and access to, Indigenous genomic data. Transcripts were analyzed and quotes representing main themes were incorporated into the introductory video to illustrate participants' perspectives. We then returned to the same groups to discuss results, ensure we had interpreted quotes correctly, and obtain approval for their inclusion in the video. **Results** Of the 30 people in the first focus groups, we were able to re-contact 20 for the follow-up sessions. Two-thirds of attendees were status First Nations, others were non-status, Métis, or Inuit. Participants concurred with our thematic interpretation: The theme of systemic racism interlaced most conversations, particularly within the theme of trust. Themes of governance emphasized privacy, fear of genetic discrimination, and discussions of genomic database structures. Some people thought Canadian Indigenous-control was essential, others recognized advantages of international databases and the artificial nature of political borders. The theme of implementation included creative ideas to collect Indigenous genomes, including using immunization and blood donor clinics, but approval from Indigenous leaders was emphasized. The final video was given to participants to use/distribute as they wish to promote awareness and discussion of genomic diagnostic inequity. This is consistent with the Truth and Reconciliation Commission's recognition that gaps between Indigenous and non-Indigenous Canadians' healthcare need to be identified and closed.

127

Retinal phenotype correction by gene therapy in the novel *Mfrp* mouse model of a human ophthalmic syndrome.

M. Voronchikhina[1], A. Chekuri[1], B. Sahu[1], V.R.M. Chavali[2], A.N. Alapati[1], J.C. Zenteno[3], M.M. Jablonski[4], R. Ayala-Ramirez[3], A. Dinculescu[5], S. Borooah[1], R. Ayyagari[1]. 1) Shiley Eye Institute,UCSD, La Jolla, CA; 2) Ophthalmology, University of Pennsylvania, Philadelphia, PA, USA; 3) Department of Genetics-Research Unit, Institute of Ophthalmology, Conde de Valenciana, Mexico City, Mexico; 4) University of Tennessee, Memphis, TN, USA; 5) Department of Ophthalmology, College of Medicine, University of Florida, Gainesville, FL, USA.

Purpose: Clinical syndrome of posterior microphthalmos-retinitis pigmentosa-foveoschisis-optic disc drusen has been reported in families harboring homozygous mutations c.498_499insC in the *MFRP* gene. The purpose of this study was to characterize ophthalmic features in a novel mouse model homozygous for the same c.498_499insC mutation in *Mfrp* (*Mfrp* $^{KI/KI}$) and to test the response of adeno-associated virus (AAV)-mediated *Mfrp* gene replacement therapy in these mice. **Methods:** Full ophthalmic examination was performed on four affected siblings (age=51,53,57,61) with color fundus photography (CFP), optical coherence tomography (OCT) and scanning laser ophthalmoscopy (SLO). Similarly, CFP, OCT and SLO were used to characterize the retinal phenotype of the *Mfrp*$^{KI/KI}$ mice (C57BL/6). Infrared imaging was done to measure the axial length to identify microphthalmia. Electroretinography and immunohistochemistry were employed to assess cone photoreceptor function and survival, respectively. Sub-retinal injections of rAAV8(Y733F)-smCBA-*Mfrp* (rAAV-*Mfrp*, 1×10^{12} genome copies/ml) were performed on one eye of a subset (n=15, age 1 month) of *Mfrp*$^{KI/KI}$ mice, with the contralateral eye injected with vehicle alone. These mice were evaluated at age 5 months. **Results:** All patients showed progression of retinitis pigmentosa over 12 years but, maintained central vision in two cases aged 51 and 53 years. *Mfrp* $^{KI/KI}$ mice exhibited an accumulation of autofluorescent spots ($p<0.001$), progressive degeneration of photoreceptors, decreased electrophysiological responses ($p<0.001$), retinal thinning ($p<0.001$), increased retinal pigment epithelium atrophy and reduced axial length ($p<0.001$) when compared to age-matched wild type controls. At 5 months, the eyes of *Mfrp* $^{KI/KI}$ mice treated with rAAV-*Mfrp* had significantly thicker retina ($p<0.01$), reduced white spots on SLO ($p<0.05$) and preserved ERG ($p<0.001$) response compared to the contralateral eye controls. **Conclusion:** We generated a homozygous *Mfrp* knock-in mouse to model human disease. The *Mfrp*$^{KI/KI}$ mice displayed progressive retinal degeneration with photoreceptor loss, reduced retinal function and decreased axial length resembling the major pathology observed in patients homozygous for the same mutation. Following sub-retinal rAAV-*Mfrp* gene delivery we found significant retinal rescue. Our study supports the use of AAV-*MFRP* gene replacement therapy to preserve vision in patients harboring mutations resulting in deficiencies of MFRP.

Copyright © 2018 American Society of Human Genetics. All rights reserved

128

One size fits all: Transcriptional upregulation of a disease modifier gene as a mutation-independent approach in muscular dystrophy. *D.U. Kemaladewi, P.S. Bassi, R. Kember, K. Lindsay, E. Hyatt, E.A. Ivakine, R.D. Cohn.* Genetics and Genome Biology, SickKids Hospital, Toronto, Toronto, ON, Canada.

Congenital muscular dystrophy 1A (MDC1A) is caused by mutations in the *LAMA2* gene encoding α2 chain of Laminin, a protein that is important for maintenance of skeletal muscle and Schwann cells stability. Consequently, MDC1A patients suffer from severe muscle wasting and peripheral neuropathy. MDC1A affects 1 in 150,000 newborns, and there are more than 350 different mutations reported to date. Due to their genetic nature, correction of the causative mutation would be a promising treatment option for MDC1A. However, the heterogeneity of mutations, combined with the rarity of the disease, hampers the development of individualized therapies in MDC1A. Therefore, we sought to develop a mutation-independent strategy via upregulation of a disease-modifying gene *Lama1* and assess its therapeutic potential. The *Lama1* gene encodes Laminin α1 protein, which shares ~80% sequence similarity to Laminin α2. Transgenic overexpression of *Lama1* ameliorates muscle wasting and paralysis in MDC1A mouse model, demonstrating its role as a protective disease modifier. We used CRISPR/dCas9 transcriptional activation system, which consists of catalytically inactive *S. aureus* Cas9 (dCas9), VP64 transactivation domains and sgRNAs targeting *LAMA1* promoter. We showed robust upregulation of Laminin α1 protein in a variety of MDC1A-derived fibroblasts, which opens up an entirely new and mutation-independent therapeutic avenue for all MDC1A patients. An important question for future therapeutic approaches for a variety of disorders concerns the therapeutic window and phenotypic reversibility. To address this, we treated MDC1A mouse model (*dy^{2J}/dy^{2J}*) at different disease stages with Adeno-associated virus serotype 9 (AAV9) carrying the CRISPR/dCas9 components. When the intervention was started early in pre-symptomatic *dy^{2J}/dy^{2J}* mice, *Lama1* upregulation successfully prevented muscle fibrosis and hindlimb paralysis. Importantly, we also demonstrated that dystrophic features and disease progression were significantly improved and partially reversed when the treatment was initiated in symptomatic 3-week old *dy^{2J}/dy^{2J}* mice with already-apparent hindlimb paralysis and significant muscle fibrosis. Collectively, our data provide the first evidence of disease reversibility and therapeutic window in this disorder. Finally, it serves as a novel therapeutic strategy that can be applied to a variety of modifier genes and may overcome medical challenges posed by numerous rare genetic diseases.

129

AAV gene therapy with artificial miRNA-mediated oligodendrocyte-specific gene suppression: Implication for the treatment of Pelizaeus-Merzbacher disease with *PLP1* duplication. *K. Inoue[1], H. Li[1], H. Okada[2], T. Okada[2], Y. Goto[1].* 1) Dept. Mental Retardation & Birth Defect Research, National Institute of Neuroscience, NCNP, Tokyo, Japan; 2) Dept. of Molecular & Medical Genetics, Nippon Medical School, Tokyo, Japan.

Adeno-associated virus (AAV) serves as a practical tool for the gene therapy of inherited neurological diseases, because of its efficient gene delivery to non-dividing cells, absence of integration into the host genome, and minimal immune response. In our attempt to develop treatments of an inherited hypomyelinaitng leukodystrophy, Pelizaeus-Merzbacher disease caused by *PLP1* whole gene duplications, we hypothesized that RNAi-mediated *PLP1* gene suppression therapy using oligodendrocyte-specific AAV delivery system may have a potential as a clinically applicable therapy for this intractable disorder. However, oligodendrocytes are difficult to establish AAV infection with high efficiency. In addition, shRNA expression requires PolIII promoter, which allows no tissue specific expression. To overcome these issues, we have developed an AAV cassette that enables highly efficient oligodendrocyte-specific *PLP1* suppression using an artificial miRNA system driven under an oligodendrocyte-specific promoter. Efficient *PLP1* knock down was confirmed in HeLa cells co-tranfection assay before AAV packaging using AAV1/2 hybrid serotypes and column purification from AAV-HEK293 cells. Direct injection of the AAV harboring *PLP1*-targeting artificial miRNA in the 3'UTR of GFP into the mouse brain at P10 revealed oligodendrocyte-specific GFP expression at high efficiency of ~80%, while almost no GFP signals was found in neuron, astrocyte, and microglia. PLP1 expression was downregulated by half at both protein and mRNA level. Together, we have constructed and evaluated an AAV vector that enables oligodendrocyte-specific *PLP1* gene suppression, which can be utilized for the development of the treatment of PMD with *PLP1* duplications.

Copyright © 2018 American Society of Human Genetics. All rights reserved

130

CRISPR/Cas9 single guide treatment of novel multi-exon duplication mouse model of Duchenne muscular dystrophy removes 139kbp duplication and restores full length dystrophin. *D. Wojtal[1,2], E. Hyatt[1], K. Lindsay[1], E.A. Ivakine[1], R.D. Cohn[1,2].* 1) Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada; 2) Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

Duchenne muscular dystrophy (DMD) is a neuromuscular disorder that leads to progressive muscle deterioration, loss of ambulation, and respiratory complications. It is caused by genetic mutations that result in the absence of dystrophin protein expression needed for muscle function. Despite significant advances in our understanding of the pathogenesis of DMD, no curative treatment has been identified to date and the disorder has a life-limiting disease trajectory. Furthermore, of the current strategies in clinical trial, none are amendable to the treatment of DMD patients with duplications. Recently, we have pioneered an approach to successfully remove large duplications in patient cells affected with several diseases including MECP2-duplication syndrome and a multi-exon (18-30) duplication in the*DMD*gene using Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR-associated Nuclease (Cas9) with a single guide. In order to test our treatment approach *in vivo,*we first generated a mouse model harboring a large duplication of 139 kb in the *Dmd*geneusing CRISPR/Cas9. This first multiexon duplication model of DMD specifically mimics a patient duplication of *DMD* Exons 18-30. Molecular and functional characterization of this model reveals dystrophin deficiency and hallmark markers of dystrophic muscle. Furthermore, using our previously described CRISPR/Cas9 single guide strategy, we have for the first time treated a large genomic duplication *in vivo*and shown successful removal of the duplication fragment leading to restoration of full-length dystrophin. Our findingsestablish the far-reaching therapeutic utility of CRISPR/Cas9, which can be tailored to target numerous inherited disorders caused by duplications.

131

Efficient cross-trait penalized regression increases prediction accuracy in large cohorts using secondary phenotypes. *W. Chung[1,2], J. Chen[3], C. Turman[1,2], S. Lindstrom[4], Z. Zhu[1,2,5], P. Loh[1,2,6], P. Kraft[1,2,7], L. Liang[1,2,7].* 1) Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA.02115, USA; 2) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; 3) Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA; 4) Department of Epidemiology, University of Washington, Seattle, WA 98195, USA; 5) Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; 6) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; 7) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA.

With the arrival of large-scale public biobanks harboring more than 500K samples, polygenic risk scores based multi-trait methods, such as multi-trait analysis of GWAS (MTAG), have been widely adopted for genetic risk prediction in practice due to computational feasibility. However, to obtain more accurate risk scores and boost predictive power, there exists a necessity to develop efficient multi-trait prediction methods using whole-genome individual-level genotypes. Here, we introduce cross-trait penalized regression (CTPR), a powerful and practical approach for multi-trait polygenic risk prediction in large cohorts. Specifically, we propose a novel cross-trait penalty function with the Lasso and the minimax concave penalty (MCP) to incorporate the shared genetic effects across multiple traits and implement it for large-sample GWAS data. Our approach has several advantages: (1) it extracts information from the secondary traits that is beneficial for predicting the primary trait but tunes down information that is not; (2) it can incorporate multiple secondary traits based on individual-level genotypes and/or summary statistics from large-scale GWAS studies; (3) our novel implementation of a distributed memory parallel computing algorithm makes it feasible to apply our methods to biobank-scale GWAS data. We illustrated our method using large-scale GWAS data (~1 million SNPs) from the UK Biobank (N=456,898) and the NHS/HPFS/PHS cohort (N=20,769) with height as the primary trait and BMI, hip circumference, waist circumference and waist-hip ratio as the secondary traits. We showed that the use of summary statistics or individual-level genotypes from the secondary traits can substantially improve the prediction accuracy for the primary trait as compared to the single-trait approach. We also showed that our multi-trait method outperforms the multi-trait genomic best linear unbiased prediction (MTGBLUP) and the recently proposed MTAG for predictive performance. The prediction accuracy for height by the aid of BMI improved from R^2=11.3% (MTAG) or R^2=11.4% (MTGBLUP) to 14.5% (MCP+CTPR) or 14.6% (Lasso+CTPR) with 30,000 training samples and from R^2=35.8% (MTAG) to 42.5% (MCP+CTPR) or 42.8% (Lasso+CTPR) with 436,898 training samples from UK Biobank. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

132

Association studies for all: A novel framework to allow for the well-calibrated genomic analysis of underrepresented admixed individuals. *E.G. Atkinson[1,2], A.X. Maihofer[3], A.R. Martin[1,2], K.C. Koenen[2,4], B.M. Neale[1,2], C.M. Nievergelt[3], M.J. Daly[1,2].* 1) Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boson, MA; 2) Stanley Center for Psychiatric Research & Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA; 3) Department of Psychiatry, University of California San Diego, La Jolla, CA; 4) Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA.

Currently, many genetics studies exclude individuals whose ancestry is not homogeneous but rather comprises components from several ancestral groups – such individuals are “admixed”. Admixed individuals are removed due to the challenges of accurately accounting for their ancestry such that population substructure can infiltrate analyses and bias results. Here, we present a framework to account for this issue and allow admixed individuals to be studied alongside homogenous ones by inferring fine-scale population structure informed by local ancestry estimates. Incorporating local ancestry information in addition to principal components takes into account individuals’ subtle differences in admixture patterns that may differ among case and control cohorts even if their genome-wide ancestry fractions are the same. We apply our framework to several admixed cohorts with high global diversity from the Psychiatric Genomics Consortium PTSD group, with a focus on African American and Latino individuals. The current lack of methods for analysis of admixed individuals is problematic for many traits, but especially for certain psychiatric disorders including PTSD, which both has a very large number of patients of admixed descent and is extremely polygenic, meaning that accounting for the local ancestral dosage is key to precisely understand the genetic contribution of the many component sites of small effect that combine to result in the disorder. We show the extent to which including admixed individuals within ancestry-specific meta-analyses of the PGC-PTSD cohorts boosts signal to discover GWAS loci. We further demonstrate that this framework gives increased precision when looking at variants across ancestry groups and improves the resolution of association signals by leveraging differences in linkage disequilibrium patterns between populations to more narrowly map where signal is coming from. This framework could be applied to solve the statistical issues related to admixture across many medical and population genetics activities, including association testing efforts and evolutionary studies such as genome-wide selection scans. In sum, this framework dramatically advances the existing methodologies for studying admixed individuals and allows for significantly better calibrated study of the genetics of complex disorders in underrepresented populations.

133

Reversing GWAS to identify and model genetic subtypes. *A. Dahl[1], N. Cai[2,3], A. Ko[4], C. Gignoux[5,6], M. Laakso[7], E. Burchard[8,9], P. Pajukanta[4], J. Flint[10], N. Zaitlen[1].* 1) Department of Medicine, University of California San Francisco, San Francisco, CA; 2) Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK; 3) European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK; 4) Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA; 5) Colorado Center for Personalized Medicine, University of Colorado, Denver, CO, USA; 6) Department of Biostatistics, University of Colorado, Denver, CO, USA; 7) Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, 70210 Kuopio, Finland; 8) Department of Bioengineering & Therapeutic Sciences & Medicine, University of California, San Francisco, CA, USA; 9) Center for Genes, Environment & Health, University of California, San Francisco, CA, USA; 10) Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA.

Genetic studies of complex disease have relied on expert classification of cases and controls. This is a known oversimplification as multiple disease subtypes, environments, and independent biological systems often coexist. High-throughput phenotype and genotype data have recently inspired attempts to disentangle and genetically validate such phenotypic subtypes. Our approach, called reverse GWAS (RGWAS), uses genetics to identify disease subtypes rather than using disease classification to find associated SNPs. Unlike previous subtyping methods, RGWAS controls for large-effect covariates, like genetic PCs. We also develop a mixed model to improve power over single-SNP tests for heterogeneity. Simulations show RGWAS is powerful when its assumptions are met and calibrated even when subtypes are absent. By contrast, standard decompositions can fail severely in the face of realistic properties, like population structure. In a major depression dataset, RGWAS recovers 5 positive- and negative-controls for SNP interactions with lifetime stress, and we further show that stress attenuates heritability ($p<2.2e-4$). We then apply RGWAS to case-only traits in a multi-ethnic asthma cohort and find 3 asthma subtypes. We tested their biological basis with known asthma SNPs: one has differential asthma ORs, and one has differential effects on lung function response to albuterol (adjusted $p<.05$). Population and ethnicity correlate differently with asthma across subtypes ($p<.01$). In the METSIM metabolic dataset, RGWAS infers 3 subtypes differing mostly in lipid and amino acid metabolites, like ratios of esterified-to-free cholesterol, histidine, omega-3 fatty acids, and polyunsaturated fats. Our polygenic model proves these metabolic subtypes differ genetically: they increase the explained heritability by 65% across traits, and per-subtype heritability differs in 13/16 traits (adjusted $p<.05$). We find differential effects for 4/93 known metabolic SNPs (adjusted $p<.05$). The statin effect on fasting blood glucose differs across subtypes ($p=5.6e-6$), so they have potential translational value. Using subtypes in GWAS uncovers 9 additional loci ($p<5e-8$). Subtypes improve predicted T2D risk in pre-diabetics after adjusting for covariates. We also examine cell-type-specific heritability to further biologically characterize subtypes. Overall, RGWAS provides the first calibrated tests for genetic interaction with inferred subtypes, and real data analyses demonstrate its utility.

Copyright © 2018 American Society of Human Genetics. All rights reserved

134

Wavelet screaming: A novel approach to analyzing GWAS data. *W. Denault[1], J. Juodakis[1,2], B. Jacobsson[1,2], A. Jugessur[1,3,4], H. Gjessing[3,4].* 1) Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo; 2) Department of Obstetrics and Gynecology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; 3) 3 Centre for Fertility and Health (CeFH), Norwegian Institute of Public Health, Oslo, Norway; 4) Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway.

The standard GWAS framework leads to an important multiple-testing burden because of the millions of tests being performed simultaneously. Additionally, it does not take into account the functional genetic effect on the response variable. Here we present a new approach to perform GWAS which takes into account the functional nature of the genome. Our method is based on a sliding-window approach that sequentially screens the entire genome for associations. We consider SNPs as a genetic signal, and for every screened region, we transform the genetic signal into the wavelet space. The association is identified at the wavelet coefficient location using a Bayesian Hierarchical model. Our method has natural connections to the Burden test. Our sliding-window approach is fast. It reduces the number of tests to be performed, enhances the detection of association signals by improving the modeling, and also provides a natural fine-mapping tool. We performed simulations to test our method using polygenic simulated phenotypes linked to real genetic data. Our method has the same power as the standard GWAS methodology for single SNP signal, but a higher power for the detection of polygenic/multiple SNP signals. In particular, it performs better than the GWAS standard methodology for small polygenic effects. Indeed, the wavelet transform allowed the detection of signals spread out at multiple independent SNPs which explained a small amount of the phenotypic (<0.1%). From our results, we consider that Wavelet Screaming is a suitable alternative methodology to perform GWAS. Moreover, this method is well-suited for the meta-analysis of multiple cohort studies. We applied our methodology to the HARVEST data set to well-studied phenotypes (e.g., gestational age). We investigated previous SNPs identified through large meta-analysis that are not detected in HARVEST using standard GWAS methodology. Wavelet Screaming detected known loci as well as loci that were previously unidentified. We will release our method via an open-source platform (e.g., CRAN).

135

PASTRY (A method to avoid Power ASymmeTRY): Achieving balanced power for detecting risk and protective alleles in meta-analysis of association studies with overlapping subjects. *E. Kim, B. Han.* Seoul National University, Seoul, South Korea.

Background: To increase samples in an analysis, meta-analysis combines summary statistics from multiple independent studies. If multiple studies in a meta-analysis utilize the same public dataset as controls, the summary statistics from these studies are become correlated. Lin and Sullivan proposed the correlation estimator based on the shared and unshared sample sizes and suggested an optimal test statistic to account for the correlations (AJHG 2010). Their method was shown to achieve similar power to the gold standard method, splitting, which refers to the method that splits shared individuals into the studies prior to meta-analysis when we have access to the genotype data. Many different methods were proposed after Lin and Sullivan, but most of these methods were based on the similar correlation estimator. **Results**: we report a phenomenon that the use of the standard method suggested by Lin and Sullivan can lead to unbalanced power for detecting protective alleles (OR<1) and risk alleles (OR>1). Specifically, when we assumed that the controls were shared, the power for detecting protective minor alleles (OR<1) were lower than the power for detecting risk minor alleles (OR>1). For example, for detecting a MAF 10% and of OR=0.85, simulating meta-analysis of 5 studies showed that the standard method only achieved 62% power whereas splitting achieved 67%. By contrast, when we flipped the effect direction (OR=1.17), the existing method conversely achieved higher power (72%) than splitting. The degree of asymmetry was exacerbated as the minor allele frequency (MAF) decreased. To our knowledge, we are the first to report this phenomenon. After investigating on this phenomenon, we identified that the power asymmetry problem occurred because the standard correlation estimator did not exactly predict the true correlation. The existing estimator was approximated under the simple assumption of the null hypothesis of no effect, but under the alternative hypothesis, the true correlation is dependent on MAF and effect size. Thus, the errors in estimator could lead to substantially unbalanced power. **Conclusions :** To overcome the power asymmetry problem, we developed a method that uses an accurate correlation estimator, called PASTRY. Our method is based on the correlation estimator that was designed to be accurate under the alternative hypothesis. We show that using our method, one can effectively achieve symmetry on power for testing risk and protective alleles.

Copyright © 2018 American Society of Human Genetics. All rights reserved

136

Meta-MultiSKAT: Region-based rare variant meta-analysis of multiple phenotypes using summary statistics. *D. Dutta[1,2], S.A. Gagliano[1,2], J. Weinstock[1,2], M. Zawistowski[1,2], F. Cucca[5,6], D. Schlessinger[7], G. Abecasis[1,2], C. Brummett[3,4], S. Lee[1,2].* 1) Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA; 2) Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA; 3) Department of Anesthesiology, University of Michigan, Ann Arbor, MI, USA; 4) Division of Pain Research, University of Michigan, Ann Arbor, MI, USA; 5) Institute for Genetics and Biomedical Research, Italy; 6) Faculty of Medicine, University of Sassari, Italy; 7) National Institute on Aging, NIH, Baltimore, MD, USA.

In rare variant association analysis, a joint test of multiple correlated phenotypes can increase power to identify sets of traits associated with variants within regions of interest. Although several common variant meta-analysis methods have been developed for multiple correlated outcomes, only limited work has been done for rare variants. Here, we develop a meta-analysis framework, Meta-MultiSKAT to test rare variants in a region of interest for association with multiple continuous phenotypes, using summary statistics from individual studies. Our approach models the heterogeneity of effects between studies using a kernel regression framework and performs a variance component test of association. To make the results robust to model misspecifications, we have developed fast and accurate omnibus tests by approximating the significance of the minimum p-value across tests. In addition, Meta-MultiSKAT accommodates situations where one or more phenotypes have not been measured in a particular study. Our method is applicable even when the contributing studies have differing correlation patterns among the phenotypes. Since Meta-MultiSKAT calculates analytical asymptotic p-values, the method is computationally feasible at a genome-wide level. Large scale numerical studies confirm that Meta-MultiSKAT can maintain a Type-I error rate at exome-wide level of 2.5 x 10^{-06}. Extensive simulations under different models of association show that Meta-MultiSKAT can improve power up to 38% on average over standard single phenotype-based meta-analysis approaches. We applied Meta-MultiSKAT to meta-analyze four white blood cell (WBC) subtype traits from the Michigan Genomics Initiative (MGI) and SardiNIA studies. Meta-MultiSKAT successfully identified *PRG2* (p-value = 3.7x10^{-08}) and *RP11-872D17.8* (p-value = 2.4x10^{-06}), which were identified by the standard single phenotype-based meta-analysis methods as well. However, Meta-MultiSKAT also identified genes that were not identified by standard meta-analysis methods or had an association signal in the individual studies, namely *IRF8* (p-value = 2.6x10^{-07}) and *CCL24* (p-value = 3.5x10^{-06}). Published reports suggest that the regions additionally detected by Meta-MultiSKAT are, indeed associated with WBC subtypes in humans. In summary, Meta-MultiSKAT can provide novel insights into the pleiotropic effects of rare variants.

137

TOPMed based imputation in minority samples. *M.H. Kowalski[1], H. Qian[2], Z. Hou[1], J.D. Rosen[1], L.M. Raffield[3], R. Kaplan[4], E. Boerwinkle[5], K.E. North[6], C. Kooperberg[7], J.G. Wilson[8], A.P. Reiner[9], Y. Li[1,3] on behalf of the TOPMed Hematology and Hemostasis Working Group.* 1) Department of Biostatistics, University of North Carolina, Chapel Hill, NC; 2) Department of Statistics and Operation Research, University of North Carolina, Chapel Hill, NC; 3) Department of Genetics, University of North Carolina, Chapel Hill, NC; 4) Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY; 5) Human Genome Sequencing Center, University of Texas Health Science Center at Houston; Baylor College of Medicine, Houston, Texas; 6) Department of Epidemiology, University of North Carolina, Chapel Hill, NC; 7) Fred Hutchinson Cancer Research Center, Seattle, Washington; 8) Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi; 9) Department of Epidemiology, University of Washington, Seattle, WA.

Background: The NIH/NHLBI Trans-Omics for Precision Medicine (TOPMed) Project generated deep-coverage whole genome sequencing (WGS) on >50,000 individuals from diverse ancestral backgrounds. We anticipated TOPMed sequencing data would improve genotype imputation, particularly for rarer variants and in minority populations. **Methods**: We performed imputation with minimac4 using TOPMed data as reference for individuals from the Jackson Heart Study (JHS, all African Americans [AA]) and Hispanic Community Health Study/Study of Latinos (HCHS/SOL, all Hispanic/Latino [HL]). For imputation with JHS subjects, we excluded them from TOPMed data; the remaining subjects were used as reference. Imputation quality was evaluated in 3082 JHS participants at all TOPMed variants not overlapping those on Affymetrix 6.0; and in 12,803 SOL individuals at all imputed MegaArray markers. We use estimated r^2 for post-imputation quality control (QC); and dosage/true r^2 (squared Pearson correlation between imputed dosages and true genotypes) for quality assessment. We compared performance when using the Haplotype Reference Consortium (HRC) or the 1000 Genomes phase 3 alone as reference. **Results:** In JHS, 51 million (M) markers were well-imputed with standard/lenient QC, including 13.1M with sample minor allele frequency (MAF) <0.05%; in SOL, 60M markers well-imputed (28M with MAF <0.05%). In contrast, approximately 25M (7M with MAF <0.05%) and 30M (8M with MAF<0.05%) markers were well-imputed with HRC and 1000G, respectively. The average dosage r^2 for markers with sample MAF <0.05% exceeded 82% (JHS) and 66% (SOL) with standard/lenient QC, and exceeded 87% (JHS) and 78% (SOL) with estimated r^2 threshold of 0.8. Towards the rare extreme, in JHS, 39% of markers with TOPMed minor allele count (MAC) 10-20 can be well imputed, with average true r^2 77% for sample/JHS singletons, and >80% (80-97%) when JHS MAC >1. Compared with standard reference panels, TOPMed resulted in many more well-imputed rare variants and in higher imputation quality for these rare variants. For example, TOPMed increased the number of well imputed variants with sample MAF <0.05% by >3x and 6x, with 17-20% and 16-24% improvement in average dosage r^2 for markers imputed by both panels, compared to 1000G and HRC, respectively. **Conclusion:** TOPMed proves a much better imputation reference panel for minority populations, in terms of both the number of variants imputable and the quality of the imputed variants. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

138

Creating population-specific reference panels for improved genotype imputation. *J.C. Carlson[1], N.L. Hawley[2], G. Sun[3], H. Cheng[3], T. Naseri[4], M.S. Reupena[5], R. Deka[3], S.T. McGarvey[6,7], R.L. Minster[1], D.E. Weeks[1,8], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium.* 1) Human Genetics, University of Pittsburgh, Pittsburgh, PA; 2) Department of Epidemiology (Chronic Disease), School of Public Health, Yale University, New Haven, CT; 3) Department of Environmental Health, College of Medicine, University of Cincinnati, Cincinnati, OH; 4) Ministry of Health, Government of Samoa, Apia, Samoa; 5) Bureau of Statistics, Government of Samoa, Apia, Samoa; 6) International Health Institute, Department of Epidemiology, School of Public Health, Brown University, Providence, RI; 7) Department of Anthropology, Brown University, Providence, RI; 8) Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA.

Isolated populations like Sāmoans are particularly useful in genomic studies due to reduced haplotype complexity and restricted allelic and locus heterogeneity. Genome-wide association studies (GWAS) in isolated populations have furthered the understanding of human biology, including the genetic architecture of complex traits. However, the existing genotyping arrays and imputation panels upon which most GWAS depend are not designed to adequately capture genetic variation in such populations. For accurate imputation in isolated populations, existing imputation reference panels need to be expanded to include haplotypes derived from population-specific whole-genome sequencing (WGS) data. Here we compared the effectiveness of a Sāmoan-specific reference panel derived from WGS of 1,195 Sāmoans to the TOPMed reference panel, a trans-ethnic panel excluding any self-reported Sāmoans of approximately 60,000 individuals from NHLBI's TOPMed Program (available through the Michigan Imputation Server). Genotypes were imputed using both panels for a separate set of 1,897 Sāmoans—the Sāmoan-specific reference panel using Minimac3 and the TOPMed reference panel using the Michigan Imputation Server. On 5q35.1 in a known body mass index (BMI) locus, we observed discordant imputed genotypes surrounding the previously implicated missense variant, rs373863828. The minor-allele frequency (MAF) of this variant was 27.63% in the Sāmoan-specific imputation and 21.5% in the trans-ethnic imputation. Imputation quality r^2 was 96.3% and 85.3%, respectively. Direct genotyping of this variant indicated that the MAF is 27.66%. We investigated the source of these discordant imputed genotypes by examining haplotype diversity in local and distant haplotypes structures around rs373863828 for both sets of imputed genotypes. As this example demonstrates, the utility of imputed genotypes depends largely on the reference panel used for imputation. Imputation performed with trans-ethnic panels, without consideration of population-specific allele and haplotype frequencies, will lead to biased allele frequencies and inaccurate results in downstream studies. This work highlights the precautions that must be taken to ensure that imputed genotypes of isolated populations are not biased by excluding relevant haplotypes.

139

The global landscape of pharmacogenomic variation. *C. Gignoux[1], E. Sorokin[2], G. Wojcik[2], G. Belbin[3], S. Bien[4], N. Abul-Husn[3], P. Norman[1], C. Hodonsky[5], J. Odgis[3], H. Highland[5], C. Avery[5], S. Buyske[6], T. Matise[7], K. Barnes[1], B. Hailu[8], J.L. Ambite[9], K. North[10], R. Loos[3], C. Haiman[10], C. Kooperberg[4], U. Peters[4], L. Hindorff[11], C. Carlson[4], C. Bustamante[2], E. Kenny[3] on behalf of the Population Architecture using Genomics and Epidemiology (PAGE) Study.* 1) Colorado Center for Personalized Medicine, University of Colorado - Anschutz Medical Campus, Aurora, CO; 2) Department of Genetics, Stanford University School of Medicine; 3) The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai; 4) Fred Hutchinson Cancer Research Institute; 5) Department of Epidemiology, University of North Carolina; 6) Department of Statistics and Biostatistics, Rutgers University; 7) Department of Genetics, Rutgers University; 8) National Institute on Minority Health and Health Disparities, NIH; 9) Information Sciences Institute, University of Southern California; 10) Department of Preventative Medicine, University of Southern California Keck School of Medicine; 11) National Human Genome Research Institute, NIH.

Pharmacogenomic variants are notable for for their common yet highly geographically structured patterns of segregation and strong effects with clinical actionability and use in precision medicine. However, patterns of pharmacogenomic variation across diverse, global populations continues to be understudied. Here, we investigate pharmacogenomic variability across 19,690 variants (2,562 with curated annotations) in 1,024 pharmacologically-relevant genes in the 51,698 individuals from 99 populations in the Population Architecture using Genomics and Epidemiology (PAGE) Study. First, we characterize global patterns of variant diversity, finding a high number of commonly-segregating variants used in recommendations from the Clinical Pharmacogenetics Implementation Committee (MAF>5% in at least one super population, N=2,181). Further, we demonstrate that pharmacogenes are highly differentiated across global populations (1,019 curated variants with a fixation index (FST)> 0.1 with Europeans in at least one global region), and exhibit systematically higher levels of differentiation than matched gene sets ($p<0.05$), indicating the importance of diversifying selection in pharmacogenomics. The unique role of pharmacogenes in interacting with exogenous compounds make them ideal candidates for the interrogation of environmental associations. To explore this, we have assembled a unique geocoded resource of over 20 climate, geographic and ecological variables from NASA, the World Wildlife Fund, GIDEON, and Berkeley Earth. We ran linear mixed effects models across the 19,960 pharmacognomically-relevant variants to identify significant differentiation associating genotypes and these environmental variables, which we term an environmental-wide association study (Enviro-WAS). Across the Enviro-WAS we determined overall significance with a combined FDR <10%. We replicate known selection signatures at *ABCB11* and *CYP3A4* (latitude, $p<1x10^{-10}$ and $1x10^{-6}$, respectively), *PPARG* (altitude, $p<1x10^{-5}$), and identify novel associations between ecological zones and *SULT1A1*, *CYP1A1*, and *FMO5* (all $p<1x10^{-9}$) among others. These associations link prehistoric evolutionary processes to modern clinical significance, and highlight the utility of population genetics to model current allele frequencies in diverse populations. Findings from PAGE and other studies impact our understanding of high-value screening candidates and improves opportunities for personalized medicine globally.

Copyright © 2018 American Society of Human Genetics. All rights reserved

140

Transcriptome-based association study in Hispanic cohorts implicates novel genes in lipid traits. *A.S. Andaleon, L.S. Mogil, H.E. Wheeler.* Loyola University Chicago, Chicago, IL.

Plasma lipid levels are risk factors for cardiovascular disease, a leading cause of death worldwide. While many studies have been conducted on lipid genetics, they mainly comprise individuals of European ancestry and thus their transferability to diverse populations is unclear. We performed genome- and transcriptome-wide association studies of four lipid traits in the Hispanic Community Health Study (HCHS) cohort (n = 11,103), with origins in Mexico, Cuba, Puerto Rico, Central America, the Dominican Republic, and South America. We tested our findings for replication in the Hispanic population in the Multi-Ethnic Study of Atherosclerosis (MESA) (n = 1,364) and compared our results to larger, predominantly European ancestry meta-analyses. In both our GWAS and TWAS, we used a linear mixed model to control for relatedness and included the first five genotypic principal components and geographic region as covariates. In our GWAS, five previously-implicated SNPs reached significance ($P < 5 \times 10^{-8}$). After predicting gene expression levels with PrediXcan software using multi-tissue models built in the Genotype-Tissue Expression Project (GTEx) and multi-ethnic monocyte models built in MESA, we tested genes for association with four lipid phenotypes (total cholesterol, HDL, LDL, triglycerides). This revealed 255 significant gene-phenotype associations (FDR < 0.05) with 84 unique significant genes, many of which occurred across multiple phenotypes, tissues and MESA populations. Of these significant genes, 36 were previously implicated within the GWAS catalog, such as *CETP*, *PSRC1*, and *DOCK7*, and 29/36 replicated in MESA and 30/36 replicated in the Genetic Lipid Global Consortium (GLGC). We found 48 of the significant genes are novel for any lipid association, including *TSNAXIP1*, which associated with HDL in eight tissues, and *C19orf52*, which associated with total cholesterol, HDL, and LDL in two tissues. Of the 58 novel gene-phenotype associations found significant, 27 replicated independently in MESA and 42 replicated independently in GLGC at P < 0.05, with 13 associations replicating in both. The largest and most diverse MESA expression prediction model, including African Americans, Caucasians, and Hispanics, had more significant genes (18, n = 1,163) compared to the Caucasian model (8, n = 578), indicating that to fully characterize the impact of genetic variation between populations, larger studies in non-European ancestry populations are needed.

141

A 100,000 Genome project haplotype reference panel of 57,786 haplotypes. *S. Shi[1], S. Hu[1], S. Myers[1,2], J. Marchini[1,2].* 1) Department of Statistics, University of Oxford, Oxford, United Kingdom; 2) Wellcome Center for Human Genetics, University of Oxford, United Kingdom.

The 100,000 Genomes Project aims to sequence 100,000 genomes from around 70,000 people from the UK. It is expected that the use of high coverage sequencing will produce an almost complete charactization of the genetic variation in the project participants and will constitute the largest human genetic variation resource ever collected in the UK, and maybe the world. One of our main research goals is to create an accurate haplotype reference panel for use in genotype imputation. We have created a preliminary haplotype reference panel using called genotypes in 28,893 participants. Sequencing data was mapped and genotypes were called using the central processing pipelines developed by 100,000 Genomes Project. This resulted in a dataset consisting of ~230 million SNPs across the autosomes. The dataset consists of a diverse set of ancestries with percentages self reporting as White, Asian, Black and Mixed ancestry of 69.2%, 8.7%, 2.3% and 2.1% respectively. Project participants consist of probands for rare diseases and their close relatives, so the dataset as a whole contains large amount of related individuals. For example, 60.67% of the 28,893 participants have at least one first degree relative also in the study, which greatly aids phasing. We used a new phasing program SHAPEIT4 to phase the genotypes at an overlapping set of 820,548 SNP sites included in the HRC reference panel on chr20. We assessed phasing performance using 200 trio parents, phased without their children, but together with 28,693 other samples. The majority (81%) of these trio parents reported White British ancestry and had a median switch error rate of 0.75%. The phasing was carried out without use of any relatedness information. We also directly compared the resulting 57,786 phased haplotypes to the HRC reference panel (64,976 haplotypes) in terms of imputation performance, by imputing genotypes into 10 individuals of European ancestry, based on genotypes on Illumina 1M-Duo3_C genotyping array, and comparing the results to genotypes derived from high-coverage sequencing. At variants with frequency 0.01% we obtained a mean imputation r^2 of 0.65 and 0.75 using the HRC and 100,000 Genomes reference panels respectively. We will also report comparisons of phasing methods that use read information and relatedness and how this translates into downstream imputation performance, and the utility of imputing the UK Biobank dataset using the 100,000 Genomes reference panel.

Copyright © 2018 American Society of Human Genetics. All rights reserved

142

Identification and characterization of adaptive regulatory variation in diverse human populations. *J.J. Vitti[1,2], S. Gosai[1,3], R. Tewhey[1,4], S. Reilly[1,2], P.C. Sabeti[1,2,3].* 1) Broad Institute, Cambridge, MA; 2) Harvard University Department of Organismic and Evolutionary Biology, Cambridge MA; 3) Harvard Medical School, Boston MA; 4) Jackson Laboratory, Bar Harbor ME.

The lens of evolution is a powerful framework with which to interpret human genetic variation data, because natural selection acts on phenotypes affecting long-term outcomes of reproduction and survival. Here, we present a survey of the 1000 Genomes Phase 3 dataset (1000G) for signals of positive natural selection within the past hundred thousand years. We have adapted the Composite of Multiple Signals (CMS) method for Bayesian identification of regions under positive natural selection with increased sensitivity for “soft sweep” scenarios, where selected mutations may segregate on multiple haplotype backgrounds and/or may fail to reach fixation. We have fit a novel demographic model for the 1000G dataset consistent with published analyses. We have trained a Convolutional Neural Network (CNN), *deepSweep*, that can localize signals of selection within target regions to tractable sets of mutations at high precision. With these new tools in hand, we identify 650 regions of positive selection in human populations, 301 of which are novel. Integrating our results with phenotypic and epigenetic insight from the ENCODE and UKBiobank Consortia identifies the importance of selective pressures from infectious disease and changes in diet. Furthermore we identify tissue-specific enrichments for putative adaptive regulatory alleles, implicating specific systems targeted by natural selection. We functionally validate many of our top-scoring candidate variants from within these regions for regulatory activity, and find mechanistic support for the hypothesis that recent adaptation in human populations has been driven primarily by mutations affecting gene regulation. We highlight adaptive hypotheses at specific loci with intent to cast light on outstanding questions regarding the differential burden of chronic diseases among contemporary ethnic groups, significantly expanding the known repertoire of variants impacting recent human evolution. In one example, we propose that cholera or other ancient infectious diseases may have driven variation at the *STX4* locus in South Asian populations while pleiotropically conferring an increased risk to metabolic syndrome.

143

Insights into effective methods for Mendelian gene discovery from family based genomic analysis of over 22,000 families from worldwide populations. *A.H. O'Donnell-Luria[1,2,3], J.E. Posey[4,5], J.X. Chong[6,7], E.E. Blue[6,8], N.L.M. Sobreira[4,9], F. López-Giráldez[10,11], J. Knight[10,12], S. Baxter[1], J.R. Lupski[4,5,13,14,15], D. Valle[4,9], A. Hamosh[4,9], K. Doheny[4,9], D. Avramoupolus[4,9], H.L. Rehm[1,16], D.G. MacArthur[1,3,16], T.C. Matise[17,18], S. Buyske[17,18,19], R.P. Lifton[10,20,21,22], M. Gunel[10,12,20], S.M. Mane[10,11], M.B. Gerstein[10,23], D.A. Nickerson[6,7], M.J. Bamshad[6,7,24], GSP Coordinating Center & Baylor Hopkins, Broad, Yale and UW Centers for Mendelian Genomics.* 1) Broad Center for Mendelian Genomics, Broad Institute of MIT and Harvard, Cambridge, MA; 2) Division of Genetics and Genomics, Boston Children’s Hospital, Boston, MA; 3) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 4) Baylor Hopkins Center for Mendelian Genomics, Baylor College of Medicine, Houston, TX and Johns Hopkins University, Baltimore, MD; 5) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 6) University of Washington Center for Mendelian Genomics, University of Washington, Seattle, WA; 7) Department of Genome Sciences, University of Washington, Seattle, WA; 8) Division of Medical Genetics, University of Washington, Seattle, WA; 9) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD; 10) Yale Center for Mendelian Genomics, Yale School of Medicine, New Haven, CT; 11) Yale Center for Genome Analysis, Yale School of Medicine, Yale University, New Haven, CT; 12) The Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 13) Department of Genetics, Yale School of Medicine, New Haven, CT; 14) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 15) Texas Children’s Hospital, Baylor College of Medicine, Houston, TX; 16) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 17) Genomic Sequencing Program Coordinating Center, Rutgers University, Piscataway, NJ; 18) Department of Genetics, Rutgers University, Piscataway, NJ; 19) Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ; 20) Department of Neurosurgery, Yale School of Medicine, New Haven, CT; 21) Department of Internal Medicine, Yale University School of Medicine, New Haven, CT; 22) Laboratory of Human Genetics and Genomics, The Rockefeller University, New York, NY; 23) Yale University Medical School, Computational Biology and Bioinformatics Program, New Haven, CT; 24) Department of Pediatrics, University of Washington, Seattle, WA.

Identifying genes and variants underlying rare diseases informs biology and medicine, yet phenotypic annotation for thousands of human genes is lacking. Technical advances in analysis of exome and genome sequence data (ES, GS) set the stage for establishment of the NIH-supported Centers for Mendelian Genomics (CMGs) and accelerated the pace of novel gene discovery for Mendelian phenotypes. The CMGs provide a collaborative framework that engages physicians and scientists from 74 countries and 1,382 organizations with resources to analyze genomic data and discover novel disease gene-phenotype relationships. Sequencing and analysis of 22,734 families with suspected Mendelian conditions has led to the discovery of 1,252 novel gene-phenotype associations (~200/year). Of the 1,252 novel discoveries to date, 503 (40.2%) meet stringent causality requirements for definition of a novel gene-phenotype association. An additional 395 discoveries are phenotypic expansions, increasing knowledge of previously delineated Mendelian conditions. The CMGs have disseminated these discoveries through 465 publications; data sharing via dbGaP, DUOS, Geno2MP, and ClinVar; and submission of candidate genes into Matchmaker Exchange through GeneMatcher, MyGene2, and matchbox. We review the methods employed by the CMGs over their first six years and highlight the continued power and cost-effectiveness of ES as a primary strategy for diagnosis and gene discovery. The CMGs have developed computational tools to improve the identification of de novo variants and recognition and interpretation of protein truncating variants. In our experience, GS of exome-negative cases has yielded very few discoveries, aside from identification of coding variants missed due to poor coverage or structural variants overlapping known disease genes. However, GS combined with RNA-sequencing can provide a functional readout of splicing and skewed allele balance that has aided interpretation of functional non-coding variation. Phenotypic annotation of all human genes, development of bioinformatic tools and analytic methods, exploration of non-Mendelian modes of human disease inheritance including reduced penetrance, multilocus variation and oligogenic inheritance, enhanced data sharing worldwide, and integration with clinical genomics, all employed by the CMGs, are critical steps on the path to realizing the full contribution of rare disease research to human biology and health.

Copyright © 2018 American Society of Human Genetics. All rights reserved

144

The genetic landscape of Diamond-Blackfan anemia. *J.M. Verboon[1,2], J.C. Ulirsch[1,2,3], S. Kazerounian[4], D. Yuan[4], L.S. Ludwig[1,2], M.H. Guo[2,5], N.J. Abdulhay[1,2], C. Fiorini[1,2], R.E. Handsaker[2], G. Genovese[2], E. Lim[2], A. Cheng[1,2], B. Cummings[2,3], K.R. Chao[2], S. McCaroll[6], A. O'Donnell[2], N. Gupta[2], S.B. Gabriel[2], D.G. MacArthur[2], E.S. Lander[2], M. Lek[2], L. Da Costa[7,8], D.G. Nathan[1], A. Korostelev[9], R. Do[10], H.T. Gazda[2,4], V.G Sankaran[1,2,11], Collaborative DBA Consortium.* 1) Division of Hematology/Oncology, The Manton Center for Orphan Disease Research, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115 USA; 2) Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; 3) Program in Biological and Biomedical Sciences, Harvard University, Cambridge, MA, USA; 4) Division of Genetics and Genomics, The Manton Center for Orphan Disease Research, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA; 5) Division of Endocrinology, Boston Children's Hospital and Department of Genetics, Harvard Medical School, Boston, MA 02115 USA; 6) Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; 7) Laboratory of Excellence for Red Cell, LABEX GR-Ex, F-75015, Paris, France; 8) University Paris VII Denis DIDEROT, Faculté de Médecine Xavier Bichat, F-75019, Paris, France; 9) RNA Therapeutics Institute, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 368 Plantation Street, Worcester, Massachusetts 01605, USA; 10) Department of Genetics and Genomic Sciences and The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 11) Harvard Stem Cell Institute, Cambridge, MA 02138, USA.

Diamond-Blackfan Anemia (DBA) is a rare blood disease characterized by a dearth of red blood cell (RBC) progenitors and precursors that occurs in 7 out of 1 million live births. Although DBA is clinically diagnosed, approximately 60% of cases are attributable to an identifiable genetic lesion, almost exclusively in proteins encoding the small and large ribosome subunits. In order to establish a more comprehensive picture of the genetic underpinnings of this disease, we recruited a cohort of 472 individuals and performed whole exome sequencing (WES). Leveraging known variation in healthy control populations and by using multiple annotations of pathogenicity, we were able to attribute 78% of cases to rare (majority singleton) and predicted damaging mutations. While most of these cases are explained by mutations in 15 of the 17 ribosome proteins genes previously implicated in DBA, we were able to nominate 8 new RP genes as putatively causal for DBA. Among the genes known to cause DBA, we were able to identify more than 30 deletions using WES coverage to infer copy number, which we could successfully validate with digital droplet PCR. Additionally, we found an enrichment of mutations in the extended splice and untranslated regions of DBA genes. In order to verify the effects of these extended splicing mutations, we created patient-derived cell lines, performed RNA sequencing and were able to identify cases resulting in exon extension, run on transcription, and other functional defects. Leveraging the size of our cohort, we also identified several robust genotype-phenotype associations, such as an increased frequency of congenital abnormalities in RPL5 and RPL11 individuals and an increased frequency of remission in RPS24 and RPL11 individuals. Finally, we were able to identify multiple cases of phenocopy, including 9 cases resulting in ADA2 deficiency caused by rare *CECR1*recessive mutations. Aside from *CECR1*, no new genes were identified at exome-wide significance by gene burden testing. Combined with power analyses, this suggests that no single gene remains undiscovered that can account for more than 5% of cases. Overall, our study provides valuable insights into the genetic architecture of this rare bone marrow failure disorder, while generally informing the design and analysis of large cohorts of a rare, Mendelian disease.

145

Quantifying the contribution of X-linked coding variants to developmental disorders. *H.C. Martin[1], N. Akawi[2], A. Sifrim[1,3], K.E. Samocha[1], J. Kaplanis[1], J.F. McRae[4], M.E. Hurles[1], the Deciphering Developmental Disorders Study.* 1) Wellcome Sanger Institute, Hinxton, Cambridgeshire, United Kingdom; 2) Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, U.K; 3) Department of Human Genetics, University of Leuven, KU Leuven, Leuven, Belgium; 4) Illumina Inc., San Francisco, USA.

Pathogenic variants in >830 X-linked genes have been robustly associated with developmental disorders. We evaluated the burden of X-linked coding variation in 5659 males and 4200 females from the Deciphering Developmental Disorders Study. Analysis of sex-specific *de novo* enrichment per gene allowed us to delineate three primary classes of X-linked disorder: recessive, dominant and semi-dominant. For example, some genes previously annotated as recessive had a significant burden of damaging *de novo* mutations in females, suggesting a dominant mode (e.g. *CNKSR2*), while others with previously ambiguous modes of inheritance were classified as likely semi-dominant and male-lethal due to having multiple *de novo* truncating mutations in females but none in males (e.g. *DDX3X*). We improved detection of new X-linked developmental disorder genes by combining individual tests with increased power to detect specific inheritance modes. Our data-driven classification of the inheritance mode of X-linked disorders provides a more accurate resource for diagnostic testing. Chromosome-wide burden testing shows that ~7% of male cases can be attributed to pathogenic X-linked coding variants, of which the vast majority are in previously associated genes and ~30% arise *de novo*. Moreover, we estimate that in male probands, only 24% of ultra-rare missense variants in known DD-associated genes are likely to be pathogenic, which implies a substantial risk of incorrect diagnosis. We find that ~6% of female probands have a pathogenic *de novo* mutation, of which 80% fall in known genes. These analyses show that the male bias in developmental disorders is not primarily due to X-linked disorders and thus needs further investigation.

Copyright © 2018 American Society of Human Genetics. All rights reserved

146

WES application in a large cohort of arthrogryposis: Evidence for oligogenic inheritance, and candidate novel disease genes. *D. Pehlivan[1,2], Y. Bayram[1,3], N. Gunes[4], A. Gezdirici[5], Z. Coban-Akdemir[1], J.M. Fatih[1], T. Yildirim[6], I.A. Bayhan[6], A. Bursali[6], E. Yilmaz-Gulec[5], E. Karaca[1], S.N. Jhangiani[7], D.M. Muzny[7], R.A. Gibbs[1,7], N.H. Elcioglu[8], J.E. Posey[1], B. Tuysuz[4], J.R. Lupski[1,9,10].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Section of Neurology, Department of Pediatrics, Baylor College of Medicine, Houston, Texas; 3) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA; 4) Department of Pediatric Genetics, Istanbul University Cerrahpasa Medical Faculty, Istanbul, Turkey; 5) Department of Medical Genetics, Kanuni Sultan Suleyman Training and Research Hospital, Istanbul, Turkey; 6) Department of Orthopedics and Traumatology, Baltalimani Bone Diseases Training and Research Hospital, Istanbul, Turkey; 7) Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas; 8) Department of Pediatric Genetics, Marmara University Medical School, Istanbul, Turkey; 9) Department of Pediatrics, Baylor College of Medicine, Houston, Texas; 10) Texas Children's Hospital, Houston, Texas.

Background: Arthrogryposis is the contracture of at least two non-consecutive joints, and represents a descriptive clinical feature that is often part of a neuromuscular condition rather than a single unifying diagnostic entity. There are more than 400 disorders found to have arthrogryposis as a feature and in excess of 200 Mendelian ‘disease genes’ have been described with this condition. Many cases remain without a molecular diagnosis despite the application of genome-wide screening assays. **Material and Methods:** We applied whole exome sequencing (WES) in a cohort of 113 families of Turkish origin with a clinical feature of arthrogryposis. In three cases for whom a CNV was suspected based on analysis of exome variant data or previously performed clinical genetics diagnostic studies such as karyotyping or low resolution chromosomal microarray (CMA), we performed an additional Agilent custom design whole genome array (Baylor Genetics laboratory, CMA). **Results:** A molecular diagnosis was identified in 73% (83/113) when including both known and candidate genes, and 50% (57/115) involving known disease genes only. Out of 83 solved families, 65 of which revealed mutations in 45 distinct known genes, further documenting the genetic heterogeneity. *CHRNG* and *ECEL1* are the most commonly mutated genes (22%) in Turkish population. Multilocus pathogenic rare variation including two or even three genes is observed in ~17% (19/113) of families. *RYR3*, *MYOM2*, *ERGIC1*, and *SPTBN4* are novel arthrogryposis genes identified in two or more families and *TMEM214*, *FGFRL1* and *FLII* are amongst the strongest candidate genes identified in single families. In 3 families, we found *de novo* CNVs and structural variants (SVs) likely contributing to the disease phenotype. We also show evidence for monoallelic and biallelic variants in the same gene in association with either clinically similar or distinct syndromes. **Discussion:** We initiated a comprehensive study of a cohort of unrelated subjects to explore the molecular etiology of arthrogryposis. Through WES, we documented the tremendous genetic heterogeneity (45 known genes in 65 families), identified multiple novel candidate genes, and showed evidence for multilocus pathogenic variation in 17% of families; the latter supporting models of oligogenic inheritance in arthrogryposis. Furthermore, we provide evidence for monoallelic and biallelic variants in the same gene causing similar and different AD or AR conditions.

147

Genetic testing of 1346 patients with cerebral palsy reveals a monogenetic cause of disease in one-third of cases, vast genetic heterogeneity, and a significant recurrence risk. *F. Millan, H.Z. Elloumi, C. Teigen, J. Scuffins, R.I. Torene, K. Retterer, D.A. McKnight.* GeneDx, Gaithersburg, MD.

Cerebral palsy (CP) is a broad diagnostic term encompassing disorders impacting movement and posture caused by changes in the fetal or infant brain. CP is a common clinical diagnosis, with an incidence of 1 in 500 births, and it has been historically attributed to pre- and perinatal complications, however, recent studies have suggested that a significant proportion of CP may be due to genetic causes. The objective of this study was to establish the positive diagnostic rate (PDR) of genetic testing for CP. Results from exome sequencing (ES) of 1346 patients with CP were retrospectively reviewed. ES yielded a positive result in 32.7% of cases (440/1346). Testing of a proband concurrently with parents (trio) significantly improved the diagnostic outcome yielding a PDR of 35.3% compared to a PDR of 23.3% for proband-only testing ($p<0.005$). Positive findings were reported in 225 different genes, indicating the vast genetic heterogeneity of CP. Most (65.2%) of the causative variants were in genes with autosomal dominant (AD) inheritance, 20.7% were autosomal recessive (AR), and 13.4% were X-linked (XL). *CTNNB1* (4.1%, 18/440) and *KIF1A* (1.8%, 8/440) were the genes most often associated with a positive result. Additionally, 12% (53/440) of cases had positive results in recently published disease genes. Pathway analysis of the positive genes revealed that signaling and metabolic pathways were enriched this cohort, with 36.5% (82/225) and 33% (74/225) of the genes respectively. Trio testing revealed that the majority of patients diagnosed with an AD or XL disorder (71.4%) had de novo variants, including 5 mosaic cases. A significant recurrence risk was revealed in 30% (130/440) of cases: 21% (92/440) were biallelic AR, 3% (13/440) were inherited XL, 4% (18/440) were inherited AD, and 1.6% (7/440) were inherited from a mosaic parent. These data support the benefits of genetic testing for patients with CP. Specifically, ES testing revealed a genetic etiology in almost a third of patients with CP and the high rate of de novo positive findings supports utilizing a trio approach for testing these individuals. Lastly, genetic testing of patients with CP can inform an accurate recurrence risk, as opposed to the general attributions to complications at birth, and also provide information for prognosis, adjusted therapies, and management options.

Copyright © 2018 American Society of Human Genetics. All rights reserved

148

Diagnostic yield from WES, WGS and RNA testing among 213 neuromuscular families: Known versus novel disease genes, coding versus non-coding variants. *L.B. Waddell[1,2], M. Lek[3], E.C. Oates[4], R. Ghaoui[5], G.L. O'Grady[6], S.A. Sandaradura[1,2,7], B.B. Cummings[8,9,10], E. Valkanas[8,9,10], R. Roxburgh[11], S. Bryen[1,2], A. Bournazos[1,2], F. Evesson[1,2,12], J.L. Marshall[8,9,10], K. Chao[8,9,10], K.J. Jones[1,2,7], L. Douglas[7], M. Rodrigues[11], M. Davis[13], N.G. Laing[14], K.N. North[15,16], N.F. Clarke[1,2], D.G. MacArthur[8,9,10,17], S.T. Cooper[1,2,12].* 1) Kids Neuroscience Centre, University of Sydney; Kids Research, Children's Hospital at Westmead, New South Wales, Australia; 2) Discipline of Child and Adolescent Health, Faculty of Medicine and Health, The University of Sydney, Westmead, New South Wales, Australia; 3) Yale University, New Haven, USA; 4) School of Biotechnology and Biomolecular Sciences, University of New South Wales, Randwick New South Wales, Australia; 5) Royal Adelaide Hospital, Adelaide, South Australia, Australia; 6) Starship Children's Health, Auckland District Health Board, Auckland, New Zealand; 7) Clinical Genetics, Children's Hospital at Westmead, Westmead New South Wales, Australia; 8) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston Massachusetts, USA; 9) Medical and Population Genetics, Broad Institute of Harvard & MIT, Boston Massachusetts, USA; 10) Center for Mendelian Genomics, Broad Institute of Harvard & MIT, Boston Massachusetts, USA; 11) Department of Neurology, Auckland DHB, Auckland, New Zealand; 12) Children's Medical Research Institute, Westmead New South Wales, Australia; 13) Department of Diagnostic Genomics, PathWest Laboratory Medicine, Perth, Western Australia, Australia; 14) Centre for Medical Research, University of Western Australia, Harry Perkins Institute of Medical Research, Perth, Western Australia, Australia; 15) Murdoch Children's Research Institute, Parkville Victoria, Australia; 16) The Royal Children's Hospital, Melbourne, Victoria, Australia; 17) Harvard Medical School, Boston Massachusetts, USA.

Aim: A review of diagnostic outcomes from complementary massively parallel sequencing techniques in a large cohort of 213 families with neuromuscular disorders (NMD). **Methods:** 85 trios and 128 individuals were subject to whole exome sequencing (WES) at the Broad Institute of Harvard and MIT or NMD panel screening at PathWest. 5/213 underwent additional whole genome sequencing (WGS; Broad Institute); 13/213 underwent additional muscle RNA sequencing (RNAseq; Broad Institute); 18/213 underwent both WGS and muscle RNAseq (Broad Institute). 15/213 had additional muscle mRNA laboratory investigations (Kids Neuroscience Centre). **Results:** To date, a genetic diagnosis has been identified in 129/213 (61%) families. Six novel disease genes have been identified, with an average time of 31 months from identification of the gene, to E-publication in a peer reviewed journal. Seven novel or expanded phenotypes for known NMD genes were also determined. One third (42/129) of diagnosed families possessed at least one splicing variant. WES alone provided a diagnosis for 92/129 (71 %) diagnosed families. Additional sequencing and functional genomics pipelines (WGS, RNAseq, mRNA studies) were required for provision of a genetic diagnosis for 37/129 (29 %) diagnosed families. 29/37 (78 %) families not diagnosed via initial WES screening, were shown subsequently to possess (at least one) splice-altering variant in known disease genes. 10/29 (33 %) of these families had essential splice variants. 19/29 (66 %) of these families possessed more complex splicing variants involving the extended splice site, intronic or exonic variants activating cryptic splice sites, or structural rearrangements. **Discussion:** Our results from a large cohort of families with NMD propose a step-wise approach to genomic analyses: initial screening via a targeted panel, triaging to WES, then WGS/RNA studies. Many cases solved by RNAseq and mRNA studies required additional insight from intronic variation provided by WGS. Novel gene discovery among a pre-screened cohort we believed would be enriched for new disease genes was 5 %; functional genomics investigation to reach publication was lengthy (average 2.5 years), expensive and highly interdisciplinary. We have yet to find a diagnosis in 84/213 (39 %) families. Our collective data confirm that variants in common disease genes occur commonly – with 'tricky variants' in known disease genes (16 %) more common than novel gene discovery (5 %).

149

Mutation rate heterogeneity and selective constraint in 28,000 deep whole genomes. *P.J. Short[1], C. Penkett[2], K. Stirrups[2], L.C. Francioli[4,5], K.J. Karczewski[4,5], K.E. Samocha[1], D.G. MacArthur[4,5], W.H. Ouwehand[3,2,1], J.C. Barrett[1], M.E. Hurles[1] on behalf of The NIHR BioResource, the 100,000 Genomes Project RD Pilot, and the gnomAD consortium.* 1) Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom; 2) NIHR BioResource–Rare Diseases, Cambridge Biomedical Campus, Cambridge, United Kingdom; 3) Department of Haematology, University of Cambridge and NHS Blood and Transplant, Long Road, Cambridge CB2 0PT, UK; 4) Analytic and Translational Genetics Unit (ATGU), Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA; 5) Broad Institute of MIT and Harvard, Cambridge, MA, USA.

Protein-coding genes identified as selectively constrained (depleted of genetic variation) in healthy individuals are highly enriched for disease-associated variants. This approach has dramatically improved identification of pathogenic variants in a number of genetic disorders. However, a large fraction of patients still remain undiagnosed after whole exome sequencing, motivating the increasing use of whole genome sequencing (WGS). The search space is much greater in WGS and functional elements are less well-defined, underscoring the need for frameworks to assess selective constraint genome-wide. Estimating constraint depends on accurate models of the germline mutation rate and background selection. We used WGS data from 15,496 individuals from the genome aggregation database, 13,049 individuals from the BRIDGE consortium, and *de novo* mutations from 1,548 WGS trios to train and validate a random forest regression model of the mutation rate. This model incorporates 32 features including CpG methylation, recombination rate, replication timing, nucleotide sequence, and germline chromatin marks and greatly outperforms established models based on sequence alone. Using this improved mutation model, we compared expected to observed variation in 1.7 million putative regulatory elements from >200 different tissues. We identify pervasive selective constraint in open chromatin regions, promoters, UTRs, and other element classes, and a strong correlation between constraint and the number of tissues in which an element is active. However, our analyses suggest that nucleotide-level evolutionary conservation metrics are more predictive of constraint, and hence likely disease relevance, than element-wide measures. We also show patterns of constraint in non-coding elements more closely match recessive than dominant disease genes, suggesting that the contribution to disease from regulatory variation may primarily be recessive or oligogenic. We find that the majority of selectively constrained nucleotides lie within poorly evolutionarily conserved elements and estimate that a large number of WGS (>1 million) will be required to robustly identify constrained non-coding elements using current methodology. We show that approaches leveraging nucleotide-level features can identify constrained bases within otherwise neutrally evolving elements and intend to present preliminary results applying this approach to identify pathogenic non-coding variation in developmental disorders.

Copyright © 2018 American Society of Human Genetics. All rights reserved

150

Widespread transcriptional scanning in testes modulates gene evolution rates. *B. Xia[1], M. Baron[1], Y. Yan[1], F. Wagner[1], S. Kim[5], D. Keefe[3], J. Alukal[3], J. Boeke[2,4], I. Yanai[1,2].* 1) Institute for Computational Medicine, NYU School of Medicine, New York, NY; 2) Department of Biochemistry and Molecular Pharmacology, NYU School of Medicine, New York, NY; 3) Department of Obstetrics and Gynecology, NYU School of Medicine, New York, NY; 4) Institute for Systems Genetics, NYU School of Medicine, New York, NY; 5) Department of Pathology, NYU School of Medicine, New York, NY.

A long-standing puzzle in molecular biology relates to why male germ cells in the testes express the largest number of genes relative to all other cell types or organs. Several hypotheses have attempted to explain this observation, including leaky transcription and extensive chromatin remodeling, however the supporting evidence of functional relevance for each has been limited. Here we used single-cell transcriptomics to reveal the gene expression dynamics of spermatogenesis in the human testes. Consistent with previous reports from bulk RNA-Seq, we detected the expression of ~87% of all protein-coding genes in the germ cells. Surprisingly, we found that these genes maintain significantly lower germline mutation rates than the unexpressed genes. Moreover, we found that transcription template strands of the expressed genes have even lower germline mutation rates than the coding strands. The fact that such an asymmetry was not observed in the unexpressed genes of the male germline strongly implicates a signature of transcription-coupled repair (TCR). Consistently, we found a striking pattern of inversed asymmetry between germline-expressed genes and their upstream sequences, as expected from bidirectional transcription events. Together, these results led us to hypothesize a model that widespread '**transcriptional scanning**' in the male germ cells functions to systematically check and remove DNA damage to safeguard its genome integrity. This novel mechanism can explain the contradictory roles that TCR and transcription-coupled DNA damage exert over the germline mutation rates throughout the genome. Interestingly, the 'transcriptional scanning' model also accounts for the long-observed higher evolutionary rates in genes related to the sensory- and immune-defense system-related genes, as well as in male reproduction genes. Collectively, our results indicate that widespread transcription in the testes maintains DNA integrity for the majority of genes, while selectively promotes variation for the remaining genes at the population level.

151

Genome-wide detection and characterization of *de novo* repeat mutations in healthy trio families. *I. Mitra[1,2], M. Gymrek[1,3].* 1) Department of Medicine, University of California San Diego, La Jolla, CA; 2) Bioinformatics Program, University of California San Diego, La Jolla, CA; 3) Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA.

Short tandem repeats (STRs), also called microsatellites, are a class of complex structural variants composed of 1-6bp repeating units. STRs exhibit mutation rates that are orders of magnitude higher than SNPs, indels, or CNVs, and thus represent one of the largest sources of human genomic variability. So far, direct studies of *de novo* STR mutations (dSTRs) have been limited to several hundred loci due to technical challenges. Here, we conduct the first population-scale analysis of *de novo* STR variation genome-wide (GW). We use recently developed tools for accurate STR profiling on 30X coverage whole-genome sequencing (WGS) data of 1000 families from the Simons Simplex Collection (SSC). We first develop a new method to identify germline dSTRs in WGS data of parent-offspring families (trios), including stringent quality filtering steps to restrict to high-confidence dSTRs. Our method incorporates genotype likelihood scores and our previously published model of the STR mutation process to determine the posterior probability of a mutation at each locus. Applying our method to the SSC samples identified 50-100 dSTRs per child and per-locus mutation rates that are highly concordant with previous studies. Similar to other class of de novo variants, dSTR rates were strongly correlated with the age of the father and showed on average a 3-fold increase in mutational burden from mothers vs. fathers. We used our GW dSTR panel to characterize STR mutation patterns and interrogate mechanisms contributing to STR mutation. The majority of mutations are additions or subtractions of a single repeat unit, with step sizes following a geometric distribution. Mutation rates were strongly correlated with repeat unit length and repeat track length, and more weakly correlated with repeat unit base composition, recombination rate, and local GC content. Intriguingly, our results confirm a previously observed bias in mutation direction dependent on the length of the parent allele: alleles longer than the population mean are far more likely to contract, whereas shorter alleles are more likely to expand. Finally, we characterize effects of recombination, local SNP mutation rate, and mutation modifier mutations on GW STR mutation processes. GW identification of STR mutations is likely to be an important tool for enabling discovery of novel pathogenic loci and provides a rich resource for interrogating mutation mechanisms at one of the largest sources of polymorphism in humans.

Copyright © 2018 American Society of Human Genetics. All rights reserved

152

Deciphering signatures of mutational processes in human germline. *V. Seplyarskiy[1,2], R. Soldatov[2], J. Goldmann[3], P. Kharchenko[2], C. Gilissen[3], W. Wong[4], S. Sunyaev[1,2].* 1) Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 2) Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA; 3) Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, the Netherlands; 4) Inova Translational Medicine Institute (ITMI), Inova Health Systems, Falls Church, VA, USA.

Stereotypic mutational processes operating in human germline are the source of genetic diversity and the cause of hereditary diseases. Patterns of germline mutations vary on different scales including existence of single nucleotide hotspots, clusters of multiple mutations at scales of ten kilobases and megabase-scale variation. Observed variation is a consequence of exposures to a combination of unknown mutational processes. However, etiology, intensities and spectrum of mutational processes in germline are almost unexplored. Current understanding of the mutational processes in human cells is primarily based on cancer data. Inference of mutational signatures extracted from cancers relies on the fact that individual tumors have dramatic variation of exposure to a mutational process. Here we made use of strong variability of mutational patterns along the genome to infer the underlying mutational processes. We assume that variability of the mutational spectra between loci is driven by the difference in relative contribution of an unknown but fixed set of stereotypic mutational processes. To formally extract mutational signatures, the genome was binned in non-overlapping windows of fixed size (e.g. 20 or 100 kilobases) and spectra were compared between windows. We use very rare polymorphisms (allele frequency below 10^{-4}) from GNomad and TOPMEDs projects served as input. Inference of signatures was formulated as a matrix decomposition problem. Using independent component analysis to perform matrix decomposition we discovered seven major mutational processes including signatures created by transcription coupled nucleotide excision repair, error-prone asymmetric bypass of DNA damages during replication, signature associated with replication timing, signature associated with repeat expansion and oocyte-specific signature. These signatures were cofirmed with *de novo* germline mutations. Oocyte-specific signature is localized in regions with disproportionally high fractions of mutations of maternal origin, which were recently discovered in studies of *de novo* mutations in human trios. This signature is active in several genomic regions comprising about 5% of the genome; and shows the highest intensity on non-transcribed strand of long genes (WWOX1, CSMD1). Also, we show that replication-associated signatures predict replication fork polarity and inter-origin distance in germline opening up an avenue of mutation-based inference of molecular features in human cells.

153

Inferring past generation times from changes in the mutation spectrum in human evolution. *P. Moorjani[1,2,6], Z. Gao[3,6], G. Amster[4], M. Przeworski[4,5].* 1) Department of Molecular and Cell Biology, University of California, Berkeley, CA; 2) Center for Computational Biology, University of California, Berkeley CA; 3) Howard Hughes Medical Institute & Department of Genetics, Stanford University, Palo Alto, CA; 4) Department of Biological Sciences, Columbia University, New York, NY; 5) Department of Systems Biology, Columbia University, New York, NY; 6) these authors have contributed equally.

Recent sequencing studies of human pedigrees have shown that the numbers of *de novo* mutations increases with ages of reproduction of both parents in a sex-specific and context dependent manner (depending on flanking sequence), reflecting characteristics of the underlying mutational mechanisms. An important implication is that changes in life history traits, notably the male and female generation times (i.e. mean age of reproduction), are expected to affect the mutation spectra at polymorphic sites. Conversely, these considerations imply that the mutation spectra of polymorphism data carries information about life history traits over millions of years. To exploit this idea, we focus on the relative proportions of mutation types seen in pedigree studies, for different combinations of male and female generation times. We show how these quantities can be related to the observed proportions of mutation types in polymorphism data in order to obtain estimates of the sex-specific mean generation times. By considering rare variants, common polymorphisms and divergent sites in turn, we estimate mean generation times over different timescales of human evolution. Our method thus provides an estimator of generation time over evolution, applicable to any species with polymorphism and *de novo* mutation data. We illustrate the method by estimating the recent and historical generation times for modern human populations, using data from the deCODE Genetics, 1000 Genomes Project, GnomAD and Simons Genome Diversity Project.

Copyright © 2018 American Society of Human Genetics. All rights reserved

154

Insights into the patterns of somatic mutation accumulation in human cancer. *K. Akdemir, A. Futreal.* MD Anderson Cancer Center, Houston, TX.

Somatic mutations arise during the life of a cell and in the generation of its progeny. Accumulation of mutations can lead to age-related diseases, and those occurring in cancer driver genes may ultimately lead to tumorigenesis and the development of clinically detectable disease. However, the spatio-temporal processes that direct mutation rates throughout cancer evolution are not fully understood. Interestingly, the hierarchical folding of genomic DNA within the nucleus is intimately linked with transcriptional regulation and DNA replication. Here, we sought to understand the effects of three-dimensional genome organization on the distribution of somatic mutations in human cancers with the aim of elucidating the potential role of genome folding on DNA damage and repair processes. We utilized data from 3000 high-coverage whole genome sequences across 47 different cancers and analyzed the distribution of more than 60 million somatic point mutations around chromatin folding domains. Our analysis revealed a strong correlation between the distribution of mutations in human cancers and the spatial organization of the genome. As a result, regional mutation rates are drastically different around the boundaries delineating transcriptionally distinct domains. Interestingly, some mutational processes can lead to notable differences in the distribution of mutations in individual tumor samples. For example, DNA mismatch repair deficiencies or APOBEC-related mutations enriched in transcriptionally-active domains while mutations occurring due to tobacco-smoke or UV light exposure enrich predominantly in transcriptionally-inactive domains. In addition, to investigate whether mutation distribution pattern is related to chromatin folding, we focused on X-chromosomal mutations, as the inactive and active X-chromosomes exhibit distinct folding structures. Indeed, the distribution pattern of X-chromosome mutations is different in female and male X-chromosomes, and reflects the differential folding structure of the inactive X-chromosome. Overall, our work highlights the role of differential mutational processes, and importantly ties the three-dimensional organization of the human genome to mutation rate variation in human cancers.

155

A small number of single cell transcriptomes stratifies Parkinsonism iPSC dopamine neurons, unravels the modelled disease process and identifies HDAC4 as a regulator of cellular phenotypes in Parkinson's dopamine neurons. *C. Webber[1,2], C. Lang[2], K. Campbell[2], B. Ryan[2], P. Carling[2], M. Attar[3], J. Vowles[4], R. Bowden[3], F. Baig[3], M. Hu[5], S. Cowley[4], R. Wade-Martins[2].* 1) Dementia Research Institute, Cardiff University, Haydn-Ellis Building, Maindy Road, Cardiff, CF24 4HQ UK; 2) Oxford Parkinson's Disease Centre, Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK; 3) The Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK; 4) Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK; 5) Oxford Parkinson's Disease Centre, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, UK.

Induced pluripotent stem cell (iPSC)-derived dopamine neurons provide an opportunity to model Parkinson's disease (PD) but neuronal cultures are confounded both by heterogeneity in cell type and asynchronicity amongst cells within the modelled disease process. Applying deep single cell transcriptomic analyses to a total of ~150 iPSC-derived dopamine neurons from Parkinson's patients carrying the *GBA-N370S* risk variants and controls, we first identified that the cells from one PD *GBA* patient clustered separately to the other cells, with signal recognition particle pathways upregulated. Clinical follow-up subsequently re-diagnosed this patient with progressive supranuclear palsy. For the remaining PD *GBA* cells, we exploited the intra-culture cellular heterogeneity to identify and validate a progressive axis of gene expression variation from more control-like to less control-like PD *GBA* cells. Analysis of the genes differentially-expressed (DE) along this axis identified the transcriptional repressor histone deacetylase 4 (HDAC4) as an upstream regulator of disease progression. HDAC4 was found to be mislocalized to the nucleus in PD iPSC-derived dopamine neurons and repressed genes early in the disease axis, leading to late deficits in protein homeostasis and endoplasmic reticulum stress. Treatment of PD *GBA* dopamine neurons with compounds known to modulate HDAC4 activity up-regulated genes early in the pseudotemporal axis and corrected all Parkinson's-related cellular phenotypes. This study demonstrates how deep single cell transcriptomics can exploit cellular heterogeneity to reveal disease mechanisms and identify therapeutic targets.

Copyright © 2018 American Society of Human Genetics. All rights reserved

156

***PARK14* genetically interacts with *PINK1* in regulating mitochondrial function and bioenergetics in a *Drosophila* Parkinson's disease model.**

S. Kalvakuri, R. Bodmer. Program in Development and Aging, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA.

Parkinson's disease (PD) is one of the two most common neurodegenerative disorders. Among the various model organisms, the *Drosophila* has emerged as an efficient system to study PD genes and to screen for genetic modifiers of PD pathology. Mutations in *Drosophila* homologues of PD genes result in phenotypes that are deemed remarkably equivalent to those observed in PD patients, including motor dysfunction, malfunctioning of dopaminergic (DA) neurons, reduced dopamine levels and mitochondrial dysfunction. Our primary goal is to find novel molecular targets/mechanisms that would ameliorate PD pathology. Towards this, we have identified iPLA2-VIA (*Drosophila* homolog of *PARK14*) in a genetic screen for novel genetic modifiers of mitochondrial dysfunction caused by *PINK1* deficiency. Knockdown of *PARK14* significantly enhanced *PINK1*-dependent phenotypes, whereas overexpression of *PARK14* suppressed *PINK1*-induced phenotypes in indirect flight muscles and in the DA neurons of *PINK1* mutants. We also find that *PARK14* mediated rescue of *PINK1* mutants is dependent on gene products that regulate cardiolipin biosynthesis. Knockdown of cardiolipin synthase (CLS) or Tafazzin (TAZ) abrogates the rescue by *PARK14* overexpression and thus highlighting the role of *PARK14* in maintaining fatty acyl side-chains and physiological functions of cardiolipin. Our data suggest that *PARK14* may regulate mitochondrial bioenergetics possibly by maintaining cardiolipin-enriched microdomains on the inner membrane of mitochondria. We propose that this helps in efficient assembly and maintenances of OXPHOS super complexes, which in turn would reduce ROS production and mitochondrial damage. We are further testing this hypothesis in our *in vivo* animal models. Interestingly, in contrast to *PINK1* overexpression, activation of *PARK14* also suppressed *PARKIN*-dependent PD phenotypes in indirect flight muscles and DA neurons, suggesting that *PARK14* may act in parallel or downstream also of *PARKIN*. Our data reveal the likely involvement of a novel molecular pathway that can modulate *PINK1/PARKIN* function in the context of cardiolipin side chain remodeling, thus providing a novel target for finding potential therapeutic avenues to treat PD patients. Further, we are also studying other phospholipases for their potentially beneficial roles in maintaining mitochondrial function/mitophagy, as an attempt to eventually mitigate the pathological phenotypes seen in *PARK14* PD patients.

157

***TP73* is an amyotrophic lateral sclerosis risk gene.**

J.M. Downie[1], S.B. Gibson[2], S. Tsetsou[3], K.L. Russell[1], M.D. Keefe[4], K.P. Figueroa[2], M.B. Bromberg[2], L.C. Murtaugh[1], J.L. Bonkowsky[4], S.M. Pulst[2], L.B. Jorde[1]. 1) Department of Human Genetics, University of Utah School of Medicine, 15 South 2030 East RM 5100, Salt Lake City, UT 84112, USA; 2) Department of Neurology, University of Utah School of Medicine, 175 North Medical Drive East, Salt Lake City, UT 84132, USA; 3) Department of Neurosurgery, Mount Sinai Hospital, Icahn School of Medicine, 1 Gustave L. Levy Place, New York, NY 10029-5674, USA; 4) Department of Pediatrics, University of Utah School of Medicine, 295 Chipeta Way, Salt Lake City, UT 84108, USA.

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease of motor neurons. Genetic factors play a role in disease pathogenesis: 68% of familial ALS and 17% of sporadic ALS (SALS) patients have an identifiable genetic risk factor. However, these genes do not account for the heritability of ALS, estimated to be ~60%. To address this, we sought to identify novel ALS risk genes. We performed a *VAAST*/*PHEVOR* burden analysis on exome sequence data from 87 SALS patients seen at the University of Utah and 324 healthy individuals from the Simons Simplex Collection. *VAAST* and *PHEVOR* identify genes with a higher number of deleterious nonsynonymous single nucleotide variants (nsSNVs) in cases versus controls. *TP73* was the second-ranked gene with five rare (minor allele frequency < 0.001) nsSNVs found among four SALS patients. Further, a rare in-frame seven amino acid deletion was found in another patient. An additional 18 rare protein-coding *TP73* variants were identified upon screening 2,800 patients from the ALS Data Browser. In total, 24 unique *TP73* rare protein-coding variants (22 nsSNVs and 2 in-frame deletions) were found among ~2,900 patients, similar to the contribution of other ALS genes. All 22 nsSNVs were predicted to be deleterious by MetaSVM. *TP73* encodes p73, a homolog of the p53 tumor suppressor transcription factor. Interestingly, aged $p73^{+/-}$ mice develop motor weakness; however, motor neurons were not directly studied in these animals. To test whether *TP73* causes motor neuron pathology consistent with ALS, an *in vitro* C2C12 myoblast growth assay was used to test the functionality of all four Sanger verified *TP73* variants from the University of Utah cohort also present in the ΔN-p73α isoform of p73. This assay confirmed that all four variants impaired or altered p73 inhibition of myoblast differentiation. Next, zebrafish embryos were injected with CRISPR/Cas9 targeted to disrupt *tp73*. We found a significant reduction in the primary axon length and number of spinal motor neurons in *tp73* mutants. Further, *tp73* mutants had a significant increase in motor neuron apoptosis. These results show *TP73* is altered or impaired in a substantial number of ALS patients. Further, loss of p73 negatively affects motor neuron development and survival. These results support a role for *TP73* as a novel ALS risk factor. This finding demonstrates a novel pathogenic basis for ALS, associated with loss of transcription factors necessary for neuronal survival.

Copyright © 2018 American Society of Human Genetics. All rights reserved

158

Simultaneous genetic analysis of 4,000 traits to identify drug targets for multiple sclerosis. *A. Fish[1], P.G. Bronson[2], R. Nagy[1], C. Vangjeli[1], J. Barrett[1], G. McVean[1], K. Estrada[2], M. Weale[1], P. Donnelly[1], S. John[3].* 1) Genomics plc, Oxford, Oxfordshire, United Kingdom; 2) Statistical Genetics and Genetic Epidemiology, Biogen, Cambridge, MA, 02142, USA; 3) Translational Biology, Biogen, Cambridge, MA, 02142, USA.

Genetic association studies can identify causal disease mechanisms and new drug targets. The joint analysis of many different diseases and traits simultaneously is substantially more powerful than focusing only on the disease of interest – it can discover subthreshold signals missed by single-trait studies and identify traits with shared genetic signals that might be biomarkers, repurposing opportunities, or safety concerns. There are two major challenges to this approach: it is technically difficult to aggregate and harmonize necessary data, and existing methods struggle to jointly analyze traits at scale. Here, we address both in a general framework applied to discover targets for multiple sclerosis (MS). We built a curated data repository of 4,069 harmonized GWAS, covering a range of complex and molecular (eQTL, pQTL, mQTL) traits. We developed a Bayesian algorithm that, for any region in the genome, determines which studies have a signal, and groups those with the same putative causal variant into clusters. The confidence of these assignments is captured by the posterior probability of cluster membership for each study. Across the five studies of MS-risk in our data repository, we identified 170 clusters containing at least one study with posterior probability > 0.5. The majority of these signals were identified by a more traditional meta-analysis of just the MS-risk studies (65%) or by a recent external meta-analysis (IMSGC Consortium) (66%); 46 signals were novel to our approach. We exemplify the power of our joint analysis to identify mechanistic pathways and biomarkers with the known risk locus *DHCR7*. A variant near *DHCR7*, an enzyme that converts the precursor of vitamin D to cholesterol, has been previously associated both to MS and to reduced levels of vitamin D. We recover these known links, showing the allele associated with decreased MS-risk colocalizes with increased vitamin D levels. We additionally identify the likely underlying mechanism, as increased expression of *DHCR7* in numerous tissues (including sun-exposed skin) also colocalizes. We also find evidence to suggest *DHRC7* may have a sex-specific effect, as taking cholesterol lowering medication in women (but not men) also colocalizes. This work illustrates the ability of joint analysis across a comprehensive collection of traits to empower drug discovery: by identifying novel signals, illuminating causal mechanisms, and detecting biomarkers.

159

iPSC-derived neurons carrying *FMR1* unmethylated full mutations show signs of neurodegeneration: Large CGG repeat expansions are required for methylation and silencing of the gene. *V. Nobile, E. Tabolacci, P. Chiurazzi, G. Neri.* Institute of Genomic Medicine, School of Medicine, Catholic University, Largo Francesco Vito 1, Rome 00168, Italy.

In the FMR1 gene, expansion of CGG triplets above 200 repeats (full mutation, FM) triggers DNA methylation (methylated full mutation, MFM), and consequent block of transcription. The resulting absence of the FMRP protein causes the Fragile X Syndrome (FXS) phenotype. There exist rare individuals of normal intelligence who are carriers of a FM, which however remains unmethylated (unmethylated full mutation, UFM), thus rescuing them from expressing the FXS phenotype. Here we show that lack of methylation of the FMR1 promoter is maintained in iPS cells derived from two unrelated UFM individuals. We also show that FMR1 does not undergo silencing during neuronal differentiation of these cells, similar to what has been described in FXS cells by Eiges et al. (2007). However, in a subset of iPSC clones in which CGG expansion exceeded 400 repeats, FMR1 became methylated and silenced, suggesting that UFM cells have a threshold for methylation higher than the typical 200 CGG repeats. These findings demonstrate that UFM individuals do not lack the cell-intrinsic ability to methylate the FMR1 gene, but require a CGG repeat expansion greater than that described for typical FXS patients. Preliminary data suggest a role of R-loops formation (hybrids between the nascent mRNA molecule and the template DNA strand present at the FMR1 locus during DNA replication) during active FMR1 transcription and of DNMT1 in maintaining both PM and UFM cell lines transcriptionally active. DNMT1 binds FMR1-mRNA in transcriptionally active cell lines preventing them from methylation, while in FXS cell lines binds to FMR1 gene resulting in silencing. Furthermore, UFM iPSC-derived neurons, in which the expression of FMR1 was preserved, rapidly showed signs of neurodegeneration, with the presence of intracellular ubiquitin containing granules, as described in premutated (PM) carriers at risk for FXTAS. UFM, while rescuing carriers from the FXS phenotype, becomes a risk factor for developing FXTAS. One of our UFM subjects, now 43 years old, shows signs of gait ataxia. On the other hand, the existence of UFM carriers with a normal phenotype hints at the possibility of treating FXS by converting a methylated full mutation into a UFM.

Copyright © 2018 American Society of Human Genetics. All rights reserved

160

Utilizing gene expression profiles to complement the analysis of genomic modifiers of the clinical onset of Huntington disease. *G.E.B. Wright[1,2,3], N.S. Caron[1,2,3], B. Ng[1,2,4], L. Casal[1,2,3], J. Ooi[5], S. Mostafavi[1,2,4], M.A. Pouladi[5,6], C.J.D. Ross[3,7], M.R. Hayden[1,2,3].* 1) Medical Genetics, University of British Columbia, Vancouver, Canada; 2) Centre for Molecular Medicine and Therapeutics, Vancouver, British Columbia, Canada; 3) BC Children's Hospital Research Institute, Vancouver, British Columbia, Canada; 4) Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada; 5) Translational Laboratory in Genetic Medicine (TLGM), Agency for Science, Technology and Research (A*STAR), Singapore; 6) Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; 7) Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, British Columbia, Canada.

Huntington disease (HD) is a neurodegenerative disorder that is caused by a trinucleotide repeat expansion in the *HTT* gene. The length of this CAG repeat explains approximately 70% of the variability in age of onset observed between patients. Recent genome-wide association studies (GWAS) have identified *trans* genetic modifiers of the disorder that contribute towards a proportion of the residual variance age of onset. However, by incorporating transcriptomic information from diverse tissues into GWAS analysis, additional candidate modifier regions of interest can be identified. We therefore assessed whether altered gene expression was associated with modifying clinical onset of HD by using S-PrediXcan to assess related GWAS data and GTEx Consortium v7 transcriptomic information. For these analyses, GWAS summary statistics from the GeM-HD (n=4,082) and the TRACK-HD (n=216) consortia were used as discovery and replication datasets, respectively. Transcriptomic gene-level association tests identified 15 genes from eight chromosomal regions significantly associated with HD age of onset that passed a false discovery rate of correction for multiple testing. These genes included those involved in DNA repair related processes, along with novel hits related to the mitochondria. A gene list analysis revealed a significant enrichment for association signals with regards to various biologically- and disease-relevant gene sets, as well as striatal co-expression modules that are mediated by CAG length. Four prioritized genes (*PMS2* P=5.1 x 10^{-5}; *EIF2AK1* P=1.1 x 10^{-5}; *SUMF2* P=3.9 x 10^{-6} and *CHCHD2* P=5.1 x 10^{-6}) showed evidence for colocalization between the GWAS and expression association signals ($P4$>0.5), as well as independent replication (P<0.05). Of note, protein truncating mutations in the mismatch repair gene, *PMS2*, have been shown to cause cancers such as Lynch syndrome. Increased expression of *PMS2* (Z=4.05) was associated with later age of HD onset and could potentially play a role in mediating the somatic instability of the CAG repeat. Further, genetic variants in *EIF2AK1* and *CHCHD2* have been previously associated with neurological phenotypes related to HD. Finally, dysregulation of a number of the prioritized genes was confirmed at the gene expression level in isogenic HD allelic human induced pluripotent stem cells and at the protein level in cortical and striatal HD patient brains. In summary, this approach expanded the number of modifier genes for HD.

161

Developmental methylomics of childhood trauma and its health consequences. *E.J.C.G. van den Oord[1], R.F. Chan[1], M. Zhao[1], L.Y. Xie[1], E.J. Costello[2], E. Copeland[2], K.A. Aberg[1].* 1) Center for Biomarker Research and Precision Medicine, Virginia Commonwealth University, Richmond, VA., USA; 2) Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, NC, USA.

By age 18, almost half of the children have suffered at least one adverse experience such as parental death, life-threatening illness, or abuse/neglect. These adversities have been robustly linked to an array of psychiatric and other medical conditions where the consequences can persist far into adulthood. It is not well understood how early adverse experiences are biologically embedded and what processes might be set into effect that would sustain long term health risks. Furthermore, as the impact of adversities can be hard to diagnose in this age group, there is a need for new diagnostic tools that can predict future health risks and guide possible intervention strategies. To address these topics we performed a prospective, longitudinal study that began in childhood and continued into adulthood and where data on adverse experiences were linked to methylation data collected before and after traumatic experiences as well as in adulthood. DNA was extracted from 1,233 bloodspots from subjects (median age=15.3, range 9.0-34.5) in the Great Smoky Mountains Study. A sequencing based approach was used that provided almost complete coverage of all 28 million CpGs in the genome. In addition to analyzing whole blood, we used cell type specific "reference" methylomes in combination with a statistical deconvolution approach to conduct methylome-wide association studies (MWAS) within constituent populations of granulocytes/T-cells/B-cells/monocytes. The impact of trauma on the methylome was pervasive, most notably for granulocytes (top MWAS P values < $1.0x10^{-15}$). We used a machine learning algorithm combined with 10-fold cross validation to obtain an unbiased estimate of the combined effect of all associated sites in whole blood, and condense this information into a single methylation risk score (MRS). Results suggested that the MRS shared over 20% of the variation with number of adverse events. The MRS made a unique contribution to the prediction of health outcomes later in life that could not be captured by clinical data or number of adverse events. For example, we used data collected at a mean age of 14.2 to predict depression symptoms 11.6 years later at age 25.8. The predictive value of the MRS was comparable to that of depression symptoms at age 14 and this contribution remained significant after a count of the number of adverse events was included in the prediction model ($P=8.0x10^{-3}$). .

Copyright © 2018 American Society of Human Genetics. All rights reserved

162

Distinguishing infection from infectious disease using methylation marks comprised within cell-free DNA. *A.P. Cheng[1], P. Burnham[1], D. Dadhania[2,3], J.R. Lee[2,3], M. Suthanthiran[2,3], I. De Vlaminck[1].* 1) Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY; 2) Division of Nephrology and Hypertension, Department of Medicine, Weill Cornell Medicine, New York, NY; 3) Department of Transplantation Medicine, New York Presbyterian Hospital - Weill Cornell Medical Center, New York, NY.

BACKGROUND. Urinary tract infections are one of the most common infections in humans. In kidney transplant recipients, UTIs are even more common, and can lead to transplant failure. There are multiple methods to diagnose UTIs, ranging from dipsticks, which cannot identify a specific pathogen, to kidney biopsies, which are highly invasive and uncomfortable for the patient. Here, we have developed a non-invasive test to inform renal tissue injury, detect the presence of viral or bacterial agents in the urinary tract, and probe the systemic host response to infection. Our method involves performing whole-genome bisulfite sequencing (WGBS) on urinary cell-free DNA to simultaneously determine the tissues-of-origin of host-derived cell-free DNA (cfDNA) and quantify cfDNA derived from viruses and bacteria.METHODS. We performed WGBS on urinary cfDNA from kidney transplant recipients. We divided transplant recipients into groups depending on whether they had recently received their transplant (n=5), had a bacterial urinary tract infection (n=11), had BK polyomavirus nephropathy (n=9), or viremia (n=8) or were otherwise free of disease (n=9). We compared cfDNA methylation patterns to the methylation patterns of individual tissues and cell types obtained from publicly available databases, and used quadratic programming to estimate relative tissue contributions for each sample. We identified viral and bacterial cfDNA in the samples and computed the representation of microbial genome copies relative to human genomes copies.RESULTS. Analysis of publicly available WGBS data from specific cell types and tissues reveals cell type specific methylation patterns. As expected, patients with kidney BK polyomavirus infections show higher abundance of kidney-derived cfDNA compared to healthy patients. Also, patients with bacterial UTIs exhibit elevated levels of leukocyte-derived cfDNA. Finally, we show that WGBS can be used to accurately detect bacterial or viral cfDNA with comparable resolution to standard sequencing, despite the reduced complexity of bisulfite-treated cfDNA.DISCUSSION. This study demonstrates a noninvasive measurement that provides a comprehensive overview of kidney injury and distinguishes infection from infectious disease by quantifying both the presence of a pathogen and its effect on the host.

163

Promoter GGX repeat expansion and exon 1 methylation of *XYLT1* in Baratela-Scott syndrome. *A.J. LaCroix[1], D. Stabley[2], M.P. Adam[1,3], M. Mehaffey[1], K. Kernan[1], C. Myers[4], C. Fagerstrom[5], G. Anadiotis[5], Y. Akkari[5], K. Robbins[2], M.B. Bober[6], A. Duker[6], D. Miller[1], M. Kircher[7], M. Bamshad[1,3], D. Nickerson[3,7], U.W. Center for Mendelian Genomics[1,7], K. Sol-Church[2,8], H.C. Mefford[1,3].* 1) Pediatrics, University of Washington, Seattle, WA; 2) Nemours Biomedical Research Department, Alfred I. duPont Hospital for Children, Wilmington, DE; 3) Brotman Baty Institute for Precision Medicine, Seattle, WA; 4) Seattle Children's Hospital, Seattle, WA; 5) Legacy Health, Portland, OR; 6) Division of Orthogenetics, Alfred I. duPont Hospital for Children, Wilmington, DE; 7) Department of Genome Sciences, University of Washington, Seattle, WA; 8) School of Medicine Genome Analysis and Technology Core, University of Virginia, Charlottesville, VA.

Baratela-Scott syndrome (BSS), also known as Desbuquois dysplasia type II, is a rare, autosomal recessive skeletal dysplasia characterized by short stature, characteristic facial features, and developmental delay. Compound heterozygous or homozygous pathogenic variants in *XYLT1* have been identified as a genetic cause of BSS. *XYLT1* encodes the xylosyltransferase enzyme XT1. The recently characterized promoter region includes a 238-bp sequence not found in the reference genome (hg19) that contains a $(GGX)_n$ repeat with 9-20 repeats in healthy individuals. We report clinical and molecular investigation of 12 individuals (10 families) with BSS. By sequencing or chromosome array, we identified biallelic variants in *XYLT1* in two individuals; however, the remaining 10 individuals either had no variants (n=3) or a single variant: a heterozygous 3.3-Mb 16p13 deletion encompassing *XYLT1* (n=5) or a heterozygous truncating variant (n=2). In probands with one or no *XYLT1* variants, bisulfite sequencing of exon 1 revealed hypermethylation of all 33 CpG sites across 300bp of exon 1 in one or both alleles, which was not present in 100 controls. Segregation testing confirmed that the methylated allele was *in trans* with the deletion or sequence variant; both alleles were methylated in the three affected individuals with no previously identified variants. We demonstrated monoallelic expression of the unmethylated *XYLT1* allele in fibroblasts from one patient with a truncating point mutation in one allele and hypermethylation of the other allele. To investigate whether hypermethylation is associated with expansion of the $(GGX)_n$ repeat in the promoter region, we performed Southern blot analysis in four families. In each case, a fragment of increased size (range 400-2500 bp) was identified, suggesting trinucleotide repeat expansion. In two families, the expansion increased in size when transmitted from unaffected parent to affected child, highlighting instability of the region. We propose that BSS represents a new trinucleotide repeat expansion disorder associated with hypermethylation of exon 1 and decreased or absent expression of the hypermethylated allele. Hypermethylation explained all “missing” disease alleles in our cohort (10/20 alleles, 50%); the 16p13 deletion accounted for 4/20 (20%) alleles. Testing for the hypermethylation variant, which traditional sequencing and array methods will miss, should be considered in individuals with a phenotype consistent with BSS.

Copyright © 2018 American Society of Human Genetics. All rights reserved

164

Circulating tumor DNA methylation haplotypes in plasma can detect cancer four years prior to conventional diagnosis. *A. Gore[1], X. Chen[2,3], J. Gole[1], J. Min[1], Q. He[1], X. Li[1], L. Cheng[1], Z. Zhang[1], H. Nu[1], Z. Li[1], Z. Xie[1], X. Yang[3,5], H. Shi[1], J. Dang[1], C. McConnell[1], J. Zhang[3], Z. Yuan[3], J. Wang[2,3], M. Lu[3,5], W. Ye[3,6], Y. Gao[1], K. Zhang[4], R. Liu[1], J. Li[2,3].* 1) Singlera Genomics, La Jolla, CA; 2) Fudan University, Shanghai, China; 3) Fudan University Taizhou Institute of Health Sciences, Taizhou, China; 4) University of California - San Diego, La Jolla, CA; 5) Qilu Hospital of Shandong University, Jinan, China; 6) Karolinska Institute, Stockholm, Sweden.

Background: Circulating tumor DNA (ctDNA) has the potential to allow early detection of cancer while patients remain asymptomatic, which could significantly improve clinical outcomes. However, ctDNA has typically been utilized to detect cancer in patients that have already been conventionally diagnosed. Through a longitudinal study of healthy individuals, we show that ctDNA methylation can identify cancer four years before conventional diagnosis. **Methods:** The Taizhou Longitudinal Study (TLS) involved an initial blood sample collection from individuals in the city of Taizhou, China, followed by indefinite monitoring of study participants for cancer diagnosis using a national registry and health insurance database. Since 2007, over 1.5 million samples have been collected from more than 100,000 participants with 5 years of follow-up on average, including blood samples from more than 700 individuals diagnosed with cancer at different time points. Blood samples from 958 healthy individuals, 363 patients with colorectal, esophageal, liver, lung, or stomach cancer, and 159 individuals who later developed one of these cancers were processed from the TLS cohort using the Singlera Genomics PAN assay, a targeted bisulfite sequencing method which identifies cancer-specific methylation haplotypes. 618 samples were used to train a classification model based on disease register results, and 862 samples were used as an independent test set to validate model performance. **Results:** In the test set, the PAN assay demonstrated a sensitivity of 86% in post-diagnosis cancer patients with a specificity of 95% in healthy patients. The PAN assay was also able to identify cancer up to four years prior to patients being diagnosed with an average sensitivity of 64%. **Conclusions:** In a longitudinal study, we have shown that ctDNA methylation can be utilized to detect cancer up to four years prior to conventional diagnosis, paving the way for a blood-based non-invasive pan-cancer screening assay for the general population.

165

Blood DNA methylation profiles are altered years before breast cancer diagnosis: Findings from a case-cohort analysis in the Sister Study. *Z. Xu[1], D.P. Sandler[1], J.A. Taylor[1,2].* 1) Epidemiology, National Institute of Environmental Health Sciences, Research Triangle Park, NC; 2) Epigenetics & Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC.

Peripheral blood DNA methylation may be associated with breast cancer, but studies of candidate genes, global, and genome-wide DNA methylation have been inconsistent. We performed a large epigenome-wide study using prospectively collected samples from the Sister Study, including 1,552 cases and a random sample of the 1224 women from the cohort. We identified 9,601 CpG markers associated with invasive breast cancer at false discovery rate $q < 0.01$ and replicated 2,095 of these in an independent dataset. Most of the top differentially methylated CpGs (dmCpGs) showed lower methylation in invasive cases and inversely correlated with time-to-diagnosis. Women who developed ductal carcinoma *in situ* had profiles in between invasive cases and non-cases. Based on ENCODE annotation, dmCpGs with lower methylation in cases occur at non-island sites enriched for the H3K36me3 histone mark, whereas sites with higher methylation in cases occur at CpG islands enriched for H3K4me3. Pathway analysis shows enrichment of breast cancer-related gene pathways, and dmCpGs are overrepresented in known breast cancer susceptibility genes. Our findings suggest that DNA methylation may be a marker of subsequent risk for invasive breast cancer.

Copyright © 2018 American Society of Human Genetics. All rights reserved

166

Genetically predicted methylation biomarkers and prostate cancer risk: A methylome-wide association study in over 140,000 European descendants. *L. Wu[1], Y. Yang[1], X. Guo[1], X.O. Shu[1], Q. Cai[1], X. Shu[1], B. Li[2,3], R. Tao[4], M.J. Roobol[5], G.G. Giles[6,7], H. Brenner[8,9,10], E.M. John[11], J. Clements[12,13], E.M. Grindedal[14], J.Y. Park[15], J.L. Stanford[16,17], Z. Kote-Jarai[18], C.A. Haiman[19], R.A. Eeles[18], W. Zheng[1], J. Long[1], the PRACTICAL, CRUK, BPC3, CAPS, PEGASUS consortia*.* 1) Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA; 2) Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN, USA; 3) Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA; 4) Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA; 5) Department of Urology, Erasmus University Medical Center, Rotterdam, the Netherlands; 6) Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, 207 Bouverie St, Melbourne, Victoria 3010, Australia; 7) Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, 615 St Kilda Rd, Melbourne, Victoria 3004, Australia; 8) Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany; 9) German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany; 10) Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany; 11) Department of Medicine (Oncology) and Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA; 12) Australian Prostate Cancer Research Centre-Qld, Institute of Health and Biomedical Innovation and School of Biomedical Science, Queensland University of Technology, Brisbane, Queensland, Australia; 13) Translational Research Institute, Brisbane, Queensland, Australia; 14) Department of Medical Genetics, Oslo University Hospital, Norway; 15) Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, USA; 16) Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA; 17) Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington, USA; 18) Division of Genetics and Epidemiology, The Institute of Cancer Research, and The Royal Marsden NHS Foundation Trust, London, UK; 19) Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA.

Background Existing epidemiologic studies on prostate cancer associated DNA methylation biomarkers are limited by relatively small sample sizes and potential biases caused by confounders, and reverse causation. To overcome these limitations, we performed a methylome-wide association study to evaluate associations of genetically predicted DNA methylation levels with prostate cancer risk. **Methods** We used genotyping and DNA methylation data obtained in white blood cells from subjects of European descent included in the Framingham Heart Study (FHS, N=1,595) and established models using the elastic net method to predict DNA methylation based on genetic variants. We selected 77,243 built methylation models with a prediction performance (R^2) of >0.01 for association analyses, using data obtained from 79,194 cases and 61,112 controls of European ancestry included in consortia PRACTICAL, CRUK, CAPS, BPC3 and PEGASUS. For CpG sites showing a significant association with prostate cancer risk, we further evaluated correlations of their methylation levels with expression levels of genes flanking these loci in blood, by using data from the Offspring Cohort of the FHS (N=1,367). Furthermore, for genes whose expression levels were correlated with DNA methylation levels, we assessed whether their genetically predicted mRNA expression levels in blood were also associated with prostate cancer risk. **Results** We identified 759 CpGs showing an association with prostate cancer risk at $P < 6.47\times10^{-7}$, a Bonferroni-corrected significance level, including 15 CpGs located >500 kb away from any risk variant reported in previous GWAS of prostate cancer. Of the 759 CpGs, consistent associations were observed for 460 at $P < 0.05$ in the UK Biobank data which included 2,495 prostate cancer cases. Among those 759 CpGs, methylation levels of 107 were correlated with expression levels of 70 adjacent genes (false discovery rate (FDR) < 0.05). Among 36 of the genes with mRNA expression prediction models built, 23 showed a significant association with prostate cancer risk (FDR < 0.05). Overall, there were 15 genes showing consistent association directions for the DNA methylation-gene expression-prostate cancer pathway. **Conclusion** In this large methylome-wide association study, we demonstrated the potential of integrating genetic variants, DNA methylation and gene expression in identifying novel biomarkers for prostate cancer. Our study provides new insights into the etiology of this common malignancy.

167

Sequence variants associating with corneal structure and diseases. *E.V. Ivarsdottir[1,2], S. Benonisdottir[1], F. Jonasson[3,4], G. Thorleifsson[1], P. Sulem[1], H. Holm[1], A. Oddsson[1], U. Styrkarsdottir[1], S. Kristmundsdottir[1], U. Thorsteinsdottir[1,3], D.F. Gudbjartsson[1,2], K. Stefansson[1,3].* 1) deCODE genetics/Amgen, Reykjavik, Iceland; 2) School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland; 3) Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland; 4) Department of Ophthalmology, Landspitali University Hospital, Reykjavik, Iceland.

The corneal endothelium is a monolayer of cells at the innermost surface of the cornea. Changes in endothelial structure and premature cell loss can result in corneal edema and loss of transparency[1]. Corneal diseases are among the most common causes of visual loss worldwide and endothelial cell failure is the leading indication for corneal transplantation[2]. A specular microscopy captures an image of the corneal endothelium and can be used to determine endothelial cell density (cells/mm^2), coefficient of cell size variation (CV), percentage of hexagonal shaped cells (HEX) and central corneal thickness (CCT). We performed a genome-wide association study on these structural measurements of the cornea in 6,125 Icelanders. We detected associations at several loci, including 7 novel variants associating with either cell density, CV, HEX or CCT. We assessed the effects of these variants on different ocular biomechanics such as corneal hysteresis (CH) and intraocular pressure (IOP), as well as various eye diseases such as glaucoma and corneal dystrophies. Most notably, an intergenic variant strongly associates with decreased cell density and accounts for 23.9% of the population variance of cell density (β=-0.77 SD, $P=1.8\times10^{-314}$) without affecting risk of corneal diseases or glaucoma in our data. Interestingly, the variant also associates with increased CH (β=0.19 SD, $P=2.6\times10^{-19}$), independently of its effect on cell density. Two missense variants associate with CCT (β=0.18 SD, $P=1.3\times10^{-10}$ and β=0.30 SD, $P=3.9\times10^{-10}$) and a pathogenic microsatellite, known to increase risk of Fuchs corneal dystrophy, associates with cell density and HEX (β=-0.38 SD, $P=1.6\times10^{-19}$ and β=-0.37 SD, $P=5.9\times10^{-18}$, respectively). Our findings shed a new light on the biology of the cornea and indicate that low endothelial cell density does not in and of itself lead to the development of disease. 1. Bourne, W. M. Biology of the corneal endothelium in health and disease. *Eye* **17,** 912–918 (2003). 2. Ruth, N. & Peralta, V. 2015 Eye Banking Statistical Report. *Eye Bank Assoc. Am. Washington, DC* **3,** 58–69 (2016).

Copyright © 2018 American Society of Human Genetics. All rights reserved

168

Biallelic loss-of-function variants in *DNMBP* cause congenital cataract and visual impairment in human and flies. *H. Chung[1,2], M. Ansar[3], R. Taylor[4], A. Nazeer[5], S. Imtiaz[6], M.T. Sarwar[5], P. Makrythanasis[3,7], S. Qureshi[6], E. Falconnet[3], M. Guipponi[8], D. Schlegel[9], S. Neuhauss[9], C.J. Pournaras[10], E. Ranza[3,8], F.A. Santoni[3,11], J. Ahmed[5], I. Shah[5], K. Gul[6,12], G. Black[4,13], H. Bellen[1,2,14,15], S.E. Antonarakis[3,8,16].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, United States of America; 3) Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland; 4) Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK; 5) Institute of Basic Medical Sciences, Khyber Medical University, Peshawar, Pakistan; 6) Department of Genetics, University of Karachi, Karachi, Pakistan; 7) Biomedical Research Foundation of the Academy of Athens, Athens, Greece; 8) Service of Genetic Medicine, University Hospitals of Geneva, Geneva, Switzerland; 9) Institute of Molecular Life Science, University of Zurich; 10) Hirslanden Clinique La Colline, Geneva, Switzerland; 11) Department of Endocrinology Diabetes and Metabolism, University hospital of Lausanne, Switzerland; 12) Department of Bio Sciences, Faculty of Life Science, Mohammad Ali Jinnah University, Karachi, Pakistan; 13) Vision Science Centre, Manchester Royal Eye Hospital, Manchester University NHS Foundation Trust, Manchester, UK; 14) Howard Hughes Medical Institute, Houston TX, United States of America; 15) Department of Neuroscience and Program in Developmental Biology, Baylor College of Medicine, Houston, TX, United States of America; 16) iGE3 Institute of Genetics and Genomics of Geneva, Geneva, Switzerland.

Bilateral congenital cataracts (CC) a broad and heterogeneous group of disorders yet numerous genes that cause recessive CC remain to be discovered. We identified three consanguineous families with congenital cataract in whom 12-individuals had residual visual impairment after cataract surgery. Using exome sequencing, we found homozygous loss of function variants in the *DNMBP* gene (OMIM#611282 NM_015221.2): a nonsense mutation c.811C>T:p.Arg271Ter in family F385 (nine affected individuals, LOD score: 5.18 at θ=0), a frameshift deletion c.2947-2948delGA in F372 (two-affected individuals) and a frameshift c.2852-2855del, p.Thr951Metfs*41 in F3 (one affected individual). The phenotypes of all affected individuals include congenital cataract; interestingly in some of the patients an ERG defect was also observed. RNAi mediated knockdown of the fly orthologue sif in Drosophila, expressed in lens secreting cells, affect the development of these cell as well as the localization of E-cadherin, alters the distribution of Septate Junctions in adjacent cone cells, and leads to a ~50% reduction in amplitudes of the ERG in young flies. DNMBP regulates the shape of tight junctions, which correspond the Septate Junctions in invertebrates, as well as the assembly pattern of E-cadherin in human epithelial cells. Importantly, E-cadherin has an important role in lens vesicle separation and lens epithelial cell survival in human. We therefore conclude that *DNMBP* loss-of-function variants cause early onset cataract and visual impairment in humans.

169

***ABCA4*-associated disease as a model for missing heritability in autosomal recessive disorders: Novel non-coding splice, *cis*-regulatory, structural and recurrent hypomorphic variants.** *M. Bauwens[1], A. Garanto[2,3], R. Sangermano[2,4], S. Naessens[1], N. Weisschuh[5], J. De Zaeytijd[6], M. Khan[2,3], F. Sadler[6], I. Balikova[6], C. Van Cauwenbergh[1,6], T. Rosseel[1], J. Bauwens[7], K. de Leeneer[1], S. De Jaegere[1], T. Van Laethem[1], M. De Vries[8], K. Carss[9,10], A. Fakin[11,12], G. Arno[12], A.R. Webster[11,12], T. de Ravel de l'Argentière[13], Y. Sznajer[14], M. Vuylsteke[15], S. Kohl[5], B. Wissinger[5], T. Cherry[16,17], R.W.J. Collin[2,3], F.P.M. Cremers[2,3], B.P. Leroy[1,6,18], E. De Baere[1].* 1) Center for Medical Genetics Ghent, Ghent University and Ghent University Hospital, Ghent, East-Flanders, Belgium; 2) Department of Human Genetics Radboud University Medical Center, Nijmegen, The Netherlands; 3) Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands; 4) Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands; 5) Molecular Genetics Laboratory, Institute for Ophthalmic Research, University of Tuebingen, Tuebingen, Germany; 6) Department of Ophthalmology, Ghent University and Ghent University Hospital, Ghent, Belgium; 7) Department of Computer Science, Free University of Brussels, Brussels, Belgium; 8) Department of Ophthalmology, Hôpital des Enfants Reine Fabiola, Brussels, Belgium; 9) Department of Haematology, University of Cambridge, NHS Blood and Transplant Centre, Cambridge ,UK; 10) UK NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, UK; 11) Moorfields Eye Hospital NHS Foundation Trust, London UK; 12) UCL Institute of Ophthalmology, London, UK; 13) Center for Human Genetics, KU Leuven and UZ Leuven, Leuven, Belgium; 14) Centre de Génétique Humaine, Cliniques Universitaires St. Luc, Université Catholique de Louvain, Brussels, Belgium; 15) GNOMIXX ltd, Statistics for Genomics, Melle, Belgium; 16) Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA; 17) Center for Developmental Biology and Regenerative Medicine, Seattle Children's Research Institute, Seattle, WA, US; 18) Division of Ophthalmology and Center for Cellular & Molecular Therapeutics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.

Purpose: Stargardt disease (STGD1) is one of the most common inherited retinal diseases (IRD), with an estimated prevalence of 1/8,000 – 1/10,000. The underlying disease gene for STGD1 is *ABCA4*. Apart from STGD1, biallelic *ABCA4* variants have been linked to a spectrum of autosomal recessive IRD phenotypes, varying from STGD1, cone-rod dystrophy, atypical retinitis pigmentosa (RP), generalized choriocapillaris dystrophy to rapid-onset chorioretinopathy (ROC), jointly named *ABCA4*-associated disease (AAD). AAD is hallmarked by a large proportion of patients with only one pathogenic variant in the disease gene *ABCA4,* suggestive for missing heritability. **Methods:** By locus-specific genotyping of *ABCA4*, combined with extensive functional studies such as *in vitro* splice assays and luciferase assays, we aimed to unravel the missing alleles in a cohort of 75 AAD patients, with one (n=73) or no (n=2) identified coding *ABCA4* variant. In addition, antisense oligonucleotide (AON)-mediated rescue was performed for a subset of deep-intronic *ABCA4* splice variants. **Results:** We identified nine (deep-)intronic *ABCA4* splice variants, seven of which are novel: two novel variants with a putative *cis*-regulatory effect, and five structural variants, of which four are novel, including the first duplications. Taken together, these variants account for the missing alleles in 31 patients (41.3%). The common hypomorphic variant c.5603A>T was found as the likely second missing allele in 29 patients (38.6%). Overall, we elucidated the missing heritability in 80% of our cohort. In addition, we successfully rescued three deep-intronic variants using AON-mediated treatment in HEK 293-T cells and for one of these also in patient-derived fibroblasts. **Conclusion**: This study demonstrates that a locus-specific integrated approach combining genomics with downstream tailored functional studies is powerful for elucidating a major portion of missing heritability in AAD. Non-coding mutations, novel structural variants and a common hypomorphic allele of the *ABCA4* gene explain the majority of unsolved cases. The discovery of novel mutations in non-coding regions and AON-mediated rescue can be envisaged for personalized therapies. Overall, this *ABCA4*-oriented study can be regarded as a model for missing heritability in other autosomal recessive diseases with a recognizable phenotype and with an incomplete molecular diagnosis.

Copyright © 2018 American Society of Human Genetics. All rights reserved

170

GWAS of African populations reveals link between glaucoma and Alzheimer's disease. *M.A. Hauser[1], R.R. Allingham[1], C.C. Khor[2], C.J. Van Der Heide[3], J.I. Rotter[4], S.H. Wang[5], P.W.M. Bonnemaijer[6], N. Risch[7,8,9], T.J. Hoffman[7,8], E. Jorgenson[9], O. Olawoye[10], A. Ashaye[10], C.C.W. Klaver[6], R.N. Weinreb[11], A. Ashley-Koch[12], J.H. Fingert[3], T. Aung[13], Eyes of Africa Consortium.* 1) Department of Ophthalmology, Duke University, Durham, NC; 2) Genome Institute of Singapore, Singapore; 3) Department of Ophthalmology, University of Iowa; 4) Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute and Department of Pediatrics, Harbor-University of California, Los Angeles Medical Center, Torrance, California; 5) Department of Pathology, Duke University, Durham, NC; 6) Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands; 7) Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA; 8) Institute for Human Genetics, University of California San Francisco, San Francisco, CA; 9) Kaiser Permanente Northern California (KPNC), Division of Research, Oakland, CA; 10) Department of Ophthalmology, University of Ibadan, Ibadan, Nigeria; 11) Department of Ophthalmology, Hamilton Glaucoma Center, Shiley Eye Institute, University of California, San Diego, La Jolla, CA; 12) Department of Medicine, Duke University, Durham, NC; 13) Singapore Eye Research Institute, Singapore National Eye Centre and Eye ACP, Duke-National University of Singapore, Singapore.

Primary open angle glaucoma (POAG) has a complex genetic etiology, and manifests as a health disparity that disproportionately affects individuals of African descent. We conducted a genome-wide association study (GWAS) and replication in 9,201 POAG patients and 16,981 control individuals with African ancestry from 11 countries. We observed significant association at amyloid β A4 precursor protein-binding, family B, member 2 (*APBB2)* rs59892895-C (P=4.0x10^{-13}; per-allele OR=1.19, 95%CI = 1.13-1.24). This association is specific for African and African diaspora populations, and no association at this locus has been identified in Caucasian or Asian GWAS studies of POAG. *APBB2* is involved in the proteolytic processing of amyloid precursor protein (APP). APP is required for normal development of the retina, but proteolytic processing of APP also produces amyloid beta (Aβ) peptides, which are toxic and which aggregate to form amyloid plaques, one of the neuropathological hallmarks of Alzheimer's disease (AD). We conducted immunohistochemical analysis of post-mortem retinal samples and primary visual cortex samples from African Americans. Retinas of individuals heterozygous for the risk allele show increased APBB2 expression levels and increased Aβ staining in the retinal ganglion cells, the ocular neurons whose death defines glaucoma. Similarly, the primary visual cortex of individuals carrying a risk allele shows increased APBB2 levels and more severe amyloid plaque pathology: there is no increase in dense core plaques, which represent end-stage plaque pathology, but rather an increase in diffuse and immature plaques, showing an increased beta-amyloid burden at earlier stages of plaque evolution. SNPs within the APBB2 locus are associated with more severe disease (vertical cup to disk ratio QTL P=3X10^{-3}). Examination of the GTex database shows strong eQTLs within the *APBB2* locus; however, these variants are not in LD with our lead SNP. Comorbidity between AD and POAG has long been suggested. Aβ is found in melanopsin retinal ganglion cells of AD patients, leading to dysfunction of the circadian rhythm, and rat models of induced ocular hypertension show retinal Aβ deposition and caspase activation. However, our findings provide the first direct evidence in humans of a shared molecular pathway in POAG and AD that converges upon increased beta-amyloid deposition. These findings implicate a common pathogenic mechanism between POAG and Alzheimer's disease.

171

Three novel hearing loss genes reveal previously unrecognized roles of their protein products in the perception of sound. *G. Bademci[1], C. Li[2], O. Diaz-Horta[1], C. Abad[1], B. Vona[3], R. Maroofian[4], A. Subasioglu[5], E. Mihci[6], O. Alper[7], B.G. Nur[6], M. Benham[8], A. Incesulu[9], F. Silan[10], S. Tokgoz-Yilmaz[11], M. Salehi[12], T. Haaf[3], F.B. Cengiz[1], S.H. Blanton[1,13,14], D. Duman[15], K. Walz[1,14], R.G. Zhai[2,16], M. Tekin[1,13,14].* 1) John P. Hussman Institute for Human Genomics, University of Miami Miller School , Miami, FL; 2) Department of Molecular and Cellular Pharmacology, University of Miami Miller School of Medicine, Miami, FL; 3) Institute of Human Genetics, Julius Maximilians University, Würzburg, Germany; 4) Genetics Research Centre, Molecular and Clinical Sciences Institute, St George's, University of London, Cranmer Terrace, London SW17 0RE, UK; 5) Department of Genetics, Izmir Ataturk Education and Research Hospital, Izmir, Turkey; 6) Division of Pediatric Genetics, Akdeniz University School of Medicine, Antalya, Turkey; 7) Department of Medical Biology and Genetics, Akdeniz University, Faculty of Medicine, Antalya, Turkey; 8) Medical Genetics Center of Genome, Isfahan, Iran; 9) Department of Otorhinolaryngology, Faculty of Medicine, Eskisehir Osmangazi University, Eskisehir, Turkey; 10) Department of Medical Genetics, Canakkale Onsekiz Mart University School of Medicine, Canakkale, Turkey; 11) Department of Audiology, Ankara University School of Medicine, Ankara, Turkey; 12) Division of Genetics and Molecular Biology, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran; 13) Department of Otolaryngology, University of Miami Miller School of Medicine, Miami, FL, USA; 14) Dr. John T. Macdonald Department of Human Genetics, University of Miami Miller School of Medicine, Miami, FL; 15) Division of Genetics, Department of Pediatrics, Ankara University School of Medicine, Ankara, Turkey; 16) School of Pharmacy, Key Laboratory of Molecular Pharmacology and Drug Evaluation (Yantai University), Ministry of Education, Collaborative Innovation Center of Advanced Drug Delivery System and Biotech Drugs in Universities of Shandong, Yantai University, Yantai, Shandong, China.

Among all sensory deficits, hearing loss (HL) is the most common, affecting about half a billion people worldwide. Genetic factors are implicated in the majority of cases with HL and lead to non-syndromic HL (NSHL) in over 70% of the cases. More than 80% of NSHL exhibits autosomal recessive transmission (ARNSHL). While recent studies have revealed a substantial portion of the genes underlying HL, the extensive genetic landscape has not been completely explored. After excluding all known deafness genes in 63 Turkish and Iranian consanguineous multiplex families with ARNSHL, analysis of whole exome and genome sequencing has identified variants in *MPZL2*, *TOGARAM2*, and *GRAP* co-segregating with the phenotype in at least two families for each gene. Identified variants are absent or very rare in public databases as well as in ethnicity-matched controls. The roles of MPZL2, TOGARAM2, and GRAP in the auditory system are unknown. Via RT-PCR, we show that each gene is expressed in the cochlea of mice during development and adulthood. In mice, immunofluorescence studies show that the protein products of *MPZL2* and *TOGARAM2* localize to auditory hair cells, and GRAP localizes to the fibers of the auditory nerve. *MPZL2* and *TOGARAM2* variants are loss of function (frameshift or nonsense), predicted to lead to nonsense mediated mRNA decay. The variant detected in *GRAP* is missense affecting a conserved amino acid residue. We show that *drk*, the *Drosophila* homologue of human *GRAP*, is expressed in Johnston's organ (JO), the fly hearing organ. Loss of drk in JO causes scolopidium abnormalities and defects in balance and locomotor behavior. Furthermore, drk highly co-localizes with synapsin at synapses, suggesting a potential role of such adaptor proteins in regulating synaptic cytoskeleton dynamics. Here we present evidence that variants in three novel genes underlie ARNSHL. This study not only advances the knowledge of HL genes by increasing the number of recognized genes, but it also unveils three previously unknown organ of Corti proteins that play fundamental roles in sound perception.

Copyright © 2018 American Society of Human Genetics. All rights reserved

172

Whole exome sequencing (WES) in infants with congenital hearing loss as a model for genomic newborn screening. *D.J. Amor[1,2,3], L. Downie[1,2,3], R.A. Burt[1,2], E. Lynch[4], M. Martyn[1,4], Z. Poulakis[1,2,3], C. Gaff[2,4], V. Sung[1,2,3], M. Wake[1,2,3], M. Hunter[5,6], K. Saunders[5], S. Lunke[1,2], J.L. Halliday[1,2].* 1) Murdoch Children's Research Institute, Melbourne, Victoria, Australia; 2) University of Melbourne, Victoria, Australia; 3) Royal Children's Hospital, Melbourne, Australia; 4) Melbourne Genomics Health Alliance, Victoria, Melbourne, Australia; 5) Monash Health, Victoria, Melbourne, Australia; 6) Department of Paediatrics, Monash University, Melbourne, Australia.

Background: The Melbourne Genomics Health Alliance is establishing systems and generating evidence to guide incorporation of genomics into the healthcare system. Through one of its clinical Flagship projects, the Alliance offered WES to all families in Victoria, Australia, who had an infant with moderate or worse bilateral hearing loss born in 2016 or 2017. Aim: The aim of this project was to define the genetic aetiology of congenital hearing loss in a population-based cohort, and to offer extended genomic analysis as a model for expanded newborn genomic screening. Methods: Families who had an eligible child were identified by the Victorian Infant Hearing Screening Program. WES with targeted gene analysis was performed in conjunction with microarray. Families that consented for WES to identify the cause of hearing loss in their child were eligible for extended genomic analysis. Parents could also choose additional analyses for selected childhood onset genetic conditions unrelated to hearing loss, with . Families were provided with a decision aid and a genetic counselling session. Gene lists were designed in collaboration with the NC-Nexus and BabySeq projects: participants could choose to receive results for childhood onset genetic conditions with a known treatment or intervention (Choice B), or all conditions, including those that do not have a treatment or intervention pathway (Choice C). The families completed evaluation surveys regarding their decision making post recruitment and at return of results. Results: Out of the 110 patients recruited, 78 results have been issued. The rate of genetic diagnosis for deafness is 63%, comprising 28% connexin mutations, 18% other non-syndromic deafness genes, and 16% have a syndromic diagnosis. Microarray has contributed to 8% of diagnoses when combined with WES results. Of the 110 patients who have consented for WES 68% have opted to receive additional findings. Of those, 39% selected Choice B and 61% Choice C. Discussion: Adding WES to the investigation pathway for this cohort has led to a genetic diagnosis in approximately two thirds of infants with congenital deafness. A majority of parents requested additional information from WES when it was offered. Conclusion: This study provides a comprehensive understanding of the genetic aetiology of congenital hearing loss in a population based cohort. The results of this project illustrate the importance of providing choice around additional findings in genomic testing.

173

Using untargeted metabolomics and Mendelian randomization to dissect causal relationships in the obesity metabolome. *Y.H. Hsu[1,2,3], C.M. Astley[1,2,3], S. Vedantam[2,3], J.M. Mercader[3,4], A. Metspalu[5], K. Fischer[5], K. Fortney[6], E.K. Morgen[6], C. Gonzalez[7,8], M.E. Gonzalez[7,8], T. Esko[5,3], J.N. Hirschhorn[1,2,3].* 1) Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America; 2) Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts, United States of America; 3) Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America; 4) Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America; 5) Estonian Genome Center, University of Tartu, Tartu, Estonia; 6) BioAge Labs, Richmond, CA, United States of America; 7) Instituto Nacional de Salud Publica, Cuernavaca, Morelos, Mexico; 8) Centro de Estudios en Diabetes, Mexico City, Mexico.

Obesity is a major health problem associated with large-scale metabolic disturbances. Understanding the mechanistic connections between obesity and its associated metabolites can uncover novel biology and inform prevention and treatment strategies. Recent studies have combined metabolite profiling with genetics to infer causality between metabolites, obesity, and related diseases using Mendelian randomization (MR). In this study, we aimed to expand upon previous research by using an MR approach to analyze both known and unknown body mass index (BMI)-associated metabolites in untargeted metabolomics datasets. This was made possible by utilizing our recently developed bioinformatics tool, PAIRUP-MS, which enabled meta and pathway analyses of unknown metabolites across multiple datasets. We identified several known metabolites that are more likely to be the cause (e.g. alpha-hydroxybutyrate and pantothenate) or the effect (e.g. isoleucine and phenylalanine) of obesity, or may have more complex bidirectional cause-effect relationship with obesity (e.g. glycine, leucine, and glutamine). Importantly, we also identified > 4 times more unknown metabolites in each group compared to the knowns, allowing us to conduct pathway analysis comparing all cause vs. effect metabolites. This revealed that the causal metabolites were enriched in the "glutathione-mediated detoxification" pathway, whereas effect metabolites were enriched in diverse pathways including "lysine catabolism", "dopamine metabolism", and "signaling by GPCR". While our results demonstrate the potential usefulness of studying causality in the obesity metabolome using untargeted profiling data and genetic instruments, future studies with meta-analysis of larger sample sizes will improve power and take full advantage of the rich datasets generated by untargeted metabolomics.

Copyright © 2018 American Society of Human Genetics. All rights reserved

174

Cross-platform, large-scale genetic discovery of small molecule products of metabolism and application to clinical outcomes. *C. Langenberg[1], C. Li[1], I. Stewart[1], L.B.L. Wittemans[1], C. Oliver-Williams[2], P. Surendran[2], E.K. Biggs[3], R. Bonelli[4], R.A. Scott[1], S. Burgess[5,6], V. Zuber[6], A. Koulman[1,3,7], F. Imamura[1], N.G. Forouhi[1], K.T. Khaw[6], J.L. Griffin[8], A.M. Wood[2], F.M. Gribble[3], F. Reimann[3], M. Bahlo[4], J. Danesh[2,9], A. Butterworth[2], N.J. Wareham[1], L. Lotta[1], MacTel Consortium.* 1) MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom; 2) Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; 3) Metabolic Research Laboratories, University of Cambridge, Cambridge, United Kingdom; 4) Population Health and Immunity Division, Walter & Eliza Hall Institute of Medical Research, Parkville, Australia; 5) MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom; 6) Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; 7) NIHR BRC Nutritional Biomarker Laboratory, University of Cambridge, United Kingdom; 8) Biochemistry Department, University of Cambridge, United Kingdom; 9) Wellcome Trust Sanger Institute, Hinxton, United Kingdom.

Background: Circulating plasma metabolites are products of metabolism that are highly heritable and have important roles for the screening and diagnosis of human diseases. Earlier genetic studies have been limited in scope by focusing on metabolites assessed using a single method. **Methods**: We conducted meta-analyses of genome-wide association studies of human plasma concentrations of 174 metabolites from 7 biochemical classes measured on the Biocrates (AbsoluteIDQ™ p180, Fenland Study), Nightingale (1H-NMR, INTERVAL Study), and Metabolon (Discovery HD4™, EPIC-Norfolk and INTERVAL Studies) platforms. We integrated large-scale, unpublished genetic “cross-platform” associations with publicly available summary statistics. Sample sizes ranged from 9,363 to 86,507 people for metabolites covered on all platforms and studies. The health consequences of identified associations were explored using disease-focused and phenome-wide analyses in up to 0.5 million people of UK Biobank. **Results**: We identified 499 locus-metabolite associations (from 47,896 SNP-metabolite associations significant at a principal-component-corrected genome-wide threshold of $p<4.9x10^{-10}$) based on 144 loci. Functional annotation revealed that largest effect sizes (0.25-1.5 standard deviation differences in log-transformed metabolites per allele) were driven by non-synonymous variants across the allele frequency spectrum. Overall, metabolite associated SNPs were highly enriched for non-synonymous variants when compared to GWAS (41.7-fold), disease- (3.7-fold) or continuous trait- (4.4-fold) associated variants. We applied independent lead mQTLs to summary statistics for selected rare and common diseases and in phenome-wide analyses and identified (a) strong genetic evidence for a protective causal association between serine and macular telangiectasia type 2, a poorly understood, rare degenerative retinal disease (odds ratio (95% confidence interval) per standard deviation genetically-higher serine 0.05 (0.03; 0.08), $p=9.5x10^{-30}$), (b) an excess of observed versus expected associations for type 2 diabetes, including a functional variant in *GLP2R* associated with citrulline, and (c) a range of multiple-test corrected associations with specific diagnoses, medical procedures or medications in phenome-wide analyses. **Conclusions**: Our cross-platform discovery gives important new insights into the genetic architecture of human metabolism and its relevance for the development of rare and common diseases.

175

Identification of FXTAS genetic modifiers by combining high-throughput metabolic profiling with fly genetics. *H.E. Kong[1], J. Lim[1], F. Zhang[1], L. Huang[1], D.L. Nelson[2], P. Jin[1].* 1) Department of Human Genetics, Emory University, Atlanta, GA; 2) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA.

Fragile X-associated tremor/ataxia syndrome (FXTAS) is an adult-onset neurodegenerative disorder that affects carriers of premutation alleles (55–200 CGG repeats) of the fragile X mental retardation 1 (*FMR1*) gene. We previously generated the FXTAS mouse model that directs expression of 90 CGG repeats in the cerebellar Purkinje neurons, and displays the key phenotypic features of FXTAS. To identify the metabolic alterations associated with FXTAS CGG toxicity and progression of disease, we performed untargeted global metabolic profiling of age-matched control and FXTAS mice at 16-20 weeks and 55 weeks. To identify the metabolic changes that result from the presence of the CGG repeats, we performed two comparisons at two different age groups, (1) young control vs. young FXTAS mice at 16-20 weeks and (2) old control vs. old FXTAS mice at 55 weeks. In addition, to examine the metabolic changes that result from the progression of the FXTAS phenotype during aging, we assessed metabolite alterations between the old FXTAS and young FXTAS mice. Out of 506 metabolites in cerebellum, we identified 89 altered metabolites ($p<0.005$) that demonstrate significant perturbations due to the presence of the $r(CGG)_{90}$ repeat, and found that these differences increase dramatically with aging. To identify the key metabolic changes for FXTAS pathogenesis, we performed a genetic screen using a *Drosophila* model of FXTAS. Out of 30 genes that we tested in the fly, 10 genes showed significant enhanced neuronal toxicity associated with rCGG repeats, 2 of which are associated with sphingolipid metabolism. Using high resolution LC/MS Lipid profiling, we further revealed a significant number of biochemicals involved in sphingolipid metabolism that are altered in the old FXTAS mice compared to the old WT counterparts and also in the old FXTAS mice compared to young FXTAS mice, signifying that sphingolipids in general are strongly affected in the onset and progression of CGG toxicity in the cerebellum, which is further supported by the strong genetic interaction between CGG repeat toxicity and key sphingolipid metabolism enzymes, such as *GBA and SPHK1*. By combining metabolic and LC/MS lipid profiling with a *Drosophila* genetic screen to identify genetic modifiers of FXTAS, we demonstrate an effective method for functional validation of high-throughput metabolomics data and reveal the role of sphingolipid metabolism in FXTAS pathogenesis. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

176

Analysis of 33,527 haploid sperm genomes from 20 individuals reveals new relationships underlying meiotic recombination and aneuploidy. *A.D. Bell[1,2], C.J. Mello[1,2], S.A. McCarroll[1,2].* 1) Genetics, Harvard Medical School, Boston, MA; 2) Medical and Population Genetics, Broad Institute, Cambridge, MA.

Many questions about meiosis could benefit from analyzing thousands of gamete genomes. For example, male meiosis appears to involve crossover interference (the tendency of consecutive crossovers to be considerably physically separated) and bias of crossover locations toward chromosome ends, but it is not known whether such properties vary across individuals, nor how they, recombination rate, and aneuploidy interrelate. To address these and other questions, we developed Sperm-seq, a droplet-based single-cell DNA sequencing technology for sperm. We analyzed 33,527 sperm from 20 human donors, sequencing 0.8-4% of each haploid genome and typing 13,000 heterozygous SNPs (median) per cell. Using these data, we phased each individual's complete set of heterozygous SNPs into chromosome-length haplotypes. We identified >850,000 haplotype transitions (crossovers) (10-48 per cell), with individuals differing (as expected) in median per-cell crossover rates (range 22-27, KW test $p<10^{-300}$). The spatial scale of crossover interference varied across the 20 individual donors, with the median distance between consecutive crossovers ranging from 68 to 90Mb (KW test $p<10^{-300}$). The tendency to place crossovers near chromosome ends also varied among individuals (range 67-77% of crossovers in the distal 50% of chromosome arms; KW test $p<10^{-300}$). Both properties negatively correlated with crossover rate (interference r=-0.97, $p=2\times10^{-12}$; end bias r=-0.92, $p=1\times10^{-8}$). Surprisingly, we found genomic regions in which individuals with low global recombination rates had substantially more crossovers than individuals with high global rates. These results suggest that recombination rate is a proxy for a complex set of recombination-related phenotypes. Recombination has been proposed to protect against aneuploidy. We inferred 521 autosomal aneuploidy events from sequence coverage, including at least one affecting each chromosome. The 20 sperm donors exhibited fourfold variation in aneuploidy rate (0.7-2.9% of cells affected). However, aneuploidy rates did not correlate with global crossover rates (r=-0.19, p=0.4) and we observed aneuploidy of chromosomes that had undergone multiple crossovers. We are currently working to characterize these and other relationships at per-cell and per-chromosome levels, and to better understand inter-individual and inter-chromosomal differences in aneuploidy frequency and crossover interference.

177

Explaining gene expression: Massively parallel in-silico testing of gene expression regulators reveals large scale patterns of shared and divergent regulation across cancers and tissues. *L. Erdman, M. Mai, D. Sokolowski, M. Wilson, A. Goldenberg.* Hospital for Sick Children, Toronto, Ontario, Canada.

Background: Changes in gene expression (GE) confer with many diseases and traits. Accurate prediction of GE can help identify likely sources of dysregulation, thus pointing to potential biomarkers and therapeutic targets. In this work, we developed integrative gene-specific GE prediction models that include transcription factor (TF) expression, miRNA, methylation (methyl), CNV and eQTLs to predict GE across 14 cancers and investigate regularities in and implications of the top predictors.Methods: Raw RNAseq reads, miRNA expression, methyl 450K, somatic SNV, and germline eQTL data were downloaded from the Cancer Genome Atlas. All data was normalized, subset to a Caucasian-only sample, and corrected for sex, batch, and the first 2 principal components of the SNP data. We then fit an elastic net model and used leave-one-out cross validation to assess our in-cancer variance explained (VE). Models were applied to normal samples in the same tissue to assess generalizability of our findings.Results: Of the 10,907 genes analyzed in all 14 cancers (average training n = 210), an average of 3323 (30%) of GE predictions resulted in over 80% of GE VE. Esophageal carcinoma (ESCA) was best explained in both cancer and normal tissue with over 80% of their VE in 83% and 60% of genes, respectively. Pancreatic of adenocarcinoma (PAAD) and kidney carcinoma are the least well predicted in cancer and normal samples, respectively, with 80% of GE VE in only 6% of genes in both.TF-only, TF+methyl and methyl-only were the most predictive regulators in both cancer and normal samples. In normal samples, 7619 genes have over 80% of their VE in at least 2 cancers. Genes grouped by top regulator were significantly different across cancers (p < 2.2e-16). Melanoma and ESCA showed the most distinct regulatory patterns among it's genes while PAAD, uterine carcinoma, pheochromocytoma and paraganglioma had the most similar regulatory patterns. Cancers from the same organ shared more top predictors and methylation was found to play a much greater role in cancers with squamous histology than other types of cancers we analyzed. The best predicted genes in both cancer and normal tissue were primarily annotated to the nucleoplasm and nucleus (p_bonf < 9e-08).Conclusion: This pioneering work broadly annotating gene regulation across the genome will allow researchers to further explain their GE biomarkers, potentially identifying causal factors of disease.

Copyright © 2018 American Society of Human Genetics. All rights reserved

178

Compositional and Time-course Aware Genetic Analysis (CTG): A novel analysis platform for high-throughput functional genetic interaction screens. *B.P. Munson[1,2], J.P Shen[2,3,4], S. Fong[1,2], A. Birmingham5[5], R. Sasik[5], J.F. Kreisberg[2,4], P. Mali[1,2,3], T. Ideker[1,2,3,4].* 1) Bioengineering, University of California San Diego, La Jolla, CA; 2) The Cancer Cell Map Initiative (CCMI); 3) Moores UCSD Cancer Center; La Jolla, CA; 4) Department of Medicine, Division of Genetics, University of California, San Diego, La Jolla, CA; 5) Center for Computational Biology and Bioinformatics, University of California, San Diego, La Jolla, CA.

The recent development of CRISPR-Cas9 technology, in which the protein Cas9 can be selectively directed to specific genomic locations using a guide RNA (sgRNA), now allows for fascicle functional manipulation of the genomes of eukaryotic cells. CRISPR technology is now frequently used to conduct high-throughput functional genomic screening experiments where up to hundreds of thousands of gRNA are pooled together and the viability effect of each gRNA is determined by measuring the change in relative abundance of each gRNA over time. Recently CRISPR technology has been modified to allow for combinatorial gene knockout, allowing for testing of genetic interaction. The pooled CRISPR knockout approach has proven immensely valuable in multiple scientific disciplines; however, this format presents several unique bioinformatics challenges. Here we present Compositional and Time-course aware Genetic analysis (CTG), a novel analysis platform designed specifically for high-throughput genetic interaction screens. Unique from prior methods, CTG accounts for the compositionality of the pool screening format and also incorporates time course data to greatly improve the precision of fitness measurements. The CTG analysis model accounts for the compositional effects inherent in pooled growth experiments by performing an initial log-ratio transformation of strain abundances and leverages time course data to improve estimation of the growth rate of individual strains. The accuracy of single gene knockout growth rates is increased by imputing fitness as the latent variables in an iterative least squares fitting of all dual knockout growth rates simultaneously. In a prior genetic interaction screen, sgRNA abundance was sampled after days 3, 14, 21 and 28 post-transduction. Using all four time points the correlation of single gene knockout fitness was highly correlated between two experimental replicates (Pearson r = 0.94, p = 1.2 x 10-37), this result was robust to downsampling to 50% of the original total read counts. However, removing the day 14 and 21 samples significantly increased the error in single gene fitness measurement. Removing the day 14 and 21 time points also increased the relative error of genetic interactions scores to a greater degree than downsampling to 50% across all four data points (0.58 vs. 0.42, p = 0.0007). To facilitate the use of the CTG method to the greater informatics community, CTG is distributed as a python package, available on GitHub..

179

Genome-wide association studies in >500,000 individuals identify novel loci for coronary artery disease and myocardial infarction. *J.A. Hartiala[1,2], Y. Han[1,2], Q. Jia[1,2], P. Huang[1,2], N.C. Woodward[1,2], J. Gukasyan[1,2], L.K. Stolze[3], Z. Kurt[4], D-A. Trégouët[5,6,7], N.L. Smith[8], M. Seldin[9,10,11], C. Pan[9,10,11], M. Mehrabian[9,10], A.J. Lusis[9,10,11], W.H.W. Tang[12,13], S.L. Hazen[12,13], X. Yang[4], C.E. Romanoski[3], H. Allayee[1,2].* 1) Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA; 2) Department of Biochemistry & Molecular Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA; 3) Department of Department of Cellular and Molecular Medicine, University of Arizona College of Medicine, Tucson, AZ; 4) Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, CA; 5) Institut National pour la Santé et la Recherche Médicale (INSERM), Paris, France; 6) Sorbonne Universités, Université Pierre et Marie Curie, Team Genomics & Pathophysiology of Cardiovascular Diseases, Paris, France; 7) Institute for Cardiometabolism and Nutrition (ICAN), Paris, France; 8) Department of Epidemiology, University of Washington, Seattle, WA; 9) Department of Medicine, David Geffen School of Medicine of UCLA, Los Angeles, CA; 10) Department of Human Genetics, David Geffen School of Medicine of UCLA, Los Angeles, CA; 11) Department of Microbiology, Immunology, & Molecular Genetics, Geffen School of Medicine of UCLA, Los Angeles, CA; 12) Department of Cardiovascular Medicine, Cleveland Clinic, Cleveland, OH; 13) Department of Cellular & Molecular Medicine, Cleveland Clinic, Cleveland, OH.

Understanding the genetic basis of coronary artery disease (CAD) and clinically significant endpoints, such as myocardial infarction (MI), has the potential to identify casual drivers of disease and novel therapeutic targets. To further define the genetic architecture of CAD and MI, we performed meta-analyses of genome-wide association study (GWAS) data with ~7.8 million SNPs in 521,504 subjects from the UK Biobank and the CARDIoGRAM+C4D Consortium. We identified 5 novel loci (*ERG, SLC44A3, ANXA4*, *PDLIM5* and *FLH5*) for CAD or MI, some of which were more strongly associated with one phenotype versus the other. Additionally, gene-based GWAS analyses identified 90 novel genes with association signals for CAD and/or MI, of which three genes overlapped with newly discovered loci in our SNP-based meta-analyses. Most notably, replication analyses in two independent datasets validated the preferential association of the *SLC44A3* locus with MI in the presence of coronary atherosclerosis. Tissue-specific analyses revealed that genes associated with CAD and/or MI were more highly connected within expression networks of endothelial cells than those of the aorta and coronary artery. Functional studies with one of the newly identified CAD loci on chr21q22.2 showed that expression of *ERG*, a gene highly expressed in endothelial cells, was downregulated by oxidized lipids and inversely correlated with proinflammatory genes. Taken together, our results support the concept that some of the biological mechanisms leading to plaque rupture and thrombus formation may be etiologically distinct from those that contribute to coronary atherosclerosis progression. Furthermore, these data identify a potentially novel protective role of *ERG* in atherogenesis and provide further evidence for the functional importance of the endothelium in the pathogenesis of CAD.

Copyright © 2018 American Society of Human Genetics. All rights reserved

180

Leveraging whole genome sequencing to identify novel determinants of platelet function. *B.A.T. Rodriguez[1], A.R. Keramati[2,3], L. Yanek[2], M.H. Chen[1], K. Ryan[4], B. Gaynor[4], J.A. Brody[5], N. Faraday[2,6], L. Becker[2,3], J. Lewis[4], A.D. Johnson[1], R.A. Mathias[2,7].* 1) National Heart, Lung and Blood Institute's The Framingham Heart Study, Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, Framingham, MA, USA; 2) GeneSTAR Research Program Department of Medicine, Division of General Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 3) Department of Medicine, Division of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 4) Division of Endocrinology, Diabetes, and Nutrition, Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA; 5) Cardiovascular Health Research Unit, University of Washington, Seattle, WA, USA; 6) Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 7) Department of Medicine, Division of Allergy and Clinical Immunology, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

Activated platelets provide the link between inflammation, thrombosis, and atherosclerotic cardiovascular disease. Platelet reactivity is highly heritable, yet the number of previously identified loci are limited and explain a relatively small portion of estimated heritability. Leveraging the scientific resources of TOPMed, we here report the first association study of platelet aggregation in response to variety of physiological stimuli using whole genome sequencing (WGS) data. Three extensively phenotyped studies of platelet function including GeneSTAR, the Framingham Heart Study and the Old Order Amish Study collaborated to (1) refine previously identified GWAS loci and (2) identify novel loci that determine platelet aggregation in response to different doses of collagen, ADP, and epinephrine. Tests for association using a 2-stage inverse-normal transformation after adjusting for age, sex and study were performed for 19 harmonized platelet aggregation phenotypes and ~69.6M variants within a multi-ethnic mega-analysis framework. We identified 19 novel, independent loci reaching genome-wide significance (P<5E-8), two of which may impact clinically actionable genes: 1p36 (P=1.04E-8, MAF=0.077, *PINK1)* and 1q31 (P=1.96E-9, MAF=0.442, *RGS18)*. Previous knock out studies in mice suggest RGS18 acts as a brake on persistent or inappropriate platelet activation. *PINK1*-null mice have previously been shown to have increased platelet reactivity and thrombosis. We developed an innovative approach for thresholding variant effect prediction in the gene-based SKAT framework to further investigate low frequency or rare coding variants and identified five genes reaching genome-wide significance: *SVEP1*, *CDNF*, *BCO1*, *NELFA*, *IDH3A*. Our results for the *SVEP1* gene, a risk locus for coronary artery disease including myocardial infarction, are driven by a missense coding variant and thus provide a testable biological mechanism for *SVEP1* in heart attack. Finally, low-frequency or rare non-coding variant SKAT of cell-lineage specific epigenetic regulatory maps identified a megakaryocyte super-enhancer region near the platelet factor gene *PEAR1*, a known locus of common non-coding variants for platelet reactivity. This shows us the *PEAR1* locus is more functionally complex than previously understood. WGS data coupled with innovative analytical strategies has resulted in new loci and better understanding of the determinants of platelet aggregation.

181

Genome-wide identification of DNA methylation quantitative trait loci in human whole blood highlights novel pathways for cardiovascular diseases. *T. Huan[1], R. Joehanes[1], C. Song[2], F. Peng[3], Y. Guo[4], M. Mendelson[1], C. Yao[1], C. Liu[1], L. Almli[5], K. Conneely[6], A. Johnson[1], M. Fornage[3], L. Liang[7], D. Levy[1].* 1) NHLBI- Framingham Heart Study, Framingham, MA; 2) Department of Medical Sciences, Uppsala University, Uppsala, Sweden;; 3) Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, TX, USA; 4) Department of Environmental Health, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA;; 5) Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA, USA;; 6) Department of Human Genetics, Emory University School of Medicine, Atlanta GA, USA; 7) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

Background: Identifying genetic variants associated with DNA methylation, known as methylation quantitative trait loci (meQTLs), and integrating them with disease-associated variants from genome-wide association studies (GWAS) may illuminate novel functional mechanisms underlying SNP-disease associations. **Methods and Results:** We analyzed genome-wide associations of genetic variants with leukocyte-derived DNA methylation assessed for over 420K CpGs in 4170 Framingham Heart Study participants. We comprehensively mapped more than 4.7M *cis*- and 630K *trans*-meQTLs targeting over 120K CpGs at Bonferroni-corrected P<0.05, representing 394K independent *cis*- and 21K independent *trans*- loci (using linkage disequilibrium r^2<0.2). External replication was performed in 963 participants from the ARIC study and 384 from the Grady Trauma Project. Next, we linked *cis*-meQTLs with GWAS results for cardiovascular disease (CVD) traits, and employed Mendelian randomization (MR) analysis to infer causal relations. We identified 93 putatively causal CpGs for CVD traits. Further integrating gene expression data revealed evidence of causal CpGs that drive gene expression to promote CVD. For example, we found that a decrease in methylation of cg12555086 is causally related to an increase in *LIPA* expression, and promotes coronary heart disease. In addition, we identified 22 *trans*-meQTL hotspots each targeting more than 30 CpGs, and found that *trans*-meQTL hotspots appear to act in *cis* on expression of nearby transcriptional regulatory genes. **Conclusions:** Our findings provide a powerful meQTL resource that can be used to highlight putative causal pathways involved in the pathogenesis of human diseases and to identify promising therapeutic targets for disease prevention and treatment.

Copyright © 2018 American Society of Human Genetics. All rights reserved

182

Exploring population-specific incidence of disease in a multi-ethnic health system reveals Native American haplotype underlying peripheral artery disease in Dominicans. *S.F. Cullina[1], G.M. Belbin[1,2], G.L. Wocjik[5], E.P. Sorokin[5], M.A. Levin[3], C.R. Gignoux[4], E.E. Kenny[1-2].* 1) Charles Bronfman Institute for Personalized Medicine, Mount Sinai, New York, NY; 2) Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 3) Department of Anesthesiology, Perioperative and Pain Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 4) Colorado Center for Personalized Medicine, University of Colorado, Denver, CO 80045; 5) Department of Genetics, Stanford University School of Medicine, Stanford, United States.

Population specific disparities in health outcomes are well documented and the inclusion of admixed populations in phenome-genome association studies provides an advantage previously under-utilised in genomics. Using an “Ancestry PheWAS” approach, identifying health disparities in an automated fashion, we examined statistical enrichment of medical billing (ICD-9) codes related to clinical phenotypes in over 21,000 participants from the Bio*Me* Biobank in the Mount Sinai Healthcare System in New York City. Ancestry PheWAS results found a significant association with 17 ICD9 codes indicating arterial disease and arteriosclerosis uniquely enriched in Dominican populations. One of the top signals was ICD9 429, encoding “Peripheral Artery Disease (PAD)”; P-value 1.1e-27; OR 2.1. PAD is characterized by an atherosclerotic narrowing of arteries in the extremities. To investigate the genetic underpinnings of PAD in Dominican Bio*Me* participants, we performed admixture mapping. On average, Dominican genetic ancestry traces to Africa (40%), Europe (50%) and the Americas (10%), and with local ancestry estimation we can test for significant deviations in patterns of local genetic ancestry between PAD cases and controls. We compiled genetic data from 2050 Dominican Bio*Me* participants genotyped on the Illumina OmniExpress (N=985) or Multi-Ethnic Genotype Array (N=1065). Local ancestry calls were made using RFMix, with European, Native American and African reference panels. A Native American ancestry haplotype at chromosome 2q35 was identified as the significant, with 12.99% vs 6.15% haplotype frequency in cases versus controls respectively, ($P=2.41x10^{-4}$; OR=2.17, 95% CI: 1.42, 3.27). To validate further, genotype data from Dominican ancestry Bio*Me* participants was imputed to the CAAPA consortium reference panel that individuals of Dominican heritage. Adjusting for sex and local ancestry haplotypes, we implicated variant rs13428326 ($P=2.36x10^{-5}$; OR=2.95, 95% CI: 1.77, 4.85). Subsequent fine mapping and functional annotation at the significant interval implicated the fibronectin gene (*FN1*), and we are currently following up the functional and clinical implications of this finding in Bio*Me*. In summary, we show how the use of an ancestrally diverse biobank in an Ancestry PheWAS approach enabled guided genomic discovery with translational utility in a relevant admixed population. .

183

Identification of structural variation associated with cardiometabolic traits in the Finnish population. *L. Chen[1,2], L. Ganel[1,2], H.J. Abel[1,3], D.E. Larson[1,3], A. Regier[1,2], I. Das[1], S.K. Service[4], X. Yin[5], A.U. Jackson[5], M.I. Kurki[6], A.S. Havulinna[7], P. Palta[8], S.K. Dutcher[1,3], V. Salomaa[7], C.W.K. Chiang[9], A. Palotie[6,8,10,11], S. Ripatti[8,12], M. Laakso[13], N.B. Freimer[4], M. Boehnke[5], A.E. Locke[1,2], N.O. Stitziel[1,2,3], I.M. Hall[1,2,3], FinMetSeq Consortium.* 1) McDonnell Genome Institute, Washington University School of Medicine, Saint Louis, MO; 2) Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA; 3) Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA; 4) Department of Psychiatry, The Jane and Terry Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine, University of California, Los Angeles, CA, USA; 5) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA; 6) Broad Institute of MIT and Harvard, Cambridge, MA, USA; 7) National Institute for Health and Welfare, Helsinki, Finland; 8) Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland; 9) Department of Ecology and Evolutionary Biology, University of Southern California, Los Angeles, California, USA; 10) Massachusetts General Hospital, Boston, MA, USA; 11) Harvard Medical School, Boston, MA, USA; 12) Public Health, Faculty of Medicine, University of Helsinki, Finland; 13) Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.

Cardiovascular disease (CVD) is a leading cause of death worldwide and has significant heritability. Most previous genome-wide studies have studied single nucleotide variants and short indels. The contribution of structural variation (SV) to CVD is largely unknown. We are midway through a large whole genome sequencing (WGS) study of coronary artery disease and CVD risk factors, undertaken by our NHGRI Center for Common Disease Genomics. Here, we describe results from an interim association analysis of SV and quantitative cardiometabolic traits in 4,935 individuals from Finland. Relative to prior work, our study benefits from deep WGS (≥20X), extensive trait measurements for METSIM and FINRISK cohorts, and the unique population history of Finland, which has resulted in an excess of low-frequency, high-impact variants that are rare or absent in other populations. We detected SV using two complementary methods: breakpoint mapping with LUMPY and CNV detection with GenomeSTRiP. We discovered 62,805 high-confidence SVs including 28,417 deletions, 15,510 duplications, 11,462 multi-allelic CNVs, 281 inversions, 2,265 mobile element insertions, and 4,870 difficult-to-classify "other" SVs. We tested SVs with MAF>0.1% for association with 137 cardiometabolic traits using a mixed model accounting for relatedness. We then attempted to replicate SVs reaching a suggestive level of statistical significance ($P\leq10^{-6}$) in an independent exome sequencing dataset of 16,135 Finnish samples analyzed with XHMM. We are currently defining SVs that meet genome-wide significance across WGS and WES data, accounting for correlation across traits, and will present these at the meeting. However, our initial analyses have identified several unequivocal and novel signals. For example, deletion of the *ALB* promoter is associated with decreased serum albumin and increased cholesterol ($p=1.6x10^{-11}$), a complex multiallelic CNV at *PDPR* is associated with pyruvate and alanine ($p=1.8x10^{-16}$), recurrent deletion of *HP* is associated with glycoprotein acetyls ($p=8.1x10^{-17}$), and recurrent gene conversion between the *CYP21A2/CYP21A1P* gene/pseudogene pair is associated with VLDL ($p=1.3x10^{-7}$). Remarkably, roughly half of trait-associated SVs are complex multi-allelic variants that are not well tagged by SNPs. Our results suggest that SVs underlie some trait associations not detectible from GWAS and have the potential to greatly expand our knowledge of genetic factors underlying CVD risk.

Copyright © 2018 American Society of Human Genetics. All rights reserved

184

Identifying novel genes influencing cardiometabolic risk factors from GWAS-identified loci for triglyceride levels using a high-throughput zebrafish screen. *B. von der Heyde[1,2], M. Masiero[1,2], A. Emmanouilidou[1,2], T. Klingström[1,2], P. Ranefall[2,3], C. Wählby[2,3], A. Larsson[4], E. Ingelsson[4,5], M. den Hoed[1,2].* 1) Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden; 2) Science for Life Laboratory, Uppsala University, Uppsala, Sweden; 3) Department of Information Technology, Division of Visual Information and Interaction, Uppsala University, Uppsala, Sweden; 4) Department of Medical Sciences, Uppsala University, Uppsala, Sweden; 5) Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, USA.

Background: High triglyceride levels are an established risk factor for coronary artery disease. In 2013, a meta-analysis of genome-wide association studies (GWAS) identified 37 previously unanticipated loci that are associated with triglyceride levels. Identifying and characterizing the causal genes in these loci remains a challenge. Results from our proof-of-principle studies show that triglyceride-levels are also associated with atherosclerosis in zebrafish larvae. Therefore, we aim to characterize positional candidate genes using a high-throughput, image-based screen in zebrafish larvae, to increase understanding of GWAS-loci and identify potential novel drug targets. **Methods**: We prioritized 37 candidate genes for functional-follow up studies. These genes together have 41 zebrafish orthologues, which were targeted in five lines of ~8 genes each using a multiplexed CRISPR-Cas9 approach. Founder mutants have been raised and 384 offspring for two of the five lines have so far been screened for genetic effects on body size as well as fluorescently labeled lipids, macrophages and neutrophils in the vessel wall at 10 days post fertilization, using a high-throughput imaging set-up. Image quantification was performed using automated, custom-written pipelines in CellProfiler and ImageJ. Additionally, whole-body lipid fractions and glucose levels were measured in each larva using enzymatic assays. CRISPR-induced mutations were quantified using paired-end sequencing, and data were analyzed using hierarchical mixed models or (zero-inflated) negative binomial regression. **Results**: Each additional disrupted allele in *lpar2a* resulted in lower triglyceride, LDLc and total cholesterol levels. Mutants for *met* were characterized by less vascular infiltration by macrophages, and less co-localization of macrophages with lipids and neutrophils. Mutants for *gatad2a* tend to have lower glucose levels, while *gmip* (a RhoA GTPase activating protein) show less co-localization of lipids and neutrophils. **Conclusion**: By thoroughly dissecting GWAS-identified loci for triglyceride levels using a high-throughput, largely image-based approach in zebrafish larvae, we identified previously unanticipated genes that influence a range of cardiometabolic risk factors. Furthermore, we show that one locus can harbor multiple genes that influence independent risk factors. This approach can help translate GWAS-findings and discover potential new drug targets.

185

Illuminating the dark genome: Evaluation of technologies to assess medically relevant dark spots. *F.J. Sedlazeck[1], H. Doddapaneni1[1], D. Kalra[1], K. Walker[1], S.N. Jhangiani[1], S. Richards[1], Y. Han[1], Q. Meng[1], G. Metcalf[1], W.J. Salerno[1], E. Boerwinkle[1,2], D. Muzny[1], R.A. Gibbs[1].* 1) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 2) 2.Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030, USA.

Since the canonical human genome was completed, efforts to sequence human exomes or genomes are now routinely performed to catalog and interpret sequence and structural variations (SVs). Multiple genomic regions elude detailed characterization due to DNA repeats, low complexity and high mutation rates resulting in substantial population diversity. We aggregated data from 22 years of genomic experience to categorize and catalogue medically-relevant sequences that are difficult to assess (i.e. “dark spots”), using literature, known fragile DNA sites (eg. Cancer) and regions that continiously underprformed (e.g. coverage) based on short-read sequencing. We ranked the top 100 out of 6,513 candidate regions based upon clinvar scores and distance to genes with associtaed disease phenotypes. We then use this catalog to evaluate emerging long-read (PacBio, Oxford Nanopore) and 10x Genomics linked-read technologies for their potential to identify and phase SNPs and SVs. For example, the GBA region that has been associated with Parkinson’s and Gauchers disease, which has a highly homologus psedogene upstream causing mappability issues and recombination-mediated SVs and thus is hard to assess with Sanger or Illumina reads. We are able to identify SNPs and SVs and phase the relevant region with long reads alone. We further use RNA-Seq to assess the impact of the predicted variation on expression and isoform usage per sample. For example, we identified a polymorphic 2.6kbp deletion upstream of HLA-F that is correlated with the expression level and exon usage of HLA-F. We assessed regional phasing within the 100 regions and long-range phasing along the chromosoes. While we were able to phase SNPs and SVs proximal to the regions with long reads alone, the best global phasing was achieved with 10x Genomics ranging up to 67 MB using our customized pipelines, which results in two phasing blocks per chromosome arm on chromosome 6 including a fully phased HLA locus. We report the performance of each technology to identify variation in these 100 regions and discuss the developed pipelines. Together these results characterize the utility of each technology to obtain an accurate assessment of clinically relevant dark spots for which traditional approaches such as Sanger and Illumina fall short and whether these mutations suggest an impact on the expression of relevant genes.

Copyright © 2018 American Society of Human Genetics. All rights reserved

186

Contribution of retrotransposition to developmental disorders. *E.J. Gardner[1], A. Sifrim[2], G. Gallone[1], E. Prigmore[1], P.J. Short[1], H.V. Firth[1,3], M.E Hurles[1] on behalf of the Deciphering Developmental Disorders study.* 1) Human Genetics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, United Kingdom; 2) Center of Human Genetics, KU Leuven, Leuven, Belgium; 3) East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, Cambridgeshire, United Kingdom.

Mobile genetic Elements (MEs) are segments of DNA which, through an RNA intermediate, can generate new copies of themselves within their host genome. Additionally, MEs can facilitate the duplication of non-ME transcripts, typically genes, through the mechanism of retroduplication. Combined, these two processes constitute what is known as retrotransposition (RT), and in humans several disorders can be attributed to such activity. However, the majority of these deleterious events have been discovered on a case-by-case basis and neither MEs nor gene duplications are routinely analysed as part of clinical sequencing. Likewise, large sequencing cohorts designed to elucidate the causes of congenital and developmental disorders (DDs) have neglected to identify pathogenic events attributable to RT-derived mutagenesis. As such, we have used computational approaches to identify RT events in 9,738 whole exome sequencing (WES) trios with DD-affected probands as part of the Deciphering Developmental Disorders (DDD) study. Through our analysis, we have discovered and genotyped 1,129 ME sites, of which ~20% directly impact coding sequence, and 576 gene retroduplication events. Of our identified ME sites, we have identified 9 *de novo* MEs, 4 of which disrupt known DD genes and are likely causative of the patient's phenotype (0.04% of probands). We have also ascertained 6 *de novo* gene retroduplications, 3 of which were discovered in a single proband and which likely represent a hypermutator RT phenotype. Beyond identifying likely diagnostic RT events, we have utilized our dataset to develop an understanding of genome-wide ME mutagenesis and constraint. To this end, we have estimated the human ME mutation rate to be ~1.8×10^{-11} mutations per bp per generation and, combined with measures of SNV constraint from the DDD study, have demonstrated that coding RT events have signatures of purifying selection equivalent to those of truncating mutations. Our study suggests that while the overall burden of RT-attributable disease is relatively low in the human population, it is nonetheless an important consideration when elucidating the genetic basis of DD in individual patients. Overall, our analysis represents the single largest interrogation of the impact of RT activity on the coding genome to date.

187

Mobile element insertions in 28,000 clinical exomes. *R.I. Torene, K. Galens, J. Scuffins, B. Friedman, E. Ryan, H. Sroka, A. Singleton, C. Teigen, L.B. Henderson, K.G. Monaghan, J. Juusola, K. Retterer.* GeneDx, Gaithersburg, MD.

It has been 30 years since the discovery of mobile element insertions (MEIs) as a cause of disease in humans (PMID: 2831458), however, only around 150 disease-causing MEIs have been identified (PMIDs: 27158268 and 29025590). We, therefore, applied MEI detection to clinical NGS data using a custom structural variant discovery tool, SCRAMble (Soft Clipped Read Alignment Mapper). SCRAMble identifies clusters of soft-clipped reads, builds a consensus sequence, and aligns to a library of mobile elements. By focusing on clipped reads at MEI breakpoints, rather than on discordant read pairs, SCRAMble has increased sensitivity when the library fragment size is small. A retrospective analysis of >28,000 individuals with clinical whole exome sequencing identified >250,000 MEIs (representing 3,293 unique elements), yet only a small fraction of these MEIs are anticipated to be clinically relevant. The MEIs show hallmarks of target primed reverse transcription (PMID: 7679954) including variable 5' truncation, median 14 bp target site duplications, and L1 endonuclease recognition sequences. Of note, we observe a depletion for MEIs within exons. In fact, 40 MEIs were in coding exons and were singleton or de novo insertions in a proband. Since September 2017, we have been prospectively identifying MEIs in our clinical samples using both SCRAMble for clinical exomes and Mobster (PMID: 25348035) for targeted gene panels. As of May 2018, 19 MEIs (18 unique events, 10 Alus, 7 SVAs, and 1 L1) have been reported as pathogenic variants representing 0.2% of all positive findings. Four of the 18 unique pathogenic MEIs result from known founder events (PMIDs: 17079174 and 29025590) while the remaining 14 appear to be novel. In one case, prenatal exome sequencing identified a novel homozygous Alu in exon 4 of the *ETFB* gene, in a fetus with bilaterally enlarged microcystic kidneys, echogenic bowel, unilateral postaxial polydactyly, and an abnormal placenta. Biallelic loss of function variants in the *ETFB* gene cause glutaric acidemia IIB. The non-consanguineous parents were confirmed to be heterozygous for the MEI, suggesting a founder event. Our results confirm that MEIs may interrupt critical genes and be disease-causing; therefore, it is important for diagnostic laboratories to pursue MEI detection as part of their routine variant analysis. By applying MEI detection to clinical exome analysis, it is possible to provide families with diagnoses that were not previously evident.

Copyright © 2018 American Society of Human Genetics. All rights reserved

188

GangSTR: Genome-wide genotyping of short tandem repeat expansions. *N. Mousavi[1], S. Shleizer-Burko[2], M. Gymrek[2,3].* 1) Electrical and Computer Engineering Dept., University of California San Diego, La Jolla, CA; 2) Dept. of Medicine, University of California San Diego, La Jolla, CA; 3) Computer Science and Engineering Dept. University of California San Diego, La Jolla, CA.

Short Tandem Repeat (STR) expansions have been associated with dozens of genetic diseases, including Huntington's Disease, Fragile X Syndrome, and cancer risk. Standard diagnostic genetic tests for genotyping STRs face several important limitations: they assay only a single locus at a time, often do not infer absolute repeat count which may be important for disease severity, and cannot be used to discover novel pathogenic loci. Next-generation sequencing (NGS) can theoretically genotype all potentially pathogenic variants simultaneously. However, a major challenge is that expanded STRs are often beyond the read length of most NGS datasets and thus are difficult to accurately genotype. Moreover, standard alignment methods fall short when dealing with large insertions or deletions caused by STR variants. Recent efforts have developed tools that use information extracted from paired-end short-read sequencing data to estimate repeat length at a target set of STRs. However these methods only use a subset of the available information, and are designed to genotype a predefined set of pathogenic STRs. We present GangSTR, a novel tool for genome-wide profiling of both normal and expanded STRs. GangSTR employs a unified model that combines multiple sources of information extracted from paired-end reads to estimate maximum likelihood diploid STR lengths. We validated GangSTR's genotyping accuracy on real and simulated STR expansions at 24 known pathogenic loci by comparing estimated genotypes and ground truth. GangSTR achieves average Root Mean Square Error (RMSE)=23.6 in genotyping simulated data, outperforming similar tools such as ExpansionHunter (RMSE=48.0) and TredParse (RMSE=86.0). Our method additionally shows on average 11 and 29 fold speedup in per locus running time compared to TredParse and ExpansionHunter, respectively. We applied GangSTR genome-wide to the NA12878 genome and identified experimentally validated repeat expansions. Finally, we applied GangSTR to a cohort of 150 individuals to profile the landscape of STR expansions in a healthy population. Our analysis revealed that each individual harbors 5-10 heterozygous repeat expansions. GangSTR is offered as a standalone open-source tool that will enable genome-wide detection of novel repeat expansions. The ability to rapidly identify genome-wide repeat expansions will likely allow for discovery of novel pathogenic loci not currently accessible using existing tools.

189

Analysis of 5'UTR variation in WGS from over 18,000 individuals identifies highly constrained regulatory elements and a role in human disease. *N. Whiffin[1,2], K. Karczewski[3,4], A. O'Donnell-Luria[3,4], S.A. Cook[1,2,5], D.G. MacArthur[3,4], J.S. Ware[1,2,4].* 1) MRC London Institute of Medical Sciences, National Heart and Lung Institute, Imperial College London, London, United Kingdom; 2) Cardiovascular Research Centre, Royal Brompton and Harefield NHS Foundation Trust, London, United Kingdom; 3) Analytic and Translational Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; 4) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 5) National Heart Centre Singapore, Singapore.

Small upstream open reading frames (uORFs) in 5' UTRs of protein-coding genes have important tissue-specific *cis*-regulatory roles on translation of the downstream protein. Half of human genes have associated uORFs, which have been shown to reduce protein levels by up to 80%. Except for a few well-studied examples, the mechanisms of *cis*-regulation are not fully understood. Isolated case reports have shown that variants that create or disrupt uORFs can cause disease, and a recent study identified loss of uORF mutations in human malignancies. To systematically explore the role of variants that perturb uORFs across the genome in human disease, and to investigate the mechanisms of translational regulation, we studied frequencies of variation in the 15,496 whole-genome samples in gnomAD. We show that variants that create an AUG upstream of the canonical protein-coding start site (i.e. uORF-creating) are severely depleted in gnomAD, with only 48.2% (95%CI 47.4-49.0%) of the expected variation, compared to 79.6% (95%CI 79.3-79.8%) across all other UTR sites. This constraint signal is stronger for AUGs formed upstream of genes that do not have an existing uORF, and is highly correlated with the strength of the surrounding Kozak consensus sequence. We also demonstrate that constraint is strongest when the resulting uORF would overlap the canonical coding sequence out-of-frame, suggesting that a major mechanism of the functional impact of novel uORFs is to reduce translation of the downstream protein. Given that uORFs typically decrease translation of the downstream protein, we hypothesised that 5' AUG creating variants would be most deleterious upstream of haploinsufficient genes. Indeed, 5' AUG creating variants upstream of known haploinsufficient genes, and genes with a high probability of loss-of-function intolerance (pLI > 0.9) are significantly more constrained than those upstream of low pLI (<0.1) and known haplosufficient genes (binomial P=6.44x10^{-143}). Similarly high constraint is seen upstream of genes known to cause both developmental disorders and congenital heart disease. Finally, we present the spectrum of inherited, germline *de novo*, and somatic 5'AUG creating variants across over 2,500 families/patients with rare early-onset phenotypes or cancer.

Copyright © 2018 American Society of Human Genetics. All rights reserved

190

Uncovering pathogenic variants in the non-coding genome through the UK 100,000 genomes project. *J.M. Ellingford[1,2], H. Thomas[1], G. Arno[3], A. Webster[3], G. Beaman[1,2], K. Webb[1], K. Ibanez-Garikano[4], D. Polychronopoulos[4], C. Odhams[4], R. O'Keefe[1], W.G. Newman[1,2], G.C.M. Black[1,2].* 1) University of Manchester, Manchester, Greater Manchester, United Kingdom; 2) Manchester Centre for Genomic Medicine, Manchester Universities Hospitals, Oxford Road, Manchester, United Kingdom; 3) Department of Genetics, UCL Institute of Ophthalmology, London, United Kingdom; 4) Genomics England, Charterhouse Square, London, United Kingdom.

Background: Whole genome sequencing (WGS) enables the analysis of genome-wide variation that can range from single nucleotide alterations to large structural genomic variants. Through the UK 100,000 genomes project, there is an opportunity to integrate the discovery of novel non-coding pathogenic variants with the clinical diagnosis and management of individuals with genomic disease within the UK National Health Service. However the capability to identify and functionally characterize pathogenic variants in the non-coding genome remains a significant challenge within mainstream diagnostics.**Methods:** We developed an integrated bioinformatics and wet laboratory analysis framework to identify candidate pathogenic variants in the non-coding genome, initially with a focus on genes established as a cause of inherited Mendelian disorders. We performed *in-vivo* and *in-vitro* functional assays to prove the effect of non-coding mutations on gene regulation and splicing.**Findings:** We analysed over 4000 genomes for families with rare disease recruited to the pilot and main UK 100,000 genomes project. We identified novel non-coding pathogenic variants as a cause of monogenic and heterogeneous Mendelian disorders, including variants resulting in misregulation of gene expression, intron inclusion and exon skipping. These analysis strategies successfully identified pathogenic non-coding variants across a range of disorders, including cystic fibrosis, respiratory disorders and inherited sensory disorders.**Interpretation:** We demonstrate the capability of a WGS analysis framework to detect pathogenic variants missed by current diagnostic methodologies, resulting in diagnostic uplift. We establish the need for routine functional assays to prove the pathogenicity of non-coding variants, and develop strategies to integrate these findings within the diagnostic workup and management of patients for a range of inherited disorders. The framework provides a methodology to rapidly identify non-coding variants underpinning cases of missing heritability in monogenic and heterogeneous Mendelian disorders.

191

Single cell transcriptomics of the human airway epithelium reveals cellular and functional changes underlying type 2-high asthma. *N.D. Jackson[1], K.C. Goldfarbmuren[1], J.L. Everman[1], S.P. Sajuthi[1], C. Rios[1], R. Powell[2], M. Armstrong[2], J. Gomez[2], C. Michel[2], N. Reisdorph[2], C. Eng[3], S.S. Oh[3], J. Rodriguez-Santana[4], E.G. Burchard[3,5], M.A. Seibold[1,6,7].* 1) Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, United States; 2) Department of Pharmaceutical Sciences, University of Colorado-AMC, Aurora, CO, United States; 3) Department of Medicine, University of California-San Francisco, San Francisco, CA, United States; 4) Centro de Neumologia Pediatrica, San Juan, Puerto Rico; 5) Department of Bioengineering and Therapeutic Sciences, University of California-San Francisco, San Francisco, CA, United States; 6) Department of Pediatrics, National Jewish Health, Denver, CO, United States; 7) Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine, University of Colorado-AMC, Aurora, CO, United States.

Asthma is a heterogeneous disease characterized by airway hyperreactivity, obstruction, and inflammation. In the dominant “type 2-high” endotype of the disease, the type 2 cytokine, IL-13, largely drives a transformation of the airway epithelium (AE) from a homeostatic state containing a balanced mix of basal, mucus secretory, club secretory, and ciliated cells to a state of mucus hypersecretion. Yet, little is known about how cell types individually respond to IL-13 in the AE, or how IL-13 shifts cellular composition. To investigate cell-specific contributions to IL-13-induced remodeling in the AE, we measured transcriptome-wide expression of 2,650 single cells from human primary mucociliary AE cultures stimulated (or not) with IL-13 for 48 hours (acute) or 11 days (chronic). We found 11 distinct AE cell states using t-Distributed Stochastic Neighbor Embedding and shared nearest neighbor cluster analysis. With acute stimulation, we found that secretory cell populations specializing in innate immunity and *MUC5B* mucin secretion at baseline were replaced with secretory cells programmed for *MUC5AC* mucin secretion. We identified four ciliated cell populations, largely capturing distinct states of multiciliogenesis, which were also shifted by IL-13 to states more characteristic of mucin secreting cells. Chronic stimulation advanced metaplasia further, reducing the frequency of all non-*MUC5AC*+ cell populations and activating an endoplasmic reticulum stress response. These transcriptome changes were supported by immunohistochemistry. Mass spectrometry of epithelial protein secretions also showed patterns similar to those from the transcriptome, with increases observed for proteins characteristic of IL-13 mucus secretory populations (e.g., *MUC5AC*, *FCGBP*, and *ITLN1*) and decreases in proteins related to innate immunity (e.g., *SCGB1A1*, *SCGB3A1*, and *BPIFA1*). Furthermore, IL-13-induced changes resulted in a near complete loss of mucociliary transport, as measured by fluorescent bead assay. Finally, we found that genes and proteins underlying IL-13-induced remodeling in culture were highly associated with type 2-high asthma status and severity of disease observed in whole transcriptomes from nasal AE brushings of 698 children in the Genes-Environments, and Admixture in Latino Americans Study. These results illustrate the power of applying single cell transcriptomics to human AE cultures for describing *in vivo* whole transcriptome expression patterns in disease cohorts.

Copyright © 2018 American Society of Human Genetics. All rights reserved

192

Single cell RNASeq of Crohn's disease tissues defines two cell subset-based signatures of inflammation, predicting treatment responses. *J.H. Cho[1,3,9], M. Giri[1,3,9], J. Martin[2], R. Ungaro[9], L.S Chuang[1,3,9], S. Nayar[1,3,9], P. Desai[7], J. Friedman[7], C. Whitehurst[8], J. Hyams[6], L.A. Denson[5], S. Kugathasan[4], A. Greenstein[10], M. Merad[2], E. Kenigsberg[3].* 1) Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Institute for Precision Immunology, Icahn School of Medicine at Mount Sinai, New York, NY; 3) Department of Genetics, Icahn School of Medicine at Mount Sinai, New York, NY; 4) Emory University, Atlanta Georgia; 5) Cincinnati Children's Hospital, Cincinnati OH; 6) Connecticut Children's Medical Center, Hartford CT; 7) Jannsen Pharmaceutical Companies, Ardmore PA; 8) Boehringer-Ingelheim, Ridgefield CT; 9) Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York NY; 10) Department of Surgery, Icahn School of Medicine at Mount Sinai, New York NY.

Background: GWAS of Crohn's disease have implicated myriad immune cells, involving numerous cytokine pathways. Blockade of pro-inflammatory cytokines, such as with anti-TNF, is a primary treatment; however, a substantive fraction of patients do not respond, highlighting the need for a deeper pathophysiologic understanding to prioritize new treatment targets, and for improved prediction. **Methods:** Single cell RNASeq of inflamed and uninflamed Crohn's tissues from 10 patients was performed using 10x Genomics Chromium on 102,717 cells taken at surgery. K-means clustering was performed jointly across patients and inflammatory states. Validation of inter-sample variability of intestinal cell clusters was confirmed by protein-based mass cytometry (CyTOF). A gene list was developed integrating transcripts most differentially expressed between cell clusters, as well as genes differentially expressed between two distinct signatures of inflammation. This gene list was projected onto five bulk RNASeq/expression microarray datasets of 469 patients, including three clinical trials and a pediatric inception cohort. **Results**: The joint model with expert annotation defined 35 cell clusters within the 20 samples, with subset predictions validated by CyTOF. Across equivalently inflamed samples, we observed two signatures of inflammation as defined by PCA of cell fractions across the 20 samples. The two signature scores were consistently observed, with significant negative correlations ($r = -0.44$, $P < 10^{-10}$) across five bulk RNA datasets from patients at various disease and treatment stages, confirming a generalizable presence of the two inflammatory signatures. Projection of the signature score onto bulk RNASeq samples taken prior to anti-TNF treatment, predicted treatment response (Kolgomorov-Smirnov $D = 0.42$, $p=0.0058$). Specifically, few anti-TNF non-responders were found in patients with low Signature 1 scores (driven by increased expression of inflammatory mononuclear phagocytes, Th17 cells, activated T cells, fibroblasts and IgG plasma cells), thereby defining specific cell subsets driving treatment non-response. **Conclusions:** As with cancer, detailed evaluation of lesional tissues provides a critical means of predicting treatment response and pathophysiologic heterogeneity. It is possible that genetic variants most associated with a complex trait (here Crohn's) may be distinct from those driving cellular heterogeneity and treatment responses.

193

Human lineage tracing enabled by mitochondrial mutations and single cell genomics. *L.S. Ludwig[1,2], C.A. Lareau[1,2,3,4], J.C. Ulirsch[1,2,4], E. Christin[1], C. Muus[1,5], A. Brack[1], T. Law[1], C. Rodman[1], O. Rozenblatt-Rosen[1], M.J. Aryee[1,3,6,7], J.D. Buenrostro[1,8], A. Regev[1,9,10], V.G. Sankaran[1,2].* 1) Broad Institute of MIT and Harvard, Cambridge, MA, USA; 2) Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School Boston, MA, USA; 3) Molecular Pathology Unit, Massachusetts General Hospital, Charlestown, MA, USA; 4) Program in Biological and Biomedical Sciences, Harvard University, Cambridge, MA, USA; 5) Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA; 6) Department of Pathology, Harvard Medical School, Boston, MA, USA; 7) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 8) Society of Fellows, Harvard University, Cambridge, MA, USA; 9) Department of Biology and Koch Institute, Massachusetts Institute of Technology, Cambridge, MA, USA; 10) Howard Hughes Medical Institute, Chevy Chase, MD, USA.

Lineage tracing provides unprecedented insights into the fate of individual cells and their progeny in complex organisms. While effective genetic approaches have been developed *in vitro* and in animal models, these cannot be used to readily interrogate human physiology *in vivo*. Instead, naturally occurring somatic mutations in the nuclear genome have been utilized to infer clonality and lineal relationships between cells in human tissues, but current approaches are limited by high error rates and scale, and provide little information about the state or function of the cells. Here, we explore the utility of somatic mutations in the mitochondrial genome (mtDNA) that provide a compelling alternative: first, the human mitochondrial genome is 16.6 kb long, providing a substantial target for genetic diversity to serve as a natural barcode, while being sufficiently small for cost-effective sequencing. Second, the reported mutation rate for mtDNA is 10- to 100-fold higher than for nuclear genomic DNA. Third, these mutations often reach high levels of heteroplasmy or even homoplasmy. Finally, the relatively high mitochondrial copy number per cell facilitates confident detection of higher frequency heteroplasmic mutations in single cells. We show that somatic mutations in mtDNA occur ubiquitously across human tissues and can be measured by single cell RNA-seq (scRNA-seq) and single cell ATAC-seq (scATAC-seq), allowing for simultaneous analysis of single cell lineage and state. We leverage somatic mtDNA mutations as natural genetic barcodes and demonstrate their use as clonal markers in primary human hematopoietic cells and relate it to expression profiles and chromatin accessibility. By applying our approach to scRNA-seq data of chronic myelogenous leukemia, we demonstrate that mtDNA "barcodes" are stably propagated over extended periods of time *in vivo* and may enhance the understanding of (sub)-clonal evolution and underlying gene expression profiles in single leukemic cells. Further we show that mitochondrial genotyping can readily deconvolve donor and recipient chimerism during hematopoietic stem cell transplantation using high throughput droplet scRNA-seq approaches. Our approach should allow clonal/ lineage tracing at a 100- to 1,000-fold greater scale than with single cell whole genome sequencing with concomitant measures of cell state, opening the way to chart detailed cell lineage and fate maps in human health and disease.

Copyright © 2018 American Society of Human Genetics. All rights reserved

194

Whole cortical scRNA-seq reveals misregulation in multiple cellular and molecular networks in Fragile X syndrome. *E. Donnard[1], H. Shu[2], K. Gellatly[1], A. Derr[1], P. McDonel[1], M. Garber[1,2].* 1) Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA; 2) Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA.

Fragile X syndrome (FXS) is the most common form of inherited intellectual disability and results from the inactivation of a single gene, *Fmr1*. FXS shows striking phenotypic differences in synaptic morphology and function, with severe impact to the cognitive abilities of patients which are replicated in animal models. Despite decades of intensive research, we still lack an overview of the molecular and biological consequences of the loss of the encoded protein FMRP in the brain. One of the better established roles of FMRP is that of a translational repressor of specific mRNAs in neurons, and its absence has also been shown to correlate with a mild increase in basal protein synthesis in the hippocampus. Recently, several studies have emerged to address the pathophysiology in glial cell types, and suggest that FXS affects multiple cell types and their interactions. Previous attempts to detect transcriptional changes using bulk RNA sequencing (RNA-Seq) of the brain have likely been hindered by the complexity and heterogeneity of the tissue. Furthermore, in cell culture approaches, even though the profiled cells are homogeneous, the transcriptional program may not accurately reflect the *in-vivo* expression resulting from the interaction between multiple cell types. To obtain a first systematic transcriptional landscape of the FXS brain, we applied single cell RNA-Seq (scRNA-Seq) to profile over 21,000 cells from the dissociated cortex of wild type and *Fmr1* knockout mice. We find that loss of FMRP resulted in downregulation of FMRP target mRNAs in neurons but not in other cell types. Furthermore, downregulated genes in excitatory neurons and inhibitory neurons are enriched for different neurological and synaptic functions, which supports the excitatory-inhibitory imbalance theory of autism. Analysis of peptide-receptor pairs involved in intercellular signaling points to a mechanism for microglia-neuron miscommunication in FXS, which may affect synaptic pruning of neurons by microglia. This may underlie the long-observed synaptic phenotype, one of the hallmarks of FXS. Our data is the first attempt to dissect the cell-type specific contributions to FXS using the power of scRNA-Seq. Our results show how the loss of FMRP affects the intricate interactions between brain cell types, which could potentially open new doors to therapeutic interventions. It showcases how a systems approach can guide future studies on the mechanisms behind complex diseases like autism.

195

Chromatin dynamics of cortical organoids at single-cell resolution. *R.M. Mulqueen[1], B.A. DeRosa[1], A.J. Fields[1], A.C. Adey[1,2], B.J. O'Roak[1].* 1) Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR; 2) Knight Cardiovascular Institute, Oregon Health & Science University, Portland, OR.

Key regulatory networks control critical aspects of human cerebral cortex formation. Cellular differentiation, migration, and cortical lamination are reflected in precise dynamic shifts in epigenetic states. Recent genetic risk data from *de novo* mutations in many neurodevelopmental disorders (NDD), converge on these regulatory networks during fetal brain development. However, our understanding of the network characteristics and how specific mutations impact brain development remains limited. This is largely due to lack of available human biological samples of relevant genotypes, cell types, and developmental stages. Induced pluripotent stem cells (iPSCs) have emerged as a powerful tool for modeling human development. Moreover, the *in vitro* differentiation of 3D cortical organoids from iPSCs is known to recapitulate the same step-wise neurodevelopmental processes that occur during *in vivo* corticogenesis. To enable mechanistic studies of these key cortical networks, we implemented a scalable cortical 3D organoid model system, which mimics the same critical fetal developmental time period implicated in NDD risk To assay these dynamic changes in epigenetic states during the maturation of cortical organoids, we sampled dozens of individual human organoids matured for 6, 30, 60 and 90 days *in vitro* (DIV) using a modified SpinΩ bioreactor protocol (Qian et al. 2016). We measured epigenomic states using a combinatorial indexing ATAC-seq assay, which allows for single-cell chromatin accessibility profile generation for thousands of single cells in one experiment (Cusanovich et al. 2015). Further, we multiplexed many organoids per experiment, allowing for valuable characterization of inter-sample heterogeneity. One of the powers of measuring chromatin accessibility is the ability to assess putative transcription factor activities based on the enrichment of accessibility at corresponding motifs. A pseudotemporal ordering analysis on the first ~3,000 single-cell profiles revealed a succession of key brain transcription factor activities involved in cortical development. For example, cells putatively expressing transcription factor EMX1 increased between 30 and 60 DIV, reflecting a cell population transition from early neuroepithelial cells to neuroprogenitors. We believe these data will be foundational for uncovering the relevant chromatin changes that occur during brain development and modeling the effect of NDD-related mutations.

Copyright © 2018 American Society of Human Genetics. All rights reserved

196

The loss of *Ret* signaling in Hirschsprung disease: Cellular consequences and penetrance. *S. Chatterjee[1,2], E. Vincent[2], D. Auer[2], G. Cannon[2], L. Goff[2], A. Chakravarti[1,2].* 1) Center for Human Genetics and Genomics, NYU School of Medicine, NY, NY; 2) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD.

Despite considerable genetic heterogeneity, ~50% of the genetic risk of Hirschsprung disease (HSCR, congenital aganglionosis), in European ancestry subjects, arises from common enhancer and rare coding variants at the receptor tyrosine kinase gene *RET*. HSCR has a male bias (4:1), evident even in cases with *RET* variants, but the underlying cause for this sex difference in penetrance is unknown. Mouse models with homozygous loss of *Ret* are an effective HSCR model. RNA-seq on developing (E10.5-E14.5) whole gut tissue in wildtype and *Ret* null mice has identified significant transcriptional changes in early transcription factors and genes affecting gut epithelial and mesenchymal development with loss of *Ret* signaling. However, these studies fail to distinguish cell autonomous from non-autonomous changes or fail to identify which enteric cells are affected, and, thereby, provide little insight into the pathophysiology or sex difference. We have now performed single cell RNA-seq on developing gut tissue from male and female wildtype and *Ret* null mice at E12.5 and E14.5, the time period of enteric neuronal development. Our results show that two distinct *Ret*-expressing cell populations, early progenitor cells expressing glial makers (*S100b, Gfap, Sox10*) and the more fate-committed enteric neuronal population (marked by genes like *Nefh, Nefl, Snap25*), are severely affected. Loss of *Ret* signaling also leads to effects on cell division and cell cycle exit (*Clspn, Aurka, Cdc45*) leading to premature differentiation of the progenitor population. Further, a subset of the neuronal cells in the female gut change fate to a dopaminergic neuronal cell type and express markers (*Th, Ddc, Isl1*) absent or weakly expressed in other cells. These female-specific dopaminergic cells exclusively express *Gfra3* (a family member of the canonical Ret co-receptor Gfra1), highlighting the use of other co-receptors and alternate signaling cascades in the absence of *Ret.* Thus, distinct subsets of *Ret*-expressing cells exist and the effect of loss of *Ret* signaling is cell-type and sex specific in HSCR. Surprisingly, some neuronal cells in females escape cell death and change fate, possibly using a non-canonical Ret signaling pathway, and may be a primary cause of the lower female HSCR penetrance. Thus, there is greater heterogeneity in Hirschsprung disease at the genetic and cellular level than previously assumed from sequencing and cell staining studies.

197

Mendelian randomization and procedure-wide associations studies to evaluate the association of obesity with surgical procedures. *J.R. Robinson[1], R.J. Carroll[1], L. Bastarache[1], Z. Mou[1], W. Wei[1], J. Connolly[2], H. Hakonarson[2], F. Mentch[2], P.K. Crane[3], S.J. Hebbring[4], D.R. Gordon[5], A.S. Gordon[6], E.A. Rosenthal[6], I.B. Stanaway[6], M.G. Hayes[7], W. Wei[8], L. Petukhova[9], B. Namjou[10], G. Zhang[11], M.S. Safarova[12], N.A. Walton[13], R.J.F. Loos[14], S.N. Murphy[15], G.P. Jackson[1], I.J. Kullo[12], G.P. Jarvik[6], E.B. Larson[16], C. Weng[17], D. Roden[1], J.C. Denny[1].* 1) Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN; 2) The Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA; 3) Department of Medicine, University of Washington, Seattle, WA; 4) Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI; 5) Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA; 6) Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA; 7) Division of Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL; 8) University of Pittsburgh Medical Center, Pittsburgh, PA; 9) Department of Epidemiology, Columbia University, New York, NY; 10) Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 11) Division of Human Genetics, Cincinnati Children's Hospital Medical Center, and the Center for Prevention of Preterm Birth, Perinatal Institute, Cincinnati Children's Hospital Medical Center and March of Dimes Prematurity Research Center Ohio Collaborative, and Department of Pediatrics, University of; 12) Department of Cardiovascular Diseases, Mayo Clinic, Rochester, MN; 13) Department of Biomedical and Translational Informatics, Geisinger Health System, Danville, PA; 14) The Charles Bronfman Institute for Personalized Medicine at Mount Sinai, The Mindich Child Health and Development Institute, New York, NY; 15) Department of Neurology, Partners Healthcare, Boston, MA; 16) Kaiser Permanente Washington Health Research Institute, Seattle, WA; 17) Department of Biomedical Informatics, Columbia University, New York, NY.

Introduction: Body mass index (BMI) is a strong predictor of comorbidities and mortality; however, the association of BMI with surgical procedures is unknown. We used Mendelian randomization (MR) and a novel method, procedure-wide association studies (ProcedureWAS), to test for associations between obesity and surgery. **Methods:** We performed a retrospective, observational study using two cohorts. The primary cohort included adults with documented, non-pregnant BMI at Vanderbilt University Medical Center. The secondary cohort was derived from 12 institutions contributing clinical and genetic data to the eMERGE consortium. In the primary cohort, ProcedureWAS were performed using pairwise logistic regressions to measure the association of BMI with 178 aggregated procedure groups translated from Current Procedural Terminology codes. In the secondary cohort, MR was performed by calculation of a weighted genetic risk score from 97 obesity-related SNPs to determine the association of genetically-determined BMI with aggregated procedure groups by ProcedureWAS. **Results:** 736,726 and 64,824 individuals were in the primary and secondary cohorts, respectively. Using a Bonferroni significance threshold, class 3 obesity (BMI>40.0 kg/m^2) was positively associated with 56 of 178 (31.5%) of the aggregated procedure categories when compared to normal BMI (18.5-25 kg/m^2). Of these, 29 were replicated in the secondary cohort using MR for obesity (Table). **Conclusions:** Obesity is strongly correlated with increased risk for invasive procedures. MR suggests obesity is a causal factor leading to many of these surgeries. This is the first study using ProcedureWAS with aggregated procedure codes and pairwise analyses to determine associations with clinical phenotypes and genetic data, validating and complementing traditional phenotype and genetic studies.

Table. Strongest Replicated Associations of Obesity with Procedures using Mendelian Randomization

Procedure	Cases (n)	OR per 1-SD (6.9 kg/m^2) BMI	MR p-value
Gastric bypass	1663	8.6	7.9×10^{-43}
Therapeutic endocrine procedures	4195	1.96	4.7×10^{-11}
Non-cardiac vascular catheterization	23529	1.36	1.1×10^{-8}
Hernia repair	2366	2.08	2.1×10^{-8}
Gastrointestinal therapeutic procedures	3140	1.84	9.5×10^{-8}
Arthrocentesis	13794	1.37	4.5×10^{-7}
Laparoscopy	952	2.76	7.3×10^{-7}
Liver biopsy	1654	2	8.5×10^{-6}
Insertion of catheter or spinal stimulator and injection into spinal canal	7790	1.34	1.3×10^{-4}
Incision and drainage, skin and subcutaneous tissue	4265	1.45	1.9×10^{-4}
Knee arthroplasty	3120	1.53	2.5×10^{-4}

Copyright © 2018 American Society of Human Genetics. All rights reserved

198

Mendelian randomization-derived priors substantially improve power of Bayesian GWAS on human lifespan. *N. Mounier[1,2], P.R.H.J. Timmers[3], K. Läll[4,5], K. Fischer[4], Z. Ning[6], X. Feng[7], A. Bretherick[8], D.W. Clark[3], X. Shen[3,6], T. Esko[4,9], J.F. Wilson[3,8], P.K. Joshi[3], Z. Kutalik[1,2].* 1) Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne 1010, Switzerland; 2) Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland; 3) Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, United Kingdom; 4) Estonian Genome Center, University of Tartu, Tartu, Estonia; 5) Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia; 6) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; 7) State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources, Key Laboratory of Biodiversity Dynamics and Conservation of Guangdong Higher Education Institutes, School of Life Sciences, Sun Yat-sen University, Guangzhou, China; 8) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, United Kingdom; 9) Broad Institute of Harvard and MIT, Cambridge, MA, USA.

To improve our understanding of the genetic architecture of complex traits (such as human lifespan) Genome-Wide Association Studies (GWASs) are nowadays often conducted in more than 1M samples. Besides ever-increasing study sizes, further power can be gained from studies of related traits. To leverage this information, we developed a Bayesian GWAS method that carefully builds informative priors for each and every single nucleotide variant (SNV). Advanced Mendelian randomization (MR) has been proven useful to derive multivariate causal effects of a set of related traits on a focal phenotype. We use them in combination with summary statistics of GWASs of the related traits to estimate prior effects for each SNV. We optimized the causal effect estimation by testing several combinations of MR parameters (instrument strength, clumping stringency and shrinkage parameter) and found that a set of (universally) optimal MR parameter settings double the out-of-sample squared correlation (from 15% to more than 30%) between the resulting prior and the observed effect, compared to standard MR settings. Capitalizing on these highly informative priors we developed a computationally efficient way to compute Bayes Factors (BFs, quantifying the evidence in favor of the prior) and found a simplified analytical formula for the null BF distribution (given the approximated prior effect distribution) in order to control type I error. Compared to a permutation-based null BF sampling, our analytical approach led to a massively reduced runtime and a more accurate estimation of P-values for large BFs. Notably, the method also provides posterior effect size estimation, which can be used to improve the precision of genetic risk score (GRS) based prediction. When applying the approach to a GWAS on more than 1M parental lifespans, we identified 12 traits significantly affecting lifespan, including BMI, smoking, education level and coronary artery diseases. The prior effects derived from these risk factors lead to the identification of 10 variants in addition to the 25 identified by standard GWAS. Out-of-sample lifespan differences between the top and bottom GRS deciles improved substantially when using the posterior effect estimates as coefficients instead of the observed effects. The method, as well as functions facilitating results visualization, have been implemented in the R package *bGWAS* (https://github.com/n-mounier/bGWAS) and can easily be applied to any GWAS summary statistics.

199

Identifying the tissue-specific influence of gene expression on neurological and psychiatric traits: A Mendelian randomisation study on gene expression within the human brain. *D.A. Baird[1], J. Liu[2], S. Sieberts[3], T. Perumal[3], J. Zheng[1], V. Haberland[1], T.G Richardson[1], M. Carrasquillo[4], M. Allen[4], J.S Reddy[4], P.L De Jager[5], N. Ertekin-Taner[4,6], L.M Mangravite[3], B. Logsdon[3], P. Haycock[1], G. Hemani[1], G.D Smith[1], K. Estrada[1], T.R Gaunt[1], AMP-AD eQTL working group.* 1) MRC Integrative Epidemiology Unit (IEU), Population Health Sciences, Oakfield House, University of Bristol, Bristol, UK; 2) Biogen, 225 Binney Street, Cambridge, MA, USA; 3) Sage Bionetworks, Seattle, WA, USA; 4) Department of Neuroscience, Mayo Clinic Florida, Jacksonville, FL, USA; 5) Center for Translational & Computational Neuroimmunology, Department of Neurology, Columbia University Medical Center, New York, NY, USA; 6) Department of Neurology, Mayo Clinic Florida, Jacksonville, FL,USA.

Although genome-wide association studies (GWAS) are able to uncover SNPs associated with disease, understanding the biological mechanisms is more elusive as it is difficult to pinpoint the genes underlying these associations. This has especially been the case for neurological and psychiatric disorders, where, until recently, obtaining large gene expression datasets in relevant brain tissue has been challenging. We therefore employed a Mendelian randomization (MR) approach to determine the genes involved in these disorders. We conducted a two-sample MR (2SMR) study using gene expression quantitative trait loci (eQTLs) collected from 1,286 samples of brain cortex tissue (as part of the AMP-AD consortium) as genetic instruments. Nine neurological/psychiatric disorders were selected from MR-Base (http://www.mrbase.org/). Single SNP MR Wald Ratio estimates were calculated. A Bonferroni corrected threshold of $P<5x10^{-6}$ was used to identify associations between gene expression probes and disorders, and colocalization analysis was conducted to verify sharing of genetic signals between traits. Analysis was conducted in TwoSampleMR and coloc R packages. A total of 40,802 MR tests on 6,145 genes was computed across traits. 80 gene expression-trait associations were detected, of which 31 showed evidence of colocalization (probability > 80%) which included: 22 eQTLs with Schizophrenia, three eQTLs with Multiple Sclerosis (*TTC34, MPV17L2, IQCB1*), two eQTLs with Parkinson's Disease (*LINC02210, LRRC37A4P*) and two eQTLs with Amyotrophic Lateral Sclerosis (*G2E3, SCFD1), RHEBL1* eQTL with Bipolar Disorder and *CR1* eQTL with Alzheimer's Disease. For example, our MR results based on the cis-acting SNP (rs229243) suggested that expression level of both *G2E3* and *SCDF1* gene are associated with ALS; increased expression of *G2E3* reduced ALS risk (OR=0.69, $P=9.7x10^{-7}$), while increased expressed of the nearby *SCFD1* gene elevated ALS risk (OR=1.36, $P=7.66x10^{-7}$). Both eQTLs strongly colocalised (probability = 91%) with ALS, implying sharing of the same causal variant between traits within the cis-region. *SCFD1* had been implicated as a potential risk locus in an ALS GWAS (van Rheenen, 2016). Our study implicated eQTLs in six neurological/psychiatric traits, which suggests a potential causal role for the genes influenced by these eQTLs in disease pathogenesis, although it remains possible that these findings may be confounded via horizontal pleiotropy.

Copyright © 2018 American Society of Human Genetics. All rights reserved

200

The neurodevelopmental 16p11.2 CNVs have a hitherto unappreciated mirror effect on sexual development in humans and animal models. *K. Mannik[1,2], M. Lepamets[2,3], A. Mikhaleva[1], K. Lepik[4,5,6], T. Arbogast[7], H. Ademi[8], Z. Kupchinsky[7], C. Attanasio[1], A. Messina[9], S. Rotman[10], E. Dubruc[10], J. Chrast[1], S. Martin-Brevet[11], T. Laisk-Podar[12], Y. Herault[13], C.M. Lindgren[14,15,16], Z. Kutalik[5,6], J.C. Stehle[10], N. Katsanis[7], S. Nef[8], B. Draganski[11], E.E. Davis[7], A. Reymond[1], R. Magi[2], The 16p11.2 European Consortium, The Simons VIP Consortium, The eQTLGen Consortium.* 1) Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; 2) Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia; 3) Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia; 4) Institute of Computer Science, University of Tartu, Tartu, Estonia; 5) Institute of Social and Preventive Medicine, Lausanne University Hospital, Lausanne, Switzerland; 6) Swiss Institute of Bioinformatics, Lausanne, Switzerland; 7) Center for Human Disease Modeling, Duke University, Durham, NC, USA; 8) Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland; 9) Endocrinology, Diabetes & Metabolism Service, Lausanne University Hospital, Lausanne, Switzerland; 10) Service of Clinical Pathology, Lausanne University Hospital, Lausanne, Switzerland; 11) LREN, Department of Clinical Neuroscience, Lausanne University Hospital, Lausanne, Switzerland; 12) Women's Clinic, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia; 13) Institute of Genetics and Molecular and Cellular Biology, Illkirch, France; 14) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA; 15) Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK; 16) The Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK.

Accumulating evidence suggests that rare CNVs are a common health problem in the population. However, investigation of their effect has been biased towards clinical (often pediatric) studies. We combined data from a comprehensive set of clinically ascertained 16p11.2 families (n=660 individuals) with population cohorts (n>470,000; UKBB, EGCUT). Further, we used animal models, as well as large-scale eQTL and GWAS data to causally follow up identified associations. We uncovered that the 16p11.2 BP4-BP5 dosage, one of the most frequent genetic causes of mental disorders, was oppositely associated with age at menarche (AaM) in the 1st wave of the UKBB ($p=7.8e^{-05}$; all corrected for BMI). Compared to controls AaM was decreased in deletion (Δ=-1.5 years, p=0.01) and increased in duplication carriers (Δ=+1.5; $p=7.8x10^{-5}$). We replicated these associations in the 2nd wave of UKBB ($p=6.5x10^{-4}$), in the EGCUT ($p=2.4x10^{-5}$) and in two unrelated cohorts of female 16p11.2 patients ($p=7.7x10^{-5}$; $p=3x10^{-3}$). A directionally consistent trend for puberty onset was observed in 16p11.2 UKBB and patient males. These features were accompanied by reproductive disorders, e.g. miscarriages (OR=2.7; p=0.02), disorders of ovary and fallopian tube (OR=7.2; $p=1x10^{-3}$), cryptorchidism/hypospadias. We validated our results in 16p11.2 female mice by detecting changes in timing of first ovulation (p<0.01), estrous cyclicity (p=0.03), uterine size ($p=8.1x10^{-3}$). Male mice showed reduced anogenital distance ($p=4x10^{-3}$), a proxy for hormone-dependent sexual maturation. Corroboratively, genes differentially expressed in 16p11.2 patient LCLs and mice cortices were enriched for urogenital disease genes. Furthermore, in mice and men, we found a significant link between the 16p11.2 dosage and volume of hypothalamus (FWE<0.05), a key structure controlling reproduction, suggesting that perturbation of the GnRH axis could contribute to the observed phenotypes. Using Mendelian Randomization we prioritized potential causal genes for AaM. We challenged them by agnostically modulating dosage of all 16p11.2 genes in *gnrh3:egfp* transgenic zebrafish larvae and quantified GnRH neuronal patterning. We identified a putative CNV driver effect for *ASPHD1*, located in the GWAS peak for AaM, in the 16p11.2 reproductive axis. Our findings highlight how identification of traits associated with rare CNVs in the population provide valuable unbiased insight into disease etiologies in terms of comorbid traits and affected genes.

201

Mendelian randomization combining GWAS and eQTL data reveals new loci, extensive pleiotropy and genetic determinants of complex and clinical traits. *E. Porcu[1,2], S. Rüeger[2,3], F.A. Santoni[4], A. Reymond[1], Z. Kutalik[2,3], eQTLGen Consortium.* 1) Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; 2) Swiss Institute of Bioinformatics, Lausanne, Switzerland; 3) Institute of Social and Preventive Medicine, CHUV and University of Lausanne, Lausanne, Switzerland; 4) Endocrine, Diabetes, and Metabolism Service, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland.

Genome-Wide association studies (GWAS) identified thousands variants associated with hundreds complex traits, but in most cases their biological interpretation remains unclear. Many of these variants overlap with expression QTLs (eQTLs), indicating their potential involvement in the regulation of gene expression. Here, we propose a summary statistics-based Mendelian Randomization (MR) approach that uses multiple SNPs jointly as instruments and multiple gene expression traits as simultaneous exposures. Such an approach is more robust to violations of MR assumptions than state-of-the-art tools (GSMR, TWAS). When applied to 43 human phenotypes and using blood eQTL,we uncovered 2,277 putative genes causally associated with at least one phenotype. Importantly, 2,681 of the resulting 5,009 gene-trait associations were missed by GWAS. Using independent association summary statistics (UKBiobank), we confirmed the majority of these loci were missed by conventional GWAS due to power issues.We found several lines of evidence for the relevance of the identified genes. For example, *PEX19* and *CDC42*, respectively height- and intelligence-associated genes, are already known to harbor rare mutations leading to monogenic short stature (OMIM#614886) and Takenouchi-Kosaki syndrome (OMIM#616737). Our analysis found 48% (1,104/2,277) of genes having pleiotropic causal effect, impacting up to 20 traits. Notably, *TSPAN14* showed causal effect on rheumatoid arthritis, Crohn's disease and inflammatory bowel disease. To evaluate the shared causal effects of gene expression on pairs of traits we computed the expression correlation of Z-scores from MR results. We compared the expression with genetic correlation and we estimated that 44% of genetic correlation can be translated to similarity at the gene-expression level in whole blood. Applying our method to eQTLs from multiple tissues (GTEx) we showed how different tissues point to different genes, highlighting the importance to identify the relevant tissues for the studied phenotype before looking for the causal gene.In particular, we confirm that *SORT1* overexpression in liver reduces LDL levels. In summary, we demonstrated that our method, through combining summary statistics from GWAS and eQTLs studies, results in substantial power advantage and identify functionally relevant genes. We believe that such approach can elucidate novel biological mechanisms underlying complex traits that could also be exploited for drug repositioning.

Copyright © 2018 American Society of Human Genetics. All rights reserved

202

Interrogating horizontal pleiotropy to infer new causal pathways to disease and explain heterogeneity in Mendelian randomization studies. *Y. Cho, P.C. Haycock, T.R. Gaunt, G. Davey Smith, G. Hemani.* MRC Integrative Epidemiology Unit, University of Bristol, Bristol, Bristol, United Kingdom.

Background: Violations in the assumptions of Mendelian randomization (MR) can introduce bias and heterogeneity in the causal estimate. A major source of heterogeneity is horizontal pleiotropy, where an instrumenting single nucleotide polymorphism (SNP) influences the outcome through pathways which bypass the exposure. Those SNPs that induce heterogeneity in MR are typically treated as a nuisance, but they could be a powerful gateway for learning novel pathways to the traits under investigation. **Methods:** Following the advice of William Bateson to “TReasure Your eXceptions”, we developed the MR-TRYX framework. Here, we begin with a single exposure-outcome hypothesis and perform radial inverse variance weighted 2-sample MR analysis. Outliers are then detected using heterogeneity statistics, and we search through the MR-Base database of GWAS summary statistics to identify other (“candidate”) traits that associate with the outliers. We then use multivariable MR analysis to test the extent to which horizontal pleiotropy with the candidate trait can explain the heterogeneity identified in the original exposure-outcome analysis. In doing so, MR-TRYX identifies novel traits influencing the outcome, and accounts for some of the heterogeneity in the original exposure-outcome analysis. **Results:** Through simulations we showed that the use of the adjusted SNP effects in MR can improve power in comparison to alternative approaches for dealing with heterogeneity. In an applied MR analysis of the association between systolic blood pressure and coronary heart disease (CHD), we found that seven of the 157 instrumenting SNPs contribute substantially towards heterogeneity and were outliers in the MR analysis. These outliers associated with 71 candidate traits in the MR-Base database which were then tested for association with the outcome. We identified some established pathways to CHD including blood cholesterol levels, adiposity (e.g. birth weight, and hip circumstance) and height. But we additionally found that pain related phenotypes (headache and self-reported intake of ibuprofen) exhibited putative protective effects on risk of CHD. After accounting for these pathways, the heterogeneity in the analysis of SBP on CHD halved. **Conclusion:** We show that incorporating broad phenotypic information to model horizontal pleiotropy in MR analysis can improve power through reducing heterogeneity and build a more detailed impression of the causal influences on complex traits.

203

Comparative genetic architectures of schizophrenia in East Asian and European populations. *H. Huang[1,2,3], M. Lam[4,5], C. Chen[1,2,3], S. Qin[6,7], P. Sham[8], N. Iwata[9], K.S. Hong[10], S.G. Schwab[11], W. Yue[12], M. Tsuang[13], J.J. Liu[5], X. Ma[14,15,16], R.S. Kahn[17], Y. Shi[6,18,19], Psychiatric Genomics Consortium - East Asia workgroup.* 1) Massachusetts General Hospital, Boston, MA.Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, 02114, USA; 2) Department of Medicine, Harvard Medical School, Boston, MA, 02115, USA; 3) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, 02142, USA; 4) Institute of Mental Health, Singapore, Singapore, Singapore, 539747, Singapore; 5) Human Genetics 2, Genome Institute of Singapore, Singapore, Singapore, 138672, Singapore. Research Division,; 6) Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education) and the Collaborative Innovation Center for Brain Science, Shanghai Jiao Tong University, Shanghai, Shanghai, 200030, China; 7) Collaborative Innovation Center, Jining Medical University, Jining, Shandong , 272067, China; 8) State Key Laboratory of Brain and Cognitive Sciences, Centre for Genomic Sciences and Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hongkong, HK China; 9) Department of Psychiatry, Fujita Health University School of Medicine, Nagoya, Aichi, 470-1192, Japan; 10) Department of Psychiatry, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, Seoul, O6351, Korea; 11) Centre for Medical and Molecular Bioscience, Illawarra Health and Medical Research Institute, Faculty of Science, Medicine and Health, The University of Wollongong, Wollongong, NSW, 2522, Australia; 12) National Clinical Research Center for Mental Disorders & Key Laboratory of Mental Health, Ministry of Health (Peking University), Peking University Sixth Hospital (Institute of Mental Health), Beijing, Beijing, 100191, China; 13) Psychiatry, University of California, San Diego, La Jolla, CA, 32093, USA; 14) Clinical Research Center for Mental Disease of Shaanxi Province, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, 710061, China; 15) Brain Science Research Center, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, 710061, China; 16) Department of Psychiatry, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi Province, 710061, China; 17) Department of Psychiatry and Behavioral Health System, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA; 18) The Biomedical Sciences Institute of Qingdao University, Qingdao Branch of SJTU Bio-X Institutes & the Affiliated Hospital of Qingdao University, Qingdao, Qingdao, 266003, China; 19) Department of Psychiatry, the First Teaching Hospital of Xinjiang Medical University, Urumqi, Xinjiang, 830054, China.

Schizophrenia is a severe psychiatric disorder with a lifetime risk of approximately 1% worldwide. Most large-scale schizophrenia genetic studies have studied individuals of primarily European ancestry, severely limiting the completeness of the knowledgebase attainable from genetics, as well as its scientific utility and applicability to most of the world’s populations. To address this gap in scientific knowledge while advancing global mental health equity, it is imperative to include all diverse populations across the worldwith sufficient sample size in psychiatric research. In this study, we assembled a schizophrenia cohort of East Asian ancestry with landmark sample size of 22,778 cases and 35,362 controls, including samples from Singapore, Japan, Indonesia, Korea, Hong Kong, Taiwan, and mainland China.We identified 21 schizophrenia associations in the East Asian ancestry. Over the genome, we found genetic effects are remarkably consistent across East Asian and European ancestries, indicating for the first time that the genetic basis of schizophrenia and its biology are broadly shared across these world populations. Based on the premise, a fixed-effect meta-analysis across individuals from East Asian and European ancestries revealed 208 genome-wide significant schizophrenia associations in 176 genetic loci (53 novel). Enabled by this sample size and the consistency of schizophrenia genetic effects across populations, we designed a novel trans-ancestry fine-mapping algorithm. This algorithm more precisely isolated schizophrenia causal alleles for 70% schizophrenia associations: an improvement essential for functional mapping of schizophrenia genetic associations, computationally and experimentally. We also demonstrated that despite of a cross-population genetic correlation indistinguishable from 1, polygenic risk models trained in one population have reduced performance in the other due to differences in allele frequency distributions and LD structures. This highlights the importance of including all major ancestral groups in genetics studies both as a strategy to improve power to find disease associations and to ensure the findings have maximum relevance for all populations.

Copyright © 2018 American Society of Human Genetics. All rights reserved

204

Exome sequencing of 23,851 cases implicates novel risk genes and provides insights into the genetic architecture of schizophrenia. *T. Singh[1,2], B.M. Neale[1,2,3], M.J. Daly[1,2,3] on behalf of the SCHEMA consortium.* 1) Analytic and Translational Genetics Unit, Center of Genomic Medicine, Massachusetts General Hospital, Boston, MA; 2) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA.

Schizophrenia, a debilitating psychiatric disorder, has a substantial genetic component with common intergenic and rare coding variants contributing to risk. Despite the discovery of hundreds of common risk loci, only a handful of associations have resulted in validated functional variants that pinpoint novel biology underlying disease pathogenesis. This central challenge is shared with most complex polygenic disorders. To address this shortcoming, sequencing studies of rare coding variants can complement existing approaches by pinpointing likely causal genes overlapping common risk loci, and complete the allelic spectrum in disease genes. However, success to this end has been hampered by power limitations. The Schizophrenia Exome Sequencing Meta-Analysis (SCHEMA) Consortium is a global effort to analyze whole-exome and genome sequencing data to advance gene discovery. We have sequenced 23,851 cases and 50,996 controls, which include individuals of European, Latin American, East Asian, Ashkenazi Jewish, and African American ancestry. We performed comprehensive quality control steps on all data jointly, with consideration of coverage differences between capture technologies. We similarly processed an additional 56,100 non-psychiatric samples from gnomAD for use as population controls. Our expanded data set consists of 23,851 cases and over 100,000 controls, one of the largest sequencing analyses to date. We first implicate protein-truncating variants (PTVs) in two novel genes, *TRIO* and *HERC1*, as conferring substantial risk for schizophrenia, and replicate a known association in *SETD1A*. After adding external controls, we identify additional novel genes at exome-wide significance, including NMDA receptor subunit *GRIN2A*, a target of psychoactive drugs. We discuss signs of convergence with common schizophrenia loci and risk genes for neurodevelopmental disorders, showing an allelic series in *GRIN2A* and an association of *de novo* mutations in *SETD1A* and *TRIO* to broader neurodevelopmental phenotypes. After excluding novel risk genes, schizophrenia cases still carry a substantial excess of rare PTVs, suggesting that more remain to be discovered. Finally, we present an online browser that displays variant- and gene-based results. In summary, analyses of whole-exomes complement those of common variants in expanding our understanding of schizophrenia, and the combined approach can serve as a roadmap for inferring the biology of disease in other disorders.

205

Cell type-specific alternation in schizophrenia and bipolar disorder. *R. Dai[1], L. Chen[2], S. Liu[1], Y. Chen[1], J. Dai[1], G. Yu[2], Y. Wang[2], C. Chen[1], C. Liu[1,3].* 1) Central South University, Changsha, China; 2) Virginia Polytechnic Institute and State University, VA, USA; 3) SUNY Upstate Medical University, NY, USA.

Background: Cell type diversity is the basis of complex brain functions, also psychiatric disorders. However how these cell types are involved in psychiatric pathogenesis is largely unknown. **Methods:** We used bulk tissue RNA-Seq data of 413 prefrontal cortex samples including controls and patients with schizophrenia (SCZ) and bipolar disorder (BD) from PsychENCODE/BrainGVEX project. We applied a novel method called 'sCAM' to deconvolute the expression profile of bulk tissue into the cell-specific expression profile for each sample. With that, we constructed co-expression networks for each cell type and further tested their relevance to the disease using MAGMA and trait association. Integrating genotype data, we identified cell-specific eQTLs and estimated their heritability in SCZ GWAS through LDSR. Further, we applied stratified LDSR to partition the estimated cell-specific heritability explained by different functional categories. Lastly, we identified possible causal genes for SCZ using SMR in a cell-specific manner. **Results:** We generated the cell-specific expression and the proportion of astrocyte, microglia, neuron, and oligodendrocyte for each sample. We found the proportions of microglia and oligodendrocytes increased in patients' brains while that of neuron decreased. After constructing co-expression networks, we identified nine cell-specific modules and they were enriched in pathways such as RNA splicing and mitochondrion, phosphorylation, cytokine, and myelination. All these nine modules showed disease-related alternations or enrichment of GWAS signal. We identified over 12 million cell-type related eQTLs, in which only ~5% were shared across cell types and ~62% were cell-specific. Using LDSR, we observed SNPs of oligodendrocyte-eQTLs explained the most heritability for schizophrenia GWAS summary statistics ($h^2 = 0.6$). Furthermore, we found in different cell types, the genetic heritability can be explained by SNPs in different functional categories, including H3K27ac, H3K4me3, and H3K4me1 region etc. Using SMR, we identified 22 casual genes for schizophrenia in a cell-specific manner. For example, *ZKSCAN3* in neuron etc. **Conclusion:** We extracted cell type-specific expression from bulk brain tissues and examined the cell-specific alternation in psychiatric disorders regarding cell composition, gene expression, genetic regulation and their functional variation. These findings provide new insights into cell-specific molecular alterations in SCZ and BD.

Copyright © 2018 American Society of Human Genetics. All rights reserved

206

Genome-wide characterization of risk genes reveals spatiotemporal heterogeneity of schizophrenia. *Y. Ji[1,2], Q. Wang[1,3], R. Chen[1,3], Q. Wei[1,3], H. Yang[1,3], B. Li[1,3].* 1) Vanderbilt Genetics Institute, Nashville, TN; 2) Human Genetics Graduate Program, Vanderbilt, Nashville, TN; 3) The Department of Molecular Physiology and Biophysics, Vanderbilt, Nashville, TN.

GWAS of Schizophrenia (SCZ) have identified more than 100 significant loci but the complex disease mechanism remains unclear. Previous studies have shown that many variants influence disease risk through perturbing gene expression. In this study we aim to leverage SCZ GWAS findings and transcriptome profiles to characterize risk genes and underlying spatiotemporal patterns of SCZ. Our approach involves two major steps. First, we identified highly confident “seed” risk genes from significant SCZ GWAS loci through integrating multiple lines of evidence. Second, we used identified “seed” genes to build a random forest model for risk gene prediction based on transcriptome profiles. Specifically, we used transcriptome datasets from Gtex, which includes a wide range of tissues, and data from Brainspan, which spans across different developmental stages, to reflect spatiotemporal patterns of genes. We first evaluated our model using cross-validation and the model achieved AUROC of 0.73 and 0.79 from Gtex and Brainspan respectively. We further confirmed the top ranked predicted genes (Gpred & Bpred to denote Gtex and Brainpan predictions) using independent genetic evidence. We assessed the heritability explained by the top ranked genes through stratified LD score regression using SCZ GWAS (N=150064). We found that top ranked genes are significantly enriched for SNP-based heritability, e.g. top 500 Gpred genes with 4.64x enrichment (p=0.0031) and top 500 Bpred genes with 5.45x enrichment (p=0.0020). We further evaluated the top genes using rare variants from published sequencing studies (N=11080) and found that top Gpred and Bpred genes are significantly enriched for missense and stop gained ultra rare damaging variants in SCZ cases. Then we investigated the spatial-temporal patterns using top Gpred and Bpred transcriptome profiles. Both top 500 Gpred and Bpred genes are specifically expressed in brain cortex, although the majority of genes do not overlap. Distinctly, Bpred genes capture patterns in prenatal stage while Gpred genes are mostly distinct in adulthood. To further understand the biological processes involved, we used functional enrichment: Gpred genes show strongest enrichment in synaptic signaling (p=4,73e-26) while Bpred genes have strongest signal in neuron projection development (p=2.15e-14). Altogether, our analysis revealed spatial-temporal distinctions of risk genes and the heterogeneity of SCZ through studying large-scale transcriptome profiles. .

207

Differential histone modifications in 250 schizophrenia cases and 330 controls. *K. Girdhar[1], G.E. Hoffman[1], Y. Jiang[2], L. Brown[2], O. Devillers[2], YC. Wang[1], H. Shah[1], E. Zharovsky[2], R. Jacobov[2], J. Wiseman[2], E. Flatow[2], R. Park[2], J.S. Johnson[2], B.S. Kassim[2], P. Sklar[1,2], P. Roussos[1,2], S. Akbarian[1,2], PsychENCODE Consortium, CommonMind Consortium.* 1) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Department of Psychiatry and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY.

Genomic dysregulation is likely to contribute to neuronal dysfunction in prefrontal cortex (PFC) and other brain regions affected in schizophrenia (SCZ), but genome-scale mapping of neuronal transcriptomes and epigenomes has not been conducted in larger cohorts. We have shown recently that the genomic landscape of open chromatin-associated histone modifications, including histone H3-acetyl-lysine 27 (H3K27ac), show in neurons significant associations with the genetic risk architecture of schizophrenia. Here we present one of the largest neuroepigenomic analysis conducted in schizophrenia postmortem brain. More specifically, we conducted chromatin immunoprecipitation sequencing (ChIP-seq) on fluorescence-activated cell sorted (FACS) nuclei from dorsolateral PFC of 149 postmortem postmortem samples from SCZ and 146 matched control brains for 2 histone modification markers: H3K4me3 (promoters) and H3K27ac (enhancers). In addition to this, we generated H3K4me3 (H3K27ac) tagged nucleosomal DNA from PFC tissue homogenate from 101 SCZ cases and 184 matched controls. The total number of samples for both histone modifications exceeds N=880 samples and has generated approximately 95 billion reads representing epigenome regulatory sequences in the PFC of cases with SCZ and controls. To further extend the analyses, genotyping for the entire cohort of 580 brains was performed. Integrating H3K27ac ChIP-seq with genotyping data, we identified thousands of new histone quantitative trait loci (hQTLs) and various neuron-specific gene categories affected by dysregulated histone acetylation in PFC neurons from subjects with schizophrenia. We performed integration of these hQTLS and expression-QTLs on same individuals obtained from CommonMind consortium with Psychgenomics consortium (PGC2) to uncover the regulatory mechanism of schizophrenia associated risk loci. Our dataset will provide unique insights into non-coding variants associated with neuronal dysfunction in schizophrenia. Our PsychENCODE sponsored resource highlights the critical role of cell-type specific signatures at regulatory and disease-associated non-coding sequences in the human frontal lobe, and provides to date one of the largest histone methylation and acetylation ChIP-seq datasets generated from prefrontal cortex of 250 schizophrenia cases and 330 controls. Supported by NIMH U01MH103392.

Copyright © 2018 American Society of Human Genetics. All rights reserved

208

CRISPR-mediated multiplex epigenomic perturbation of functional noncoding sequences of schizophrenia in hiPSC models. *S. Zhang[1,2], H. Zhang[1], M. Streit[1], J. Shi[3], AR. Sanders[1,2], Z. Pang[4], PV. Gejman[1,2], X. He[5], J. Duan[1,2].* 1) Center for Psychiatric Genetics, NorthShore University HealthSystem, Evanston, IL 60201, USA; 2) Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL 60637, USA; 3) Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA; 4) Department of Neuroscience and Cell Biology and Child Health Institute of New Jersey, Rutgers University, New Brunswick, NJ 08901, USA; 5) Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA.

While genome-wide association studies (GWAS) on schizophrenia (SZ) have identified more than 100 risk loci, functional interpretation of noncoding GWAS SNPs and their regulated genes remains challenging. Our recent proof-of-concept study of open chromatin regions (OCRs) in human induced pluripotent stem cell (hiPSC) models suggested OCRs could be used to prioritise putatively functional SZ GWAS risk variants. Here, using allele-specific open chromatin (ASoC, i.e., allelic difference of chromatin accessibility of a heterozygous SNP in the same sample) as a functional readout of regulatory variants, we systematically identified functional non-coding SZ risk variants. We further validated their perturbed genes by multiplex CRISPR epigenomic editing of the putative functional non-coding sequences in the hiPSC model. We first carried out global OCR profiling by the assay for transposase-accessible chromatin sequencing (ATAC-seq) in hiPSC-differentiated neural progenitor cells (NPCs), cortical glutamatergic neurons (CNs), GABAergic neurons, and dopaminergic neurons of 20 individuals. We found that ASoC variants are prevalent and cell type-specific. ASoC variants in CNs showed the strongest enrichment (~7 fold) of SZ GWAS SNPs. We next used CRISPR/dCas9-mediated epigenome editing (i.e., transcriptional inhibition by dCas9-fused KRAB) in NPCs, targeting the sequences flanking the strongest ASoC variant, which was adjacent to *VPS45*. Both qPCR and RNA-Seq results support a high level of transcriptional repression (60~80%) of *VPS45* upon CRISPR epigenome editing, which was accompanied by expression changes of genes with the similar molecular functions of *VPS45*, such as ER lumen and vesicle. To further systematically identify genes regulated by all the SZ variants showing ASoC, we then performed a multiplexed CRISPR epigenome editing assayed by 10x Genomics single-cell RNA-Seq (scRNA-Seq) in hiPSC-derived NPCs. To minimise potential experimental variations when assaying multiple individuals, we have shown the feasibility of co-culturing of neuronal cells of different individuals by de-duplexing an arbitrary mixture of datasets of HEK293 cells and brain organoids. We are currently analysing the scRNA-seq data of > 8,000 cells and will present the results at the meeting. Our experiments suggest a novel strategy for identifying and validating the putative functional noncoding SZ GWAS risk variants and the effector genes *en masse* at single-cell resolution in hiPSCs.

209

Local and global chromatin interactions are altered by large genomic deletions associated with human brain development. *C. Purmann[1,2], X. Zhang[1,2], Y. Zhang[3], X. Zhu[1,2], M.S. Haney[1,2], T. Ward[1,2], J. Yao[4], S.M. Weissman[5], A.E. Urban[1,2].* 1) Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA; 2) Department of Genetics, Stanford University School of Medicine, Stanford, CA; 3) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 4) Department of Cell Biology, Yale University School of Medicine, New Haven, CT; 5) Department of Genetics, Yale University, New Haven, CT.

Large copy number variants (CNVs) in the human genome are strongly associated with common neurodevelopmental, neuropsychiatric disorders such as schizophrenia and autism. Using Hi-C analysis of long-range chromosome interactions, including a novel approach for haplotype-specific Hi-C analysis, and ChIP-Seq analysis of regulatory histone marks, we studied the epigenomic effects of the prominent heterozygous large deletion CNV on chromosome 22q11.2, with replication analyses for the CNV on 1q21.1 [BioRxiv 182451]. There are local and global gene expression changes as well as pronounced and multilayered effects on chromatin states, chromosome folding and topological domains of the chromatin, that emanate from the large CNV locus. Regulatory histone marks are altered in the deletion flanking regions, in opposing directions for activating and repressing marks. Histone marks are changed along chromosome 22q and genome wide. Chromosome interaction patterns are weakened within the deletion boundaries and strengthened between the deletion flanking regions. The long-range folding contacts between the telomeric end of chromosome 22q and the distal deletion-flanking region are increased. Using haplotype specific Hi-C analysis we determined that on the chromosome 22q with deletion the topological domain spanning the CNV boundaries is deleted in its entirety while neighboring domains interact more intensely with each other. Finally, there is a widespread and complex effect on chromosome interactions genome-wide, i.e. involving all other autosomes, with some of this effect tied to the deletion region on 22q11.2. These findings suggest novel principles of how such large genomic deletions can alter nuclear organization and affect genomic molecular activity.

Copyright © 2018 American Society of Human Genetics. All rights reserved

210

Human-specific NOTCH2NL genes affect notch signaling and cortical neurogenesis. *F.M.J. Jacobs[12], G.A. Lodewijk[2], M. Mooring[1], C.M. Bosworth[1], A.D. Ewing[1], G.L. Mantalas[3], A.M. Novak[1], A. van den Bout[2], A. Bishara[4], J.L. Rosenkrantz[5], J. Lorig-Roach[1], A.R. Field[3], M. Haeussler[1], A. Bhaduri[6], T.J. Nowakowski[6], A.A. Pollen[6], M.L. Dougherty[7], X. Nuttle[8], M.C. Addor[9], S. Zwolinski[10], S. Katzman[1], A. Kriegstein[6], E.E. Eichler[11], S.R. Salama[5], I. Fiddes[1], D. Haussler[13].* 1) UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA; 2) University of Amsterdam, Swammerdam Institute for Life Sciences, Amsterdam, the Netherlands; 3) UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA; Molecular, Cell and Developmental Biology Department, UC Santa Cruz, Santa Cruz, CA, USA; 4) Department of Computer Science, Stanford University, Stanford, CA, USA; 5) UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA; Howard Hughes Medical Institute, UC Santa Cruz, Santa Cruz, CA, USA; 6) Department of Neurology and the Eli and Edythe Broad Center for Regeneration Medicine and Stem Cell Research at the University of California, San Francisco, San Francisco, CA, USA; 7) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA; 8) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; Department of Neurology, Harvard Medical School, Boston, MA, USA; Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA; 9) Service de génétique médicale, Lausanne, Switzerland; 10) Department of Cytogenetics, Northern Genetics Service, Institute of Genetic Medicine, Newcastle upon Tyne, UK; 11) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA; Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA; 12) University of Amsterdam, Swammerdam Institute for Life Sciences, Amsterdam, the Netherlands. UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA; Electronic address: F.M.J.Jacobs@uva.nl; 13) UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA; Howard Hughes Medical Institute, UC Santa Cruz, Santa Cruz, CA, USA. Electronic address: haussler@ucsc.edu.

Genetic changes causing dramatic brain size expansion in human evolution have remained elusive. Notch signaling is essential for radial glia stem cell proliferation and a determinant of neuronal number in the mammalian cortex. In this study, we describe the evolutionary history, function and potential association to human neurodevelopmental disorders of four paralogs of human-specific NOTCH2NL genes. NOTCH2NL genes emerged in hominin genomes after a complex series of partial segmental duplications involving NOTCH2, followed by gene-conversion events which happened after the split with chimpanzee. NOTCH2NL genes encode a secreted protein containing only the six N-terminal-most EGF-like domains of NOTCH2. Three of the NOTCH2NL paralogs are highly expressed in ventricular and outer radial glia cells. We found that NOTCH2NL directly interacts with NOTCH receptors on the extra-cellular domain, and functional analysis reveals different paralogs of NOTCH2NL have varying potencies to enhance Notch signaling. Consistent with a role in Notch signaling, NOTCH2NL ectopic expression delays differentiation of neuronal progenitors, while deletion of NOTCH2NL genes by CRISPR/Cas9 in human cortical organoids accelerates neuronal differentiation. Suggestive of an important role for NOTCH2NL genes during normal human cortical development in vivo, we found that NOTCH2NL genes provide the breakpoints in typical cases of 1q21.1 distal deletion/duplication syndrome, where duplications are associated with macrocephaly and autism, and deletions with microcephaly and schizophrenia. Indeed, our analysis in patient genomes indicates that deletions and duplications in the 1q21.1 locus are recurrently linked to differences in NOTCH2NL copy number. Our data suggest that the creation of NOTCH2NL genes during hominin evolution may have contributed to the rapid evolution of the larger hominin neocortex. Ironically, this may have been at the expense of genomic stability at the 1q21.1 locus, leading to an increased susceptibility of humans to 1q21.1 associated neurodevelopmental disorders.

211

Tissue-specific molecular signatures associated with 16p11.2 reciprocal genomic disorder. *P. Razaz[1,2,3], S. Erdin[1,3], D.J. Tai[1,2,3], T. Aneichyk[1,2,3], T. Arbogast[4], A. Ragavendran[1], K. Mohajeri[1,2,3], A. Stortchevoi[1,2], B.B. Currall[1,2,3], C.E.F. de Esch[1,2,3], E. Morini[1,2], W. Ma[1,2], R.J. Kelleher[1,2], C. Golzio[4,5], N. Katsanis[4], J.F. Gusella[1,2,3,6], M.E. Talkowski[1,2,3,6].* 1) Center for Genomic Medicine and Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA; 2) Department of Neurology, Harvard Medical School, Boston, Massachusetts, USA; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 4) Center for Human Disease Modeling, Duke University Medical Center, 300 N Duke Street, Durham, NC, USA; 5) Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch, France; 6) Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.

Reciprocal genomic disorders (RGDs) are a recurrent class of copy number variants (CNVs) that collectively comprise a major contributor to neurodevelopmental disorders (NDD) and altered anthropometric traits. Dissecting the specific mechanisms by which they confer NDD risk represents a biological “black box”—as their non-allelic homologous recombination (NAHR) genomic-architecture alters precisely the same multi-genic segments. We systematically dissected the functional networks associated with 16p11.2 RGD from tissue-specific transcriptome analyses of 102 mice with reciprocal CNV of the syntenic 7qF3 region across cortex, striatum, and cerebellum, as well as liver, white and brown adipose tissues in a subset of 16 mice (n=354 samples). We integrated these data with brain tissues from a *Kctd13* mouse model (a putative driver of 16p11.2 neuroanatomical phenotypes, n=50), and CRISPR-engineered, isogenic 16p11.2 iPSC-derived NSCs (n=31) and induced neurons (n=26). The greatest magnitude of local CNV effects was observed across brain regions in comparison to non-brain tissues (cortex 7qF3 region average p = 8.80E-35; non-brain p = 0.0013), reflecting the ~3x higher differences in basal expression. Global transcriptome profiling of differentially expressed genes (DEGs) across brain tissues revealed convergence on NDD-associated pathways highly enriched for constrained genes (ExAC pLI≥0.9), ASD-associated genes from recent exome sequencing studies, as well as early- and late-development co-expression networks derived from BrainSpan. These effects were recapitulated in 16p11.2 RGD neuronal models in a manner reflective of the stage of development that they represent: NSC DEGs were highly enriched for early development chromatin remodeling pathways, whilst iNs were enriched for later development synaptic transmission terms. Effects specific to brain regions were also observed, with cerebellum consistently yielding more DEGs for duplications than deletions. Coexpression network analyses of brain tissues and iPSCs isolated a recurrent module of 16p11.2 genes, as well as constrained networks highly enriched for NDD genes, BrainSpan-derived pathways, CHD8 targets, and neurological phenotypes. DEGs from the *Kctd13* mouse coalesced into these networks, suggesting overlap in altered transcriptional networks between full length CNV and deletion of *KCTD13* alone. These analyses identify a tissue-specific impact of 16p11.2 RGD on pathways critical for human neurodevelopment.

Copyright © 2018 American Society of Human Genetics. All rights reserved

212

Transcriptome network modeling in cerebral organoids identifies critical cell types in autism-associated copy number variants. *E.T. Lim[1,2], Y. Chan[1,2], G.M. Church[1,2].* 1) Harvard Medical School, Boston, MA; 2) Wyss Institute for Biologically Inspired Engineering, Boston, MA.

De novo and rare copy number variants (CNVs) have been robustly associated with complex neuropsychiatric diseases such as autism spectrum disorders (ASD). Recent studies in animal models have provided evidence for critical cell types and driver genes that are perturbed in these CNV loci. On the other hand, cerebral organoids demonstrate great promise for modeling biological processes in human neurological diseases, and discoveries from cerebral organoids can help complement discoveries from animal models. There are several challenges in achieving robust and reproducible phenotyping of complex neuropsychiatric diseases using cerebral organoids. Here, we sequence RNA from 1,420 cerebral organoids from 26 individual donors, and systematically evaluate the sources of variability in the data. We describe two statistical methods (*CellScore* and *GeneScore*) to prioritize critical cell types and candidate driver genes for two CNV loci associated with ASD (15q11-13 duplication and 16p11.2 deletions). We identify dopaminergic neurons and proliferative progenitor cells as critical cell types that are perturbed in 15q11-13 duplications and 16p11.2 deletions respectively, and discover candidate driver genes that were previously reported using animal models, as well as novel candidate driver genes that were previously unreported. Our work presents an unbiased quantitative framework for phenotyping cerebral organoids to facilitate disease modeling of risk loci in complex neurological diseases, and our approach can be extended to quantitative phenotyping of organoids for other tissues in complex diseases.

213

Functional assessment of *Drosophila* and *Xenopus* models shows how 16p12.1 deletion leads to neurodevelopmental defects and is modulated by rare variants in the genetic background. *L. Pizzo[1], M. Lasser[2], P. Ingraham[1], D. Mayanglambam[1], E. Huber[1], M. Jensen[1], A. Krishnan[3], M. Kelly[1], A. Weiner[1], M. Rolls[1], L. Lowery[2], S. Girirajan[1].* 1) Pennsylvania State University, University Park, PA; 2) Boston College, Chestnut Hill, MA; 3) Michigan State University, East Lansing, MI.

We recently identified a deletion on chromosome 16p12.1 that is mostly inherited and associated with variable expressivity, where severely affected probands carry an excess of rare pathogenic hits elsewhere in the genome compared to mildly affected carrier parents. We hypothesized that the 16p12.1 deletion sensitizes the genome for neuropsychiatric disease, but other hits in the genetic background modulate the ultimate phenotypic trajectory. To test this model, we examined 25 developmental, neurological, and cellular phenotypes of 16p12.1 homologs using *Drosophila melanogaster* and *Xenopus leavis*. Neuronal knockdown of 16p12.1 genes in *Drosophila* showed developmental delay, early lethality and decreased brain size for *POLR3E* ($p<0.002$) and *C16orf52* ($p<0.04$), seizures phenotypes for *UQCRC2* ($p=0.006$) and decreased complexity of dendritic arbors for *C16orf52* ($p<0.0001$). These results were validated by morpholino knockdown in *Xenopus*, with decreased forebrain sizes for *POLR3E* and *C16orf52* ($p<0.01$) and axon outgrowth defects for *EEF2K* and *C16orf52* knockdowns ($p<0.01$). We further tested 450 pairwise interactions between 16p12.1 homologs with 100 core developmental genes, using the fly eye as a model. We identified additive interactions between 16p12.1 homologs and core developmental genes (75%), with other synergistic (20%) and suppressive interactions (5%). For example, we observed synergistic interactions between *POLR3E* and the autism-associated *CHD8*, as well as *C16orf52* with the intellectual-disability-associated *SETD5* ($p<0.001$). Similarly, RNA sequencing of *Drosophila* brain samples with knockdown of 16p12.1 homologs showed altered expression of multiple developmental genes (such as *SCN8A*, *DSCAM*, *CHRNA7*, *ASPM* and *LAMC3*) and compromised pathways associated with 16p12.1-deletion-phenotypes, such as response to heat (*UQCRC2*, $p<0.001$) and muscle contraction processes (*POLR3E*, $p=0.013$). Further, proliferation and apoptosis defects ($p=0.01$) in the developing fly brain suggested specific cellular mechanisms through which the effects of 16p12.1 genes could be potentially modulated. Analysis of a human brain-specific gene interaction network showed increased genome-wide connectivity of 16p12.1 genes ($p<0.05$), further evidencing how other variants could modulate the effects of individual genes. Our results show that multiple 16p12.1 genes sensitize the genome to disease, and their interactions with other genetic variants ultimately contribute to a severe manifestation.

Copyright © 2018 American Society of Human Genetics. All rights reserved

214

Detection of copy number variations in epilepsy using exome data. *N. Tsuchida[1,2], M. Nakashima[1,3], M. Kato[4,5], Y. Uchiyama[1], E. Imagawa[1], T. Mizuguchi[1], T. Atsushi[1], N. Miyake[1], H. Nakajima[2], H. Saitsu[3], S. Miyatake[1], N. Matsumoto[1].* 1) Yokohama City University Graduate School of Medicine, Yokohama, Japan; 2) Department of Stem Cell and Immune Regulation, Yokohama City University Graduate School of Medicine, Yokohama, Japan; 3) Department of Biochemistry, Hamamatsu University School of Medicine, Hamamatsu, Japan; 4) Department of Pediatrics, Yamagata University Faculty of Medicine, Yamagata, Japan; 5) Department of Pediatrics, Showa University School of Medicine, Tokyo, Japan.

Epilepsies are common neurological disorders and genetic factors contribute to their pathogenesis. Copy number variations (CNVs) are increasingly recognized as an important etiology of many human diseases including epilepsy. Whole exome sequencing (WES) is becoming a standard tool for detecting pathogenic mutations and has recently been applied to detecting CNVs. Here, we analyzed 294 families with epilepsy using WES, and focused on 168 families with no causative single nucleotide variants (SNVs) in known epilepsy-associated genes to further validate CNVs using two different CNV detection tools using WES data. We confirmed 18 pathogenic CNVs, and two deletions and two duplications at chr15q11.2 of clinically unknown significance. Of note, we were able to identify small CNVs less than 10 kb in size, which might be difficult to detect by conventional microarray. We revealed two cases with pathogenic CNVs that one of the two CNV detection tools failed to find, suggesting that using different CNV tools is recommended to increase diagnostic yield. Considering a relatively high discovery rate of CNVs (18 out of 168 families, 10.7%) and successful detection of CNV with <10 kb in size, CNV detection by WES may be able to surrogate, or at least complement, conventional microarray analysis.

215

GWAS and pathway analyses in childhood and adult-onset asthma in UK Biobank identifies novel loci, effector genes and pathways that highlight distinct aetiology between subtypes. *K. Song[1], T. Lee[1], J. Hoffman[1], R. Scott[2], L. Yerges-Armstrong[1], S. Ghosh[1].* 1) Target Sciences, GlaxoSmithKline, Collegeville, PA; 2) Target Sciences, GlaxoSmithKline, Stevenage, UK.

Genome-wide association studies (GWAS) have been widely used to identify novel genetic associations for both childhood and adult-onset asthma. However, using these results to better understand asthma pathogenesis remains a formidable challenge. We performed two genome-wide association studies of asthma in UK Biobank unrelated European participants self-reporting either (i) pediatric age of onset (≤16 years of age; n= 11,443) or (ii) adult-onset asthma (onset >16 years of age; n=26,756). We used the different 165,158 controls for each study. Fine-mapping and colocalization were done to refine the association signals and obtain a more objective insight into distinct signals between both studies, respectively. Next, we identified effector genes by joint modeling of GWASs and expression/protein quantitative trait loci (e/pQTL) in 125 tissues from 26 studies using an approximate colocalization approach called PICCOLO. Pathway analysis comparing results from the two age of onset groups was done using Ingenuity Pathway analysis (IPA) by the fold change of predicted individual transcript levels for the effector genes in childhood asthma and adult-onset asthma groups using PrediXcan. We identified 62 and 30 genome-wide significant loci ($p < 5.0E-08$) for childhood and adult-onset asthma, respectively. Fourteen loci were present in both age of onset groups. We replicated known childhood asthma loci including *C11orf30* and *GSDMB*/*ORMDL3* and found novel loci in both studies. Using PICCOLO, 59 and 25 effector genes colocalized with childhood and adult-onset association signals, 15 of these genes present in both age of onset groups. Pathway analyses identified stronger enrichment of immune response in childhood asthma (eg. Hematopoiesis of mononuclear leukocytes(p-value in childhood (Pc)=6.1E-09, p-value in adult (Pa)=ns), whereas tissue/cell growth were significant in both groups but more strongly associated in adult-onset asthma (eg. hyperplasia of epithelial(Pc=1.1E-03, Pa=1.3E-06) and goblet(Pc=2.0E-03, Pa=3.8E-06) tissue/cells). In summary, utilizing self-report age of onset data, we have identified different genetic associations in childhood and adult-onset asthma. Pathway analysis of predicted effector gene expression identified by colocalization revealed distinct genetic mechanisms providing support that childhood and adult-onset asthma may have different underlying mechanisms.

Copyright © 2018 American Society of Human Genetics. All rights reserved

216

Integration of GWAS summary statistics and miRNA-target gene network with tissue-specific miRNA expression profile identified novel pathogenesis of complex human traits implicated in tissue specificity.

S. Sakaue[1,2,3], K. Yamamoto[2], Y. Okada[1,2,4]. 1) Graduate School of Medicine, Osaka University, Osaka, Japan; 2) RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; 3) Graduate School of Medicine, the University of Tokyo, Tokyo, Japan; 4) Immunology Frontier Research Center (WPI-IFReC), Osaka University, Osaka, Japan.

MicroRNA (miRNA) modulates the post-transcriptional regulation of specific target genes and is related to biology of complex human traits, but genetic landscape of miRNAs remains largely uncovered. We introduce a method to quantitatively evaluate enrichment of genome-wide association study (GWAS) signals on miRNA-target gene network (MIGWAS). Given the strikingly tissue-specific miRNA expression profiles, this enrichment signal should be emphasized within the disease-relevant tissue. Our approach integrates the high-throughput miRNA expression sequencing data in 180 human cell types with GWAS summary statistics to test whether polygenic signals enrich in miRNA-target gene network and whether they fall within a specific tissue. We applied MIGWAS to 49 GWAS summary statistics (n_{Total} = 3,520,246), and successfully identified biologically relevant tissue in each trait (e.g. immune cells for Lupus and Graves' disease, fat cells for LDL cholesterol, and immune, lung and bone cells for rheumatoid arthritis [RA]). Further, MIGWAS could point the miRNA and target genes as candidate biomarkers of the trait as well as novel associated loci with significant genetic risk. As an illustrative example, we identified potential causal miRNAs associated with RA, which were *in silico* replicated by the novel GWAS meta-analysis adding two RA GWAS of Japanese (3,308 RA cases and 8,357 controls) to the trans-ethnic cohorts (19,234 RA cases and 61,565 controls), and further *in vivo* validated by case-control analysis of differentially expressed miRNAs between RA patients and healthy controls (n = 63). Our result highlighted that miRNA-target gene network contributes to human disease genetics in the context of cell type-specific expressions, and provided a potential for discovering miRNAs as promising biomarkers.

217

A fast linkage method for large GWAS cohorts with related individuals.

G.J.M. Zajac, S.A. Gagliano, G.R. Abecasis. Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI.

Linkage analysis, a class of methods for detecting co-segregation of genomic segments with a trait within families, mapped disease-causing genes for decades before genotyping arrays provided cheap genome-wide coverage to allow GWAS in unrelated individuals. GWAS cohorts often contain related individuals, but the segregation of alleles within families is rarely used because traditional linkage methods are computationally inefficient for GWAS-scale datasets.Here we propose a scalable linkage approach that can handle thousands of quantitative traits in a large GWAS dataset where pedigree structure is unknown prior to analysis. First, we estimate kinship and identical-by-descent segments (IBD) between all pairs of individuals. Next, we estimate variance-components for IBD sharing, kinship, and pure error using cross-product Haseman-Elston regression. The resulting estimates of the variance attributed to IBD sharing at each locus and its standard error are then used to test the hypothesis of genetic linkage.We tested our method in 6,602 individuals with genotypes at 18,754,911 sites from the SardiNIA project. We found that 5,230 individuals (79%) had at least one relative 3rd degree or higher. All individuals shared an IBD segment with another individual. Using these kinship and IBD estimates, we tested 120 inverse-normalized traits for linkage at 4,693 sites (the unique endpoints of all estimated IBD segments).Our linkage analysis identified 11 genome-wide significant loci with a LOD > 3. These overlap 10 of 89 significant SardiNIA GWAS loci for the same traits. IBD sharing near the alpha-globin gene cluster on chromosome 16 explained 24% of the variance in mean cell hemoglobin (LOD = 5.47), our largest effect size. This overlaps a significant GWAS variant in the region (rs141494605) which explained 9.3% of the trait variance. Another notable linkage signal was for LDL cholesterol levels and the *APOE* locus (LOD = 3.34, variance explained = 5.0%). This replicates a previous finding that the missense variant rs7412 in *APOE* reduced LDL levels and accounted for 2.4% of variation in circulating LDL.In summary, we show that our method can successfully perform linkage analysis on 1000s of individuals and accommodate arbitrary pedigree structures. While most of our signals overlapped with known GWAS loci, running linkage analysis at this scale has the added potential of identifying effects that are non-additive or at variants not directly genotyped or sequenced.

Copyright © 2018 American Society of Human Genetics. All rights reserved

218

Genome wide meta-analysis of parent-of-origin effects of asthma, atopy and airway hyperresponsiveness in four cohorts. *A. Eslami[1], L. Akhabir[1], J.M. Vonk[2], A.B. Becker[3], A.L. Kozyrskyj[4], A.J. Sandford[1], G.H. Koppelman[2], C. Laprise[5,6], D. Daley[1].* 1) University of British Columbia, Vancouver, Canada; 2) Groningen Research Institute for Asthma and COPD (GRIAC), University Medical Center Groningen, University of Groningen, Groningen, Netherlands; 3) Department of Pediatrics and Child Health, Faculty of Medicine, University of Manitoba, Winnipeg, MB, Canada; 4) Department of Pediatrics, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada; 5) Université du Québec à Chicoutimi, Saguenay, QC, Canada; 6) Centre intégré universitaire de santé et de services sociaux de Saguenay (CIUSSSS), QC, Canada.

Background: The main genetic effects of the common SNPs identified by Genome-Wide Association Studies (GWAS) do not fully explain the heritability of asthma. Genomic imprinting (parent-of-origin effects) is a potential mechanism which may explain this missing heritability. **Aim**: Identify candidate genomic regions for imprinting in asthma and related phenotypes (atopy and airway hyperresponsiveness (AHR)) by using existing GWAS data from four family-based studies. **Methods:** We used GWAS data from four family-based studies (two parents and one affected child). These studies are: 1) the Canadian Asthma Primary Prevention Study (CAPPS), a high-risk asthma birth cohort, 2) the Study of Asthma Genes and Environment (SAGE), a population-based asthma birth cohort, 3) the Saguenay-Lac-Saint-Jean Québec Familial Collection (SLSJ), a founder population of French-Canadians, and 4) The Dutch Asthma GWAS (DAG), a cohort from the Netherlands. We used a likelihood-based variant of the Transmission Disequilibrium Test. Parent-of-origin effects were analyzed by including parental sex as a modifier in the analysis, which determines whether the asthma risk is modified by the parental origin of the allele. An odds ratio for parent-of-origin effects is determined by dividing maternal odds ratio by paternal odds ratio. Meta-analysis was conducted using the parent-of-origin effects results of SLSJ, DAG, and the joint analysis of the two birth cohorts CAPPS and SAGE (CAPPS/SAGE), weighted by the number of informative transmissions for each study. **Results**: Meta-analysis for asthma, using results of SLSJ (251 trios), DAG (316 trios), and CAPPS/SAGE (141 trios), resulted in 5 independent SNPs with significant parent-of-origin effects with $P \leq 1.49\times10^{-5}$ (suggestive P-value threshold). Meta-analysis for atopy, using results of SLSJ (229 trios), DAG (312 trios) and CAPPS/SAGE (217 trios) resulted in 2 independent SNPs. Meta-analysis for AHR using results of SLSJ (132 trios), DAG (260 trios) and CAPPS/SAGE (219 trios) resulted in 7 independent SNPs. Of the significant results, 11 out of 14 of the SNPs were in or near long non-coding (lnc)RNA genes. **Conclusion**: Meta-analysis of four family-based studies yielded several SNPs with significant parent-of-origin effects ($P \leq 1.49\times10^{-5}$) for asthma and the related phenotypes. These SNPs were mostly located in or near lncRNA genes. LncRNAs are known to be involved in genomic imprinting and gene regulation.

219

Simultaneously modeling host genetics and microbiome composition reveals the heritability and the proportion of variance explained due to the microbiome of immune-related traits. *E.R. Davenport[1], T.D. Spector[2], R.E. Ley[3], A.G. Clark[1].* 1) Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY; 2) Department of Twin Research and Genetic Epidemiology, King's College London, London, UK; 3) Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen, Germany.

Both host genetics and environmental factors determine human immune-related phenotypes and disease, including asthma, allergies, and rheumatic disorders. Although many known environmental exposures contribute to immune disease risk and severity, the extent to which the human microbiome is responsible for variation in these phenotypes remains largely unknown. Additionally, whether gene-by-environment interactions with the microbiome influence these phenotypes remains uncharacterized, as well as which specific genetic variants and microbes play roles in this process. To address these gaps, we examine both the proportion of variation explained by host genetics (PVE-G, or heritability) and gut microbiome composition (PVE-M) in a unified framework for ~30 immune-related phenotypes using the large TwinsUK cohort, for which microbiome, genetic, and phenotypic data are available for 2,500 individuals. First, we examine the relative contributions of PVE-M and PVE-G for each trait using a linear mixed model framework, by incorporating random effects that account for the extent of sharing of both genotypes (genetic relatedness matrix) and microbes (beta diversity matrix) among individuals. We find that PVE-M accounts for up to 44.8% of non-genetic variation in traits such as body mass index (BMI). Second, we explore whether explicitly including gut microbiome composition in genome-wide association study (GWAS) models improves power to detect genetic variants associated with immune phenotypes. To do so, we compare linear mixed models both with or without microbiome random effects, applied to 1.3 million genotyped variants for each phenotype. Including the microbiome in the model does not increase power to detect genetic variants associated with certain phenotypes (including height, asthma, and eczema), but does for other phenotypes (BMI, white blood cell counts, and level of vitamin D in blood). For example, in the case of BMI, including the microbiome term pushes the significance of the top associated SNP to surpass the genome-wide Bonferroni threshold (rs1036286, $P = 8.6 \times 10^{-8}$) as compared to the model without the microbiome term ($P = 1.8 \times 10^{-8}$). Using these data, we have conducted one of the first GWA studies to explicitly model the microbiome, demonstrating that the microbiome is capable of explaining a large proportion of variation due to non-genetic effects and can improve power to detect genetic variants associated with phenotypes.

Copyright © 2018 American Society of Human Genetics. All rights reserved

220

A little data goes a long way: Finding target genes across the GWAS Catalog by colocalizing GWAS and eQTL top hits. *C. Guo[1], M.R. Nelson[1], J. Esparza-Gordillo[2], M.R. Hurle[1], T. Johnson[2], K.B. Sieber[1].* 1) Target Sciences, GlaxoSmithKline, Collegeville, PA; 2) Target Sciences, GlaxoSmithKline, Stevenage, UK.

Identifying effector genes from genome-wide association studies (GWAS) is a key step in translating genetic associations to disease understanding and successful medicines. This task is challenging given that GWAS for complex traits can yield hundreds of loci, each harboring up to dozens of genes. Moreover, most associations cannot be explained by protein coding changes. There are many methods of mapping these non-coding associations to effector genes. One method is testing for colocalization of disease-related genetic associations with those associated with changes in gene expression (i.e. expression quantitative trait loci, eQTLs). However, the utility of this is approach is hampered by the lack of available full summary statistics from most previously published GWAS and eQTL studies. To overcome the need for summary statistics, we developed PICCOLO, a colocalization test of PICS (Probabilistic Identification of Causal SNPs) credible sets. PICCOLO estimates the colocalization of GWAS and eQTL signals using lead SNPs and p-values. Compared to colocalization testing with full summary statistics, we estimate PICCOLO to have moderate sensitivity and high specificity. Using PICCOLO, we estimated the colocalization of all genome-wide significant associations in the EBI GWAS Catalog with eQTLs from GTEx and 28 additional studies. In total, we identified ~26,000 colocalizations within 3,400 associations for 350 traits and 2,500 genes. About 30% of mapped genes were pleiotropic, with 10% mapping to more than four traits. To determine if PICCOLO was identifying putative causal genes, we tested for enrichment of mapped genes within targets of approved drugs. PICCOLO colocalized genes were 3.3-fold enriched for successful drug targets, and limiting the analysis to loci with single gene colocalizations increases the enrichment to 4.4-fold. Additionally, the enrichment for targets of drugs reaching Phase III or approval were about twice those that only reached Phase I. In summary, we present a new method that can identify potential effector genes at GWAS loci without the need for complete summary statistics, thereby greatly expanding the number of GWAS and eQTL datasets that can be tested. These data offer the most comprehensive evidence to date that colocalization testing can help hone in on potential drug targets. As such, we anticipate that this approach will improve the identification of effector genes from the growing wealth of genetic association data.

221

Understanding the molecular basis of a novel *ADPRHL2*-mediated neuropathology. *S.G. Ghosh[1], K. Becker[2], H. Huang[1], T.D. Salazar[1], G. Chai[1], H. Wang[2], G. Haddad[1], M. Karakaya[2], B. Wirth[2], J.M. van Hagen[3], N.I. Wolf[4], R. Maroofian[5], H. Houlden[6], S. Cirak[2], J.G. Gleeson[1].* 1) University of California, San Diego, La Jolla, CA 92093, USA; 2) Institute of Human Genetics, Center for Molecular Medicine, and Center for Rare Diseases, University of Cologne, Cologne, Germany; 3) Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands; 4) Department of Child Neurology, VU University Medical Center, and Amsterdam Neuroscience, Amsterdam, The Netherlands; 5) Molecular and Clinical Sciences Institute, St George's, University of London, Cranmer Terrace, London SW17 0RE, UK; 6) Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK.

ADP-ribosylation, the addition of poly-ADP ribose (PAR) onto proteins, is a response signal to cellular challenges, such as excitotoxicity or oxidative stress. This process is catalyzed by a group of enzymes, referred to as Poly(ADP-ribose) polymerases (PARPs). As accumulation of proteins with this modification results in cell death, its negative regulation restores cellular homeostasis: a process that is mediated by poly-ADP ribose glycohydrolases (PARGs) and ADP-ribosylhydrolase proteins (ARHs). Using genome-wide linkage analysis and exome sequencing or genome sequencing, we identified recessive inactivating mutations in *ADPRHL2,* encoding for ARH3, in six families. Affected individuals exhibited a pediatric-onset neurodegenerative disorder with progressive brain atrophy, developmental regression, and seizures that correlated with periods of stress such as infections. Loss of the *Drosophila* paralogue *parg* showed lethality in response to oxidative challenge that was rescued by human *ADPRHL2,* suggesting functional conservation. Previous studies have identified ARH3 as the only active glycohydrolase present in mitochondria; however, its importance and role in the mitochondria remain understudied. Patient cells lack ARH3 protein, display increased basal and stress-associated PAR levels, and enhanced susceptibility to cell death following insult. PARP inhibitors, already in clinical trials, rescued lethality, suggesting that this class of drugs may be used to treat this lethal disorder. The goal is to describe this clinical condition as a new syndrome cause of neurodegeneration and study oxidative-stress induced mechanisms by which loss of ADPRHL2 promotes cell death both in vitro and in vivo.

Copyright © 2018 American Society of Human Genetics. All rights reserved

222

Argininosuccinate lyase is required for central regulation of catecholamines synthesis. *S. Lerner[1], R. Eilam[2], M. Prigge[3], M. Tsoory[2], Y. Kuperman[2], E. Anderzhanova[4], O. Yizhar[3], A. Chen[3,4], A. Erez[1].* 1) Department of Biological Regulation, Weizmann, Rehovot, Israel; 2) Department of Veterinary Resources, Weizmann, Rehovot, Israel; 3) Department of Neurobiology, Weizmann, Rehovot, Israel; 4) Department of Stress Neurobiology and Neurogenetics, Max Planck Institute of Psychiatry, Munich, Germany.

Argininosuccinic-aciduria (ASA) is a urea cycle disorder (UCD) caused by germ line mutations in the argininosuccinic-lyase (ASL) gene. In spite of diagnosis by newborn screening and early initiation of therapy, and even in the absence of documented hyperammonemia, ASA patients have a yet unexplained higher incidence of neurocognitive abnormalities and hypertension. We previously reported that ASL is essential for the synthesis of nitric oxide (NO) and that NO deficiency in blood vessels of ASA patients contributes to hypertension. Importantly, we demonstrated that the high blood pressure can be normalized with supplementation of NO donors. Interestingly, an ASA patient treated with NO donor for his hypertension, showed a significant cognitive improvement. To dissect a potential role for ASL in the central nervous system, we first assayed its expression in wild-type mouse brains. Surprisingly, we found that ASL is strongly expressed in the locus coeruleus (LC), a brain stem region that is the main source of norepinephrine synthesis and is involved in coordinating both neurobehavioral cognitive circuits and blood pressure. Using a neuroblastoma cell line, we found that ASL knock-down decreases the expression of Tyrosine Hydroxylase (TH), the rate-limiting enzyme in the synthesis of catecholamine (CA). The decrease in TH levels was associated with a concomitant decrease in NO levels and was rescued following treatment with NO-donors. Using a novel conditional knockout model of ASL in the LC ($ASL^{f/f}$;$TH\ Cre^{+/-}$), we were able to confirm that deletion of ASL in the LC decreased the expression of TH. In turn, we also observed reduction of NO levels and in nitrosylation of proteins in the CA synthesis pathway. We found that ASL deficiency in the LC alters CA levels and their turnover rate in the brain. Phenotypically, under stress conditions, $ASL^{f/f}$;$TH\ Cre^{+/-}$ mice had significantly higher blood pressure and hyperactivity. Finally, functional acute slice imaging revealed that following Carbachol induction, NO production by neurons in the LC is significantly decreased in $ASL^{f/f}$;$TH\ Cre^{+/-}$ mice. Our results suggest a metabolic role for ASL in regulating catecholamines secretion in the LC *via* NO levels. Further deciphering the role of ASL in the LC may shed light on both the hypertension and neurobehavioral delays in ASA. More broadly, our study identifies ASL as a new player in the LC, involved in the central regulation of blood pressure, cognition and behavior.

223

Genetic and functional approaches highlight a novel ciliary complex implicated in Joubert syndrome. *J.C. Van De Weghe[1], T.D. Rusterholz[2], B. Latour[3], A. Gomez[1], UW Center for Mendelian Genomics[4], M. Bamshad[1], D. Nickerson[5], R. Roepman[3,6], R. Bachmann-Gagescu[2,7], D. Doherty[1,8].* 1) Pediatrics, Genetic Medicine Division, University of Washington, Seattle, WA; 2) Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland; 3) Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands; 4) University of Washington, Seattle, WA; 5) Department of Genome Sciences, University of Washington, Seattle, WA; 6) Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands; 7) Institute of Medical Genetics, University of Zurich, Zurich, Switzerland; 8) Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA, USA.

Joubert syndrome (JS) is a neurodevelopmental disorder defined by a distinctive brain imaging finding called the “molar tooth sign.” Subsets of individuals with JS have progressive retinal, kidney and liver involvement; therefore, a therapeutic window exists for these issues that arise over decades. Variants in >35 genes cause JS, and all gene products are required for normal primary cilium function, making JS a ciliopathy. Ciliopathies are disorders rooted in aberrant ciliary function that exhibit overlapping clinical features and have a combined prevalence of 1/1000. Many organ systems are affected in ciliopathies, because the primary cilium is a microtubule-based projection found on nearly every cell that functions like an antenna to detect extracellular signals. Despite great progress on defining the genetic causes of JS, how variants in >35 ciliary-related genes cause the disorder remains an open question. Using targeted sequencing of >500 families affected by JS, we recently published *ARMC9* variants as a new cause of JS. Thereafter, we used ARMC9 as bait in protein-protein interaction screens, and identified a novel ciliary complex implicated in JS. This complex is composed of known JS-associated proteins and some not yet linked to any human disorder. We mined our JS exome data and targeted sequencing data for genes encoding ARMC9 complex members, and found individuals with causal variants in one of them, a new gene that was not previously linked to JS. The cognate protein directly binds to ARMC9, and preliminary data in zebrafish indicate that CRISPR/Cas9-induced truncations of both complex members cause decreased numbers of cilia and typical ciliopathy phenotypes, including curved body shape, retinal degeneration and pronephric cysts. Patient and engineered cell lines have decreased post-translational modifications of ciliary microtubules and decreased cilium stability. Most of the JS-proteins function to regulate ciliary protein content at the transition zone, a domain at the base of the cilium that partitions it from the rest of the cell. This work defines a new complex of non-transition zone JS-proteins that may regulate cilium stability, bringing us one step closer to understanding the mechanisms underlying JS that will facilitate developing future targeted therapies.

Copyright © 2018 American Society of Human Genetics. All rights reserved

224

The loss of fragile X mental retardation protein alters the development of human forebrain organoid. *Y. Kang[1], F. Zhang[1], Y. Li[1], Z. Wen[2], P. Jin[1].* 1) Department of Human Genetics, Emory University, atlanta, GA; 2) Department of Cell Biology, Emory University, atlanta, GA.

Fragile X syndrome (FXS) is the most common inherited form of intellectual disability and a leading genetic cause of autism. FXS is caused by the loss of functional fragile X mental retardation protein (FMRP). FMRP is an RNA-binding protein forming a messenger ribonucleoprotein complex with polyribosomes in the regulation of protein synthesis at synapses. Three-dimensional (3D) aggregate culture of human-induced pluripotent stem cells (iPSCs) has evolved from embryoid body cultures, quite faithfully following human organogenesis, and provides a new platform to investigate human brain development in a dish, otherwise inaccessible to experimentation. To determine whether the loss of FMRP could alter the development of human brain organoids, we have generated forebrain organoids from three FXS male patients and three healthy male controls. We observed reduced proliferation of neural progenitor cells and premature neural differentiation as well as perturbed cell cycle progression in FXS forebrain organoids. There is also a deficit in the production of GABAergic neurons as well as an altered balance between the number of excitatory and inhibitory neurons in FXS organoids. Interestingly these deficits were not observed with FXS mouse model. To compare the differential gene expression caused by the loss of FMRP between human and mouse, we then performed RNA-seq to identify the differentially expressed genes using both mouse embryonic brain cortex and human forebrain organoids at the comparable developmental stages. We detected very few genes differentially expressed in the absence of Fmrp in mouse. However, we identified 200 genes downregulated and 126 genes up-regulated in human FXS organoids, indicating human-specific impact caused by the loss of FMRP. These results together suggest that the loss of FMRP could cause neurodevelopmental deficits specifically in human, and fragile X organoids could provide a unique platform to study the molecular pathogenesis of FXS and identify human-specific druggable targets for FXS and autism in general.

225

Investigating the role of rare DPP6 missense variants in dementia: Insights from *in vitro* investigation of protein stability. *R. Cacace[1,2], B. Heeman[1,2], C. Van Broeckhoven[1,2].* 1) VIB Center for Molecular Neurology, University of Antwerp, Antwerp, Belgium; 2) Institute Born-Bunge, University of Antwerp, Antwerp, Belgium.

Using whole genome sequencing in an unresolved autosomal dominant Alzheimer disease (AD) pedigree linked to chromosome 7q36 (Rademakers et al., 2005) and supported by genetic, genomic as well as expression studies in brain tissue of patients, we identified dipeptidyl peptidase 6 (*DPP6*) as novel candidate gene in neurodegenerative brain diseases (Cacace et al., unpublished data). DPP6 is a single pass type II transmembrane protein which forms a multimeric complex with the potassium channel $K_v4.2$, regulating the voltage-dependent gating properties and the surface expression of $K_v4.2$ in the brain (Pongs and Schwarz, 2010). DPP6 re-sequencing in dementia patients identified both rare/novel premature termination codon (PTC) variants (p.E79Gfs*9 and p.Q23O*) as well as missense variants. Which were enriched in both AD (n=558, SKAT-O p-value=0.03) and frontotemporal dementia (FTD; n=614, SKAT-O p-value=0.006) patients when compared to control individuals (n=755). Mutation modelling suggested that DPP6 protein stability could be compromised by the mutations located in the extracellular domain of the protein. To model the stability of DPP6 when mutated, we developed a specific *in vitro* protein stability assay using the Nano-Glo® HiBit Extracellular detection System (Promega). HiBit-tagged DPP6 wild-type and mutant proteins, including DPP6 fused to a PEST sequence as control, were expressed *in vitro.* Cycloheximide was then used to block the protein synthesis and to assess DPP6 stability and half-life. As expected, the extracellular p.Q230* PTC variant, showed a complete loss of expression on the cell membrane. While the majority of the patients specific missense variants showed a spectrum of reduced expression indicating destabilization of DPP6. The DPP6 protein stability assay represents an easy tool to understand the effect of DPP6 missense variants of unknown significance and to further select deleterious variants. The overall results of this study provides a mechanistic link that DPP6 loss of function is the underlying mechanism in neurodegenerative dementia.

Copyright © 2018 American Society of Human Genetics. All rights reserved

226

Genetic variation in the AnkyrinG interactome causes a range of neurological disorders. *F. Kooy[1], I.M. van der Werf[1], S. Jansen[2], B.B.A. de Vries[2], G. Vandeweyer[1].* 1) Department of Medical Genetics, Universiteit Antwerpen, Antwerp, Belgium; 2) Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands.

Numerous genes are involved in the pathogenesis of neurodevelopmental disorders. Several studies showed that a significant proportion of all putative disease genes converge on a relatively limited number of molecular pathways or protein networks. Thusfar, the identification of the disease-networks followed the identification of the individual genes. To what extend this type of selection bias affects the detection of the disease-related pathways is unknown. We here took an unbiased approach by selecting a well-defined protein-protein interaction network and investigating the mutational load therein in a large cohort of patients with various neurodevelopmental disorders along with their parents. For our study, we selected the AnkyrinG PPI network. The AnkyrinG protein binds simultaneously to the spectrin cytoskeleton as well as to multiple membrane proteins. *ANK3* encodes the AnkyrinG protein that is enriched at the axon initial segment and the nodes of Ranvier of myelinated neurons in the central nervous system. Mutations in *ANK3* have been implicated in many neurodevelopmental disorders, including ID, ASD and behavioral problems, whereas multiple SNPs have confirmed associations with bipolar disorder, schizophrenia and post-traumatic stress disorder. Upon an extensive literature search, we identified a total of 17 proteins that interact with AnkG. Direct interactors include cell adhesion molecules, cytoskeleton-component β4-spectrin and specific sodium and potassium channel components. The other proteins in this PPI network are specific sodium and potassium channels as well as components of Casein kinase 2. MIP sequencing of all genes in the *ANK3* PPI network was performed for over 1009 patient-parent trios. For this cohort, all nonsynonymous *de novo* variants that passed quality parameters, were selected for Sanger sequencing validation. A total number of 14 confirmed *de novo* variants were identified in multiple *ANK3* PPI network members. Considering the average rate of base-substitutional mutations and size of the encoded genomic region by the network genes, a priori only one to two *de novo* mutations are expected in our screen assuming no selective pressure on the included genes. Thus our cohort, we detected a significant overrepresentation of *de novo* mutations in the AnkyrinG PPI network (p=0.007, Fisher's exact test). Our finding stresses the importance of the AnkyrinG PPI network in neurodevelopmental disorders.

227

***SLC10A7* mutations in human and mouse cause a skeletal dysplasia with amelogenesis imperfecta mediated by GAG biosynthesis defects.** *J.M. Dubail[1], C. Huber[1], S. Chantepie[2], S. Sonntag[3], B. Tüysüz[4], E. Mihci[5], C.T. Gordon[1], E. Steichen-Gersdorf[6], J. Amiel[1], B. Nur[4], I. Stolte-Dijkstra[7], A.M. van Eerde[8], K.L. van Gassen[8], C.C. Breugem[8], A. Stegmann[9], C. Lekszas[10], R. Maroofian[11], E.G. Karimiani[11,12], A. Bruneel[13], N. Seta[13], A. Munnich[1], D. Papy-Garcia[2], M. De La Dure-Molla[1,14], V. Cormier-Daire[1].* 1) Institut Imagine INSERM U1163, Paris, France; 2) Université Paris-Est Créteil, Créteil, France; 3) PolyGene AG, Rümlang, Switzerland; 4) Istanbul University, Istanbul, Turkey; 5) Akdeniz University Faculty of Medicine, Antalya, Turkey; 6) Department of Paediatrics I, Medical University of Innsbruck, Innsbruck, Austria; 7) University of Groningen, Groningen, The Netherlands; 8) University Medical Center Utrecht, The Netherlands; 9) Radboud University Medical Center, Nijmegen; Maastricht University Medical Center, Maastricht, The Netherlands; 10) Julius Maximilians University, Würzburg, Würzburg, Germany; 11) St George's University of London, London, UK; 12) Next Generation Genetic Clinic, Mashhad, Iran; 13) Hôpital Bichat, Paris, France; 14) Centre de Recherche des Cordeliers, INSERM UMRS 1138, Paris , France.

Skeletal dysplasias with multiple dislocations are a group of severe disorders characterized by dislocations of large joints, scoliosis, short stature and a variable combination of cleft palate, heart defects, intellectual disability and obesity. With the help of massively parallel sequencing technologies, the majority of these rare disorders have been linked to pathogenic variants in genes encoding glycosyltransferases ("linkeropathies"), sulfotransferases, epimerases or transporters, required for glycosaminoglycan (GAG) biosynthesis. These findings support the existence of a new group of inborn errors of development defined by impaired GAG biosynthesis. However, several findings suggest that GAG synthesis is more complex than previously described and that there are a number of partners of unknown function still to be identified. Especially, correct GAG biosynthesis is dependent on a tightly regulated Golgi pH and ion homeostasis. Using exome sequencing, we identified homozygous mutations in *SLC10A7* in five individuals with a skeletal dysplasia with dislocations and amelogenesis imperfecta. Common features were severe growth retardation <-3SD, cleft palate, yellow/brown teeth, knee dislocations, spine anomalies and advanced carpal ossification. *SLC10A7* encodes a 10-transmembrane-domain transporter located at the plasma membrane, with a yet unidentified substrate. Functional studies *in vitro* demonstrated that *SLC10A7* mutations were loss-of-functions mutations reducing SLC10A7 protein expression. We generated a *Slc10a7*$^{-/-}$ mouse model which displayed short long bones, growth plate disorganization and tooth enamel anomalies, recapitulating the human phenotype. Furthermore, we identified decreased heparan sulfate levels in *Slc10a7*$^{-/-}$ mouse cartilage and patient fibroblasts. We also found an abnormal N-glycoprotein electrophoretic profile in patient blood samples. Finally, we demonstrated an increased intracellular calcium intake in patient fibroblasts, suggesting that SLC10A7 acts as a negative regulator of intracellular calcium homeostasis and could affect the Golgi divalent ion homeostasis. Together, our findings support the involvement of SLC10A7 in glycosaminoglycan synthesis and specifically in skeletal and tooth development.

Copyright © 2018 American Society of Human Genetics. All rights reserved

228

A homozygous missense variant in *NUDT6* is responsible for an autosomal recessive form of osteogenesis imperfecta. *O. Essawi[1,2], S. Symoens[1], B. Guillemyn[1], D. Syx[1], F. Malfait[1], B. Callewaert[1], A. Willaert[1], T. Essawi[2], P. Coucke[1].* 1) Center for Medical Genetics,Ghent University, Ghent, Belgium; 2) Master Program in Clinical Laboratory Science,Birzeit University, Birzeit, Palestine.

Osteogenesis Imperfecta (OI) is a rare monogenic connective tissue disorder that is characterized by increased susceptibility to bone fractures. OI is usually (>90%) inherited as an autosomal dominant (AD) disorder, mainly caused by mutations in the genes encoding type I collagen. In recent years, mutations in several genes have been identified that cause autosomal recessive OI. Here we present a consanguineous Palestinian family with three affected individuals who were clinically diagnosed as OI patients based on the presence of recurrent fractures (>20) that were mainly located in the extremities, accompanied with other skeletal manifestations including short-limb dwarfism, mild frontal bossing, bowing of legs and scoliosis. Additional clinical features are limitations of joint function, joint hypermobility and progressive muscle development. Notably, three infant deaths were reported in this family. Targeted next-generation sequencing of the coding region and flanking introns of all known recessive OI genes did not reveal any causal variant. On the other hand, exome sequencing of the proband DNA revealed a novel homozygous missense variant (NM_007083: c.308G>T, NP_009014: p.Arg103Leu) in exon 2 of the *NUDT6* gene, encoding Nudix Hydrolase 6. *NUDT6,* known as an antisense gene, is an expression regulator for the *FGF2* gene, encoding Fibroblast growth factor-2, which is involved in many physiologic and pathologic processes, including limb development, wound healing, tumor growth and angiogenesis. Segregation analysis revealed that the 3 patients harbor this homozygous missense variant whereas 6 non-affected family members were either heterozygous carriers or non-carriers. RT-qPCR analysis of RNA extracted from the proband's dermal fibroblasts revealed a respective down and upregulation of *NUDT6* and *FGF2* gene expressions. Transient overexpression of the wild type and mutant *NUDT6* gene in different cell lines, revealed significantly decreased levels of mutant *NUDT6* RNA expression, indicating that the missense variant results in decreased gene expression. Protein analysis and the characterization of an in house developed zebrafish nudt6 knockout model are currently ongoing and will be useful to further explore the underlying disease mechanism. Our findings illustrate that the homozygous missense variant identified in the *NUDT6* gene is most likely responsible for the OI phenotype segregating in this family. This finding further expands the panel of OI genes.

229

Mutations in *PIK3C2A* cause a novel ciliopathy characterized by syndromic short stature associated with cataracts and skeletal abnormalities. *D.A. Buchner[1,2,3], D. Tiosano[4,5], A. Chen[2], M. Schueler[6], M.M. Hitzert[7], A. Wiesener[8], A. Berguaz[9], A. Mory[10], A. Yuan[11], F. Gulluni[12], P. Rump[7], H. van Meer[13], D.A. Sival[13], V. Haucke[14], K.X. Knaup[6], A. Reis[8], N.N. Hauer[6], E. Hirsch[12], B.M. McDermott[1,15], B.D. Perkins[16], R. Roepman[17], R. Pfundt[17], C.T. Thiel[9], M.S. Wiesener[8], M.G. Aslanyan[17], H.N. Baris[4,10].* 1) Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH; 2) Department of Biochemistry, Case Western Reserve University, Cleveland, OH 44106, USA; 3) Research Institute for Children's Health, Case Western Reserve University, Cleveland, OH 44106, USA; 4) Division of Pediatric Endocrinology, Mayer Children's Hospital, Rambam Medical Center Haifa 30196, Israel; 5) Rappaport Family Faculty of Medicine, Technion - Israel Institute of Technology, Haifa 30196, Israel; 6) Department of Nephrology and Hypertension, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany; 7) Department of Genetics, University of Groningen, University Medical Center Groningen, PO Box 30001 9700 RB Groningen, The Netherlands; 8) Institute of Human Genetics, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany; 9) Department of Ophthalmology, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany; 10) The Genetics Institute, Rambam Health Care Campus, Haifa 3109601, Israel; 11) Cole Eye Institute, Cleveland Clinic, Cleveland, OH 44195, USA; 12) Department of Molecular Biotechnology and Health Sciences, Molecular Biotechnology Center, University of Turin, Via Nizza 52, 10126, Torino, Italy; 13) Department of Pediatrics, Beatrix Children's Hospital, University of Groningen, University Medical Center Groningen, PO Box 3001 9700 RB Groningen, The Netherlands; 14) Leibniz-Institut für Molekulare Pharmakologie, Robert-Rössle-Strasse 10, 13125 Berlin Faculty of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, 14195 Berlin, Germany; 15) Department of Otolaryngology, Case Western Reserve University, Cleveland, OH 44106, USA; 16) Department of Ophthalmic Research, Cole Eye Institute, Cleveland Clinic, Cleveland, OH 44195, USA; 17) Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands.

PIK3C2A is a class II member of the phosphoinositide 3-kinase (PI3K) family that catalyzes the phosphorylation of phosphatidylinositol (PI) into PI(3)P and the phosphorylation of PI(4)P into PI(3,4)P2. PIK3C2A is critical for the formation of primary cilia and for receptor mediated endocytosis, among other biological functions. We identified homozygous loss-of-function mutations in *PIK3C2A* in five children between the ages of 8 and 21 from three independent consanguineous families. The children presented with short stature, coarse facial features, cataracts with secondary glaucoma, and multiple skeletal abnormalities. Other recurrent features included hearing loss, short stature, stroke, developmental delay, and nephrocalcinosis. Cellular studies of patient-derived fibroblasts were consistent with loss of PIK3C2A function as evidenced by the lack of PIK3C2A protein, impaired cilia formation, and reduced proliferative capacity. Additionally, *pik3c2a* deficiency in zebrafish also led to cataract formation. Thus, the genetic and molecular data collectively implicate that mutations in *PIK3C2A* cause a new Mendelian disorder of PI metabolism. This represents the first Mendelian disorder due to a mutation in a class II PI3K and thus sheds light on their critical role in growth, vision, skeletal formation and neurological development. In particular, the considerable phenotypic overlap, yet distinct features, between this syndrome and Lowe's syndrome, which is caused by mutations in the PI-5-phosphatase *OCRL*, highlight the key role of PI metabolizing enzymes in specific developmental processes and demonstrate the unique non-redundant functions of each enzyme. This discovery sheds further light on what is known about disorders of PI metabolism and will stimulate future research to unravel the role of *PIK3C2A* and other class II PI3Ks in health and disease.

Copyright © 2018 American Society of Human Genetics. All rights reserved

230

FAM92A underlies non-syndromic postaxial polydactyly in humans and an abnormal limb and digit skeletal phenotype in mice. *S. Leal[1], I. Schrauwen[1], A.P.J. Giese[2], A. Aziz[3,4], D.T. Lafont[5], I. Chakchouk[1], R.L.P. Santos-Cortez[1], K. Lee[1], A. Acharya[1], U.W. C.M.G.[6], D.A. Nickerson[6], M.J. Bamshad[6,7], G. Ali[8], S. Riazuddin[2], M. Ansar[3], W. Ahmad[3], Z.M. Ahmed[2].* 1) Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA; 2) Otorhinolaryngology-Head & Neck Surgery, School of Medicine University of Maryland, Baltimore, Maryland, USA; 3) Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan; 4) Department of Computer Science & Bioinformatics, Khushal Khan Khattak University, Karak, KPK, Pakistan; 5) Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SA; 6) Department of Genome Sciences, University of Washington, Seattle, Washington, USA; 7) Department of Pediatrics, University of Washington, Seattle, Washington, USA; 8) Department of Biotechnology, University of Azad Jammu and Kashmir, P.O. Box 13100, Muzaffarabad, Pakistan.

Polydactyly is a common congenital anomaly of the hand and foot. Postaxial polydactyly (PAP) is characterized by one or more posterior or post-axial digits. In a Pakistani family with autosomal recessive non-syndromic postaxial polydactyly type A (PAPA), we performed genome-wide genotyping, linkage analysis and exome and Sanger sequencing. Exome sequencing revealed a homozygous nonsense variant [c.478C>T, p.(Arg160*)] in the *FAM92A* gene within the mapped region on 8q21.13-q24.12, that segregated with the PAPA phenotype. The identified variant leads to the loss of the FAM92A/Chibby1 complex that is crucial for ciliogenesis. We found that FAM92A is highly expressed in mice limbs during development and is highly expressed in the skin, the chondrocytes and the mesenchyme of the rat embryonic forelimb. In addition, we show that *Fam92a*$^{-/-}$ homozygous mice exhibit distinct abnormalities on the deltoid tuberosity of their humeri and abnormal digit morphology, including metatarsal osteomas and polysyndactyly. In conclusion, we present new a non-syndromic PAPA ciliopathy due to a loss-of-function variant in FAM92A.

231

Mutations in *PERP* cause novel forms of recessive and dominant keratoderma with hair abnormalities. *L.M. Boyden, J. Zhou, R. Hu, Y.H. Lim, R.P. Lifton, K.A. Choate.* Yale University School of Medicine, New Haven, CT.

Discovery of genes underlying diseases of skin and hair has advanced understanding of epidermal biology. Via exome sequencing of a large cohort of subjects with keratinization disorders, notable for substantial phenotypic and genotypic heterogeneity, we show that mutations in *PERP* (p53 Effector Related to PMP22, an apoptosis mediator and component of desmosomes and other cell junctions) cause two novel forms of inherited keratoderma. Homozygosity for an N-terminal frameshift mutation causes erythrokeratoderma affecting virtually all skin, and wiry but normally colored hair. Heterozygous C-terminal nonsense mutations, found in multiple unrelated kindreds and clustered within one codon, cause severe periorificial and palmoplantar keratoderma, disfiguring caries in adult teeth, and wiry hair that is strikingly yellow, even in those with otherwise darker pigmentation, which is likely a distinguishing feature of this disorder. The dominant C-terminal truncating mutations are in the terminal exon and consequently should escape nonsense-mediated decay and be expressed, while the recessive N-terminal truncating mutation is expected to have unequivocal loss of function with ablated expression; quantitative RT-PCR confirms this. Immunostaining of subject skin displays expansion of epidermal differentiation markers, including keratin 14 and filaggrin, and nuclear mislocalization of loricrin. Staining for c-Kit and the proliferation marker Ki67 demonstrates increased mast cell counts in the dermis and greater percentages of proliferating cells in the basal layer of the epidermis. Electron microscopy reveals desmosomes lacking the electron-dense midline, indicative of an immature state usually associated with wound healing, tumor invasion, or keratinocyte mitosis. Investigation of desmosomal adhesive function by dispase dissociation assay of keratinocytes shows increased fragmentation after mechanical stress. The identification of *PERP* mutations in two unique forms of keratoderma featuring specific structural abnormalities and adhesion defects adds to the understanding of desmosomal skin disease and establishes the integral importance of *PERP* in human skin. The considerable but not total phenotypic overlap of the dominant and recessive disorders, and absence of phenotype in heterozygous carriers of the recessive N-terminal truncating mutation, suggests that the tightly clustered C-terminal truncating mutations in *PERP* have both dominant-negative and neomorphic effects.

Copyright © 2018 American Society of Human Genetics. All rights reserved

232

Biallelic loss-of-function variants in the calcium-binding protein encoding gene *CCDC47* cause a novel disease characterized by woolly hair, liver dysfunction, pruritus, dysmorphic features, and global developmental delay. *M. Morimoto[1], E. Maguire[2], Z. Ammous[3], X. Song[4], D. Pehlivan[4], C. Lau[1], E. Karaca[4], H. Waller-Evans[2], C.R. Holst[5], X. Chepa-Lotrea[1], E. Macnamara[1], T. Tos[6], S. Isikay[7], M. Nehrebecky[1], C. Gonzaga-Jauregui[8], J.D. Overton[8], K.W. Brigatti[9], M. Klein[5], T.C. Markello[1], J.E. Posey[4], D.R. Adams[1,10,11], E.G. Puffenberger[9], K.A. Strauss[9], E. Lloyd-Evans[2], J.R. Lupski[4,12,13,14], W.A. Gahl[1,10,11], M.C.V. Malicdan[1,10,11].* 1) National Institutes of Health Undiagnosed Diseases Program, Common Fund, Office of the Director, National Institutes of Health, Bethesda, MD; 2) School of Biosciences, Cardiff University, Cardiff, Wales, United Kingdom; 3) The Community Health Clinic, Topeka, IN; 4) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 5) BioElectron Technology Corporation, Mountain View, CA; 6) Department of Medical Genetics, Dr. Sami Ulus Research and Training Hospital of Women's and Children's Health and Diseases, Ankara, Turkey; 7) Department of Physiotherapy and Rehabilitation, Hasan Kalyoncu University, School of Health Sciences, Gaziantep, Turkey; 8) Regeneron Genetics Center, Regeneron Pharmaceuticals Inc., Tarrytown, NY; 9) Clinic for Special Children, Strasburg, PA; 10) Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 11) Office of the Clinical Director, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 12) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 13) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 14) Texas Children's Hospital, Houston, TX.

Introduction: Calcium (Ca^{2+}) signaling is vital for various cellular processes including synaptic vesicle exocytosis, muscle contraction, regulation of secretion, gene transcription and cellular proliferation. The endoplasmic reticulum (ER) is the largest intracellular Ca^{2+} store, and dysregulation of ER Ca^{2+} signaling and homeostasis contributes to the pathogenesis of various complex disorders and Mendelian disease traits. We describe 4 unrelated patients with a distinctive, multisystem disease, and identify a gene encoding a Ca^{2+}-binding ER transmembrane protein associated with this novel genetic disease. **Methods:** Clinical evaluation was performed on each of the probands at 3 clinical centers: The National Institutes of Health, Baylor College of Medicine, and The Community Health Clinic. Whole exome sequencing and family-based genomics were performed, and candidate variants were validated by Sanger sequencing. Expression analyses were performed on patient dermal fibroblasts or lymphoblastoid cells using quantitative PCR, western blot, and immunocytochemistry. Functional assays of Ca^{2+} signaling, oxidative stress susceptibility, and ER stress response were performed on patient dermal fibroblasts. **Results:** Four probands presented with a multisystem clinical phenotype characterized by woolly hair, liver dysfunction, pruritus, dysmorphic features, and global developmental delay. These patients exhibited striking dysmorphic features that included coarse facies, a downturned mouth with full lips, bitemporal narrowing, brachycephaly and/or plagiocephaly, distal arthrogryposis, hypoplastic nipples, and overlapping toes. Biallelic nonsense or frameshift variants were identified in *CCDC47* in all four probands segregating with disease. Characterization of patient cells with predicted likely damaging alleles showed decreased *CCDC47* mRNA expression and protein expression, decreased total ER Ca^{2+}, and impaired Ca^{2+} signaling of two major Ca^{2+} release channels, IP_3R and RyR. **Conclusion:** The aggregate results, together with the previously described role of CCDC47 in development and Ca^{2+} signaling, provide evidence consistent with the contention that biallelic loss-of-function mutations in *CCDC47* delineate a novel human disease and foment further insights into the biological consequences of perturbation of Ca^{2+} signaling.

233

Meta-analysis of 1.2 million individuals for blood lipid levels. *I. Surakka[1], S.E. Graham[1], G. Abecasis[2], G. Peloso[3], S. Kathiresan[4,5,6,7], C.J. Willer[1,8,9], Global Lipids Genetics Consortium.* 1) Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI; 2) Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI; 3) Department of Biostatistics, Boston University School of Public Health, Boston, MA; 4) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; 5) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA; 6) Department of Medicine, Harvard Medical School, Boston, MA; 7) Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA; 8) Department of Human Genetics, University of Michigan, Ann Arbor, MI; 9) Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI.

Circulating cholesterol and triglyceride levels are important, causative markers of cardiovascular disease. Approximately 250 genetic loci contributing to variation in lipid levels have been identified thus far, but these variants explain only a portion of the variability in lipid levels attributed to genetic factors. As part of the Global Lipids Genetics Consortium, we have meta-analyzed lipid association results from up to 1.2 million individuals for association with total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), nonHDL cholesterol (nonHDL-C), and triglycerides (TG). Variants were imputed within each cohort using 1000 Genomes Phase 3 with European cohorts also imputed using the Haplotype Reference Consortium. In total, we have identified ~125 novel loci each for HDL-C, TC, and TG, ~75 novel loci for LDL-C, and ~50 novel loci for nonHDL-C. The most significant index variants within the identified novel loci range in frequency from rare to common and include both SNPs and indels. For example, 24 of the 443 unique index variants within Europeans had a minor allele frequency of less than 1% and approximately 3% were insertion/deletion variants. In addition, we have identified variants showing sex and ancestry specific effects. For example, variants within the *IFNL4* gene previously associated with Hepatitis C clearance are associated with reduced LDL-C levels in African American individuals within our cohort, but are not significantly associated in other ancestries. Moreover, participating cohorts in the present meta-analysis are more ancestrally diverse than previous studies; approximately 25% of individuals are of non-European ancestry and > 90,000 are of African American ancestry. This allows for improved fine-mapping of identified loci and construction of ancestry-specific genetic risk scores. As one of the largest meta-analyses to date, our results yield substantial insight into the genetic contributions to blood lipid levels and their association with cardiovascular and other related diseases.

Copyright © 2018 American Society of Human Genetics. All rights reserved

234

Genome-wide study of statin usage and related adverse events using national drug prescription registries from Estonia and Finland. *K. Krebs[1,2], P. Helkkula[3], V. Kukuškina[1,2], R. Mägi[1], S. Ripatti[3,4], L. Milani[1,5].* 1) Estonian Genome Centre, University of Tartu, Tartu, Estonia; 2) Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia; 3) Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland; 4) Department of Public Health, University of Helsinki, Finland; 5) Science for Life Laboratory, Department of Medical Sciences, Uppsala University, Uppsala, Sweden.

Nationwide electronic health and drug prescription registries provide information on major health events for epidemiological and genetic association studies. Although statins are the number one lipid lowering drug globally, a considerable subset of individuals suffer from adverse reactions to statin use. Genome-wide association studies have identified genetic variants and loci modifying the risk of adverse events, but these explain only a small fraction of cases where individuals change statins or stop their usage altogether. To study the genetics of statin use and identify possible causes of side effects of statins, we surveyed usage and adherence to statin therapy among participants of the Estonian (EST) and Finrisk (FIN) biobanks by analysing the electronic drug prescription data of the subjects (available over the periods of 2003-2017 and 1995-2015, respectively). In the EST cohort 9608 individuals had been prescribed and purchased statins, and in the FIN cohort the respective number was 8579. First, we investigated the genetic profile of statin users in separate genome wide analyses of 35,696 Estonians and 22,710 Finns. By meta-analyzing the results, we found strong associations with genes that are well known in lipids metabolism (*APOE*; *PCSK9*; *LDLR*; *HMGCR*, etc) as well as a new locus in *ALDH1B1* (rs3043, p-value=2.76*10^{-17}). In addition, we identified a potentially interesting triallelic variant in *CYP2C8* (rs1058930, p-value=4.22*10^{-58}) in the FIN cohort. For the genome-wide analysis of possible statin side effects, we conducted a case-control study of individuals who had either discontinued statin therapy or been prescribed another statin during 2010-2015, while excluding all of the one-time buyers (632 in EST; 327 in FIN). Among our initial findings in EST we report a novel variant in the *ABCG2* gene (p=3.63*10^{-8}, OR=4.72) that is known to be involved in the transport of simvastatin. We also detected a significant signal in the *MCTP2* gene (p=2.68*10^{-8}, OR=0.50), which is important in cardiac outflow and has been associated with drug-induced liver injury from flucloxacillin. Taken together, we illustrate that health registry data can be used to survey genetic causes of drug side-effects resulting in non-adherence and thereby in poor treatment outcomes. Identifying variants that are predictive of elevated risk of non-adherence may provide further guidance for prescription of statins.

235

Monogenic and polygenic predictors of plasma lipid extremes from whole genome sequencing in diverse ancestries: The NHLBI TOPMed program. *G.M. Peloso, NHLBI TOPMed Lipids Working Group.* Dept. of Biostatistics, Boston University, Boston, MA.

Plasma lipid levels are heritable risk factors for coronary heart disease. Deleterious coding variants in known lipid genes are present in a very small fraction of individuals with extreme quantitative lipid levels. Whole genome sequencing (WGS) permits simultaneous assessment of Mendelian coding variants and the entirety of the genome for comprehensive evaluation of genetic drivers of extreme lipid levels. We estimated the contribution of monogenic and polygenic determinants of the extremes (upper and lower 5th percentiles) of LDL, HDL, and triglycerides in 5,910 white and in 4,380 black participants from the Framingham Heart Study (FHS), Jackson Heart Study (JHS), and the Multi-Ethnic Study of Atherosclerosis (MESA). Subject samples underwent deep-coverage WGS as part of the NHLBI TOPMed program. We catalogued presence of rare coding deleterious variants in known lipid Mendelian genes and calculated polygenic risk scores (PRS) utilizing 2M variants from prior plasma lipid GWAS summary statistics (Willer CJ et al, Nat Genet 2013). We defined individuals as having high polygenic risk if their PRS was in the upper 5th percentile of the distribution. Among whites, carrying a monogenic deleterious variant was associated with a 30 mg/dl increase in LDL (P=2x10^{-4}); further, having a high PRS was associated with a 33 mg/dl increase in LDL (P=1x10^{-57}). In blacks, those with a monogenic deleterious variant had 41 mg/dl higher LDL (P=2x10^{-7}); further, those having a high PRS had 17 mg/dl greater LDL (P=6x10^{-10}). Having a monogenic variant resulted in a larger increase in LDL in blacks than observed among whites, yet having a high PRS yielded a smaller increase in LDL in blacks than in whites. These results translate to individuals carrying a monogenic variant having an increased odds of extremely high LDL in both whites (10.9; 95% CI: 3.7-32.1) and blacks (7.4; 95% CI: 3.0-18.4); further, subjects having a high PRS is associated with an elevated odds of high LDL, greater in whites (7.7; 95% CI: 5.6-10.5) than in blacks (3.2; 95% CI: 2.1-4.9). In conclusion, we found that both rare coding variants and a PRS composed of common variants contribute to high LDL. In particular, a PRS comprised of >2M common variants substantially altered the odds of having an extreme lipid value. These results will be extended to larger sample sizes (>10,000 for whites and >9,000 blacks) for validation and other ancestries (Hispanics and Pacific Islanders) for transferability.

Copyright © 2018 American Society of Human Genetics. All rights reserved

236

A missense variant in *B4GALT1* reduces low-density lipoprotein and fibrinogen. *M. Montasser[1], A. Howard[1], R. McFarland[1], C. Van Hout[2], G. Della Gatta[2], B. Shen[2], N. Li[2], G. Tzoneva[2], N. Gosalia[2], A. Economides[2], B. Mitchell[1], M. Healy[2], J. O'Connell[1], E. Streeten[1], N. Zaghloul[1], C. Sztalryd-Woodle[1], S. Taylor[1], A. Shuldiner[1,2].* 1) Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, Maryland USA; 2) Regeneron Genetics Center, Regeneron Pharmaceuticals, Inc, Tarrytown, NY.

Elevated low-density lipoprotein cholesterol (LDL) and fibrinogen are major independent risk factors for cardiovascular disease (CVD). Understanding their genetic basis may identify novel therapeutic targets to lower their levels and treat or prevent CVD. Isolated founder populations can enable discovery of novel diseaseassociated variants enriched in these populations through genetic drift. Although very rare in general populations, these novel variants can inform biology relevant to all humans. We identified a strong novel association (p = 3.3E-18) between a missense SNP (N352S) in *B4GALT1* and LDL in the Amish population. Each 352S allele is associated with a 14.7 mg/dl lower LDL, and has a frequency of 6% in the Amish while extremely rare in the general population. In addition, this SNP was associated with a 20% lower fibrinogen level (p= 5.0E-4). *B4GALT1* encodes a glycosyltransferase responsible for adding galactose to maturing glycan chains of glycoproteins. Knockdown of the *B4GALT1* orthologue in zebrafish resulted in significantly lower LDL compared to control (p=0.02). Co-expression of wild type human *B4GALT1* mRNA rescued the LDL phenotype, while co-expression of mutated human *B4GALT1* mRNA resulted in a 15% lower rescue of the LDL phenotype, suggesting only a partial defect in function introduced by the missense variant. To assess the impact of *B4GALT1* N352S on glycosylation, the carbohydrate deficient transferrin test was performed using serum samples from 24 subjects from the 3 genotype groups. Wild type homozygotes had normal glycosylation, while 352S homozygotes had abnormally high levels of carbohydrate deficient transferrin; heterozygotes were intermediate (p=7.6 E-10). These in vivo data strongly support the hypothesis that the N352S variant decreases the total enzymatic activity of B4GALT1 under physiological conditions. We identified a novel gene and variant that is associated with lower LDL and fibrinogen which may be cardioprotective. Evidence from cell and animal based experiments as well as human data indicate that the variant causes decreased protein glycosylation. Further understanding of the underlying mechanisms may provide new insights and therapeutic strategies for CVD.

237

Evaluating the contribution of cell-type specific alternative splicing to variation in lipid levels. *K.A.B. Gawronski[1], W. Bone[1], E. Pashos[1], Y. Park[1], X. Wang[2], W. Yang[3], D. Rader[1,4,7], K. Musunuru[1,4], B. Voight[1,5,6], C. Brown[1].* 1) Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; 2) Cardiovascular Institute, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; 3) Institute for Regenerative Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; 4) Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; 5) Department of Pharmacology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; 6) Institute for Translational Medicine and Therapeutics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; 7) Division of Translational Medicine and Human Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA.

Blood lipid levels are heritable traits associated with cardiovascular disease risk and previous genome-wide association studies (GWAS) have identified 150+ loci associated with these traits. However, the relevant cell types and the genetic mechanisms underlying most of these loci are not well understood. Recent research indicates that changes in the abundance of alternatively spliced transcripts are important mechanisms contributing to complex trait variation. Consequently, identifying genetic loci that associate with alternative splicing (i.e., sQTLs) in disease-relevant cell types such as hepatocytes and determining the degree to which these loci are informative for lipid biology is of broad interest. We present results from an analysis of transcript splicing in 84 sample-matched iPSC and hepatocyte-like cell (HLC) lines (n=168), as well as an analysis of an independent collection of primary liver tissues (n=96). The regulation of transcript splicing is highly cell-type specific: 12,298 genes are differentially spliced upon iPSC differentiation (FDR 5%). Genes that are differentially spliced between iPSCs and HLCs are enriched for insulin signaling and lipid metabolism pathway annotations. We further identify 2,562 intron-level HLC sQTLs and 3,088 intron-level iPSC sQTLs at a false discovery rate (FDR) of 5%. Replication analysis indicates that HLC sQTLs more closely represent primary liver sQTLs compared to iPSC sQTLs, suggesting that HLCs are a tractable cell-based model for the mechanistic characterization of liver gene regulation. To evaluate the contribution of our sQTLs to variation in lipid levels, we present the results of colocalization analysis using blood lipid GWAS data from the Global Lipids Consortium (Global Lipids Genetics Consortium et al., 2013). We identify 35 lipid GWAS loci that co-localize with an HLC eQTL or sQTL. We identify strongly colocalized sQTL-genes with both previously known (e.g., *PGS1*) and hitherto undescribed effects on lipid biology. Importantly, 17 of these loci only colocalize with an HLC sQTL, demonstrating that a substantial fraction of GWAS effects may be mediated by genetically determined changes in transcript splicing and are not discoverable through analysis of steady-state gene expression alone. In sum, our efforts provide an important foundation for future efforts that use iPSC and iPSC-derived cells to evaluate genetic mechanisms influencing both cardiovascular disease risk and complex traits in general.

Copyright © 2018 American Society of Human Genetics. All rights reserved

238

Leveraging functional genomic data for etiologic insight from GWAS summary statistics. *C. Quick, G. Abecasis, M. Boehnke, X. Wen, H.M. Kang.* Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI.

Background. Genome-wide association studies (GWAS) have identified thousands of genetic loci associated with hundreds of complex traits. However, the biological mechanisms underlying these associations are often poorly understood. Transcriptomic and epigenomic projects have provided greater insight into the functional effects of noncoding variation and complex trait etiology. Here we describe an empirical Bayes framework to identify causal genes and pathways by integrating functional genomic data and GWAS summary statistics. **Approach**. We first estimate functional weights between regulatory variants and genes using data from GTEx, ENCODE, and FANTOM5. We then use these functional weights to compute stratified gene-based Bayes factors (BFs) from GWAS association statistics across regulatory and coding variants. Next, we fit a penalized prior enrichment model to compute the probability that each gene is causal given gene ontology annotations. Finally, we update the enrichment model, functional weights, and gene-based posterior probabilities in an E-M algorithm. We illustrate that this approach favors configurations of functionally interrelated causal genes across the genome. **Results**. We applied our approach to GWAS of lipid traits, atrial fibrillation, and type 2 diabetes. We find enrichment of heart-specific regulatory effects for atrial fibrillation (p=2e-5) and liver-specific regulatory effects for lipid traits (p=2e-4), as well as enrichment in relevant biological pathways for each trait. We show that our approach has substantially higher concordance with known genes from OMIM than conventional criteria based on gene distance from GWAS hit or gene-based test statistics. Finally, we introduce a publicly available software implementation to enhance interpretation and biological insight from GWAS summary statistics.

239

Genomic, transcriptomic, and clinical determinants associated with aberrant clonal expansions. *MJ. Fave[1,2], E. Bader[1,3], P. Mehanna[1,2], V. Bruat[1,2], P. Awadalla[1,3].* 1) Ontario Institute for Cancer Research, Toronto, Ontario, Canada; 2) Montreal Hearth Institute, Montreal, Canada; 3) University of Toronto, Toronto, Canada.

Somatic mosaicism of hematopoeitic cells is a common phenomenon in older individuals, however its proximate causes and the mechanisms underlying its impacts remain to be elucidated. An increase the in risk of both hematologic cancers and atherosclerosis is observed in individuals with aberrant clonal expansions, illustrating its range of consequences and clinical importance. Yet, the genome-wide profile of somatic clonal expansions remains poorly documented, and the mechanisms underlying its association with diseases other than cancer are not well understood. Here, we combine genotyping data of more than 18,000 participants from the Canadian Partnership for Tomorrow's Project (CPTP) with whole-transcriptome, exome sequencing, cytokine profiling, blood parameters, and more than 500 health phenotypes and environmental exposures. We characterize the genome-wide profile of somatic structural variants (SVs) in CPTP, and in a subset of 1,000 participants, we also detect somatic SNVs using exome and RNA-sequencing. We document a non-random distribution of somatic SVs across the genome, and the presence of exonic somatic SNVs in more than 75% of the participants, including highly deleterious variants. Using coinertia analyses, we found associations between the presence of a somatic SV overlapping either of *JAK2*, *PCSK9*, *ASXL1* or *ABCG5* and a sharp increase in risk for cardiovascular conditions (e.g high arterial stiffness, stroke). In addition, we find instances of somatic duplications increasing the expression level of overlapping genes (e.g. *PRKAB2*), and for which the expected phenotypic effect is detected in the participant (e.g. low HbA1c), revealing a clinical impact of somatic mutations. Finally, by comparing whole-transcriptome profiles of 1,000 participants with and without detectable clonal expansions, we find a dysregulation of pathways involved in cell cycle regulation, tumorigenesis, and lipid and drug transport (e.g WNT signalling, ABC transporters), after controlling for age and other covariates. WNT signalling is a critical regulator of stem cells often mutated in cancer, is involved in the development of the cardiac muscle, and has a role in atherosclerosis. Collectively, these results show clinical impacts of somatic clonal expansions in healthy individuals, and reveal that a dysregulation of critical cell-cycle determinants may underly the pleiotropic nature of associations between clonal expansions, cancer risk, and cardiovascular conditions.

Copyright © 2018 American Society of Human Genetics. All rights reserved

240

APOBEC3A or APOBEC3B? Causes and consequences of APOBEC mutagenesis in human cancers. *R. Banday, O. Onabajo, S. Lin, A. Obajemu, J. Vargas, L. Prokunina-Olsson.* Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA.

Identification of mutational processes that shape tumor genomes holds promise for development of better cancer treatment, diagnosis and prevention options. Activity of the deamination enzymes, APOBEC3A (A3A) and APOBEC3B (A3B) is considered the major source of mutations in several cancer types. However, it remains unclear when and in which biological context these enzymes get activated and induce mutations. Based on integrated analysis of germline genotype, somatic mutation, RNAseq, methylation and clinical data for ~10, 000 TCGA samples representing 32 cancer types, we show that A3A and A3B-induced mutagenesis is associated with several intrinsic and extrinsic factors. Specifically, we found that *A3A* mRNA expression is associated with tumor microenvironment enriched with M1 macrophages but *A3B* mRNA expression is associated with DNA damage response pathways. These associations appear cancer-specific, such as the relationship between M1 macrophages and *A3A* expression is predominant in head and neck cancer (HNSC) and cervical cancer (CESC), while the relationship between *A3B* expression and DNA damage is predominant in bladder cancer (BLCA). However, both relationships exist in breast cancer (BRCA) and lung adenocarcinoma (LUAD), highlighting that A3A and A3B are equal drivers of mutagenesis in these two cancers. We also identified alternative splicing of *A3A* and *A3B* as an intrinsic mechanism that modulates APOBEC-induced mutagenesis. We show that A3A has two protein isoforms – one catalytically active and one catalytically inactive, and A3B has four isoforms – one catalytically active and three catalytically inactive. Expression of transcripts for these catalytically inactive non-mutagenic isoforms at above 10% of total *A3A*/*A3B* expression levels, which was observed in a quarter of all TCGA tumors, was associated with up to a three-fold decrease in APOBEC-signature mutation burden ($P<10^{-115}$). We propose that among all intrinsic and extrinsic factors, alternative splicing of *A3A* and *A3B* can be modulated to control APOBEC-induced mutagenesis. In conclusion, both A3A and A3B cause mutations in cancer genomes. Induction and the mutagenic effects of A3A or/and A3B in different cellular environments depends on the combination of extrinsic and intrinsic factors.

241

Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *A. Fujimoto[1,2], M. Fujita[2], Y. Shiraishi[3], K. Maejima[2], K. Nakano[2], T. Tsunoda[2], S. Imoto[3], S. Miyano[3], N. Matsubara[4], N. Tomita[4], H. Nakagawa[2].* 1) Kyoto University, Kyoto, Kyoto, Japan; 2) RIKEN Center for Integrative Medical Science, Japan; 3) Human Genome Center, Institute of Medical Sciences, The University of Tokyo, Japan; 4) Division of Lower Gastrointestinal Surgery, Department of Surgery, Hyogo College of Medicine, Nishinomiya, Japan.

Microsatellites are repeats of 1-6bp units and ~10M microsatellites have been identified across the human genome. Microsatellites are sensitive to mismatch errors, therefore microsatellite markers have been used to detect cancers with mismatch repair deficiency. Some microsatellite mutations could be actionable markers to detect hyper-mutation phenotypes for immune-checkpoint-blockade therapy. To reveal the mutational landscape of the microsatellite repeat regions in whole-genome level, we analyzed approximately 9.2 million microsatellites in ICGC PanCancer samples across 21 tissues. First, we developed an insertion and deletion (indel) caller that considers error patterns of different types of microsatellites. Among the 2,717 PanCancer samples, our analysis identified 31 samples with higher microsatellite mutation rate, which we defined as microsatellite instability (MSI) cancers. As expected, colorectal, uterus, and stomach cancers had a larger number of MSI samples, but MSI was also observed in a minority of samples in liver, pancreas, ovary, esophageal, and skin cancers. Next, we selected 19 microsatellites which mutated frequently in MSI-positive cancers in the PanCancer samples, and defined a novel microsatellite marker set to detect MSI cancers with high sensitivity. Validation of the marker set in an independent cancer cohort showed the efficiency of the marker set. Third, we found that replication timing and DNA shape were significantly associated with the mutation rate of the microsatellites. Analysis of the germline variation of the microsatellites suggested that similar mechanisms are working on the germline polymorphisms. Lastly, analysis of mutations in the mismatch repair genes show that somatic point mutations and short indels had larger functional impact than germline mutations and structural variations. Our analysis provides a comprehensive picture of mutations in the microsatellite regions, and reveal possible causes of mutations, and provides a useful marker set for MSI detection.

Copyright © 2018 American Society of Human Genetics. All rights reserved

242

Mismatch-repair signature mutations activate gene enhancer activity across colorectal cancer epigenomes. *S. Hung[1], A. Saiakhova[1], Z. Faber[1], C. Bartels[1], D. Neu[1], G. Dhillon[1], E. Hong[1], I. Bayles[1], M. Kalady[3], S. Markowitz[1,2], P. Scacheri[1].* 1) Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH; 2) Department of Medicine, Case Western Reserve University, Cleveland, OH; 3) Lerner Research Institute, Cleveland Clinic, Cleveland, OH.

The search for cancer driver mutations has largely focused on the 2% of the human genome that codes for genes. Commonly mutated genes have been found for many cancers, but far less is known about the prevalence of mutations in cis-regulatory elements. We leveraged an approach that exploits gains in enhancer activity in tumor versus normal, in combination with mutation detection from H3K27ac ChIP-seq data, to pinpoint potential activating mutations in enhancer elements in colorectal cancer (CRC). Analysis of a genetically diverse cohort of CRC specimens revealed that samples of MSI subtype have a particularly high rate of indel mutation in active enhancers. In support of a functional role, enhancers with indels show evidence of positive selection and their target genes show elevated expression. Moreover, a subset of the enhancer indels is highly recurrent. The indels arise in short homopolymer tracts of A/T and generate sequences that closely resemble consensus motifs for the FOX family of pioneer transcription factors. We show the capacity of the noncoding indels to modulate enhancer activity through CRISPR/Cas9 inactivation of the *MLH1* gene followed by H3K27ac ChIP-seq studies. Our results suggest that indel mutations in noncoding poly(A/T) tracts, previously presumed benign, frequently augment enhancer activity in the epigenomes of mismatch repair-deficient CRC tumors and provide a selective advantage.

243

Longitudinal monitoring of AML tumors with high-throughput single-cell DNA sequencing reveals rare clones prognostic for disease progression and therapy response. *D.J. Eastburn[1], C.M. McMahon[2], A. Viny[3], R. Bowman[3], L. Miles[3], R. Durruthy-Durruthy[1], R.L. Levine[3], M. Carroll[2], C.C. Smith[4], A.E. Perl[2].* 1) Mission Bio, Inc., South San Francisco, CA; 2) Division of Hematology-Oncology, University of Pennsylvania, PA; 3) Memorial Sloan Kettering Cancer Center, New York, NY; 4) Division of Hematology-Oncology, University of California, San Francisco, CA.

AML (acute myeloid leukemia) is increasingly being treated with precision medicine. To better inform treatment, the mutational content of patient samples must be determined. However, current tumor sequencing paradigms are inadequate to fully characterize many instances of the disease. A major challenge has been the unambiguous identification of potentially rare and genetically heterogeneous neoplastic cell populations, capable of critically impacting tumor evolution and the acquisition of therapeutic resistance. Standard bulk population sequencing is unable to identify rare alleles and definitively determine whether mutations co-occur within the same cell. Single-cell sequencing has the potential to address these key issues and transform our ability to accurately characterize clonal heterogeneity in AML. Previous single-cell studies examining genetic variation in AML have relied upon laborious, expensive and low-throughput technologies that are not readily scalable for routine analysis of the disease. We applied a newly developed platform technology to perform targeted single-cell DNA sequencing on over 140,000 cells and generated high-resolution maps of clonal architecture from AML tumor samples. Single-cell sequencing of multiple patient samples demonstrated that relapse clones acquired oncogenic *RAS* mutations. We utilized the high-throughput and sensitivity of our single-cell approach to more definitively assess where in the course of treatment these *RAS* mutated clones were acquired. Oncogenic *RAS* harboring clones, comprising between 0.4%, and 0.1% of tumor populations, were identified in patient samples either prior to or shortly after onset of treatment. Significantly, these *RAS* variant alleles were not detectable with targeted bulk sequencing. Throughout the course of treatment with the FLT3 inhibitor gilteritinib, the *RAS* mutant clones selectively expanded and were responsible for resistance to therapy and relapse. These findings point to the presence of underlying genetic heterogeneity in AML and demonstrate the utility of sensitively assaying clonal architecture to better inform patient stratification and therapy selection.

Copyright © 2018 American Society of Human Genetics. All rights reserved

244

Genome-wide CRISPR screening and integrative genomic analyses identify novel mechanisms of glucocorticoid resistance in acute lymphoblastic leukemia. *R.J. Autry[1,2], S.W. Paugh[1], R.M. McCorkle[1], R.A. Carter[1], J. Liu[1], L. Shi[1], D.C. Ferguson[1], C.E. Lau[1], J.C. Panetta[1], E.J. Bonten[1], J.A. Beard[1], K.R. Crews[1], W. Yang[1], J.D. Diedrich[1], D. Pei[1], S.E. Karol[1], C. Smith[1], K.G. Roberts[1], S. Pounds[1], S.M. Kornblau[3], C.G. Mullighan[1,2], J.J. Yang[1,2], D. Savic[1,2], S. Jeha[1], C.H. Pui[1], C. Cheng[1], J. Yu[1], C. Gawad[1], M.V. Relling[1,2], W.E. Evans[1,2].* 1) St. Jude Children's Research Hospital, Memphis, TN; 2) The University of Tennessee Health Science Center, Memphis, TN; 3) The University of Texas MD Anderson Cancer Center, Houston, TX.

Glucocorticoids are essential components of combination chemotherapy to treat acute lymphoblastic leukemia (ALL), and resistance to glucocorticoids is associated with a poor prognosis in patients with ALL. To understand the mechanisms underlying glucocorticoid resistance, we analyzed the prednisolone (PRED) sensitivity of primary leukemia cells from 225 pediatric patients with B-lineage ALL. We interrogated multiple genomic and epigenomic features in these patients including mRNA expression, miRNA expression, DNA methylation, single nucleotide variants (SNVs) and copy number alterations (CNAs). Our analysis of these data resulted in individual signatures (254 mRNAs, 49 miRNAs, 203 CpG sites, 380 SNVs and 25 CNAs) for each feature based on its ability to discriminate PRED resistant vs. sensitive leukemia. Features were then aggregated at the gene level using a novel statistical approach to assess the association of all genes with PRED resistance. Gene-level aggregation of all genomic features identified 98 overlapping genes. As an orthogonal method to validate our findings, we performed genome wide CRISPR/Cas9 knockout screening in the Nalm6 human Pre-B leukemia cell line. Resistant cells were selected by treatment with multiple concentrations of PRED (10, 100 and 500 μM). We found 17 genes that were significant in all three methods, 15 of which have not been previously linked to glucocorticoid resistance. Collectively, our approach identified 28 of 37 (76%) genes and miRNAs known to confer glucocorticoid resistance and uncovered 15 previously undiscovered glucocorticoid-resistance genes significant by all methods. In addition to validating previously reported mechanisms of resistance [e.g. decreased expression of GR (*NR3C1*) and *SMARCA4*]; our top novel candidate gene (*CELSR2*) was highly significant in all analyses, and its effects were recapitulated by manipulating CELSR2 in human ALL cell lines. These findings were also confirmed via single cell RNA sequencing of primary leukemia cells and by network analyses (NetBID). Our genomewide integrative genomic analysis has uncovered new genomic mediators of glucocorticoid resistance, revealed novel mechanisms of glucocorticoid resistance and identified a novel drug combination (prednisolone and venetoclax) to mitigate CELSR2 mediated resistance.

245

High-throughput identification of genes involved in single and multi-drug resistance in pancreatic cancer with pooled CRISPR screening. *R.C. Ramaker, A.A. Hardigan, E. Gordon, R.M. Myers, S.J. Cooper.* Genetics, University of Alabama at Birmingham/HudsonAlpha Institute for Biotechnology, Huntsville, AL.

Despite the use of multi-drug cocktails as first-line therapy, pancreatic ductal adenocarcinoma (PDAC) harbors one of the worst prognoses of all common cancers with five year survival rates remaining below 10% for the last three decades. Large-scale genome sequencing and transcriptome profiling of PDAC patients has yet to produce any successful targeted therapies for PDAC treatment. We performed genome-wide CRISPR activation (CRISPRa) and CRISPR knock-out (CRISPRko) screens for mechanisms of resistance to four cytotoxic chemotherapies (gemcitabine, 5-fluorouracil, irinotecan, and oxaliplatin) commonly used in PDAC patients using two PDAC cell lines (BxPC3 and Panc1). We identified both drug-specific and multi-drug resistance genes using this approach. The most consistent multi-drug resistance mechanism identified across drugs and cell lines was activation of *ABCG2* (FDR<0.05), a cell efflux pump previously associated with resistance to multiple drugs. Integration of our screen results with PDAC tumor expression data revealed activation of several members of the hemidesmisome complex and transcriptional repressor complexes to be potentially clinically relevant mechanisms of multi-drug resistance. We also describe an approach for integrating our screen data with PDAC tumor or cell line transcriptomic data to predict drug sensitivity. In sum, we have used genome-wide CRISPRa and CRISPRko to generate a resource capable of nominating important mechanisms of drug resistance and predicting response to chemotherapy for PDAC cell lines and tumors.

Copyright © 2018 American Society of Human Genetics. All rights reserved

246

An integrative functional genomic approach identifies genotype-specific therapeutic vulnerabilities in lung cancer. *A.H. Berger[1], A. Vichas[1], N. Nkinsi[1], F. Mundt[2], F. Piccioni[2].* 1) Fred Hutchinson Cancer Research Center, Seattle, WA; 2) Broad Institute, Cambridge, MA.

Genotype-informed treatment of non-small cell lung cancer (NSCLC) has transformed clinical practice and improved patient survival. However, targeted therapeutic options are lacking for some genetic subtypes such as those with mutated *KRAS* or *RIT1*. The goal of this study was to identify downstream effectors of KRAS and RIT1 that might be targeted for therapeutic benefit in *KRAS*- and *RIT1*-mutant lung cancer. We integrated multiple genome-scale CRISPR/Cas9 knockout screens with global protein and phosphoprotein profiling by mass spectrometry to identify pathways and proteins that are modulated by KRAS/RIT1 and essential for survival of *KRAS/RIT1*-mutant cells. CRISPR/Cas9 screening was performed in five isogenic human cell lines engineered to express vector control or lung cancer oncogenes $KRAS^{G12V}$, $RIT1^{M90I}$, $EGFR^{T790M/L858R}$, and $PIK3CA^{E545K}$. Each of the oncogenes confer resistance to cell death induced by the EGFR inhibitor, erlotinib. In the presence of erlotinib, cell survival is exquisitely dependent on the function of the expressed oncogene. Cells were transduced with the Brunello library (Broad Institute), comprised of 74,442 sgRNAs targeting 19,363 human genes and sgRNA abundance in the presence or absence of erlotinib assessed by Illumina sequencing. Through integrative analysis of essential genes across each of the five cell lines, we identified synthetical lethal relationships unique and shared in each isogenic background. Among the dependencies identified was the requirement of Aurora kinase pathway genes *AURKA*, *AURKB*, and *HASPIN* for survival of *RIT1*-mutant cells. Mass spectrometry revealed significant overlap between RIT1- and KRAS- modulated proteins and phosphorylation sites, with 66% and 50% of $RIT1^{M90I}$-regulated proteins and phospho-sites shared with those modulated by $KRAS^{G12V}$ or $KRAS^{Q61H}$. While MYC activation was induced by both KRAS and RIT1, modulation of the Aurora kinase pathway was unique to $RIT1^{M90I}$ cells. We then performed a small molecule screen using 160 small molecules with clinical utility. Cell viability of *RIT1*- and *KRAS* -mutant isogenic cells was highly correlated (R^2 = 0.88), but Aurora kinase inhibitors alisertib and barasertib were selectively lethal to *RIT1*-mutant cells (Z score = 2.2 and 4.5). These data nominate Aurora kinase inhibition as a therapeutic strategy for *RIT1*-mutant lung cancer. Continued integrative analysis of multiple functional 'omic datasets will improve identification of therapeutic targets in cancer. .

247

Deconvolution of CRISPR tiling screen data for the discovery and dissection of functional non-coding elements. *J.Y. Hsu[1,2], C.P. Fulco[3,4], M.A. Cole[5], M.C. Canver[5], D. Pellin[5], F. Sher[5], R. Farouni[2], K. Clement[2], J.A. Guo[2], L. Biasco[5], S.H. Orkin[5,8], J.M. Engreitz[3], E.S. Lander[3,4,6], J.K. Joung[2,7], D.E. Bauer[5], L. Pinello[2,3,7].* 1) Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; 2) Molecular Pathology Unit, Center for Cancer Research, Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, Massachusetts, USA; 3) Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 4) Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA; 5) Division of Hematology/Oncology, Boston Children's Hospital, Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Stem Cell Institute, Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA; 6) Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; 7) Department of Pathology, Harvard Medical School, Boston, Massachusetts, USA; 8) Howard Hughes Medical Institute, Boston, Massachusetts, USA.

The advent of programmable genome editing using CRISPR-based technologies has allowed for high-throughput functional interrogation of non-coding elements throughout the genome. Functional mapping can be achieved by densely tiling single guide RNAs (sgRNAs) across a non-coding region of interest, where each sgRNA enables linking of a unique genomic location to an observable phenotype. These tiled sgRNAs can be used with CRISPR-Cas nucleases, CRISPR interference (CRISPRi), or CRISPR activation (CRISPRa) to introduce mutations, repress a target element, or activate a target element, respectively. The ability to perform CRISPR tiling screens allows for the systematic and quantitative assessment of causal links between non-coding regulatory elements and biological phenotypes of interest, highlighting its substantial potential to elucidate the regulatory architecture underlying gene expression. Although experimental protocols are available for tiling a genomic region with genetic or epigenetic perturbations, analyzing the resulting sequencing data is not trivial. The computational challenges include dealing with (i) the variability in sgRNA targeting efficiencies, (ii) the non-uniform spacing of sgRNAs due to PAM requirements, (iii) the extent of shared information amongst neighboring sgRNAs, and (iv) different properties of the CRISPR technologies. We developed CRISPR-SURF (Screening Uncharacterized Region Function) to address these challenges and to provide an open-source command line tool and interactive web-based application for the analysis of CRISPR tiling screen data. We performed two matched CRISPR tiling screens on the *BCL11A* locus using CRISPR-Cas9 (SpCas9) and CRISPRi (dCas9-KRAB) to characterize different screening modalities and CRISPR-SURF performance. CRISPR-SURF reliably identifies non-coding regions regulating the expression of transcription factor *BCL11A*, a potent repressor of fetal haemoglobin (HbF) levels, and also elucidates important properties of the CRISPR technologies used. CRISPR-SURF is a user-friendly open-source software that can be used to design sgRNAs for tiling screens, to analyze new CRISPR tiling screen data, and to explore several pre-computed datasets at http://crisprsurf.pinellolab.org/. CRISPR-SURF provides the capability to discover functional non-coding regions such as enhancers and to dissect their critical elements enabling a powerful characterization of genetic variants involved in traits or diseases.

Copyright © 2018 American Society of Human Genetics. All rights reserved

248

Interrogation of human hematopoiesis at single-cell and single-variant resolution. *J.C. Ulirsch[1,4,5,6], C.A. Lareau[1,2,3], E.L. Bao[1,4,5,7], L.S. Ludwig[1,4,5], M.H. Guo[1,8,9,10], C. Benner[11,12], A.T. Satpathy[13], R. Salem[1,8,9,10], J.N. Hirschhorn[1,8,9,10], H.K. Finucane[1], M.J. Aryee[1,2,3], J.D. Buenrostro[1,14], V.G. Sankaran[1,4,5].* 1) The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; 2) Department of Biostatistics, Harvard T.H. Chan School of Public Health; 3) Department of Pathology, Massachusetts General Hospital; 4) Division of Hematology/Oncology, The Manton Center for Orphan Disease Research, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA; 5) Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA; 6) Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA; 7) Harvard-MIT Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA; 8) Division of Endocrinology, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115; 9) Department of Genetics, Harvard Medical School, Boston, MA 02115; 10) Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA 02115; 11) Institute for Molecular Medicine Finland, University of Helsinki, 00014 Helsinki, Finland; 12) Department of Public Health, University of Helsinki, 00014 Helsinki, Finland; 13) Department of Pathology, Stanford University School of Medicine, Stanford, CA, 94305; 14) Harvard Society of Fellows, Harvard University, Cambridge, MA 02138, USA.

Genome-wide association studies (GWAS) have identified thousands of genomic loci linked to complex phenotypes such as blood cell traits, but a major challenge has been the identification of causal genetic variants and relevant cell types underlying the observed phenotypic associations. In order to identify causal variants across 16 blood cell traits, we performed genome-wide association studies in the UK Biobank and fine-mapped 2,178 genome-wide significant regions allowing for up to 5 causal variants per region. Overall, we identified 36,919 variants with >1% posterior probability (PP) of being causal for a trait association and 1,110 regions (51%) contained at least one variant with >50% PP. We observed strong evidence of multiple signals occurring in a single region, as the posterior expected number of independent causal variants was >2 for 36% of regions and >3 for 13% of regions. In many instances, we determined that multiple signals were within 500 nucleotides. Strikingly, we found that when we tested variant pairs in open chromatin using reporter assays, each variant affected regulatory activity additively and the effect direction agreed with the direction of the phenotypic effects of individual diplotypes. Next, in order to identify relevant cell types, we developed and validated a new analytic approach (g-chromVAR) that takes into account both annotation strength (ATAC-seq peak counts) and variant confidence (fine-mapped posterior probability). We show that this approach has better control of false positives (*i.e.* lineage specificity) than other methods when applied to closely related cell types, such as those that comprise the hematopoietic tree, and further show that it can be applied to score GWAS enrichment at the single cell level. Applying g-chromVAR to chromatin accessibility profiles from 2,034 bone marrow-derived hematopoietic progenitors, we observe enrichments along pseudotime-inferred developmental trajectories and discover significant heterogeneity within classically defined common myeloid and megakaryocyte erythroid progenitors. Finally, we identify high confidence target genes and postulate consensus regulatory mechanisms for hundreds of fine-mapped variants in open chromatin. In several cases, we determined that pleiotropic fine-mapped variants act within progenitor-restricted regulatory elements. Our overall approach will allow for the interrogation of other complex traits and diseases at the single-variant and single-cell level.

249

Macromap: A systematic map of condition-specific genetic effects on expression in immune response based in induced Pluripotent Stem Cell (iPSC)-derived macrophages. *N.I. Panousis, A. Knights, C. Gomez, L. Boquete-Vilarino, A. Tsingene, D. Gaffney.* Wellcome Trust Sanger Institute, Hinxton, United Kingdom.

Immune disorders are triggered by irregular under or over activation of the immune system. Inappropriate overreaction of immune response can lead to autoimmunity while immunodeficiency increases susceptibility to infection. It is well established that genetic variation among individuals can modulate immune response and risk for certain immune related diseases. However, identification and characterisation of immune-modulating variants remains challenging because many have no observable impact of gene expression in naive cells. Additionally, mapping variant effects in more than a handful of stimulated cell states using primary cells is extremely challenging, due to the logistical problems in obtaining sufficient cells from a large cohort of individuals. Here, we use a high-throughput *in vitro* system based on a panel of human induced pluripotent stem cell-derived macrophages from 100 donors derived by the HIPSCI Consortium to study the effects of common variation on gene expression in 24 different cell states. Our stimulus panel included multiple immunologically active compounds, including agonists for a range of Toll Like Receptors (TLRs), compounds that mimic macrophage T-cell cross talk, chronic and acute inflammatory and metabolic stimuli, and multiple myeloid differentiation timepoints. We mapped hundreds of previously hidden response QTLs at the whole gene and exon level and use these to identify candidate causal genes for a wide range of diseases. We identify cases where disease risk alleles appear to cluster in particular response pathways and examine how variant effects propagate over time during cell activation. Our study highlights for the true diversity of condition-specific effects on expression across a broad range of states for a single cell type and provides a rich resource for dissecting the functions of common immune disease risk alleles.

Copyright © 2018 American Society of Human Genetics. All rights reserved

250

Blood cell genetic architecture and phenotype prediction. *D. Vuckovic[1], H. Ponstingl[1], P. Akbari[2,1], T. Jiang[2], W.J. Astle[2], A.S. Butterworth[2], M. Inouye[3,4,5], N. Soranzo[1,6].* 1) Department of Human Genetics, Wellcome Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; 2) MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 3) Cambridge Baker Systems Genomics Initiative, Cambridge, UK; 4) Dept Public Health & Primary Care, University of Cambridge, Cambridge, UK; 5) Baker Heart & Diabetes Institute, Melbourne, Australia; 6) Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.

Blood cells are essential for basic physiological processes such as oxygen transport and immune responses (Jensen et al., J of Exp Biol 2009; Jenne et al., Int J Lab Haematol 2013). Variation in quantity and quality of blood cells has been associated to severe disorders such as anaemias and bleeding (Routes et al., J Clin Immunol 2014), predisposition to cancer and systemic diseases (Schneider et al., Clin Genet 2015). Hence, understanding the causal link between genetic variation and blood indices can offer important insights in healthcare. Recent research highlighted the largest number of loci discovered by a GWAS and the allelic landscape of blood cell genetic variation (Astle et al., Cell 2016). Here we exploit large-scale population studies, i.e. the full release of UK Biobank (N~500k) and INTERVAL (N~40k) to discover up to 8,000 new associations, to explore properties of the allelic architecture and the prediction potential for blood cell composition. We fine-map 599 independent associated signals to a unique genetic variant with high posterior probability of causality (>95%). We observe a strong enrichment for coding variants, providing a direct link to the genes involved (OR=2.5, $p=3x10^{-11}$). We then explore the extent to which phenotypic variance is captured by genome-wide variation compared to a selection of SNPs determined from fine-mapping analysis of having high probability of causality. We use discovery estimates derived from the statistically powered UK Biobank to build genetic scores (GS) in a validation cohort (INTERVAL). We find that fine-mapping based GS is more predictive than genome-wide GS, showing up to 28% of phenotype correlation (Pearson's r) for red cell distribution width (after adjustment for covariates) compared to only ~23% for genome-wide LD-pruned GS. We further integrate these data with data from rare hematological diseases and cancers, with a view to partition disease-causing genes from genes underpinning normal phenotypic variation and to assess the assumptions of the recently proposed omnigenic model of disease. In conclusion, we now have the means to efficiently explore different models of genetic architecture for blood cell traits. These findings will not only complement the biological state-of-the-art knowledge about the haematopoietic system, but also help re-thinking diagnostic and prognostic criteria for blood-related diseases, taking us closer to a personalized healthcare informed by genetic variation.

251

High throughput characterization of genetic effects on DNA:protein binding and gene transcription. *C. Kalita[1], C. Brown[2], A. Freiman[1], X. Wen[3], F. Luca[1], R. Pique-Regi[1].* 1) Wayne State University, Detroit, MI; 2) University of Pennsylvania, Philadelphia, PA; 3) University of Michigan, Ann Arbor, MI.

The majority of the human genome is composed of non-coding regions containing regulatory elements such as enhancers, which are crucial for controlling gene expression. Many variants associated with complex traits are in these regions, and may contribute to an individual's phenotype by disrupting gene regulatory sequences. Consequently, it is important to not only identify functional enhancers, but also to test if a variant within a binding site affects gene regulation. We developed a new streamlined protocol, BiT-STARR-seq (Biallelic targeted STARR-seq), designed to identify allele-specific expression (ASE) while directly accounting for PCR duplicates through unique molecular identifiers (UMI) incorporation. We tested 75,501 oligos corresponding to 43,500 SNPs and identified 2,720 SNPs with significant ASE (FDR 10%). To validate disruption of binding as one of the mechanisms underlying ASE, we performed high throughput EMSA (BiT-BUNDLE-seq) for NFKB-p50. We tested the same oligo library used in BiT-STARR-seq and identified 2,951 SNPs with significant allele-specific binding (ASB) (FDR 10%). Of the SNPs with ASB, 173 also had ASE (OR=1.97, p-value=0.0006). When we focused on variants associated with complex traits in GWAS, we identified 1,531 SNPs with ASE in the BiT-STARR-seq and 1,662 SNPs with ASB in the BiT-BUNDLE-seq assay. We characterized the mechanism whereby the alternate allele for variant rs3810936 increases risk for Crohn's disease through increased NFkB binding and consequent altered gene expression. We are now using this high-throughput system to characterize gene regulatory sequences and allelic variants across multiple environmental contexts. We identified ASE in cells treated with: dexamethasone (723 ASE), retinoic acid (1272 ASE), selenium (102 ASE), and caffeine (201 ASE). This approach is ideal for testing the regulatory effect of rare genetic variants of pharmacogenetic importance for which standard QTL mapping would not have sufficient power..

Copyright © 2018 American Society of Human Genetics. All rights reserved

252

Identifying novel regulatory elements using RELICS, a statistical framework for the analysis of CRISPR regulatory screens. *P.C. Fiaux[1,2], H. Chen[2], I. Luthra[3], G. McVicker[2].* 1) Bioinformatics and Systems Biology Graduate Program, University of California San Diego, San Diego, CA, USA; 2) Salk Institute of Biological Studies, San Diego, CA, USA; 3) Simon Fraser University, Vancouver, BC, Canada.

High-throughput CRISPR/Cas9 screens are a powerful new tool for the systematic discovery of regulatory elements in the human genome. In these regulatory screens, thousands of guide RNAs (gRNAs) are delivered to cells to target potential regulatory sequences for mutation, activation or inhibition. The cells are then sorted into high- and low-expression pools based on the expression of a target gene. While, these screens have the potential to perform unbiased discovery of regulatory elements, they generate noisy data and the performance of analysis methods has not been rigorously assessed. Here we describe RELICS, a statistical framework for **R**egulatory **El**ement **I**dentification in **C**RISPR **S**creens. RELICS models the observed guide counts in different expression pools with a generalized linear mixed model. This approach is very flexible, can jointly model multiple expression pools (beyond just high and low), incorporate variability across guides, and accommodate over-dispersion. RELICS outperforms existing analysis methods on simulated data and we have applied it to identify regulatory elements in several published datasets. In addition, we have applied RELICS to data from a paired-guide regulatory screen that we performed for *GATA3* in Jurkat T cells. We identify a total of 44 putative regulatory elements (FDR < 0.1) within the 2MB targeted region surrounding *GATA3*. Notably 26 of the identified elements lie within the same topological associating domain as *GATA3*, but only 2 overlap enhancers predicted by ChromHMM.

253

Functional interpretation of genetic variants using deep learning predicts impact on epigenome. *G.E. Hoffman[1,2,3], P. Roussos[1,2,3,4].* 1) Pamela Sklar Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY; 3) Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY; 4) Mental Illness Research, Education, and Clinical Center (VISN 2 South), James J. Peters VA Medical Center, Bronx, NY, U.S.A.

Identifying causal variants underling disease risk and the adoption of personalized medicine are currently limited by the challenge of interpreting the functional consequences of genetic variants. Predicting the functional effects of disease- associated protein-coding variants is increasing routine. Yet the majority of risk variants are non-coding, and predicting the functional consequence and prioritizing variants for functional validation remains a major challenge. Non-coding variants contribute to disease risk by regulating gene expression, and this effect is driven in large part by the genetic regulation of transcription factor binding and histone modification. Predicting the functional effect of non-coding variants on the epigenome will distinguish benign variants from variants with the potential to confer disease risk. Deep convolutional neural networks have recently been used to develop predictive models linking the genome sequence to splicing, protein binding, and the discrete presence or absence of a signal from epigenomics assays. Here we introduce a deep learning framework for functional interpretation of genetic variants (DeepFIGV). We develop predictive models of quantitative epigenetic variation in DNA-seq and H3K27ac, H3K4me3, and H3K4me1 histone modifications from 75 lymphoblastoid cell lines (LCL). Modeling quantitative variation, integrating whole genome sequencing, and training the models on many experiments from the same cell type and assay improves the power to identify variants with functional effects on the epigenome. By integrating with external datasets, we demonstrate that variants which are QTLs for gene expression, chromatin accessibility and histone modifications are enriched for strong DeepFIGV scores. Moreover, disease risk variants, variants with allelic effects in a massively parallel reporter assay, rare variants near gene expression outliers, and somatic variants driving gene expression changes in cancer are also enriched for strong DeepFIGV scores. Nucleotide-level functional consequence scores for non-coding variants from DeepFIGV can refine the mechanism of known causal variants, identify novel risk variants and prioritize downstream experiments. We show how DeepFIGV can elucidate the functional mechanism variants conferring risk to inflammatory bowel disease. Finally, we have developed a public resource of the DeepFIGV predicted functional scores for 500 million variants.

Copyright © 2018 American Society of Human Genetics. All rights reserved

254

Disease heritability enrichment of regulatory elements is concentrated in elements with ancient sequence age and conserved function across species. *M.L.A. Hujoel[1], S. Gazal[2], F. Hormozdiari[2], B. van de Geijn[2], A.L. Price[1,2].* 1) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; 2) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA.

Regulatory elements, e.g. enhancers and promoters, are known to be enriched for disease and complex trait heritability. We investigated how this enrichment varies with both the age of the underlying genome sequence and the conservation of regulatory function across species. We constructed genomic annotations based on genome-wide alignments of 100 vertebrates (Marnetto et al. 2018 AJHG) and predicted enhancer or promoter function across 10 mammals (Villar et al. 2015 Cell). We estimated heritability enrichment by applying stratified LD score regression to summary statistics from 41 independent diseases and complex traits (average *N*=320K) and meta-analyzing results across traits. We reached 3 main conclusions. First, disease enrichment of human regulatory elements is concentrated in elements with older sequence age, assessed via alignment with other species irrespective of conserved functionality. Human enhancer elements with ancient sequence age (older than the split between marsupial and placental mammals; 16% of enhancers) were 8.8x enriched (compared to 2.5x for all human enhancers; P=3e-14 for difference), and human promoter elements with ancient sequence age (28% of promoters) were 13.5x enriched (compared to 5.1x for all human promoters; P=6e-16 for difference). Second, disease enrichment of human regulatory elements is larger in elements whose regulatory function is conserved across species. For example, human enhancers that are also enhancers in ≥5 of 9 other mammals (16% of enhancers) were 4.6x enriched (P=8e-12 for difference vs. all enhancers). Third, disease enrichment of human promoters is larger in promoters of loss-of-function (LoF) intolerant genes (3,230 genes that are strongly depleted for protein-truncating variants; ExAC Consortium, Lek et al. 2016 Nature; 16% of promoters): 12.0x enrichment (P=1e-14 for difference vs. all promoters). The mean value of quantitative measures of negative selection (nucleotide diversity, predicted allele age and McVicker B statistic) within these genomic annotations mirrored all of these findings. Notably, all of these heritability enrichments were jointly significant conditional on each other and on our baseline-LD model (Gazal et al. 2017 Nat Genet), which includes a broad set of coding, conserved, regulatory and LD-related annotations and quantitative measures of negative selection. Our findings further elucidate the role of regulatory elements in the genetic architecture of diseases and complex traits.

255

Rare disease diagnosis by integrating RNA sequencing in the Undiagnosed Diseases Network. *S. Chen[1*], H. Lee[2*], L. Fresard[4*], YC. Lee[1], A. Eskin[3], N. Hanchard[1], N. Stong[5], M. Wheeler[4], B. Dawson[1], D. Murdock[1], L. Burrage[1], J. Rosenfeld[1], V. Shashi[7], D. Bonner[8], C. Esteves[6], D. Goldstein[5], S. Montgomery[4], S. Nelson[2,3], B. Lee[1], Undiagnosed Diseases Network Members.* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Department of Pathology and Laboratory Medicine, University of California, Los Angeles, CA; 3) Department of Human Genetics, University of California, Los Angeles, CA; 4) Stanford University School of Medicine, Stanford, CA; 5) Columbia University Medical Center Institute for Genomic Medicine, New York, NY; 6) Harvard Medical School, Harvard University, Boston, MA; 7) Department of Pediatrics, Duck University, Durham, NC; 8) Center for Undiagnosed Diseases, Stanford University, Stanford, CA.

Next-generation sequencing has revolutionized how rare genetic disorders are diagnosed; however, most clinical laboratories report an overall diagnostic rate of 30-40%, reflecting a restricted focus of analysis that is largely limited to coding variation. To improve this diagnostic rate, the Undiagnosed Diseases Network (UDN) has made significant efforts to understand the effects of coding and non-coding variation through the application of RNA sequencing (RNA-seq). RNA-seq allows for investigation of the role of alternative splicing, allele-specific expression, nonsense-mediated decay and gene expression signatures in clinical diagnoses. RNA was obtained from patient and parental samples, including whole blood, skin fibroblasts, muscle, or bone marrow. Enriched polyA(+) mRNAs were sequenced by the Illumina platform generating ~30 million paired-end reads, which were aligned to hg19 reference genome followed by gene-based expression quantification. The junction reads were extracted and compared to those from controls to identify rare and abnormal junctions indicating alternative splicing. To date, UDN sites have analyzed 59 RNA-seq samples (32 whole blood, 24 skin fibroblasts, 2 muscle and 1 bone marrow) from 46 affected individuals. For 32 (69%) affected individuals, we were able to prioritize one or more variants that have a deleterious impact on a transcript. From a total of 120 variants analyzed through RNA-seq pipelines, we examined 83 (69%) splicing or intronic variants, 17 (14%) frameshift indel or stop-gain, 11 (9%) missense variants, 5 (4%) synonymous variants, 3 (3%) in-frame variants and 1 (1%) CNV. We found 38 (31%) variants induced anomalous transcripts including aberrant splicing (n=18), allelic-specific expression (n=3) or dysregulated mRNA degradation (n=16). Incorporating exome and/or genome sequencing analyses, we found 11 (28.9%) out of these 38 variants supported or identified the diagnostic genes (n=9) or strong candidates (n=2) for the diseases. Our findings suggest that probing non-coding human variation through RNA-seq has the potential to identify deleterious variants responsible for rare genetic disorders. The discovery of these candidate pathogenic variants yields a wealth of new diagnostic and therapeutic targets. Finally, RNAseq can complement candidate gene prioritization identified by either whole exome or whole genome sequencing.

Copyright © 2018 American Society of Human Genetics. All rights reserved

256

Clinical utility of combining blood-based monocyte assay and RNA sequencing with gene-panel testing: Higher diagnosis of 306 Dysferlinopathy-suspected US patients. *S. Chakravorty[1], S. Shenoy[1], K. Berger[2], D. Arafat[2], B. Nallamilli[1], L. Rufibach[3], S. Shira[3], G. Gibson[2], M. Hegde[1].* 1) Human Genetics, Emory University School of Medicine, Atlanta, GA; 2) Georgia Institute of Technology, Atlanta, GA; 3) Jain Foundation Inc., Seattle, WA.

Limb girdle muscular dystrophy type 2B (LGMD2B) and Miyoshi myopathy (MM), together called Dysferlinopathy, are two distinct autosomal recessive neuromuscular diseases caused by the same human *DYSF* gene, encoding a membrane repair protein called dysferlin. Dysferlinopathy, though one of the most common adult-onset myopathies, are underdiagnosed due to overlapping phenotypes with other myopathies and lack of assays using muscle biopsies which are painful, not preferred by patients. Therefore less invasive alternate methods are desired. Dysferlin is the only muscle protein found in blood monocytes that is shown to be affected similarly to muscle dysferlin in disease. Here we used a less-invasive blood monocyte assay for dysferlin protein estimation in 306 clinically-suspected dysferlinopathy patients recruited using automated LGMD diagnostic assistant (ALDA) tool across US, followed by NGS-based transcriptome sequencing (RNAseq) to re-classify the variants of uncertain significance (VUSs). We also performed LGMD 35 gene-panel testing. Our results show that combining monocyte assay and subsequent RNAseq along with genotype information provides a higher 75% diagnostic yield; compared to 36% and 37% when ALDA tool clinical prediction was combined with monocyte assay and RNAseq with or without genotype information respectively. This showed ALDA prediction, albeit cost-effective, is not a reliable diagnostic parameter for dysferlinopathy. We also identified correlation of dysferlin levels with segregation of a pathogenic intronic splice variant (c.4886+1249G>T) in a large family with dysferlinopathy history. Moreover, we identified multiple cases with dominant-like *DYSF* variants that cause pathogenic allelic expression imbalance (AEI) suggesting dysferlinopathy carrier-testing should be combined with functional assays. Interestingly, we identified 41 patients with reduced/absent Dysferlin but either with confirmed diagnosis of a LGMD gene other than *DYSF*, or harboring single pathogenic variants or VUSs in ≥ 2 LGMD genes (such as *DYSF* and *COL6A2*). This suggests the genotype-proteotype-phenotype spectrum is broader in dysferlinopathy and possibly includes multi-genic contribution. While variant detection provides the exact molecular diagnosis and associated genotype, we show that wherever possible a functional assessment such as monocyte assay and RNAseq in the clinical setting is required for true genotype:phenotype assessment to implement personalized medicine.

257

Integrating whole genome sequencing and RNA sequencing through allele specific expression analysis in the COPDGene study. *M.M. Parker[1], S. Lee[1], R.P. Chase[1], D. Qiao[1], E.K. Silverman[1,2], C.P. Hersh[1,2], M.H. Cho[1,2], P.J. Castaldi[1,3], NHLBI TOPMed Investigators, COPDGene Investigators.* 1) Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA; 2) Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA; 3) Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, Massachusetts, USA.

Background Multi-omics data integration is one of the major challenges of the precision medicine era. One method to integrate genetic and transcriptomic data is allele-specific expression (ASE) analysis, which quantifies the variation in gene expression observed between the two haplotypes of an individual. We aimed to assess the role of ASE in identifying functional variants from whole genome sequencing (WGS) and RNAseq. **Methods** To assess genome-wide ASE, we combined WGS from the TOPMed project and whole-blood RNA sequencing from 1,100 individuals in the COPDGene Study. For subjects passing standard quality control, allelic counts at heterozygous sites were quantified using GATK's ASEReadCounter. To define significant ASE, we performed binomial tests with a 5% FDR p-value correction across all heterozygous sites with at least 15x RNAseq coverage. Variant sites were annotated using WGS Annotator. Non-sense mediated decay (NMD) was predicted using SNPEff. **Results** Using WGS to define heterozygous variants for ASE resulted in > 12 million tested sites (mean sequencing coverage = 38x). This represents 2.16x and 2.77x more variants than using imputed array data in non-Hispanic White and African American subjects, respectively. Overall 3.9% of sites showed significant ASE, with 91,896 sites (0.7%) showing complete ASE. Variants predicted to trigger NMD were significantly more likely to show ASE as compared to nonsynonymous/synonymous (p-value< $6x10^{-8}$, proportion of NMD variants with ASE = 0.13, proportion of synonymous = 0.04, proportion of nonsynonymous = 0.03). Among the predicted NMD variants with significant ASE (n= 243), 21 were extremely rare (singletons/doubletons in TOPMed) and our ASE analysis confirmed their predicted function (ASE q-value < 0.05). Additionally, our analysis identified a well-characterized NMD-triggering variant (rs11549407 in the *HBB* gene) and confirmed its function (proportion of reads with reference allele= 99.3%, ASE q-value = $2.4x10^{-115}$) **Conclusions** We integrated WGS and RNAseq data in a large sample of COPDGene subjects. Our findings document extensive evidence of ASE in whole blood and confirm bioinformatic predictions of NMD in both known and novel variants. These findings may help determine the impact of rare variants in complex disease.

Copyright © 2018 American Society of Human Genetics. All rights reserved

258

OUTRIDER and FraseR: Statistical methods to detect aberrant events in RNA sequencing data. *C. Mertes[1], F. Brechtmann[1], A. Matusevičiūtė[1], V.A. Yépez[1,2], Z. Avsec[1,2], M. Herzog[1], D.M. Bader[1], L. Kremer[3], H. Prokisch[3,4], J. Gagneur[1,2].* 1) Department of Informatics, Technical University Munich, Boltzmannstr. 3, 85748 Garching, Germany; 2) Quantitative Biosciences Munich, Gene Center, Department of Biochemistry, Ludwig-Maximilians Universität München, Feodor-Lynen-Strasse 25, 81377 München, Germany; 3) Institute of Human Genetics, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany; 4) Institute of Human Genetics, Klinikum rechts der Isar, Technical University Munich, 13 Ismaninger Str. 22, 81675 München, Germany.

RNA sequencing (RNA-seq) is gaining popularity as a complementary assay to genome sequencing for precisely identifying the molecular causes of rare disorders. An obvious and powerful approach is to identify aberrant gene expression levels as potential pathogenic events. Using RNA-seq, another compelling strategy is the detection of alternative splicing events long known to cause Mendelian disorders. However, existing methods for detecting aberrant events in RNA-seq data either lack assessments of statistical significance, so that establishing cutoffs is arbitrary, or they rely on subjective manual corrections for confounders. Here, we present OUTRIDER (OUTlier in RNA-seq fInDER) and FraseR (Find Rare Aberrant Splicing Events in RNA-seq data), algorithms developed to address these issues. OUTRIDER uses an autoencoder to model read count expectations according to the co-variation among genes resulting from technical, environmental, or common genetic variations. Given these expectations, the RNA-seq read counts are assumed to follow a negative binomial distribution with a gene-specific dispersion. Likewise, FraseR controls for co-variation by utilizing an autoencoder and models the spliced reads by a beta binomial distribution with junction specific shape parameters. By adapting the percent spliced in metric to unspliced reads FraseR is able to call intron retention based on the same models. Outliers are then identified as read counts that significantly deviate from the respective distribution. The autoencoder models are automatically fitted to achieve the best correction of artificially corrupted data. Precision-recall analyses using simulated outlier read counts demonstrated the importance of correction for co-variation and of significance-based thresholds. Additionally, we show that both methods were able to recall all known disease-causing events within a previously analyzed rare disease cohort. The R packages OUTRIDER and FraseR include functions (i) for filtering out genes not expressed in a data set, (ii) for identifying outlier samples with too many aberrantly expressed genes or splice junctions, and (iii) for the P-value-based detection of aberrant events, with false discovery rate adjustment. Overall, OUTRIDER and FraseR provide computationally fast and scalable end-to-end solutions for identifying aberrant gene expression, alternative splicing, and intron retention events, suitable for the use by rare disease diagnostic platforms.

259

Biological factors strongly impact cell type abundances in human tissues. *M. Oliva[1,2], B.E. Stranger[1,2], S. Kim-Hellmuth[3], F. Aguet[4].* 1) Institute for Genomics and Systems Biology , The University of Chicago, Chicago, IL; 2) Department of Genetic Medicine, The University of Chicago, Chicago, IL; 3) New York Genome Center, New York, NY; 4) The Broad Institute, Cambridge, MA.

Human tissues are complex and composed of multiple cell types. Accurate tissue cell-type heterogeneity profiles and knowledge of the extent to which they vary with biological contexts (e.g. sex and age) are lacking for low-abundance cell types and non-blood tissues. Here, we apply computational cell-type deconvolution methods to the comprehensive catalogue of tissue expression profiles generated by the Genotype-Tissue Expression (GTEx) Consortium (17,382 RNA-seq samples from 838 individuals, v8 release) to characterize cellular composition across human tissues. Cell-type enrichment analysis was conducted to quantify the inter-sample variability of 64 immune and non-immune cell types in 49 GTEx tissues and identify numerous tissue-specific associations between cellular abundances and biological factors (donor sex, age and body mass index (BMI)). Donor sex is a major determinant of breast cell type composition. Stromal cells (smooth muscle, endothelial) are 7-27% more abundant in males, and epithelial cells are 500% more abundant in females. Donor age is associated with cellular composition of several tissues; in tibial artery, osteoblasts decrease and chondrocytes increase with age reflecting bone-vascular interactions that shape the artery structure. The immune profile is strongly affected too; we observe neutrophils decreasing in blood and spleen and eosinophils increasing in tibial nerve significantly with age. Aging also shapes adipose tissue: adipocyte content is negatively correlated with age even after adjusting for BMI, although this observation might also be compatible with low adipocyte transcriptional activity or larger adipocytes in older tissue. We find an independent impact of donor BMI on adipose tissue composition, particularly on the immune profile which include previously-described and novel obesity-associated effects. Interestingly, several of the age-related trends are non-linear after 50-60 years and/or manifest in a sex-specific manner, potentially unraveling post-menopause or late-life tissue composition changes. Importantly, many of the observed trends replicate using an orthogonal method and in independent datasets. This work characterizes biological factors associated with tissue cell-type heterogeneity and provides a resource of age-, sex-, and tissue-specific cell-type abundance profiles for the interpretation of tissue composition changes in the context of health and disease.

Copyright © 2018 American Society of Human Genetics. All rights reserved

260

Using multi-tissue gene expression to estimate individual- and cell-type-specific gene expression via deconvolution. *J. Wang[1], B. Devlin[2], K. Roeder[1].* 1) Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA; 2) Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA.

Quantification of gene expression in tissues can be a critical step towards characterizing the etiology of complex phenotypes and diseases. For instance, such a characterization, often called bulk tissue gene expression data, has been key to identifying expression quantitative trait loci (eQTLs), which in turn have been linked to disease risk. Hence many data sets from bulk tissue have been generated. To understand etiology fully, however, cell-level gene expression is arguably more informative than bulk tissue and indeed many recent studies have sought to characterize expression at the level of cell types. This is potentially critical for brain tissue, which harbors myriad cell types whose function is not fully resolved. While there are many strengths to cell-specific expression, the data tend to be expensive and quite noisy. Moreover, the cells can be difficult to access, especially from the brain. We wondered if it might be possible to estimate individual- and cell-type-specific gene expression from existing bulk tissue resources, specifically multi-tissue gene expression derived from the same subjects. In this work, we develop a new deconvolution method that borrows information across multiple tissue types collected from the same individuals. This enables the estimation of individual-level cell-type-specific gene expression for a large number of individuals. We first estimate the cell type composition by integrating bulk tissue data with reference samples of purified-cells or single-cell expression data. Then, we assume the cell-type-specific gene expression to be random and calculate their empirical Bayes estimates through a computationally efficient EM (Expectation-Maximization) algorithm. Simulations and data analyses demonstrate the advantages of our novel method, which is used to analyze multiple brain tissue types from the GTEx (Genotype-Tissue Expression) project and the BrainSpan atlas of the developing human brain. The results complement single-cell expression data and provide new insights into the interpretation of bulk tissue expression data, for example, via cell-type-specific analysis of eQTLs and co-expression networks. By aggregating information from genes associated with autism spectrum disorder (ASD), we find that ASD genes are more likely to have eQTLs in the pyramidal neurons and have more connections in the co-expression network of the fetal quiescent newborn neurons than non-ASD genes.

261

Addressing confounding artifacts in reconstruction of gene co-expression networks. *P. Parsana[1], C. Ruberman[2], A.E. Jafee[2,3,4,5,6], M.C. Schatz[1,7], A. Battle[8], J.T. Leek[2,6].* 1) Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA; 2) Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; 3) Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, USA; 4) Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; 5) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 6) Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA; 7) Department of Biology, Johns Hopkins University, Baltimore, MD, USA; 8) Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA.

Gene co-expression networks can capture biological relationships between genes, and are important tools for predicting gene function and understanding disease mechanism. True *in vivo* interactions between genes are not fully characterized for most organisms, and therefore reconstruction of co-expression networks is of interest in many contexts including diverse tissues and diseases. However, accurate reconstruction of networks remains a challenging problem. Widely used network learning methods such as weighted gene correlation network analysis (WGCNA) and graphical lasso typically do not account for batch effects or other confounding factors known to routinely affect gene expression measurements. Correlations introduced by these confounders are therefore incorrectly inferred as biological relationships between genes, leading to inaccurate networks and erroneous biological conclusions. Many studies do not employ any form of correction of gene expression data prior to network reconstruction, and there is not a standard, widely-adopted approach for doing so. In this study we show that network reconstruction algorithms are susceptible to known technical and latent confounders. We first demonstrate theoretically that principal component correction of gene expression measurements prior to network inference can reduce false discoveries. Then, using gene expression data from the GTEx project in diverse tissues we show that this approach improves precision in networks reconstructed both with WGCNA and graphical lasso, especially when extensive records of potential confounders are not available. We find that improved network reconstruction is more evident in tissues where known artifacts explain a large proportion of expression variance but accounting for commonly recorded factors such as RIN, and exonic rate is insufficient. When a wide-range of annotated technical factors are measured and available, correcting gene expression data jointly with multiple covariates (batch, sequencing, mapping artifacts) also improves network reconstruction. Our study demonstrates that it is critical to correct gene expression for unwanted variation before applying standard network learning methods. We show that principal component correction is a straightforward and effective approach for removing artifacts and improving network reconstruction, particularly when sample covariate annotations are limited.

Copyright © 2018 American Society of Human Genetics. All rights reserved

262

Brain region- and developmental time-specific genes are enriched in psychiatric genetics signals. *C. Jiao[1], QT. Meng[1], KL. Wang[1], C. Chen[1,2], CY. Liu[1,3].* 1) Central South University, Changsha, Hunan, China; 2) National Clinical Research Center for Geriatric Disorders, Central South University, Changsha, Hunan, China; 3) Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA.

Objective: Psychiatric disorders such as autism, schizophrenia (SCZ) have been related to early brain development and particular brain regions. To fully understand the relationship between psychiatric disorders and developmental stages and brain regions, we re-analyzed the existing BrainSpan data of spatiotemporal gene expression profiling in normal human brain to identify genes highly expressed in special developmental period or brain regions, and their relationships with psychiatric disorders. **Materials and Methods:** We analyzed the published BrainSpan data from 468 samples of 40 postmortem brains, which includes 16 brain regions from 8PCW (Post Conceptual Week) to 40yrs (*HJ Kang et al. Nature, 2011*). After data normalization using TPM (Transcripts Per Million, log2 changed), we performed spatiotemporal differential expression analysis (FDR<0.05 and |Fold Change|>2), spatiotemporal specific high-expression analysis, and weighted gene co-expression network analysis to detect the brain region- and developmental time-specific genes. GO, KEGG, PPI network enrichment, and Molecular complex Detection analysis were applied to explore their primary functions and molecular complexes. Additional enrichment analysis was performed to test whether above-listed time- or region-specific genes are associated with risk genes of different psychiatric disorders. **Results:** We found genes differentially expressed between three adjacent periods enriched for psychiatric risk genes: P2-P3 (8~10 vs. 12~13PCW), P6-P7 (19~24 vs. 24~38PCW), and P9-P10 (6~12M vs. 1~6yrs). These periods are respectively related to neurons differentiation and migration, gliogenesis, and synaptic pruning. We also observed protein complexes of CENPM, C4B, and CHRM4 (SCZ GWAS signals) were respectively detected as P2, P7, and striatum specific high-expression genes. For particular disorders, major depression signals mainly enriched in P6, P7, P8 stages, and striatum, orbital prefrontal cortex regions. Bipolar disorder signals enriched in midfetal and late infancy to adolescence stages, and frontal cortex, temporal cortex regions. Gene co-expression implicates a network relevant for these psychiatric disorders. The hub genes in network are *ADNP2*, *CLU*, *ARID2*, *GABRD*, *FAT4*, and *SLCO5A1*, which are closely connected with genes highly related to development. **Conclusion:** Our study distinctly presents a systematic landscape of associated network of psychiatric disorders with specific regions and developmental times.

263

Morning sickness and hyperemesis gravidarum: Genetic and functional analysis of placenta and appetite hormone GDF15 supports causality. *M.S. Fejzo[1,2], P.A. Fasching[3], M.O. Schneider[3], J. Schwitulla[3], M.W. Beckmann[3], E. Schwenke[3], D. Arzy[1], R. Tian[1], K.W. MacGibbon[4], P.M. Mullin[2].* 1) Division of Hematology-Oncology, David Geffen School of Medicine, Jonsson Comprehensive Cancer Center, University of California at Los Angeles, Los Angeles, CA, USA; 2) Department of Maternal-Fetal Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; 3) Department of Gynecology and Obstetrics, University Hospital Erlangen, Erlangen, Germany; 4) Hyperemesis Education and Research Foundation, Damascus, OR, USA.

Most women experience nausea and vomiting of pregnancy (NVP) and 18% take medication to treat it. The most severe form, Hyperemesis Gravidarum (HG), occurs in 2% of pregnant women. HG is associated with weight loss, electrolyte imbalance, and ketonuria. It accounts for 285,000 hospitalizations in the US annually and can cause Wernicke's encephalopathy, renal and liver abnormalities, esophageal rupture, and post-partum post-traumatic stress. HG is associated with a 4-fold increased risk of adverse fetal outcome including low birth weight, IUGR, preterm birth, fetal and neonatal death, and a 3-fold increased risk of neurodevelopmental delay. It is highly heritable. Our recent GWAS identified the placenta and appetite gene *GDF15* as the strongest genetic signal ($2.4x10^{-41}$) associated with NVP and HG. GDF15 is expressed at the highest levels in the placenta, is believed to play important roles in implantation and placental development, and decreases prior to miscarriage. Herein we show GDF15 serum levels are significantly higher in women with HG compared to controls at 12 weeks, and are no longer significantly different at 24 weeks, when symptoms typically decline. Whole-exome sequencing of 5 HG families revealed risk variants segregated with disease. In addition, the risk allele is significantly linked to having a recurrence. Finally, the GDF15 receptor gene, *GFRAL*, is also significantly linked to HG, providing further evidence implicating this pathway in the etiology of HG. GDF15 drives cancer cachexia, a condition with similar symptoms to HG, and which kills 20% of cancer patients. Lerner et al. blocked GDF15 in an animal model of cancer cachexia and restored body weight and appetite, making this a promising strategy for treating HG. Conversely, GDF15 analogs are under development by Amgen Inc. and others to treat obesity, and have effectively reduced appetite and caused weight loss in animal models. It has long been assumed that the pregnancy hormone hCG is the cause of morning sickness and HG. Our study provides no evidence to support this and instead, suggests a major role for the hormone GDF15. Its dual roles in placentation and appetite provide a molecular explanation for the age-old paradox as to why NVP, which can reduce reproductive fitness, has not been selected out in nature. This work paves the way for a promising new area for research into the development of tools for prediction, diagnosis, and treatment for those suffering from NVP and HG.

Copyright © 2018 American Society of Human Genetics. All rights reserved

264

Causative genes and mechanism of androgenetic hydatidiform moles. *N.M.P. Nguyen[1], Z. Ge[1], R. Reddy[1], S. Fahiminiya[1,2], P. Sauthier[3], R. Bagga[4], F. Sahin[5], S. Mahadevan[6], M. Osmond[1,2], M. Breguet[3], K. Rahimi[7], L. Lapensee[8], K. Hovanes[9], R. Srinivasan[10], I. Van den Veyver[6,11], T. Sahoo[9], A. Ao[12], J. Majewski[1,2], T. Taketo[13,14], R. Sllim[1,12].* 1) Department of Human Genetics, McGill University Health Centre, Montréal, Québec, Canada; 2) Genome Québec Innovation Center, Montréal, Québec, Canada; 3) Department of Obstetrics and Gynecology, Centre Hospitalier de l'Université de Montréal (CHUM), Centre Intégré de Cancérologie du CHUM, Registre des Maladies Trophoblastiques du Québec; 4) Department of Obstetrics & Gynecology, Post Graduate Institute of Medical, Education and Research, PGIMER, Chandigarh, India; 5) Department of Medical Genetics, Faculty of Medicine, Baskent University, Ankara, Turkey; 6) Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston, Texas, 77030, USA; 7) Department of Pathology, Centre Hospitalier de l'Université de Montréal, Montréal, Québec, Canada; 8) Ovo Clinic, Montréal, Québec, Canada; 9) Invitae, Irvine, CA 92618, USA; 10) Cytology & Gynecological Pathology, Post Graduate Institute of Medical Education and Research PGIMER, Chandigarh, India; 11) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, 77030, USA; 12) Department of Obstetrics and Gynecology, McGill University Health Center, Montréal, QC, Canada; 13) Department of Surgery, 14Department of Biology, McGill University, Montréal, Québec, Canada; 14) Department of Biology, McGill University, Montréal, Québec, Canada.

Hydatidiform mole (HM) is an aberrant human pregnancy that manifests in the first trimester and is characterized by early embryonic arrest and excessive trophoblastic proliferation. Recurrent hydatidiform moles (RHM) are defined by the occurrence of at least two hydatidiform moles in the same patient. Three genotypic types of HM have been found to recur: diploid biparental (one copy of the maternal genome and 1 copy of the paternal genome), diploid androgenetic (2 copies of the paternal genome and no maternal genome), triploid dispermic (2 copies of the paternal and 1 copy of the maternal genomes). There are two genes, *NLRP7* and *KHDC3L*, responsible for RHM and these two genes explain the etiology of RHM in 60% of patients. HM tissues from these patients are all diploid biparental. However, no genes have been found for patients with recurrent androgenetic or triploid dispermic HM. To identify novel genes responsible for RHM, we performed whole exome sequencing on patients with RHM and then targeted sequencing of candidate genes in larger cohorts of patients with milder phenotypes. We identified biallelic deleterious mutations in three genes, *MEI1*, *TOP6BL*/*C11orf80*, and *REC114*, with roles in meiotic double-strand breaks formation in five unrelated patients and three affected siblings with recurrent androgenetic HM, miscarriages, and infertility. We demonstrated that their HM tissues have androgenetic monospermic genomes. All three genes are highly conserved during evolution and known to play roles in homologous chromosome pairing and recombination in mice. Mutations in two out of the three identified genes were previously reported to cause infertility in both male and female mice. We investigated the possible occurrence of androgenesis in *Mei1*-decificient mice and discovered that *Mei1* null mice produce "empty" oocytes that lose all their chromosomes by extruding them into the first polar body. We demonstrate that *Mei1* null oocytes are capable of fertilization and lead to androgenetic zygotes. Thus, we identified three novel causative genes for recurrent androgenetic HM, miscarriages, infertility in humans and uncovered, for the first time in mammals, a mechanism for the genesis of androgenetic zygotes.

265

Expanded targeted exome sequencing in fetuses with ultrasound abnormalities reveals an important fraction of cases with associated gene defects. *C.G. Pangalos[1,2], B. Hagnefelt[1], S. Karapanou[1], C. Konialis[1,3].* 1) InterGenetics, Athens, Attiki, Greece; 2) Genomis Ltd, London, UK; 3) ClinGenics Ltd, London, UK.

Objective: Abnormal ultrasound findings are detected in 3-5% of all pregnancies, 10-15% of which are due to the presence of chromosomal abnormalities, while in majority of the remaining cases the genetic cause remains undiagnosed, leading to an inability to provide a precise diagnosis and accurate reproductive and fetal risk assessment. Starting in 2015 we designed and implemented an expanded exome sequencing-based test targeting gene disorders presenting with abnormal prenatal ultrasound findings. We present and discuss the latest data from the overall application of this approach (the *Fetalis*® test) in 51 pregnancies with ultrasound findings, leading to the diagnosis of many complex and unsuspected genetic diseases in the embryos. **Methods:** Testing was preceded by consultation with the attending physician and by genetic counseling and informed consent of the parents. All cases involved euploid fetuses ascertained by prior prenatal CMA. Fetal DNA was subjected to Next Generation Sequencing (NGS), followed by variant prioritization utilizing a custom analysis pipeline-algorithm targeting ~760 genes associated with known genetic disorders which may present with abnormal fetal ultrasound findings. Variant reporting included only known pathogenic or obligatory pathogenic gene mutations, overlooking other types of variants. **Results:** Exome sequencing results were available within 7±3 days. Overall, pathogenic mutations associated with a known genetic disorder were detected in 37% of cases (19/51) involving 3 abortuses and 16 on-going pregnancies. No pathogenic mutations were detected in 32 cases and in at least 18 of these cases an apparently healthy child was born. The genetic disorders diagnosed in the affected fetuses included Noonan syndrome, Nemaline myopathy, X-linked myopathy with excessive autophagy, Bartter syndrome, congenital myasthenic syndrome, etc. **Conclusion:** The expanded targeted exome sequencing-based approach described herein indicates that at least 40% of euploid fetuses with troubling sonographic abnormalities harbor gene mutations associated with known debilitating genetic disorders. It should be stressed that this targeted testing strategy overcomes the problems and limitations associated with clinical wide-scale WES testing in a prenatal setting, by reporting only highly confident clinically actionable results and avoiding the problems associated with the interpretation and communication of uncertain findings in the course of pregnancy.

Copyright © 2018 American Society of Human Genetics. All rights reserved

266

Pre-natal inflammation primes the neonatal intestine for endotoxin tolerance. *E. Banfield[1], W. Fulton[2], I. Burd[3], C. Sodhi[2], D. Hackam[2].* 1) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) Department of Surgery, Johns Hopkins University School of Medicine, Baltimore MD; 3) Department of Gynecology and Obstetrics, Johns Hopkins University School of Medicine, Baltimore, MD.

Neonatal sepsis occurs in roughly 9.7/1000 live births in the United States. The risk of sepsis drastically reduces as children age and their immune systems mature, suggesting that augmenting immune maturity may prevent neonatal sepsis. Endotoxin tolerance, in which an initial exposure to bacterial endotoxin leads to hypo-responsiveness to subsequent exposures, is associated with immune protection in adults yet whether this occurs in neonates remains unknown. This is especially true in the context of prenatal infection and subsequent postnatal exposure. We hypothesize that in utero exposure to bacterial endotoxin will lead to increased inflammatory responsiveness and decreased cell proliferation and healing of the neonatal intestine during subsequent bacterial exposure, leaving the neonate at high risk of secondary infection. To mimic endotoxin tolerance, we injected 2 mg/kg of endotoxin, also termed lipopolysaccharide (LPS), prenatally into the uterine horns of pregnant CD-1 mice, inducing an inflammatory response mimicking bacterial infection. We then injected LPS at postnatal day 1 and collected tissue for analysis at P2. Analyses included qPCR for pro-inflammatory cytokines as well as immunohistochemistry for cell proliferation. Relevant controls (e.g. saline injection) were used at both time points. Our immunohistochemistry results showed a reduction in cell proliferation in the pre- and postnatal LPS group compared to all control groups. Additionally, we saw a significant increase in iNOS and IL6, proinflammatory cytokines involved in immunity, expression between the group given pre- and postnatal LPS injections and groups given only prenatal LPS, prenatal LPS and postnatal PBS, prenatal PBS, and postnatal LPS, respectively (iNOS p=0.04, 0.02; IL6 p=0.01, 0.01). We saw the most significant differences in Lipocalin 2 expression, a gene involved in innate immunity, between the pre- and postnatal LPS and prenatal PBS, prenatal LPS, and prenatal LPS and postnatal PBS groups (p=0.001, 0.006, and 0.006, respectively). These results demonstrate that in the presence of prior exposure to LPS, there is a reduction in cell proliferation and a significant increase in expression of proinflammatory cytokines and Lipocalin 2 when a second exposure to endotoxin occurs early in life. This suggests that in utero exposure to inflammation primes the intestine for hypo-responsiveness to further inflammation after birth, increasing the likelihood of secondary infection.

267

Predicting incident cardiometabolic and cancer events using highly polygenic risk scores. *N. Mars[1], J. Koskela[1], P. Ripatti[1], T. Kiiskinen[1], A.S. Havulinna[1,2], L. Groop[1,3], A. Palotie[1,4,5,6], M. Daly[1,4,5,6], V. Salomaa[2], E. Widén[1], S. Ripatti[1,4,7].* 1) Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Helsinki, Finland; 2) National Institute for Health and Welfare, Helsinki, Finland; 3) Diabetes and Endocrinology, Lund University Diabetes Centre, Malmö, Sweden; 4) Broad Institute of MIT and Harvard, Cambridge, MA, USA; 5) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 6) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 7) Department of Public Health, University of Helsinki, Helsinki, Finland.

The true clinical utility of polygenic risk scores (PRS) cannot be evaluated without considering them jointly with routinely used clinical risk factors. Therefore, we sought to evaluate how well the comprehensive genetic profiles predict incident disease cases beyond the routinely used clinical prediction tools. We utilized Finnish population-based biobank data with genome-wide genotyping and decades of nationwide hospitalization, prescription drug purchase and causes of death information to prospectively capture common complex disease events, including coronary heart disease (CHD), type 2 diabetes (T2D), and breast and prostate cancer. We estimated how well PRSes based on >6M common variants predicted incident disease events. High polygenic risk was associated with developing the disease not only in CHD (HR=3.31, 95% CI [2.76,3.98] for lowest versus highest quintile, $p<2\times10^{-16}$), but also in breast cancer (HR=2.46, [1.77,3.43], $p=1.02\times10^{-7}$), prostate cancer (HR=2.93, $p=2.87\times10^{-11}$), and T2D (HR=3.18, [2.76,3.66], $p<2\times10^{-16}$). In the highest PRS quintile, one in three had developed T2D by age 65, and in those above the 97.5th percentile, one in two. For T2D, adding BMI to the regression model had no effect on the dose response of T2D PRS. Self-reported first-degree family history of diabetes, however, reduced the risk conferred by the PRS by ~10% (HR=2.92, [2.47,3.44], $p<2\times10^{-16}$ for T2D PRS adjusted with family history). For T2D, with 20-80% as the reference, there was a strong protective association in very low-risk individuals (<2.5th percentile, HR=0.39, $p=6.04\times10^{-6}$) and a risk increase in high-risk individuals (>97.5th, HR=2.96, $p<2\times10^{-16}$). We observed similar results for CHD (HR=0.44, [0.28,0.71], p=0.0007 and HR=2.94, $p<2\times10^{-16}$). For CHD, adjusting for the clinical ASCVD risk score had no effect on CHD PRS (HR 3.30, [2.74,3.96], $p<2\times10^{-16}$), while self-reported family history of myocardial infarction had only a minor effect (HR 3.18, [2.64,3.82], $p<2\times10^{-16}$). In the highest PRS quintile for CHD, similar to average-risk individuals, 21.8% had BMI≥30kg/m^2 at baseline, 76.1% had hypercholesterolemia and 40.2% hypertension. For several common diseases, high polygenic risk was associated with developing the disease and the risk appeared to be independent of the many clinical risk factors. Polygenic risk scores complement clinical risk assessment and provide important new tools to identify individuals at particularly high or low risk for a given disease.

Copyright © 2018 American Society of Human Genetics. All rights reserved

268

Genome-wide association meta-analysis identifies 12 common risk alleles for Brugada syndrome, a rare cardiac arrhythmia disorder. *R. Redon[1], J. Barc[1], R. Tadros[2], C. Dina[1], F. Manevy[2], J.B. Gourraud[1], Y. Mizusawa[2], K.E. Odening[3], C. Antzelevitch[4], M.P. van den Berg[5], M.M. Borggrefe[6], P.D. Lambiase[7], J. Saenen[8], E. Schulze-Bahr[9], T. Robyns[10], O. Campuzano[11], E. Arbelo[12], A. Leenhardt[13], S.G. Priori[14], L. Crotti[15], P. Mabo[16], C. de Asmundis[17], D.F. Giachino[18], E. Behr[19], M. Haissaguerre[20], J.R. Gimeno[21], J.J. Schott[1], V. Probst[1], A.A.M. Wilde[2], C.R. Bezzina[2], The Rythmogene Working Group, The BrS Genetics Consortium.* 1) L'institut du thorax, Inserm, CNRS, Univ. Nantes, CHU Nantes, Nantes, France; 2) Academic Medical Center, Amsterdam, Netherlands; 3) University Heart Center Freiburg, Freiburg, Germany; 4) Lankenau Institute for Medical Research, Wynnewood, PA; 5) University Medical Center Groningen, Groningen, Netherlands; 6) Univ of Heidelberg, I Medizin Klinik, Mannheim, Germany; 7) The Heart Hospital, Cardiology, Edgware, United Kingdom; 8) UZA, Edegem, Belgium; 9) Institute for Genetics of Heart Diseases, Muenster, Germany; 10) University of Leuven, Leuven, Belgium; 11) University of Girona-IDIBGI, Girona, Spain; 12) Hospital Clinic De Barcelon, Cardiology Department, Barcelona, Spain; 13) Bichat University Hospital, France, France; 14) University of Pavia & ICS Maugeri, Pavia, Italy; 15) Center for Cardiac Arrhythmias of Genetic Origin, Milan, Italy; 16) CHU Rennes, Rennes, France; 17) Univ Hospital of Brussels, Dept of Cardiology, Brussels, Belgium; 18) University of Torino, Department of Clinical and Biological Sciences, Torino, Italy; 19) St George's University of London, London, United Kingdom; 20) CHU Bordeaux, IHU Liryc, Bordeaux, France; 21) Virgen de la Arrixaca University Hospital, Cardiology Department, Murcia, Spain.

The Brugada syndrome (BrS) is an inherited cardiac disorder associated with high risk for sudden cardiac death. Rare genetic variants in the *SCN5A genes* are causally related to ~20% of cases. By a genome-wide association study (GWAS) conducted on 312 index cases with BrS, we had previously identified three common risk alleles at the *SCN5A-SCN10A* and *HEY2 loci.* Here, aiming at dissecting further the genetic architecture underlying BrS, we have conducted a case-control GWAS in an extended set of 2,112 BrS patients of European descent *versus* 6,731 ancestry-matched controls. We could identify 12 independent risk haplotypes reaching genome-wide significance ($p<5x10^{-8}$). The 3 previously reported signals at *SCN5A-SCN10A* (rs10428132; rs4130467) and *HEY2* (rs980014) were confirmed. Conditional analysis at the *SCN5A-SCN10A* locus revealed 5 additional independent signals, while four additional hits were detected at new loci harbouring genes coding for transcription factors expressed during cardiac development. Overall, the estimated odds ratios (ORs) range between 1.26 and 2.43 per risk allele. As a result, we observed an unexpectedly high cumulative effect size for the 24 risk alleles on BrS susceptibility, indicating putative clinical utility of the resulting polygenic risk score. The existence of multiple independent risk haplotypes near *SCN5A* further underscores its preponderant role in BrS pathophysiology, while all novel loci indicate the role of transcriptional regulation. Replication analysis of the novel hits, as well as 12 additional signals displaying sub-threshold statistical significance ($p<1x10^{-6}$), is in progress. We acknowledge the following consortia for providing control sample sets: GoNL, KORA, UK10K, D.E.S.I.R., Project Mine, Psychiatric Genetics Unit (Barcelona, ES) and Hypergenes.

269

High-throughput functional genomics of cardiac sodium channel variants. *A.M. Glazer[1], B.M. Kroncke[1], K.A. Matreyek[2], T. Yang[1], J. Salem[1], D.M. Fowler[2], D.M. Roden[1].* 1) Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; 2) Department of Genome Sciences, University of Washington, Seattle, WA.

Ion channels are an important class of disease genes; when mutated they can cause a broad spectrum of phenotypes, including cardiac, neuronal, and muscular diseases. Ion channel variants are conventionally studied in low to medium throughput with patch clamp electrophysiology. We report here the development and validation of a Deep Mutational Scanning (DMS) method to simultaneously characterize hundreds or thousands of variants in *SCN5A*, which encodes the main voltage-gated cardiac sodium channel. Loss of function and gain of function variants in *SCN5A* can cause Brugada Syndrome and Long QT Syndrome, respectively, and the method was designed to identify both functional defects. A mixture of three drugs (veratridine, brevetoxin, and ouabain) was used to promote lethal sodium influx and discriminate between cells with normal, loss of function, and gain of function channels. When tested on previously characterized variants in HEK293T cells, the drug challenge successfully discriminated between wild-type channels (35% normalized cell survival) and four previously studied variants: a nonsense variant (W822X, 98% normalized cell survival), a partial loss of function variant (G752R, 71%), a gain of function variant (N1325S, 16%), and a variant with wild-type like function (R620H, 38%). We then created a comprehensive mutant library (98% of possible mutants) in a 12 amino acid/36 base pair region of *SCN5A* and stably integrated the library into HEK293T cells, with one variant expressed per cell. High-throughput sequencing of the pre- and post-drug challenge pools was used to count the prevalence of each variant and thereby identify those with abnormal function. The DMS functional scores strongly distinguished nonsense from synonymous variants ($p<0.001$). The scan also identified dozens of missense variants with predicted abnormal function. The scan identified V1624E as a complete loss of function variant and V1624S as a gain of function variant, and these were validated by patch clamping. These experiments demonstrate a method for a high-throughput *in vitro* screen to detect both loss and gain of function variants in *SCN5A* and potentially other sodium channel genes.

Copyright © 2018 American Society of Human Genetics. All rights reserved

270

Receiving personal genome-based disease risk information motivates individuals to take action to prevent cardiovascular disease (CVD). *E. Widen[1], J. Aro[1], P. Pollanen[2], K. Hotakainen[3], J. Partanen[4], S. Ripatti[1].* 1) Institute for Molecular Medicine Finland, Helsinki, University of Helsinki, Finland; 2) CAREA Kymenlaakso Social and Health Care Services, Kotka, Finland; 3) Mehiläinen Oy, Helsinki; 4) Finnish Red Cross Blood Service, Helsinki, Finland.

The systematic use of genomics to guide clinical decision is expected to significantly improve both risk assessment and prevention of common multifactorial disease. Hitherto this opportunity remains largely untapped, partly due to insufficient knowledge of the underlying genetic architecture, but also due to lack of tools for interpreting and communicating complex genomic information. To facilitate the translation of recent genomic discoveries and to empower individuals to use this information we have developed a novel interactive graphical interface, KardioKompassi®, to help patients and doctors to estimate an individual's 10-year CVD risk based both on traditional and genomic risk data (polygenic risk score (PRS) from ~49,000 common genetic variants). We collected a prospective study cohort, GeneRISK, including 7,328 randomly sampled middle-aged individuals from Southern Finland, and communicated personal CVD risk information back to all study participants using our new web-tool. While follow-up of the full cohort is underway, we now report prospective results for 3,049 subjects. At baseline, 23.8% of study participants had high CVD risk (>10% 10-year risk). 40.3% of them were smokers, but only 17.3% received statin therapy. 12.1% had been reclassified from a lower risk category because of a high PRS. Of these, 36.6% were women. When reassessed 1.5 years after baseline, with an e-questionnaire and a clinical visit, 90% of the study subjects reported having received useful information, and 88.1% indicated that their personal risk information inspired them to take better care of their health. While 13.3% had achieved sustained weight loss (on average -2.3 kg), 15.0% of smokers had quit smoking compared to the 4% annual cessation rate in the general population. 34.5% of subjects at high CVD risk had undertaken risk-reducing intervention (weight loss, smoking cessation or visiting a physician) vs 20.4 % of subjects at lower risk. An increased PRS associated with successful intervention in the high risk-group ($p<0.005$). While 28.7% of all study participants thought their disease risk information was worrisome, this information also inspired to undertake lifestyle changes ($p<0.001$). Our data show that integrating genomic and traditional health information for the prevention of complex chronic disease and communicating the information back to individuals can support lifestyle changes. We expect to have completed the follow-up visits by the time of the conference.

271

Mechanistic and therapeutic interrogation of a novel mouse model of vascular Ehlers-Danlos syndrome. *G. Rykiel[1,2], C.J. Bowen[1,2], J.C. Giadrosic[1,2], M. Helmers[1,2], H.C. Dietz[1,2,3,4].* 1) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, MD; 3) Division of Pediatric Cardiology, Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD; 4) Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD.

Patients with vascular Ehlers-Danlos syndrome (vEDS) caused by heterozygous mutations in *COL3A1* experience spontaneous vascular dissection. The disease has historically been challenging to study given the lack of predictive biomarkers, high intrinsic variability in disease severity, and the lack of relevant phenotypes in existing haploinsufficient mouse models. We created two knock-in mouse models of vEDS, *Col3a1* G209S/+ and *Col3a1* G772D/+, that manifest highly penetrant spontaneous aortic dissection and rupture (median survival of ~1 year and 6 weeks, respectively). Most dissections occur in the proximal descending thoracic aorta in the absence of prior aortic dilatation or histologic evidence of altered aortic wall architecture. In contrast to Marfan syndrome (MFS) or Loeys-Dietz syndrome (LDS), expression profiling of vEDS aorta does not show a synthetic repertoire typical for high TGFβ signaling, however we do see evidence of increased ERK activation (similar to MFS and LDS). A small human study suggested that the beta-blocker celiprolol has the potential to delay adverse events in patients with vEDS, while angiotensin receptor blockers such as losartan afford dramatic protection in mouse models of MFS or LDS. Informatively, we do not see evidence of protection from dissection or death in losartan-treated vEDS mice, while celiprolol associates with significant acceleration of dissection and death in both vEDS models; both drugs achieved the predicted lowering of hemodynamic stress. While early data show that hydralazine, which inhibits the PLC/IP3/PKC/ERK axis, affords some protection in vEDS mice, these data highlight the need for discovery-based methods to reveal unanticipated therapeutic strategies. We introduced our vEDS mutations onto pure 129S6/SvEvTac (129) and C57BL6/J (BL6) backgrounds to assess for modulation of phenotypic severity. Both mutations are associated with early death due to aortic dissection on the BL6 background. Quite remarkably, the 129-background leads to complete protection from dissection for both vEDS genotypes, with a fully normal lifespan. Rescue associates with normalization of gene expression profiles for the aortic wall. These data provide rationale and incentive to perform genetic studies to identify the source and mechanism of modification in vEDS mice, with the hope and intention to mimic nature's successful strategies using pharmacologic agents.

Copyright © 2018 American Society of Human Genetics. All rights reserved

272

Effect of soluble activin receptor type IIB-Fc on hindlimb skeletal muscle contractile and mitochondrial function in the osteogenesis imperfecta model (*oim*) mouse. *Y. Jeong[1], V. Gremminger[1], G.M. Meers[2], R.S. Rector[2], C.L. Phillips[1,3].* 1) Biochemistry, University of Missouri, Columbia, MO; 2) Nutritiona and Physiology and Medicine-GI, University of Missouri; Harry S Truman Memorial VA Hospital, Columbia, MO; 3) Child Health, University of Missouri, Columbia, MO.

Osteogenesis Imperfecta (OI) is a heritable connective tissue disorder primarily due to mutations in the COL1A1 and COL1A2 genes. The major clinical manifestation in OI is skeletal fragility. Thus, the majority of current treatment options are directly targeted to bone. However, our previous studies have demonstrated that homozygous *oim* (*oim/oim*; an OI mouse model with a functional null Col1a2 mutation) mice exhibit reduced skeletal muscle mass and contractile function as compared to wildtype (WT) mice. Muscle weakness likely contributes to the compromised bone properties in OI through the mechanosensitive characteristics of bone. Citrate synthase activity was also found to be reduced in isolated mitochondria of *oim* gastrocnemius muscles, suggesting potential mitochondrial dysfunction in *oim* mice as well. Myostatin, a member of TGF-β superfamily, signals through activin receptor type IIB to negatively regulate muscle fiber growth. To investigate the potential of myostatin inhibition as a therapeutic strategy to enhance both muscle and bone properties, we utilized a soluble activin receptor type IIB decoy molecule (sActRIIB-mFc) to inhibit myostatin binding to its endogenous cellular receptors in WT and *oim* mice. sActRIIB-mFc treated WT mice exhibited increased muscle mass, but reduced hindlimb muscle contractile function as compared to vehicle treated counterparts. Whereas sActRIIB-mFc treated *oim* mice demonstrated increased muscle mass and enhanced muscle contractile function compared to vehicle treated counterparts. In addition, vehicle treated *oim* mice exhibited increased protein content of TFAM and PGC1α, markers of mitochondrial biogenesis, relative to WT mice, which sActRIIB-mFc treatment restored back to WT levels. Protein content of the mitophagy markers, bnip3 and parkin, were equivalent between WT and *oim*, and remained unchanged by sActRIIB-mFc treatment. Alterations in mitochondrial function in response to sActRIIB-mFc may have contributed to improved muscle contractile function in the *oim* mouse, although further investigations are required to elucidate the molecular mechanisms of the impact of myostatin inhibition on mitochondrial function and muscle contractile function. .

273

Safety and efficacy of low-dose sirolimus in the PIK3CA-related overgrowth spectrum: An open-label study in adults and children. *L. Faivre[1], V. Parker[2], K. Keppler-Noreuil[3], M. Luu[4], N. Oden[5], L. De Silva[2], J. Sapp[3], K. Andrews[2], M. Bardou[4], K. Chen[6], T. Darling[7], B. Goldspiel[8], S. Hadj-Rabia[9], J. Harris[2], G. Kounidas[2], P. Kumar[6], M. Lindhurst[3], R. Loffroy[10], L. Martin[11], A. Phan[12], K. R&oumlther[6], B. Widemann[13], P. Wolters[13], C. Coubes[14], L. Pinson[14], M. Willems[14], C. Vincent-Delorme[15], P. Vabres[1], R. Semple[16], L. Bieseker[3].* 1) Centres de références Anomalies du Développement et Anomalies Dermatologiques Rares, Equipe GAD UMR1231 et FHU TRANSLAD, CHU Dijon-Bourgogne et Université de Bourgogne, Dijon, France; 2) Institute of Metabolic Science, Cambridge University Hospitals NHS Trust, UK; 3) Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA; 4) Centre d'Investigation Clinique INSERM 1432, Centre Hospitalier Universitaire de Dijon, Dijon, FR; 5) The EMMES Corporation, Rockville, MD; 6) Section on Pediatric Diabetes and Metabolism, National Institute of Diabetes, Digestive, and Kidney Diseases, National Institutes of Health, Bethesda, MD; 7) Department of Dermatology, Uniformed Services University of the Health Sciences, Bethesda, MD; 8) Pharmacy Department, NIH Clinical Center, National Institutes of Health, Bethesda, MD; 9) Department of Dermatology and Reference Center for Genodermatoses and Rare Skin Diseases (MAGEC), Université Paris Descartes - Sorbonne Paris Cité, INSERM U1163, Institut Imagine, Institut Imagine, Hôpital Universitaire Necker-Enfants Malades, Paris, FR; 10) Department of Interventional Radiology, Dijon University Hospital, Dijon, FR; 11) Department of Dermatology, University Hospital Center of Angers, Angers, FR; 12) Department of Dermatology, Claude Bernard-Lyon 1 University and Hospices Civils de Lyon, Lyon, FR; 13) NCI, CCR, Pediatric Oncology Branch, National Institutes of Health, Bethesda, MD; 14) Département de Génétique Médicale, Maladies Rares et Médecine Personnalisée, CHU de Montpellier, Montpellier, FR; 15) Service de Génétique médicale, Hôpital Jeanne de Flandre, CHRU de Lille, Lille, FR; 16) Centre for Cardiovascular Science, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK.

Introduction: The PIK3CA-Related Overgrowth Spectrum (PROS) encompasses a range of debilitating conditions defined by asymmetric overgrowth caused by mosaic activating PIK3CA variants. The variable anatomy and natural history of overgrowth in PROS pose considerable challenges for study design, and raise concerns about reporting bias in case reports and small case series. We designed a molecularly-targeted therapeutic study of patients with PROS. PIK3CA encodes the p110alpha catalytic subunit of Class 1A phosphatidylinositol-3-kinase (PI3K), a critical transducer of growth factor signaling. As mTOR mediates the growth-promoting actions of PI3K, we hypothesized that the mTOR inhibitor sirolimus would slow or reverse pathological overgrowth. **Methods and results:** Thirty-nine participants with PROS and progressive overgrowth were enrolled into parallel, open label, single arm pilot studies with run-in control across three centers, and results were pooled. For the primary outcome measure, tissue volumes (fat and lean) at affected and unaffected sites were measured by dual-energy X-ray absorptiometry during 26 weeks of untreated run-in and 26 weeks of sirolimus treatment and statistical analysis was undertaken using paired t-tests. Thirty participants completed 26 weeks of therapy. Sirolimus led to a negative change in the mean percentage total tissue volume of -7.2% (SD 16.0, p=0.04) at affected sites, but not at unaffected sites (+1.7%, SD 11.5, p=0.48) (n=23 evaluable). Twenty-eight of 39 (72%) participants had at least one adverse event (AE) related to sirolimus therapy of which 37% were grade 3 or 4 in severity and 7/39 (18%) participants were withdrawn consequently. Limitations of the study included the open-label design and unequal study visits which may have led to an imbalance in AE reporting between the run-in and treatment phases. **Conclusion:** This study suggests that low-dose sirolimus can modestly reduce overgrowth, but cautions that the side-effect profile is significant, mandating individualized risk-benefit evaluations prior to sirolimus treatment in patients with PROS. Our study also holds lessons for the design of future therapeutic studies, and suggests the development of trial agents to target the PIK3CA gene product directly. Study Registration: France (NCT02443818), UK (EudraCT: 2014-000484-41), US (NCT02428296).

Copyright © 2018 American Society of Human Genetics. All rights reserved

274

The combinative effect of testosterone treatment and the timing of diagnosis on neurocognitive abilities and ADHD in 47,XXY (Klinefelter Syndrome). *C. Samango-Sprouse[1,2,3], P. Lasutschinkow[1], S. Chea[1], T. Sadeghin[1], A. Gropman[4,1].* 1) The Focus Foundation, Davidsonville, MD; 2) George Washington University, Washington, DC; 3) Florida International University, Miami, FL; 4) Children's National Health System, Washington, DC.

47,XXY is the most frequently occurring X & Y chromosomal disorder (1:660). These boys may exhibit motor planning deficits, language-based learning disabilities, ADHD & executive dysfunction. Testosterone replacement has been shown to mitigate some neurodevelopmental differences. Few studies have documented the possible combinative effect of prenatal diagnosis & testosterone treatment on intellectual capabilities and behavior.288 males with 47,XXY were evaluated using the Child Behavior Checklist, Weschler Intelligence Scale for Children, and Leiter International Performance Scale (LIPS). 78.5% were prenatally diagnosed (PRE) & the rest were postnatally diagnosed (POST). 71.5% received testosterone treatment & 28.5% received none. Treatment included early hormonal treatment (E) before five years, hormonal booster treatment (B) between 5-10 years, testosterone treatment after 10 years (T) & combinations of the three. Treatment was based on the patient's pediatric endocrinologist's assessment of the size of phallus in comparison to neurotypical boys of the same age.Boys who received both E & B had significantly reduced ADHD symptoms in comparison to boys with just B (P=0.007). Within the treated group, PRE boys performed better than POST in Performance IQ (PIQ) (P=0.038) & Processing Speed (PSI) (P=0.002). PRE boys that were treated with some type of testosterone performed better in Verbal IQ (VIQ) (P=0.005), Perceptual Reasoning (PRI) (P=0.003), PSI (P=0.023) & Working Memory (WMI) (P=0.001) than untreated boys. PRE boys who received all levels of testosterone treatment (E, B & T) did significantly better than PRE untreated boys (P=0.012) on VIQ & PRI. On the LIPS, boys who received E, B & T performed significantly better when compared to untreated boys (P=0.020) & boys with some but not all testosterone treatment (P=0.019).This study has further expands our knowledge of the positive impact of testosterone on the neurodevelopmental outcome of boys with 47,XXY and suggests that receiving multiple treatments (E, B, and T) results in the most positive outcome. For the first time, a combinative effect of hormonal treatment and diagnosis suggests that boys with 47,XXY may need treatment at several life stages to optimize their outcome. Further exploration of the most advantageous timing and dosage is warranted. Additional study is required to investigate the relationship between repeated treatments of testosterone during childhood and maximum outcome.

275

Discovery of over 50 novel dominant developmental disorders from exome sequencing of 22,518 parent-child trios. *M.E. Hurles[1], J. Kaplanis[1], K.E. Samocha[1], Z. Zhang[2], S.H. Lelieveld[3], G. Gallone[1], H.G. Brunner[3], C. Gilissen[3], K. Retterer[2] on behalf of the Deciphering Developmental Disorders study.* 1) Wellcome Sanger Institute, Hinxton, United Kingdom; 2) GeneDx, Gaithersburg, MD 20877, USA; 3) Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands.

Despite the rapid discovery of novel dominant developmental disorders (DD) in recent years, analyses of exome-wide burden of damaging *de novo* mutations (DNMs) has indicated that there are many more dominant DDs to be discovered. Following clear consent practices, and only using aggregate, deidentified data, we pooled parent-child trio exome data from GeneDx, the Deciphering Developmental Disorders (DDD) study, and Radboud University Medical Center to generate a dataset five-times larger than the largest previously published study of DD trios (n = 22,518 trios). Having called DNMs with similar sensitivity and specificity across all three datasets, we identified a total of 32,969 coding and splicing DNMs, and observed highly similar levels of exome-wide burden of damaging DNMs across each individual cohort. We also observe similar prevalences of known dominant developmental disorders in each cohort, demonstrating similar clinical ascertainment across the cohorts. We devised a novel simulation-based method to test for a statistically significant enrichment of damaging DNMs in individual genes. This method scores all classes of variants (e.g. nonsense, missense, splice site) on a unified severity scale based on the empirically-estimated positive predictive value of being pathogenic, and incorporates a gene-based weighting derived from the deficit of protein truncating variants in the general population. We show that this method increases power to detect DD-associated genes. Applying this novel method to the DNMs from the 22,518 trios, we identified over 250 significantly enriched genes exome-wide including over 50 novel DDs. We will describe these newly identified disorders, their functional concordance with known disorders and analyses of how power to discover novel DDs scales with increasing sample sizes.

Copyright © 2018 American Society of Human Genetics. All rights reserved

276

Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *B. Wang[1], P. Sample[1], D. Reid[2], V. Presnyak[2], I. McFadyen[2], D. Morris[3], G. Seelig[1,4].* 1) Department of Electrical Engineering, University of Washington, Seattle, WA; 2) Moderna Therapeutics, Cambridge, MA; 3) Department of Biochemistry, University of Washington, Seattle, WA; 4) Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA.

The sequence of the 5′ untranslated region (5′ UTR) is a primary determinant of translation efficiency. While many *cis*-regulatory elements within human 5′ UTRs have been characterized individually, the field still lacks a means to accurately predict protein expression from 5′ UTR sequence alone, limiting the ability to estimate the effects of genome-encoded variants and the ability to engineer 5′ UTRs for precise translation control. Massively parallel reporter assays (MPRAs) – methods that assess thousands to millions of sequence variants in a single experiment – coupled with machine learning have proven useful in addressing similar voids by producing quantitative biological insight that would be difficult to achieve through traditional approaches. We report the development of an MPRA that measures the translation of hundreds of thousands of randomized 5′ UTRs in an *in vitro* transcribed mRNA library via polysome profiling and RNA sequencing. We then use the data to train a convolutional neural network (CNN) to build a predictive model that relates 5' UTR sequence to translation, and the model could explain 93% of the variation. The same approach can be extended to chemically modified RNA (pseudouridine and 1-methylpseudouridine), an important feature for applications in mRNA therapeutics and synthetic biology. We then tested 35,212 truncated human 5' UTRs and 3,577 naturally-occurring variants from ClinVar and showed that the model accurately predicts ribosome loading of these sequences. We also investigated the model's ability to predict the translation level change between pairs of wild-type and single-nucleotide variant(SNV)-containing 5′ UTR sequences, and our model could explain 55% of the change. Finally, we provide evidence of 47 SNVs associated with human diseases that cause a significant change in ribosome loading and thus a plausible molecular basis for disease. As an example, one of the ClinVar variants with significant differences in translation level in our assay, rs121908813, is found in the 5′ UTR of the TMEM127 gene. It introduces an out-of-frame upstream AUG in the 5′ UTR of TMEM127 and is implicated in familial pheochromocytoma, a condition characterized by tumors found in the neuroendocrine system that excrete high levels of catecholamines. TMEM127 acts as a tumor suppressor and decreased expression of it could explain the observed pathogenicity of this variant.

277

Biallelic variants in *TONSL* cause SPONASTRIME dysplasia and an expanded spectrum of skeletal dysplasia phenotypes. *L. Burrage[1], J.J. Reynolds[2], N.V. Baratang[3], J. Phillips[4], J. Wegner[4], A. Christiansen[5], M.R. Higgs[2], C. Logan[6], D. Chitayat[7,8], I. Chinn[9,10], D. Muzny[11], R. Gibbs[11], W. Bi[12], J.A. Rosenfeld[1], J. Postlethwait[4], M. Westerfield[4], M. Dickinson[5], J. Orange[9,10], M. Adeli[13], D. Cohn[14], D. Krakow[15], A. Jackson[16], M. Lees[17], A.C. Offiah[18], C. Carlston[19], J.C. Carey[20], G. Stewart[2], C.A. Bacino[1], P.M. Campeau[3], B.H. Lee[1], Members of the Undiagnosed Diseases Network.* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX USA; 2) Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK; 3) CHU Sainte-Justine Research Center, University of Montreal, Montreal, QC, Canada; 4) Institute of Neuroscience, University of Oregon, Eugene, Oregon, USA; 5) Department of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, TX, USA; 6) MRC Institute of Genetics & Molecular Medicine, The University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK; 7) The Prenatal Diagnosis and Medical Genetics Program, Department of Obstetrics and Gynecology, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada; 8) Department of Pediatrics, Division of Clinical and Metabolic Genetics, the Hospital for Sick Children, University of Toronto, Toronto, ON, Canada; 9) Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA; 10) Texas Children's Hospital, Division of Pediatric Immunology, Allergy, and Rheumatology, Houston, TX, USA; 11) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA; 12) Baylor Genetics, Houston, TX, USA; 13) Department of Allergy and Immunology, Sidra Medicine /Hamad Medical Corporation / Weill Cornell Medicine - Qatar, Doha, Qatar; 14) Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, Los Angeles, CA, USA; 15) Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; 16) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK; 17) North East Thames Regional Genetics Service, Great Ormond Street Hospital, London, UK; 18) Department of Oncology and Metabolism, Academic Unit of Child Health, University of Sheffield, Sheffield, UK; 19) Department of Pathology, University of Utah, Salt Lake City, UT, USA; 20) Department of Pediatrics, Division of Medical Genetics, University of Utah, Salt Lake City, UT, USA.

SPONASTRIME dysplasia is an autosomal recessive spondyloepimetaphyseal dysplasia named for characteristic clinical and radiographic findings including spine abnormalities (spondylar), midface hypoplasia with depressed nasal bridge, and striation of the metaphyses. Typically, patients with SPONASTRIME dysplasia also have disproportionate short stature. Additional features include scoliosis, coxa vara, limited elbow extension, hypogammaglobulinemia, childhood cataracts, and short dental roots. Because multiple sibships have been reported, an autosomal recessive inheritance pattern has been suspected, but no gene has been associated with this disorder. In 5 subjects with SPONASTRIME dysplasia from 4 families, we identified biallelic variants in *TONSL*, which encodes the Tonsoku Like DNA repair protein. Likewise, biallelic variants in this gene were identified in 4 subjects with varied skeletal dysplasia diagnoses including microcephalic osteodysplastic primordial dwarfism and spondylometaphyseal dysplasia with immunologic abnormalities (hypogammaglobulinemia and neutropenia) but no apparent metaphyseal striations. Immunoblotting performed using patient-derived fibroblasts suggested reduction in TONSL protein compared to control cells confirming a suspected loss of function mechanism. The pathogenicity of *TONSL* deficiency is further supported by the finding of embryonic lethality in a *Tonsl* knock out mouse and significantly reduced length and shortened lifespan in a knock out zebrafish model. TONSL functions in a complex with MMS22L and is necessary for the repair of replication-associated DNA damage. Although the TONSL-MMS22L complex is reported to bind to all replication forks, increased binding is noted on stalled or collapsed replication forks and facilitates the binding of RAD51 to ssDNA to promote homologous recombination. Functional studies in patient-derived fibroblasts showed significantly increased levels of spontaneous replication fork stalling and fewer camptothecin (CPT)-induced Rad51 foci. Overall, our functional studies in mouse, zebrafish, and patient-derived cell lines confirm that loss of function pathogenic variants in *TONSL* cause impairment of homologous recombination leading to a wide spectrum of skeletal dysplasia phenotypes ranging from spondylometaphyseal dysplasia with discrete short stature to SPONASTRIME dysplasia and severe growth restriction associated with hypogammaglobulinemia and neutropenia.

Copyright © 2018 American Society of Human Genetics. All rights reserved

278

Large genome-wide analysis of sexual orientation identifies for the first time variants associated with non-heterosexual behavior and reveals overlap with heterosexual reproductive traits. *A. Ganna[1,2,3,4], K.J.H. Verweij[5], F.R. Day[6], M.G. Nivard[7], R. Maier[1,2,3], R. Wedow[8,9,10], A.S. Busch[11,12], A. Abdellaoui[5], S. Guo[13], F. Sathirapongsasuti[14], P. Lichtenstein[4], H. Larsson[4], S. Lundström[15], N. Långström[4], D.A. Hinds[14], G.W. Beecham[13], E.R. Martin[13], A.R. Sanders[15,16], B.M. Neale[1,2,3], J.R.B. Perry[6], B.P. Zietsch[17], 23andMe Research Team.* 1) Analytic and Translational Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 3) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; 5) Department of Psychiatry, Academic Medical Center, University of Amsterdam, Meibergdreef 5, 1105 AZ Amsterdam, The Netherlands; 6) MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK; 7) Department of Biological Psychology, VU University, 1081 BT, Amsterdam, The Netherlands; 8) Department of Sociology, University of Colorado, Boulder, Colorado 80309-0483, USA; 9) Health and Society Program and Population Program, Institute of Behavioral Science, University of Colorado, Boulder, Colorado 80309-0483, USA; 10) Institute for Behavioral Genetics, University of Colorado, Boulder, Colorado 80309-0483, USA; 11) Department of Growth and Reproduction, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark; 12) International Center for Research and Research Training in Endocrine Disruption of Male Reproduction and Child Health (EDMaRC), Rigshospitalet, Copenhagen, Denmark; 13) Department of Human Genetics, University of Miami, Miami, Florida 33136, USA; 14) 23andMe, Inc. Mountain View, CA 94041, USA; 15) Department of Psychiatry and Behavioral Sciences, NorthShore University HealthSystem Research Institute, Evanston, Illinois 60201, USA; 16) Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, Illinois 60637, USA; 17) School of Psychology, University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia; 18) Centre for Ethics, Law and Mental Health, University of Gothenburg, Sweden.

Twin and family studies have shown that sexual orientation is in part genetically influenced (~40% narrow-sense heritability), but previous efforts to identify the specific genes involved have been unsuccessful due to a lack of power. To better understand the genetics underlying sexual orientation and the overlap with other traits and, in particular, with heterosexual behavioral and reproductive traits (e.g. lifetime number of same-sex partners) we performed genome-wide association analyses on 493,001 individuals enrolled in 5 studies from the UK, USA and Sweden. We found 4 genome-wide significant loci for non-heterosexual behavior and 40 loci for the number of opposite-sex partners in heterosexual. We estimated that, in aggregate, common genetic variants account for 8-20% of variation in non-heterosexual behavior and further analyses suggested an overlap with genes underlying sex hormone regulation and olfactory processes. We detected a substantial degree of heterogeneity in the genetic basis of sexual behavior as these effects were only partially shared among women and men and across different definitions of non-heterosexual behavior. We found that variants predisposing to non-heterosexual behavior are, among heterosexuals, positively associated with having more self-reported lifetime sexual partners and, in heterosexual males, with being judged more to be physically attractive. This is consistent with the hypothesis that genetic variants predisposing to non-heterosexual behavior confer a mating advantage to heterosexual carriers. Recognizing the sensitivity of the topic, we have employed best practices for communicating our results including the preregistration of the study plan and the creation of a website to report our findings following feedback obtained through public engagement.

279

Integrative analysis of diverse transcriptomic alterations to identify cancer-relevant genes across 27 histotypes. *N.R. Davidson[1,7], D. Demircioglu[2], N.A. Fonesca[4], A. Kahles[1], K.V. Lehmann[1], F. Liu[6], L. Urban[4], J. Goke[2], R.F. Schwarz[5], A. Brooks[3], G. Rätsch[1], Z. Zhang[6], A. Brazma[4], PanCancer Analysis of Whole Genomes and Transcriptomes Working Group (PCAWG-3), PCAWG Consortium.* 1) ETH Zürich, Zürich, Switzerland; 2) Genome Institute of Singapore, Singapore; 3) Department of Biomolecular Engineering, UCSC, Santa Cruz, CA; 4) European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom; 5) Max-Delbrück-Centrum für Molekulare Medizin (MDC) Berlin, Deutschland; 6) BIOPIC, School of Life Sciences, Peking University, China; 7) Tri-Institutional Program in Computational Biology and Medicine, Weill Cornell Medicine, New York.

Previous multi-cancer genomic studies have focused on somatic mutations as the driver of phenotypic changes. Here, we integrate both RNA and DNA-level changes to account for the role of the transcriptome in tumorigenesis. We present a novel analysis that 1) identifies cancer relevant genes through a recurrence analysis over diverse RNA alterations 2) identifies RNA alteration patterns in 1,188 samples across 27 histology types (HT) as part of the PanCancer Analysis of Whole Genomes (PCAWG) of the International Cancer Genome Consortium. Alteration types in this analysis are: expression outliers, copy-number alterations (CNA), alternative splicing outliers (AS), gene fusions (GF), alternative promoters (AP), non-synonymous variants (NSV), RNA-editing, and allele-specific expression.To identify cancer relevant genes, we created a recurrence analysis method for RNA features. Our method has 3 strengths: flexibility to handle any number/type of alteration; sensitivity to alteration frequencies; and prioritization of genes with heterogeneous alterations. We performed >1M permutations to identify an informative cut-off which yielded 1,012 genes with an empirical p-value<0.05. These genes show a 2.82-fold enrichment (p-value: 5x10-26) for cancer census genes and driver genes (from PCAWG-9); evidence our analysis identifies cancer-relevant genes.CDK12 is in the top 5% of our ranked genes and is impacted by multiple alterations in a protein kinase domain (PKD) associated with dysregulation of DNA repair. 87 samples have an alteration in the PKD, with 74% samples having only an RNA alteration. AP (the most frequent alteration), GF and AS, all lead to disruptions of the PKD. CDK12 exemplifies the value of integrating RNA and DNA alterations.We also compare alteration patterns across HTs and pathways. Kidney-ChRCC and Skin-Melanoma have significantly different frequencies of NSVs (t-test; p-adj.: 1.42x10-5), CNAs (p-adj.: 6.70x10-4), GFs (p-adj.: 1.56x10-4), and ASs (p-adj.: 7.05x10-10). In contrast, similar cancers like Kidney-RCC and Kidney-ChRCC only differ in the amount of NSVs (t-test; p-adj.: 5.50x10-25). Across cancer-specific pathways, we find TOR and metabolism pathways more impacted by RNA alterations. Furthermore, we find 22.7% of the 578 samples with an altered p53 pathway, typically associated with NSVs, carried **only RNA** alterations. This is evidence that neglecting transcriptomic alterations could underestimate the degree of cancer pathway disruption.

Copyright © 2018 American Society of Human Genetics. All rights reserved

280

Transcriptome sequencing illuminates the extent of alternative splicing in multiple myeloma in regards to *SF3B1* mutations and identifies a potential MM specific splicing pattern. *M.A. Bauer, C. Ashby, R.G. Tytarenko, C.P. Wardell, P. Patel, S. Deshpande, F. van Rhee, M. Zangari, S. Thanendrarajan, C.D. Shinke, F.E. Davies, G.J. Morgan, B.A. Walker.* University of Arkansas for Medical Sciences 4301 W. Markham, Little Rock AR 72205.

Alternative splicing of RNA transcripts is common in hematological malignancies, especially in chronic lymphocytic leukemia and myelodysplastic syndromes where *SF3B1*, a spliceosome component, is mutated. RNA splicing has not been investigated in multiple myeloma (MM), even though risk stratification is commonly based on gene expression profiling. In this study we took a multifaceted approach to understand the extent of alternative splicing in MM. In a dataset of 1273 newly diagnosed MM samples we identified mutations in *SF3B1*in 1.7% (22/1273) of samples, of which 5 were in the hotspot codons of K666 and K700. Using RNA-sequencing from the matched exome samples (n=615) we compared the splice junction usage of *SF3B1*mutants against normal samples matched for key MM molecular sub-types. Differential splicing was detected in the *SF3B1*mutant samples and the genes with most significant levels of alternative splicing were *DYNLL1*, *TMEM14C*, *CRNDE*, *BRD4*and *BCL2L1*, several of which are also seen in other cancers with mutated *SF3B1*. Taking a more agnostic approach, we detected novel splice junctions in all samples and ranked them accordingly. The dataset was split, so that the top and bottom deciles could be compared to the middle group. The top decile was investigated for defining genomic and clinical features and we found this group had an association with a higher number of non-silent mutations in *TP53*, *RASGRF2*, and *CHD3*, and a lower frequency of mutations in *NRAS*, *IGLL5,*and *DIS3*, and a lower frequency of t(4;14). The group with the most novel splice junctions was also enriched for the high risk clinical outcome group, ISS III. The low novel splice junction group had both a significantly increased progression free survival (541.8 days vs. 402.7 days, P=0.00395) and overall survival (631.8 days vs. 459.5 days, P=0.00739) over the medium groups, which was recapitulated in the high group. Our results show that analysis of alternative splicing has the potential to further refine MM risk stratification, where samples with increased alternative splicing are associated with high risk factors. Identification of the genes undergoing alternative splicing, as well as the mechanisms leading to it, may present novel therapeutic targets.

281

Cancer eQTL profiles can be recovered from bulk tumor gene expression data by modeling tumor purity. *P. Geeleher[1], .F Wang[1], Z. Zhang[1], R.L. Grossman[1], C. Seoighe[2], R.S. Huang[3].* 1) Department of Medicine, University of Chicago, Chicago, IL; 2) School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland; 3) Department of Experimental and Clinical Pharmacology, University of Minnesota.

Genome-wide association studies (GWAS) have identified hundreds of inherited genetic variants affecting cancer risk. Most of these variants are in non-coding regions of the genome and modulate risk by affecting gene regulation. Thus, determining how inherited genetic variation affects gene expression in cancer is critically important to understanding disease development. Consequently, by leveraging large genomics datasets like The Cancer Genome Atlas (TCGA), previous studies have mapped expression quantitative trait locus (eQTLs) using tumor expression data. However, tumors are mixtures of both cancer and normal cells, for example, immune cells and stroma. We have developed a new approach that can accurately account for the effect of tumor-infiltrating normal cells on cancer eQTLs. The approach involves first estimating the proportion of tumor-infiltrating normal cells (tumor purity) using a combined estimate from genomics data and H&E staining. Then, we developed a statistical model that can account for the effect of tumor purity on eQTLs by modeling the interaction of the tumor purity estimate and genotype. Intuitively, this models how the magnitude of the association between gene expression and genotype changes as a function of tumor purity and extrapolates this effect to 100% cancer cells. We applied this approach to the TCGA breast cancer cohort. Remarkably, while 57,189 eQTLs were identified in TCGA breast cancer patients using a conventional model, only 8,833 (15.4%) could be attributed to cancer cells—when tumor purity was accounted for. Analysis of the Genotype-Tissue Expression (GTEx) data suggested almost 50% of tumor eQTLs inferred using a conventional approach affect gene expression in tumor-infiltrating immune cells and fibroblasts, not cancer cells. We also investigated the eQTL profiles variants identified in a meta-analysis of GWAS data for breast cancer risk. Strikingly, for 33% of breast cancer risk variants identified as eQTLs using an uncorrected approach, we found strong evidence of an effect in tumor-infiltrating normal cells, but no evidence of an effect in cancer cells. This suggests cancer risk is mediated by the effect of inherited genetic variation on gene regulation in the cells of the tumor microenvironment, as well as in cancer and pre-cancer cells. Our findings challenge the current interpretation of inherited genetic regulation in cancer and should be considered in functional validation of all cancer risk GWAS. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

282

Mapping tumor-immunity-specific expression QTL (tis-eQTL) in cancer. *X. Wang[1], B.L. Fridley[1], X. Yu[1], Y.A. Chen[1], B. Li[2], C.H. Chung[1], S. Antonia[1], J.R. Conejo-Garcia[1].* 1) Moffitt Cancer Center, Tampa, FL; 2) UT Southwestern.

Introduction: Immunotherapy has produced promising results in treating cancers. Recent studies that aim to predict which patients can benefit from an immune checkpoint blocker have been largely focused on tumor molecular profiling such as tumor mutation burden and T cell infiltrations. However, whether the immunity landscape in tumor can be favorably or adversely affected by polymorphisms carried in the germline remains unknown and understudied. **Methods**: Here we conduct the largest investigation of tumor-immunity-specific expression QTL to date, or 'tis-eQTL', to systematically identify germline genetic variants that affect immune landscape in tumor. We develop a reliable statistical method for tis-eQTL mapping using RNA-seq data from heterogeneous tumor samples. We analyze genomic data from 10,380 cancer patients from 33 cancer types to reveal interactions between genetic variants (derived from germline SNP array and WES data) and immune-phenotypes derived from tumor RNAseq data. These phenotypes include immune gene expression (such as CD3E, GZMA, CXCR3 and PRF1), T-cell receptor (TCR) clonality, and antigen presenting scores (APM). **Results:** We observed that stratifying patients by the prioritized pathogenic germline polymorphisms exposed distinct tumor immune landscape, implicating new prognostic factors or risk genes of cancer. We further prioritize SNPs that fulfill two criteria: (a) with differential overall survival (b) have differential expression pattern in immunogenic signatures. In support of these findings, we show that the top-ranked SNPs are associated with treatment responses to anti-PD-1 therapy in melanoma and non-small cell lung cancer. This study creates a unique and pioneering resource of prognostic germline variants with the potential to modulate immune therapy response and cancer risk, opening new avenues for developing next generation therapeutics and for personalizing risk assessment.

283

Germline GATA3 variants influence chromatin topology and pathogenesis of acute 2 lymphoblastic leukemia. *H. Yang[1], H. Zhang[2,3], T. Liu[1], K. Roberts[4], M. Qian[2], Z. Zhang[1], W. Yang[2], V. Perez-Andreu[2,5], Y. Luan[1,7], J. Xu[1], S. Iyyanki[1], D. Kuang[8], H. Xu[9], S. Reshmi[10,11], J. Gastier-Foster[10,11], C. Smith[2], C. Pui[12], W. Evans[2], S. Hunger[13], M. Relling[2], C. Mullighan[4], M. Loh[14], F. Yue[1,6], J. Yang[2].* 1) Dept. Biochemistry and Molecular Biology, Pennsylvania State University College of Medicine, Hershey, PA; 2) Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, Tennessee, USA; 3) Department of Pediatric Hematology/Oncology, 12 Guangzhou Women and Children's Medical Center, Guangzhou, Guangdong, China; 4) Department of Pathology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA; 5) Internal Medicine Department, MountainView Hospital, University of Reno, Las Vegas, Nevada, USA; 6) Bioinformatics and Genomics Program, The Pennsylvania State University, University Park, Pennsylvania, USA; 7) Key Laboratory of Agricultural Animal Genetics, Breeding 17 and Reproduction of Ministry of Education and Key Laboratory of Swine Genetics and Breeding of Ministry of Agriculture, Huazhong Agricultural University, Wuhan 430070, China,; 8) Department 19 of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA; 9) Department of Laboratory Medicine, National Key Laboratory of Biotherapy/Collaborative 21 Innovation Center of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu, China; 10) Department of Pathology and Laboratory Medicine, Nationwide Children's Hospital, Columbus, Ohio, USA; 11) Department of Pediatrics, Ohio State University School of Medicine, Columbus, Ohio, USA; 12) Department of Oncology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA,; 13) Division of Oncology and the Center for Childhood Cancer 26 Research, Children's Hospital of Philadelphia and the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA; 14) Department of Pediatrics, Benioff Children's 28 Hospital and the Helen Diller Comprehensive Cancer Center, University of California, San 29 Francisco, San Francisco, California, USA.

For many cancers, inherited genetic variants in the non-coding region of the genome confer significant disease susceptibility, often times specifically to tumor subtypes with defined somatic alterations (e.g., GATA3 variants associated with Philadelphia chromosome-like acute lymphoblastic leukemia [Ph-like ALL]). However, the molecular processes by which germline variants influence the acquisition of somatic lesions and contribute to tumorigenesis in general are poorly understood. Focusing on Ph-like ALL as a model system, we comprehensively identified leukemia risk alleles in GATA3 by targeted sequencing of 5,008 patients and discovered a critical non-coding regulatory variant. To characterize the function of this variant in hematopoietic cells, we created an isogenic cellular model by CRISPR-Cas9 editing, and observed that the leukemia risk variant induced a strong enhancer that significantly upregulated GATA3 transcription in cis, which in turn reshaped the global chromatin accessibility landscape. Hi-C experiments showed that GATA3 risk allele led to large scale 3-dimensional genome reorganization, causing hundreds of genes to switch between active and repressive compartments. Remarkably, this genotype switch resulted in a selective increase in promoter-enhancer interactions in the CRLF2 oncogene, analogous to the super enhancer hijacking at this locus somatically acquired in Ph-like ALL. Finally, we showed that increased GATA3 transcription directly upregulated CRLF2, influencing oncogenic effects of JAK-STAT signaling. Taken together, our results point to transcription factor-mediated epigenomic and 3-dimensional genome reprogramming as an important mechanism of oncogene activation, particularly in the context of cancer risk variants in the germline genome.

Copyright © 2018 American Society of Human Genetics. All rights reserved

284

Using allelic expression to extending rare disease diagnosis beyond coding mutations. *P. Mohammadi[1,2,3,4], S.E. Castel[3,4], B. Cummings[5,6], D. MacArthur[5,6], T. Lappalainen[5,6].* 1) Dept of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA; 2) The Scripps Translational Science Institute, La Jolla, CA; 3) New York Genome Center, New York, NY; 4) Dept of Systems Biology, Columbia University, New York, NY; 5) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 6) Broad Institute of MIT and Harvard, Cambridge, MA.

Despite the abundance of data from sequencing-based assays from individuals with severe disease, discovering potential pathogenic regulatory variants from these data has been challenging. Allelic expression (AE) data is a potentially valuable resource to address this problem, specifically due to the fact AE data captures the net effect of all *cis*-regulatory variants in an individual and it is minimally obscured by other confounding factors. However, widespread use of this valuable data source has been hampered by its complexity and lack of theoretical frameworks to enable advanced analyses. In this work we first describe a mathematical model of *cis*-regulatory genetic variation, and demonstrate that AE data distribution emerges as a constrained form of Binomial Logistic Normal (BLN) distribution function. We further show how this model can be used for quantifying regulatory variation in human population using our new method called ANalysis of Expression VAriance (ANEVA). Building upon this we provide methods for testing extreme allelic imbalance and dosage outliers. We apply our model to estimate per-gene natural spectrum of cis-regulatory genetic variation for close to 20,000 genes in the Genotype-Tissue Expression (GTEx) project v7 data. We then use these estimates to identify genes with potentially pathogenic regulatory mutations in rare genetic disease data. Applying dosage outlier test to AE data from 33 patients diagnosed with rare muscle disorders from Cummings *et al.* 2017, we found on average 21 such dosage outlier genes at 5% FDR. We were able to find the causal disease gene in 73% of the cases where the previously identified underlying mutation was expected to induce allelic imbalance, demonstrating high sensitivity. Notably, in 63% of these cases the causal gene was identified among the top-5 outlier genes by p-value. Our work demonstrates how the modeling of genetic regulatory variation in the general population can be used to interpret potential disease-causing variation in patient transcriptomes.

285

Systematic characterization of genetic variants associated with type 1 diabetes for differential binding of 530 transcription factors. *P. Benaglio[1], J. Yan[2,3], J. Chiou[4], N. Nariai[5], M. Sander[1], B. Ren[2,5], K. Gaulton[1], K. Frazer[1,5].* 1) Department of Pediatrics, UC San Diego, La Jolla, CA; 2) Ludwig Institute for Cancer Research, La Jolla, CA; 3) Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden; 4) Biomedical Sciences Graduate Program, UC San Diego, La Jolla, CA; 5) Institute of Genomic Medicine, UC San Diego, La Jolla, CA.

Type 1 diabetes (T1D) affects 1.25 million individuals in the United States and develops as a result of autoimmune destruction of insulin-secreting pancreatic beta cells. To date, GWAS studies have identified 58 T1D-risk loci; however, the causal variants and functional mechanisms at each locus are mostly unknown. As the majority of risk variants are non-coding, we sought to investigate their function through systematic testing of differential transcription factor (TF) binding. To test all potential causal variants associated to date with T1D, we selected 86,076 variants, including: credible set SNPs from fine mapping of 36 loci (Immunochip), all SNPs in LD (r^2>0.2) with index GWAS variants at the remaining loci, all SNPs with MAF >0.5% in regulatory elements within 250 kb of each T1D risk locus, and 8,000 random SNPs as negative controls. Using high-throughput SELEX sequencing, we assessed binding of 530 non-redundant *E. coli*-expressed TF proteins for a total of 94,076 oligonucleotides (44-nt) containing different alleles of the selected variants. Overall, we found 18,196 distinct SNPs with preferential binding (pbSNPs) to one of the alleles (χ^2 FDR <0.05) in at least one TF, with a median of one TF per SNP and of 78 SNPs per TF. Preferential binding was consistent between replicates of the same TF (all SNPs Pearson r=0.5; pbSNPs r=0.91) and of different TFs from the same DNA binding family (r=0.21; r=0.75); and showed good correlation with predicted motif alterations (PWM models) (r=0.52; r=0.78) and predicted TF binding effect from DeepSEA (r=0.31; r=0.59). These results demonstrate that SELEX identified pbSNPs in a consistent and reproducible fashion. To prioritize putative causal variants at T1D loci, we conducted a GWAS meta analysis from 15k cases and controls available from different studies, followed by fine-mapping using Bayesian factors. We observed that pbSNPs within active regulatory regions of immune cells were enriched in the 99% credible set (Fisher test OR=1.3; p=0.004), suggesting that they are likely functionally involved in T1D risk. Using FGWAS, we integrated our fine mapping results with both chromatin and SELEX annotations, and identified 8 loci where pbSNPs were prioritized, providing promising candidates for additional functional assays. We anticipate that this comprehensive annotation of T1D risk variants for differential TF binding will serve as a unique resource for elucidating novel disease mechanisms.

Copyright © 2018 American Society of Human Genetics. All rights reserved

286

A novel type 1 diabetes susceptibility locus affects *CFTR* regulation in pancreatic ductal cells. *J. Chiou[1,2], R. Geusz[1,2], M. Okino[2], S. Huang[2], M. Sander[2,3], K. Gaulton[2].* 1) Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA; 2) Department of Pediatrics, University of California San Diego, La Jolla, CA; 3) Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA.

Type 1 diabetes (T1D) is a complex disorder that affects millions of individuals, but its etiology is poorly understood. While autoimmune destruction of pancreatic beta cells is a major factor in the pathogenesis of T1D, recent evidence points to additional risk factors such as beta cell fragility. To identify novel risk mechanisms involved in T1D, we performed a genome-wide scan of T1D in 20,228 individuals of European ancestry imputed into 39M variants in the HRC reference panel. Among 34 genome-wide significant loci, we identified a novel locus on 7q31 near *CFTR* (P=9.2x10^{-9}). T1D-associated variants at this locus were non-coding and mapped in regulatory elements active in pancreas and pancreatic islets, but not in immune cells or other tissues. Through integration with eQTL data from GTEx and other sources, we determined that T1D risk alleles were correlated and co-localized with decreased *CFTR* expression in pancreas (eQTL P=5.4x10^{-5}; coloc. Prob.=85%) but not in islets (eQTL P=0.98; coloc. Prob.=5%) or any other tissues. Through computational deconvolution of pancreas eQTLs with single cell RNA-seq signature genes, we mapped the *CFTR* eQTL signal to pancreatic ductal cells (eQTL P=2.9x10^{-9}). We fine-mapped causal variants with histone modification and transcription factor ChIP-seq from pancreatic ductal adenocarcinoma (PDAC) cell lines. A single variant (rs7795896) mapped in a ductal cell-specific enhancer region and an HNF1B binding site 33kb upstream of the *CFTR* promoter, and the risk allele had significantly reduced luciferase reporter activity in two PDAC lines (both P<0.05). Using UK Biobank phenotype associations, we determined that rs7795896 was also associated with insulin treatment within a year of diabetes diagnosis (T-stat=3.4), increased risk of fibroblastic disorders (T-stat=4.2), and death from acute pancreatitis (T-stat=2.2) among other traits, but not with other autoimmune diseases. We are currently using CRISPR/Cas9 to delete this enhancer in PDAC cells to validate functional effects on TF binding, *CFTR* expression, and cellular function. Together these results reveal a novel T1D risk locus that affects activity of a pancreatic ductal cell enhancer and decreases ductal expression of *CFTR*. More broadly, our findings support a previously unknown role for the exocrine pancreas and ductal cells in T1D pathogenesis and a possible mechanistic link between T1D and *CFTR*-related disorders such as pancreatitis and cystic fibrosis-related diabetes.

287

Evaluating geographic differences in polygenic risk in Finland. *S. Kerminen[1], J. Koskela[1], A.S. Havulinna[1,2], I. Surakka[1,3], A.R. Martin[4,5,6], A. Palotie[1,4,5,7,8], M. Perola[1,2], V. Salomaa[2], M.J. Daly[1,4,5,6], S. Ripatti[1,9], M. Pirinen[1,9,10].* 1) Institute for Molecular Medicine Finland, Helsinki Institute of Life Sciences, University of Helsinki, Helsinki, Finland; 2) National Institute of Health and Welfare, Helsinki, Finland; 3) Department of Internal Medicine, University of Michigan, Ann Arbor, MI, US; 4) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, USA; 5) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, USA; 6) Program in Medical and Population Genetics, Broad Institute, Cambridge, USA; 7) Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, USA; 8) Department of Neurology, Massachusetts General Hospital, Boston, USA; 9) Department of Public Health, University of Helsinki, Helsinki, Finland; 10) Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland.

Interpreting genetic risk differences with polygenic risk scores (PRS) is challenging because complex genetic structure within and between the training and testing data can bias the results. Identifying such bias and correcting for it is crucial for a robust prediction of individual- and population-level risk for complex diseases and traits. Here, we study PRS differences of several complex diseases and traits between Western (WF) and Eastern (EF) Finland using 2,376 geographically well-defined samples from the National FINRISK study. Using height as a model trait, with a clear 1.6 cm difference between WF and EF, we: 1) compare differences predicted by three PRS based on three independent GWAS, 2) introduce an approach to detect bias accumulation in PRS, and 3) evaluate the effect of related and overlapping individuals between the GWAS and target samples. First, we built PRS for height using GWAS summary statistics from three sources: the GIANT consortium, the UK Biobank, and the population specific FINRISK study (2,376 target individuals excluded). The following table summarizes the proportion of variance of height in target samples explained by PRS (Target R^2) and shows how strongly the predicted height difference varies between different PRS.

Source GWAS	GWAS n	Variants in PRS	Target R^2	Predicted WF-EF difference in cm (95%CI)
GIANT	250,000	27,000	14%	3.52 (3.14, 3.90)
UK Biobank	300,000	113,000	22%	0.64 (0.39, 0.89)
FINRISK	25,000	51,000	15%	1.35 (1.14, 1.58)

Second, to detect the possible accumulation of biases, we generated several additional PRS with different numbers of seemingly unassociated variants (GWAS p-value > 0.5). This approach showed that the GIANT height PRS accumulates geographic differences from such variants more strongly than the other two PRS, explaining its unrealistically large prediction of WF-EF difference. Third, we showed that overlap and relatedness between GWAS samples and target samples have only a small effect on the geographic difference accumulation in our height data. In conclusion, our results demonstrate the importance of careful evaluation of possible biases in PRS, offer an approach to study bias accumulation, and propose that geographic height differences have a genetic background in Finland. We also report the geographic PRS distributions for other phenotypes, such as coronary artery disease and schizophrenia, using the framework we developed with height as our model trait.

Copyright © 2018 American Society of Human Genetics. All rights reserved

288

Polygenic adaptation signals for height are confounded by population structure. *R. Maier[1,2,3], M. Sohail[4,5,6], A. Ganna[1,2,3,7], A. Bloemendal[1,2,3], A. Martin[1,2,3], M. Daly[1,2,3], N. Patterson[8,9], B. Neale[1,2,3], I. Mathieson[10], D. Reich[8,9,11], S. Sunyaev[5,6].* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4) Systems Biology PhD Program, Harvard Medical School, Boston, MA, USA; 5) Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 6) Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA; 7) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; 8) Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 9) Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA; 10) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, USA; 11) Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts, USA.

In the past six years, numerous publications have reported directional selection signals of polygenic selection on height. Here, we show that many of these results are severely confounded by uncontrolled population stratification. In particular, signals of polygenic adaptation based on summary statistics from the from the GIANT consortium meta-analysis are dramatically reduced in magnitude and, in many cases no longer statistically significant when using summary statistics derived from the UK Biobank. This polygenic adaptation signal was apparently independently confirmed by a tests based on the singleton density score statistic (SDS). However, we show that this signal too is only present when using GIANT but not UK Biobank summary statistics. Specifically, the Spearman correlation between p-value and SDS statistic is 2e-65 using GIANT statistics but only 0.077 using UK Biobank. We further show that correlations between effect size estimates and allele frequency differences between North- and South- European populations underlie most of these discrepancies. The confounding with population stratification appears to be most severe for less significant SNPs; restricting the analyses to genome-wide significant SNPs results in a higher concordance between GIANT and UK Biobank data. While stratification-free within-family estimates suggest that the phenotypic north-south height gradient in Europe is indeed paralleled by genetically predicted height as reported before, the magnitude of this effect had been greatly overestimated. Height is widely used as a model for polygenic traits, which raises the question whether other methods using similar summary statistics might suffer from the same kind of confounding. Existing methods that aim to identify the presence of population stratification in GWAS summary statistics are not always applicable, and we provide simple suggestions that can supplement these methods in order to detect uncontrolled stratification and avoid resulting biases.

289

Polygenic risk scores perform poorly across populations. *B.D. Bitarello, I. Mathieson.* Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

The vast majority of genome-wide association studies (GWAS) are performed in cohorts of European ancestry. Systematic differences in polygenic risk scores (PRS) between European and non-European ancestry populations are believed to be largely spurious. However, it is not clear whether they are completely inaccurate nor how much individual-level predictive power is lost by applying PRS based on European-ancestry GWAS to non-European ancestry populations. Finally, a quantitative understanding of the biological or statistical basis for the poor performance of PRS in non-European-ancestry populations is lacking. To test this, we explored how well PRS predict a well-studied and highly polygenic trait: height. We calculated PRS using 41 sets of independent SNPs based on distance (physical or genetic) or LD clumping and pruning methods. We first compared PRS based on effect sizes from two independent GWAS: GIANT and UK Biobank (UKB), both of which were performed in individuals of European ancestry. We replicate previous observations of population-level differences in PRS, but these results are significantly different depending on datasets and clumping strategies. Depending on clumping strategy, the average difference between 1000 Genomes European and African PRS varies from 0.57-6.75 standard deviations (SD) using GIANT and 0.48-2.16 SD using UKB summary statistics. This dependence on clumping strategy supports the idea that most of these differences are spurious. We then investigated individual-level prediction in ~7,300 African American (AA) and ~7,400 European American (EA) individuals. Using UKB effect sizes, we found that PRS explain ~1.7% of height variation in AA individuals, compared to about 5.5% of variation in EA individuals. In both AA and EA, PRS explains more of the variance in height in individuals with a higher proportion of European ancestry. Interestingly, although we find a significant positive correlation (r=0.262, p=7.3e-115) between PRS and European ancestry among AA individuals, the correlation between height and European ancestry is extremely low (r=-0.02, p=0.056) and genome-wide ancestry explains only 0.05% of the variance in height, confirming that cross-population differences in PRS do not correlate to phenotypic differences. Finally, we evaluated whether local ancestry improves prediction for non-European populations, investigated dependence on other genomic features and extended our model to other phenotypes. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

290

Polygenic risk prediction for the world: A powerful approach for multi-ancestry meta-analysis across globally diverse populations. *A.R. Martin[1,2,3], P. Turley[1,2,3], R.K. Walters[1,2,3], H. Li[4], C.E. Carey[1,2,3], M. Kanai[1,2,3,5], H. Huang[1,2,3], C.Y. Chen[1,2,3,6], M. Lam[7], D. Palmer[1,2,3], M. Zacher[8], J. Koskela[9], G. Wojcik[10], M. Akiyama[5], C.R. Gignoux[11], E.E. Kenny[12], Y. Okada[5,13], Y. Kamatani[3,14], D. Cesarini[15,16], D. Benjamin[16,17,18], S. Ripatti[9], A. Palotie[9], B.M. Neale[1,2,3], M.J. Daly[1,2,3].* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; 2) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; 3) Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; 4) Department of Economics, Harvard University, Cambridge, MA; 5) Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; 6) Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; 7) Research Division, Institute of Mental Health, Singapore; 8) Department of Sociology, Harvard University, Cambridge, MA; 9) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki 00014, Finland; 10) Department of Genetics, Stanford University, Stanford, CA 94305, USA; 11) Colorado Center for Personalized Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO 80045, USA; 12) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 13) Department of Statistical Genetics, Osaka University Graduate School of Medicine, Osaka 565-0871, Japan; 14) Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan; 15) Department of Economics and Center for Experimental Social Science, New York University, New York, NY, USA; 16) National Bureau of Economic Research, 1050 Massachusetts Ave, Cambridge, MA 02138, USA; 17) Behavioral and Health Genomics Center, Center for Economic and Social Research, University of Southern California, 635 Downey Way, Los Angeles, CA 90089, USA; 18) Department of Economics, University of Southern California, 635 Downey Way, Los Angeles, CA 90089, USA.

The vast majority of genome-wide association studies (GWAS) are performed in individuals of European descent, raising questions about their applicability to other populations. Among the problematic consequences of ancestrally biased studies, we have previously shown that polygenic risk scores (PRS) computed from European GWAS explain many-fold less variation with increasing genetic distance observed in other populations. The fact that PRS offer far greater predictive value in individuals of recent European ancestry than others is perhaps the primary *ethical* and *scientific* challenge preventing their clinical implementation. To address this, we develop a novel multi-ancestry meta-analysis (MAMA) method. Unlike existing genetic risk prediction approaches, MAMA considers summary statistics from multiple populations. Our method recalibrates effect size estimates from GWAS for each population by modeling corresponding linkage disequilibrium (LD) patterns within and between these populations. This additional consideration is important because although causal variants are mostly shared across populations, the correlation between neighboring SNPs differs with population history. Thus, MAMA improves predictors for each population compared to existing approaches. In addition to testing our approach in simulations, we have assembled and/or harmonized summary statistics from four large, ancestrally diverse biobanks, including the UK Biobank (361,194 British European individuals), the FINRISK cohort (29,286 Finnish individuals), the Biobank Japan Project (162,255 Japanese individuals), and the Population Architecture for Genetic Epidemiology study (49,839 non-Europeans, mostly Hispanic/Latino and African American individuals). We apply MAMA to 16 quantitative phenotypes, including anthropometric, blood pressure, glycemic, inflammatory, lipid, and lifestyle traits. We also evaluate genetic risk of schizophrenia with data from the Psychiatric Genomics Consortium across European, East Asian, African American, and Hispanic/Latino populations. Here, we provide an analytical approach that makes better use of globally diverse GWAS to ensure that genetics does not exacerbate health disparities for those already most in need. This approach and translational genetics generally will be aided by further prioritization of larger and more diverse GWAS.

291

Deletion of expanded CGG repeats lowers *Fmr1* mRNA and increases FMRP levels in *Fmr1* knock-in mice. *C.M. Yrigollen[1], L. Ohl[1], E. Lim[1], S. Zheng[1], K. Brida[1], Y. Chen[1], A. Mas Monteys[1], B.L. Davidson[1,2].* 1) Raymond G. Perelman Center for Cellular and Molecular Therapeutics, The Children's Hospital of Philadelphia, Philadelphia, PA; 2) Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA.

The CGG trinucleotide repeat in the 5' untranslated region of the Fragile X Mental Retardation 1 (*FMR1*) gene is normally 5-44 CGG repeats in length. Premutation (55-200 CGG repeats) alleles of this triplet repeat can result in Fragile X-associated Tremor/Ataxia Syndrome (FXTAS), a late onset neurodegenerative disorder. Whereas a full mutation (greater than 200 CGG repeats) is the predominant cause of Fragile X Syndrome (FXS), a neurodevelopmental disorder that is the most common single gene cause of intellectual disability and autism. This study evaluates the use of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/Cas9 to delete the expanded CGG repeat for therapeutic benefit. We tested multiple gRNA sequences for cutting efficiency in HEK 293 cells prior to packaging two gRNAs and Cas9 into two AAV1 vectors. *Fmr1* knock-in mice harboring a premutation sized allele upstream of *Fmr1* were injected in the striatum with the optimized AAV1-CRISPR constructs and three weeks post injection the mice were euthanized. The striatum was isolated and DNA, RNA or protein was evaluated. PCR amplification of genomic DNA followed by sequencing showed partial CGG repeat deletion along with 3-48 nucleotides upstream and downstream of the repeats. Sequencing revealed an intact transcriptional start site and start codon. In contrast to control-treated mice, where *Fmr1* transcripts are elevated approximately 3-fold, AAV1-CRISPR treated mice had *Fmr1* mRNA levels similar to WT levels. The *FMR1* protein, FMRP, which is reduced in the *Fmr1* knock-in mice by approximately 50-75% reverted back to WT levels in mice injected with the CRISPR construct. These results are the first *in vivo* report of editing the *Fmr1* trinucleotide repeat with CRISPR and shows rescue of abnormal mRNA and protein expression.

Copyright © 2018 American Society of Human Genetics. All rights reserved

292

Deletion of *Hdac9* protein coding exons that also function as transcriptional enhancers, leads to *Twist1* haploinsufficiency and results in limb and craniofacial phenotypes. *R.Y. Birnbaum[1,2], N. Hirsch[1,2], F. Shemuluvich[1,2], T. Kaplan[3], D.G Lupiáñez[4,5].* 1) Ben-Gurion University of the Negev, Beer Sheva, Beer Sheva, Israel; 2) Department of Life Sciences, Faculty of Natural Sciences, The Ben-Gurion University of the Negev, Israel; 3) School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel; 4) Max Planck Institute for Molecular Genetics, RG Development and Disease, Berlin, Germany; 5) Institute for Medical and Human Genetics, Charité-Universitätsmedizin Berlin, Berlin, Germany.

The transcription factor *TWIST1* plays a vital role in mesoderm development, particularly in limb and craniofacial formation. *TWIST1* haploinsufficiency during development could lead to craniosynostosis and limb malformation such as Saethre-Chotzen syndrome. However, the transcriptional regulatory mechanism that controls *TWIST1* expression during development is yet to be elucidate. Here, we characterized active enhancers in the *TWIST1-HDAC9* locus that control transcription in the developing limb and branchial arches. Using p300 and H3K27ac ChIP-seq data, we identified 12 enhancer candidates encompass protein coding exons of *Histone deacetyase 9 (HDAC9*). Using zebrafish and mouse enhancer assays, we showed that 8 candidates have limb/fin and branchial arch enhancer activity that recapitulate *Twist1* expression. Each enhancer showed discrete activity pattern that together compile a spatiotemporal transcriptional regulation of *Twist1* in the developing limb/fin and branchial arches. Furthermore, we showed that *Twist1* enhancers are regulated by limb-expressed transcription factors, including Lmx1b and Tfap2 that bind and regulate their activity. Using 4C-seq, we showed that *Twist1* promoter interacts with *Hdac9* exons 18-19 that function as enhancers in the limb bud and branchial arch of mouse embryos at day 11.5. Using CRISPRCas9, we deleted a DNA sequence encompassing these *Hdac9* exons that altered *Twist1* expression (but not *Hdac9*) and led to polydactyly phenotype as seen in *Twist1* protein coding deficient mouse. In addition, deletion of *Hdac9* exons 4-5 that encompasses a CTCF site interacting with *Twist1* promoter region, altered *Twist1* expression and led to similar polydactyly phenotype. Thus, our study elucidated essential components of *TWIST1* transcriptional machinery, suggesting that alteration of *HDAC9* coding exons could lead to a phenotype that has nothing to do with *HDAC9* protein function but rather a disruption of *TWIST1* transcription regulation that lead to similar phenotypic outcome as *TWIST1* coding mutations.

293

Downregulation of *SNCA* expression by targeted editing of DNA-methylation: A potential strategy for precision therapy in PD. *B. Kantor, L. Tagliafierro, J. Gu, M. Zamora, E. Ilich, C. Grenier, Z. Huang, S. Murphy, O. Chiba-Falek.* Duke University.

SNCA gene has been associated with Parkinson's disease (PD) and accumulating evidence suggest that elevated levels of wild-type *SNCA* are pathogenic. On the other hand, robust reduction of *SNCA* level showed neurotoxicity, demonstrating that normal physiological levels of *SNCA* are needed to maintain neuronal function. Thus, there is an unmet need to develop new therapeutic strategies targeting the regulation of *SNCA* expression. DNA methylation at *SNCA* intron 1 contributes to the regulation of *SNCA* transcription, and differential methylation levels at *SNCA* intron 1 were found between PD and controls. These evidences established DNA-methylation at *SNCA* intron 1 as an attractive therapeutic target mediated by manipulation of *SNCA* levels. In this study, we developed a system that comprises of an all-in-one lentiviral vector for targeted editing the methylation in the CpG islands along the *SNCA* intron 1. The system is based on CRISPR/deactivated-Cas9 nuclease (dCas9) fused with the catalytic domain of the DNA methyltransferase 3A (DNMT3A). Applying the system to human induced pluripotent stem cells (hiPSC)-derived dopaminergic neurons from a PD-patient with the triplication of *SNCA* locus, resulted in targeted DNA-methylation of *SNCA* intron 1 that enabled fine-tuned downregulation of *SNCA*-mRNA and protein. Furthermore, we showed that the reduction in *SNCA*-mRNA levels by the gRNA-dCas9-DMNT3A system rescued cellular disease related phenotypes characteristics of the *SNCA*-triplication hiPSC-derived dopaminergic neurons, *e.g.* mitochondrial ROS production and cellular viability. Our findings established that DNA hypermethylation at particular CpG islands within *SNCA* intron 1 allows an effective and sufficient tight-downregulation of *SNCA* expression levels, suggesting the potential of this target sequence combined with the CRISPR/dCas9 technology as a novel epigenetic-based therapeutic approach for PD.

Copyright © 2018 American Society of Human Genetics. All rights reserved

294

Therapeutic modulation of epigenetic memory at a novel TGFβ2 enhancer in systemic sclerosis. *J.Y. Shin[1], J. Beckett[1], A. Shah[1], Z. McMahan[1], J. Paik[1], M. Sampedro[1], E. MacFarlane[1], M.A. Beer[1], D. Warren[1], F. Wigley[1], H.C. Dietz[1,2]*. 1) Johns Hopkins University School of Medicine, Baltimore, MD; 2) Howard Hughes Medical Institute, Bethesda, MD.

Systemic sclerosis (SSc) is a mysterious disease, in which adults acquire an inflammatory prodrome with subsequent fibrosis of the skin and viscera. Previously, we found that primary dermal fibroblasts (PDFs) of SSc patients maintain a fibrotic synthetic repertoire (FSR) in culture due to sustained histone acetyltransferase (HAT-EP300) activity and H3K27ac at a novel, distal enhancer of *TGFB2*, encoding a dominant pro-fibrotic cytokine. Sequencing did not reveal meaningful sequence variation within the enhancer. Though HAT inhibition normalizes TGFβ2 expression, in addition to H3K27ac and EP300 occupancy at the *TGFB2* enhancer in SSc PDFs, full rebound was promptly observed upon removal of the drug. These data suggested a form of epigenetic memory that maintains enhancer activity in SSc. We posited that this epigenetic memory—analogous to the regulation of inflammatory super-enhancers—might be initiated by inflammatory effectors, such as NF-kB, and enforced by the recruitment of BRD4 (a BET family member). In support of this hypothesis, we found significantly high NF-kB and BRD4 occupancy at the *TGFB2* enhancer in SSc PDFs. Interestingly, TNFa stimulation (a potent activator of NF-kB signaling) induced *TGFB2* expression in both control and SSc fibroblasts in a NF-kB and BRD4-dependent manner. In keeping with our hypothesis, treatment with NF-kB or BRD4 inhibitor normalized *TGFB2* expression in addition to H3K27ac and BRD4 occupancy levels at the *TGFB2* enhancer, all of which were now refractory to drug removal. We hypothesized that BRD4 inhibition with consequent suppression of TGFβ2 expression would normalize the FSR of SSc PDFs. Indeed, RNAseq confirmed that BRD4 inhibition significantly mitigated transcriptomic differences between control and SSc fibroblasts, resulting in productive modulation of gene expression related to extracellular matrix abundance and organization. Finally, organ-culture of control or SSc lesional skin biopsies was performed to test the *in vivo* efficacy of BRD4 inhibition. Ten-day incubation with the BRD4 inhibitor normalized *TGFB2* and collagen expression, with striking histological evidence for fibrotic regression in patient skin. These data reveal a complex pathogenic mechanism for SSc that is initiated by inflammatory events and reinforced by epigenetic dysregulation, and identify therapeutic targets that warrant further clinical investigation.

295

Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *I.E. Eres[1], K. Luo[1], Y. Gilad[1,2]*. 1) Department of Human Genetics, University of Chicago, Chicago, IL; 2) Section of Genetic Medicine, University of Chicago, Chicago, IL.

Over the last several decades, a growing body of evidence has suggested that variation in gene expression plays a crucial role in both speciation and tissue differentiation. However, a comprehensive functional understanding of the mechanisms regulating variance in gene expression remains elusive. Numerous studies have suggested that much of this regulation is driven by cis-regulatory elements (CREs), relatively short stretches of DNA that interface with transcription factors to make contact with gene promoters, thereby affecting expression. While chromatin conformation capture technologies have enabled a genome-wide quantification of these CRE-promoter contacts, relatively little interspecies research has been done. Of the few comparative studies examining the 3D genome, most focus on distantly related species and differences in broad-scale genomic structures, such as topologically associating domains (TADs). A higher-resolution evaluation of 3D genome divergence between more closely related species is necessary to understand how much differential contacts affect differential expression. To address these issues, we probed 3D regulatory divergence between humans and chimpanzees by performing Hi-C on induced pluripotent stem cells (iPSCs) from both species. Initial analysis of Hi-C data in iPSCs revealed that contacts were most different between humans and chimpanzees on chromosomes with large-scale structural rearrangements between the species. In order to assess how much variance in CRE-gene contacts is concomitant with gene expression divergence between species, we integrated our data with orthogonal RNA-seq data from the same individuals. Analyzing this joint dataset, we found that differentially contacting loci and differentially expressed genes were significantly more likely to be involved in a contact that crosses TAD boundaries in one species but not the other . We also found that as much as 12% of the interspecies variance seen in gene expression could be explained by interspecies variance in CRE-gene contacts. In addition, we quantified the overlap between species-divergent Hi-C contacts and published human iPSC histone mark data. We observed strong enrichment for both active and repressive marks in loci involved in contacts that were species specific, that overlapped a differentially expressed gene, or that were both. Overall, our data demonstrates that, as expected, 3D genome reorganization is key to explaining regulatory evolution in primates.

Copyright © 2018 American Society of Human Genetics. All rights reserved

296

3D GNOME 2.0: Three-dimensional GeNOme modeling engine of human genome structure at the population scale. *D. Plewczynski[1,3], M. Wlasnowolski[3], M. Sadowski[1,3], P. Szalaj[1], A. Kraft[1,3], Z. Tang[2], P. Michalski[2], Y. Ruan[2].* 1) Centre of New Technologies, University of Warsaw, Warsaw, mazovia, Poland; 2) 2The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA; 3) Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland.

Multiple evidences from the recent literature demonstrate the importance of genome spatial organization for gene regulation both in health and disease [1]. Several experimental examples were shown of structural variants (SVs) emerging in noncoding regions but disrupting local 3D genome organization, which altered gene transcription and led to malicious phenotypic effects. However, this area of research is dominated by the identification of spatial nucleome structure for selected human cell lines, lacking the population scale that is of great importance for the 1000 Genomes Project [4], and other cohort studies. We present here 3D GeNOme Modeling Engine 2.0 (3D-GNOME 2.0), a web service that generates *in silico* three-dimensional conformations of Human genome [3] from the list of structural variants (SVs) and reference 3D ChIA-PET data [2] (GM12878 cell line), and visualizes changes in chromatin higher order organization after introducing deletions, duplications, inversions, and insertions. 3D-GNOME 2.0 provides novel tools to visually inspect 3D models of chromatin loops, genomic domains (chromatin contact domains CCDs, or topologically associated domains TADs), or whole chromosomes, which differ in genomic 1D sequence. Heatmaps, contact diagrams and statistical plots are displayed for comparison between reference 3D structure and the one that is affected by given SVs. Deletions, duplications and other types of structural variants reorganize chromatin contacts by removing DNA segments participating in contacts, duplicating them or inverting the directionality of binding motifs for proteins bringing the segments together, as it can happen in case of CTCF protein. SVs that miss the interacting segments in most cases will only result in shortening or extending the corresponding chromatin loops. 3D-GNOME 2.0 server given a set of SVs alters an arrangement of chromatin loops provided as a reference (GM12878) and generates 3D structures of a chosen genomic region: the reference structure and the structure changed by contacts alternation, including Trios from 1kGP [5]. **References** [1] Tang et al. *Cell*. **2015** Dec 17;163(7):1611-27. [2] Szałaj P, et al. *Genome Res.* **2016** Dec;26(12):1697-1709. [3] Szalaj P et al. *Nucleic Acids Res.* **2016** Jul 8;44(W1):W288-93. [4] 1000 Genomes Project Consortium et al. Nature. **2015** Oct 1;526(7571):68-74. [5] Mark J.P. Chaisson et al. BioRxiv **2017** Sept 23 https://www.biorxiv.org/content/early/2017/09/23/193144. Submitted to Nature 2018.

297

Using multiplex non-coding CRISPR deletion libraries to individually perturb CTCF binding sites in the human genome. *S.K. Reilly[1,3], J. Xue[1,3], R. Tewhey[2], M. Bauer[1,3], P.C. Sabeti[1,3,4].* 1) The Broad Institute, Cambridge, MA., USA; 2) The Jackson Laboratory, Bar Harbor, ME. USA; 3) Harvard University, Cambridge, MA. USA; 4) Howard Hughes Medical Institute, Cambridge, MA.

The ubiquitously expressed CCCTC-binding factor (CTCF) is of particular interest as it organizes the 3-D organization of the genome, defines the boundaries of interacting chromatin, and directs enhancer-promoter DNA looping events. However, functionally characterizing each of its thousands of binding sites individually remains technically infeasible. Here we describe a novel non-coding CRISPR screening method and its use in investigating 100,000 CTCF sites individual, through the use of variable regions of the CTCF motif and flanking DNA sequence. We perturb nearly all sites directly bound by CTCF or in a CTCF-directed DNA loop in K562 cell lines. We identify that relatively few CTCF sites (5.3%) are essential for viability, and even fewer (.4%) that cause positive growth phenotypes in K562s, suggesting that most CTCF binding sites are individually dispensable. As expected, we find that guides targeting the core CTCF motif cause the strongest phenotypes and that essential CTCF sites lie closer to essential genes. These sites anchor fewer chromatin loops on the average anchor and these domains are smaller (average=170kb) than most loops (average=310kb). In a pan-cancer analysis we find that CTCF deletions causing growth defects are enriched in regions amplified in cancer, while deletions promoting growth are enriched in common cancer deletions. CTCF deletions promoting growth are also enriched near tumor suppressors. Intriguingly, pro-growth CTCF deletions are enriched in somatic mutations from the Pan-Cancer Analysis of Whole Genomes above previously reported rates of mutations in CTCF sites. Furthermore, we find enrichment of essential CTCF sites near known immortalizing translocations in the K562 line and in K562 inversions. Interestingly, a comparison of CRISPRko and CRISPRi screens against these CTCF sites shows that most CTCF sites are not acting solely as enhancers at these loci, suggesting that these sites are acting via looping or repression. We further describe in depth epigenetic, expression, and structural analysis for 30 individual growth-effecting CTCF loci, where we find notable changes in the structure of the local genome and expression of important growth-related genes. Together, this screen provides a novel CRISPR single guide targeting method, new insights in to epigenetic modulation of genes following CTCF deletion, and mechanistic examples of CTCF sites contributing to disease and cellular function. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

298

Structural variation and its impact on 3D genome structure in cancer cells. *J. Xu[1], J.R. Dixon[2], V. Dileep[3], Y. Zhan[4], F. Song[1], V.T. Le[2], G.G Yardımcı[5], A. Charkraborty[6], D.V. Bann[1], J.R. Broach[1], R. Kual[7], L. Zhang[1], T. Sasaki[3], J.C. Rivera-Mulia[3], H. Ozadam[4], B.R. Lajoie[4], J.A. Stamatoyannopoulos[7], S. Hadjur[8], D. Pezic[8], C. Ernst[9], D.T. Odom[9,10], R.C. Hardison[11], F. Ay[6,12], W.S. Noble[5], J. Dekker[4,13], D.M. Gilbert[3], F. Yue[1].* 1) College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA; 2) Salk Institute for Biological Studies, 10010 N Torrey Pines Rd. La Jolla, CA 92130, USA; 3) Florida State University, 319 Stadium Drive, Tallahassee, Florida 32306-4295, USA; 4) University of Massachusetts Medical School, Worcester, MA 01605, USA; 5) University of Washington, Seattle, USA; 6) La Jolla Institute for Allergy and Immunology, La Jolla, California 92037, USA; 7) Altius institute for Biomedical Sciences 2211 Elliott Avenue, Suite 410, Seattle, WA 98121, USA; 8) Cancer Institute, University College London, London, UK; 9) Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom; 10) German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany; 11) Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 12) School of Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA; 13) Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, MD 20815-6789, USA.

We present an integrative framework for comprehensively identifying structural variation and investigating their effect on 3D genome structure. For the first time, we applied next-generation optical mapping, high-throughput chromosome conformation capture (Hi-C) techniques, and whole genome sequencing to detect SVs in up to 35 of normal and cancer cell lines. We compiled a list of high-confidence SVs, and we resolve complex SVs and phase multiple SV events to a single haplotype. We further studied their impact on genome organization, including the formation of novel topological associating domains (TADs) and enhancer hijacking events. Furthermore, we observed that the disruption of TADs in cancer genomes is associated with changes in the gene expression, including many essential oncogenes. Our results underscore the importance of comprehensive structural variant identification and indicate that non-coding structural variation may be an underappreciated mutational process in cancer genomes.

299

Pitfalls of clinical exome and gene-panel testing: Alternative transcripts. *D. Bodian[1], P. Kothiyal[1], J. Schreiber[2], T. Vilboux[1], N. Hauser[1].* 1) Inova Translational Medicine Institute, Inova Health System, Falls Church, VA; 2) Pediatric Specialists of Virginia, Falls Church, VA.

Clinical gene-panel and exome sequencing can provide molecular diagnoses for patients with rare Mendelian disorders, but for many patients these tests are non-explanatory. Negative results have been attributed to factors including incomplete knowledge of disease architecture, a focus on exonic variation, challenges in variant pathogenicity interpretation, and technical limitations influencing variant calling. Incomplete consideration of alternative transcripts can also cause pathogenic variants to be missed. Using neonatal epilepsy as an example, we investigated whether interrogation of alternative transcripts in known disease genes could provide answers for additional patients. We integrated alternative transcripts for known neonatal epilepsy genes with RNA-Seq data to identify brain-expressed coding regions that are not evaluated by popular neonatal epilepsy clinical gene-panel and exome tests. We found brain-expressed alternative coding regions in 89 (30%) of 292 neonatal epilepsy genes. The 147 regions encompass 15,713 bases that are noncoding in the primary transcripts analyzed by the clinical tests. Alternative coding regions from at least 5 genes carry previously reported pathogenic mutations. Furthermore, we performed trio-based whole genome sequencing incorporating alternative transcripts for a patient presenting with intractable seizures at 2 weeks of age, for whom gene panel testing was unrevealing. We identified a de novo mutation, NM_001323289.1: c.2828_2829delGA, in an alternative transcript of the known epilepsy-associated gene *CDKL5*. The mutation was undetected by gene panel sequencing due to its intronic location in the *CDKL5* transcript typically used to define the exons of this gene for clinical sequencing (NM_003159). This is the first report of a patient with a mutation in an alternative transcript of *CDKL5*. These results demonstrate that assessment of alternative transcripts in exon-based clinical genetic tests, including gene-panel, exome, and whole-genome sequencing, may provide diagnoses for patients for whom standard testing is unrevealing.

Copyright © 2018 American Society of Human Genetics. All rights reserved

300

Curating clinically relevant transcripts for the interpretation of sequence variants. *M.T. DiStefano[1], S.E. Hemphill[1], B.J. Cushman[1], M.J. Bowser[1], E. Hynes[1], A.R. Grant[1], R.K. Siegert[1], A.M. Oza[1], M.A. Gonzalez[2], S.A. Amr[1,3], H.L. Rehm[1,4,5], A.N. Abou Tayoun[2,6].* 1) Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA; 2) Division of Genomic Diagnostics, The Children's Hospital of Philadelphia, The University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 3) Department of Pathology, Brigham & Women's Hospital, Harvard Medical School, Boston, MA; 4) Center for Genomic Medicine, Massachusetts General Hospital, Boston MA; 5) The Broad Institute of MIT and Harvard, Cambridge, MA; 6) Al Jalila Children's Specialty Hospital, Dubai, UAE.

Proper analysis of genomic variants is critical for patient care. The ACMG/AMP has set forth guidelines for the interpretation of sequence variants (Richards et al. 2015, PMID 25741868); however, these guidelines depend on defining the biologically relevant transcripts to accurately annotate the impact of variation on gene function. We have developed a systematic strategy for designating primary transcripts for variant interpretation and applied it to 109 hearing loss-associated genes from the OtoGenome™ Test (GTR000509148.8) at the Laboratory for Molecular Medicine (LMM). Genes were divided into 3 categories. Category 1 (C1) genes (N=38) had a single transcript, Category 2 (C2) genes (N=33) had multiple transcripts, but a single transcript sufficiently represented all exons, and Category 3 (C3) genes (N=38) had multiple transcripts with unique exons. Transcripts were curated with respect to gene expression reported in the literature and the Genotype-Tissue Expression (GTEx) Project. In addition, exonic loss-of-function (LoF) variants with a frequency over 0.3% were queried from the Genome Aggregation Database (gnomAD). All variants classified as pathogenic or likely pathogenic in ClinVar or as DM in the Human Gene Mutation Database were pulled and evaluated for each exon. These data were used to classify exons. "Clinically significant" exons lacked high frequency LoF variants or were supported by literature, "Uncertain significance" exons were spliced of out major transcripts, had no data in the literature, or, contained one high frequency LoF variant, and "Clinically insignificant" exons had non-supporting expression data or had multiple high frequency LoF variants. Interestingly, 6% of all exons were of "uncertain significance", yet contained >124 variants reported as clinically significant, questioning their accurate interpretation. Finally, we used exon-level next generation sequencing quality metrics generated across exome samples analyzed at LMM and the Children's Hospital of Philadelphia (CHOP) to identify a total of 43 exons in 20 different genes that had inadequate coverage and/or homology issues which may lead to missed or false variant calls. We have demonstrated that transcript analysis plays a critical role in accurate variant interpretation. Transcript curation such as this can greatly improve the quality of variant interpretation and patient care.

301

Transcript expression-aware annotation increases power for rare variant discovery in Mendelian and complex disease. *B.B. Cummings[1,2], K.J. Karczewski[1,2], J.A. Kosmicki[1,2], F.K. Satterstrom[1,2], T. Poterba[1,2], F. Aguet[1], C. Seed[1,2], M.J. Daly[1,2], D.G. MacArthur[1,2].* 1) Broad Institute, Cambridge, MA; 2) Analytical and Translational Genetics Unit. Massachusetts General Hospital, Boston, MA.

Alternative mRNA splicing enables individual exons of a gene to be expressed at very different levels across cell types. Consequently, genetic variants in different exons may have markedly different levels of tissue expression, and thus differential impact on biology and disease risk. Currently, no existing annotation tool systematically incorporates expression information into variant interpretation. Here we describe a transcript-level expression metric using isoform quantifications from RNA-seq of over 11,000 human samples across 53 tissue types from the Genotype Tissue Expression project (GTEx). Our method annotates the expression of every possible consequence of a variant in a VCF for each tissue, which can be used to summarize expression in any combination of tissues of interest. We validate this approach by showing it differentiates clearly between weakly and highly evolutionarily conserved exons, a proxy for functional importance ($p < 2.2$ e-16). In addition, we use data from the Genome Aggregation Database (gnomAD) to show that genetic variation in high-expressed exons is skewed towards low allele frequencies, suggesting that variation arising in such exons tends to be more deleterious. We show that our method filters falsely annotated loss-of-function (LOF) variants, removing 25% of LOF variants found in haploinsufficient developmental disease genes in gnomAD, while only filtering less than 1% of pathogenic variants in the ClinVar database in the same genes. We apply our expression filter to analysis of rare variants in autism patients and show that LOF variants in weakly expressed regions have effect sizes similar to synonymous variants [OR LOFs = 0.9, p = 0.63; OR syn = 1.0, p = 0.55] whereas LOF variants in highly expressed regions have much larger effect sizes [OR = 2.27, p = 2.7 x e^-10] than if no filtering was used [OR = 1.56, p = 1.7 e^-8], thus demonstrating that expression information boosts power in rare variant analysis. Our annotation is fast, flexible and generalizable. It takes under an hour to annotate gnomAD (>120,000 individuals) with the GTEx dataset. While this initial analysis utilizes GTEx, our method can be used to annotate any VCF file with any isoform expression dataset. We foresee this metric being applied in rare disease diagnosis, rare variant burden analyses in complex disorders, and prioritization of variants for recall-by-genotype studies.

Copyright © 2018 American Society of Human Genetics. All rights reserved

302

Converging on transcript annotation from Ensembl/GENCODE and RefSeq. *J.E. Loveland[1], S. Pujar[2], A. Astashyn[2], R. Bennett[1], C. Davidson[1], O. Ermolaeva[2], C. Farrell[2], L. Gil[1], M. Kay[1], K. McGarvey[2], A. McMahon[1], J. Morales[1], S. Rangwala[2], G. Threadgold[1], F. Cunningham[1], A. Frankish[1], T. Murphy[2].* 1) European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridgeshire, United Kingdom; 2) National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, U.S.A. 20894.

Accurate identification and description of the genes in the human genome provides the foundation for high quality analysis of data informing both genome biology and clinical genomics. Over the last two decades, the RefSeq project at NCBI and the GENCODE project led by EMBL-EBI have produced independent, high-quality, reference datasets describing the human gene complement. In 2005, the two groups, in collaboration with HGNC, UCSC, and others, initiated the CCDS project to develop a consensus set of protein-coding genes and coding regions with matching start, stop, and exon coordinates, which has resulted in a common definition for over 34,000 unique CDSes for over 95% of protein-coding genes. Building on this success, GENCODE and RefSeq have initiated a new project to extend this collaboration and 1) identify a representative transcript that captures most information about each protein-coding gene and 2) revise the annotation of that selected transcript in the RefSeq and GENCODE sets to completely match in overall exon structure, precise 5' and 3' transcript boundaries, and at the nucleotide sequence level. Our goals for 2018/2019 include convergence on key high value annotations to provide a common minimal set of transcripts per gene. Representative transcripts are identified using independent in-house computational pipelines complemented with joint manual review by annotation experts at EMBL-EBI and NCBI. The pipelines variously utilize evidence of conserved functional potential, RNAseq expression levels, and clinical significance to inform selection. Once a transcript has been identified, we define precise 5' and 3' transcript ends to use for both GENCODE and RefSeq using FANTOM5 CAGE data and polyA site data from conventional and nextGen sequencing methods. The result is an exactly matching annotation at the level of genome feature and transcript and protein sequence, with synonymous RefSeq and GENCODE identifiers. While a single transcript per gene will not capture the full spectrum of biological complexity at many loci, we envision this dataset serving as a unified high-value reference set for use in comparative genomics, clinical reporting, and basic research. We encourage the community to continue working with us to have input into this future goal. This work was supported in part by the intramural research program of the National Library of Medicine (NIH), and grants from the Wellcome Trust, and the National Human Genome Research Institute.

303

Patient reported strategies for managing uncertainty related to variants of uncertain significance. *S. Makhnoon, B. Shirts, H. Meischke, D. Bowen.* University of Washington, Seattle, WA.

Background: Variants of uncertain significance (VUS) are a well-recognized source of uncertainty in genomic medicine with clinical management challenges, yet our understanding of uncertainty management is limited. Since ability to deal with uncertain genetic test information is likely important determinant of patients' confidence in decision making, information seeking may help patients cope with their VUS result and aid subsequent clinical decisions. **Methods:** We conducted a mixed-methods study to understand uncertainty management strategies used by patients at the Seattle Cancer Care Alliance. Qualitative semi-structured interviews were conducted with 11 patients to understand their experiences of receiving VUS, reflections about their result and thoughts regarding implications of the result. We used a combined method of thematic analysis - the data was deductively compared against Han's taxonomy of uncertainties in genome sequencing and inductively analyzed to generate themes not specified in Han's taxonomy. Quantitatively, we surveyed 46 patients about their VUS information-seeking and management behavior. **Result:** Patients' primary concern with VUS involved personal and practical issues as they directly inform health care decisions. Patients demonstrated good understanding of the epistemic nature of VUS uncertainty. However, discordant provider explanations of the implication of this epistemic uncertainty for patients' diagnosis, prognosis, and therapy was a major contributor to overall uncertainty. Strategies for uncertainty management involved periodically checking back for reclassification and receiving concordant recommendation from providers. Other proactive strategies of uncertainty management such as information seeking and reading genetic test reports were not helpful. 52.4% of patients sought information about VUS after receiving test results - providers were preferred and trusted sources whereas family were not. Most did not undergo surgery (61.8%) or screening (62.5%) based on VUS. 46.7% asked family members to get genetic test because of their VUS result but 69.5% never checked back for VUS reclassification. Information-seeking was not associated with these behaviors. **Discussion:** We organized the various provider and patient level management strategies into a framework of uncertainty management strategies. These findings offer insight into patient experiences of VUS-related uncertainty which can be used to guide clinical management.

Copyright © 2018 American Society of Human Genetics. All rights reserved

304

Barriers to cascade testing: Impact on accessibility of a no-additional cost family genetic testing program for hereditary cancer risk. *E. Esplin, R. Miller, R. Truty.* Invitae, San Francisco, CA.

Background: To realize the full benefit of genetic testing in preventing disease, testing for pathogenic familial variants must be accessible to unaffected family members. Cascade family variant testing (FVT) uptake is limited by various barriers, including cost. In order to minimize barriers to genetic testing, Invitae implemented a policy including FVT at no additional charge for first-degree relatives (FDR) of probands who test positive for a pathogenic or likely pathogenic (P/LP) variant within 90 days after the initial test report is released. We assessed its impact on access to genetic testing by examining FVT uptake before and after policy implementation. **Methods:** We studied de-identified data from probands with P/LP variants in any of 80 hereditary cancer genes, and their relatives. 4,617 probands had P/LP variants identified in Jan-July of 2017, when the charge for FVT was $200 for FDR. 6,076 probands had P/LP variants identified in July 2017-Jan 2018, when FVT was no additional charge for FDR. **Results:** Overall uptake of FVT in hereditary cancer testing increased 81% in the six months post policy implementation. This included testing for P/LP variants in genes such as A*PC, ATM, BRCA1, BRCA2, CHEK2, MLH1, MSH2, MSH6, MUTYH, PALB2, PMS2, PTEN, RET, SDHA, SDHB, SMAD4, STK11*, and *TP53*. After accounting for the increase in overall proband testing, there was a significant increase in the percentage of probands with at least one relative tested during the no-additional charge time frame (19.6% vs. 23.2%, p=7.8e-6). Among the families that underwent FVT, an average of 1.9±0.1 (95% CI) relatives per proband had FVT requiring payment, while an average of 2.2±0.1 relatives per proband had FVT during the no-additional charge time frame. **Conclusion:** These data demonstrate a significantly increased uptake of FVT in the 6 months after removal of the cost barrier, both in breadth of families tested and depth of testing within a single family. Genes tested in this study have management guidelines with implications for disease prevention in the tested family members. It suggests the potential to further increase the overall uptake of cascade FVT as other barriers are removed. Additional research is needed to identify solutions for the remaining barriers, such as proband communication of test results with FDRs, in order to maximize the impact of genetic information-based precision medicine on the healthcare management of probands and their families.

305

Are we ready for genetic testing in all women with breast cancer? *P. Salyer[1], M. Bray[1], R. Reddy[2], J. Chen[1], S.L. Fisher[1], F.O. Ademuyiwa[1], L.J. Bierut[1].* 1) Washington University School of Medicine, St. Louis, MO; 2) The University of Texas at Dallas, Richardson, TX.

Introduction: Mutations in BRCA1, BRCA2, and other genes account for 5% to 10% of breast cancer diagnoses. Genetic testing to screen women for mutations that increase the risk for breast cancer could be a useful public health intervention. With advanced technologies, genetic testing is now relatively inexpensive, approximately the cost of a mammogram. This study examined how many African American women with breast cancer who were eligible for genetic testing according to the National Comprehensive Cancer Network (NCCN) guidelines, actually received testing. Methods: From 2016-2018, 250 African American women in the St. Louis area with a diagnosis of breast cancer completed comprehensive interviews regarding their personal health history and family history, and they received genetic testing through Color Genomics. We reviewed interviews against the NCCN guidelines for genetic testing to see how many women met the criteria. We compared these results to the Color Genomics results and to medical records to ascertain how many women had actually received previous genetic testing. Results: Approximately half of the women who met the criteria for genetic testing according to NCCN guidelines had received it in their clinical care. Women who were diagnosed with breast cancer at a younger age were more likely to receive genetic testing. 72% of women 45 and younger who qualified for genetic testing received it whereas only 38% of women who were older than 50 years and qualified for testing, received testing. Approximately 8% of those who qualified for genetic testing had a genetic mutation associated with higher risk for breast cancer. In addition, approximately 6% of women who did not meet NCCN guidelines for genetic testing had genetic mutations associated with breast cancer. Conclusions: Implementation of genetic testing for predisposing mutations associated with breast cancer remains a challenge, and many women who should be tested are not. In addition, many women who did not qualify for the NCCN guidelines for genetic testing had a predisposing genetic mutation identified (6%). Given the simplicity of universal testing and the increasingly low cost of genetic testing through companies like Color Genomics (currently $249 for testing of 12 genes associated with risk for breast cancer), it is time to consider genetic testing for all women with breast cancer. Funding Source: The Foundation for Barnes-Jewish Hospital, Siteman Cancer Center.

Copyright © 2018 American Society of Human Genetics. All rights reserved

306

Exploring choices among women undergoing panel-based genetic testing: A mixed-methods study. *J. Shuldiner[1], G. Rodin[1,2], J. Knight[1,3], Y. Bombard[1,4], K. Metcalfe[1,5], J. Kotsopoulos[1,5], S. Ferguson[1,2], A. Tone[2], J. McCuaig[2], M. Bernardini[2].* 1) University of Toronto, Toronto, Canada; 2) Princess Margaret Cancer Center, Toronto, Canada; 3) Lunenfeld-Tanenbaum Research Institute, Toronto, Canada; 4) Li Ka Shing Knowledge Institute, Toronto, Canada; 5) Women's College Research Institute, Toronto, Canada.

Background: The inclusion of a growing number of mutations in panel-based genetic testing (PBGT) raises many critical challenges for effective informed decision making about a wide range of potential findings. The growing use of PBGT for hereditary cancers in clinical practice means that more patients will face decisions regarding what genetic information they want to learn. Understanding panel choice is needed to develop guidelines specific to PBGT and guide clinical decision making. The present study examined how individuals undergoing PBGT decide how much genetic information to receive. METHODS: This mixed-methods study explored PBGT decisions among 380 women with >1 first-degree relatives with ovarian cancer. Participants could receive genetic information from some or all four panels. Quantitative data were collected on decisional conflict, risk perception, knowledge, anxiety, and cancer-related distress and PBGT choice. Qualitative data were collected through semi-structured interviews with purposeful sampling based on panel choice. Transcripts were analyzed using content analysis. Data from the questionnaires and interviews were integrated during the analysis phase. RESULTS: 348 women participated, 90% of whom chose to receive genetic information from all panels. Participants who selected all panels experienced less decisional conflict than those who chose specific panels. Qualitative analysis demonstrated those that selected all panels were interested in clinical and personal utility, and most did not consider potential emotional consequences from genetic test results. In contrast, those that chose specific panels were interested in only clinical utility, and the decision process included consideration of the implications of receiving genetic information and possible emotional consequences.Conclusion: Our results highlight the importance of the informed consent process for PBGT to ensure that individuals consider the impact of all types of genetic information and have realistic expectations regarding actionability of results.

307

Integrative epigenomics analyses identified a novel gene *ELF1* associated with lung cancer. *Y.X. Chen, Y.Y. Duan, H.M. Nui, Y. Rong, S. Yao, S.S. Dong, Y. Guo, T.L. Yang.* Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University.

Lung cancer is one of the most common causes of cancer-related deaths. Several genome-wide association studies (GWASs) have identified over 70 susceptibility SNPs associated with lung cancer. However, most susceptibility SNPs reported by GWASs reside within non-coding regions of genome and their mechanisms remain elusive. To evaluate functional enrichment of the lung cancer susceptibility SNPs on non-coding regulatory elements, we gathered the tag SNPs of lung cancer and identified 1223 proxy SNPs in LD with the tag SNPs across Asian and European populations. We performed enrichment analyses for the susceptibility SNPs in the transcription factor (TF) binding sites in lung adenocarcinoma related cell lines. The most significantly enriched TF was *ELF1*. We further evaluated the potential mechanism of *ELF1*, by integrating functional genomics, transcriptome and interactome datasets. Gene expression profiling analyses were performed in four independent datasets of lung adenocarcinoma and revealed that *ELF1* was significantly down-regulated ($P < 0.05$) in lung cancer samples. Additionally, more individuals of lung adenocarcinoma have the deletion of *ELF1* gene ($P = 3.29\times10^{-10}$) in the Cancer Genome Atlas (TCGA) data sets. For investigating coordinated genes of *ELF1* in lung cancer, we performed the co-expression analysis in lung adenocarcinoma samples, with the combination of differential gene expression analysis ($P < 0.05$). A total of 110 differentially expressed genes were positively co-expressed with *ELF1* ($P < 0.05$ and $rho > \mu+2\sigma$). Gene set enrichment analyses revealed that these genes were significantly enriched in cancer-related pathways (FDR < 0.05), including KEGG cancer pathway (FDR = 1.98×10^{-5}), pancreatic cancer pathway (FDR = 5.86×10^{-5}), and small cell lung cancer pathway (FDR = 2.41×10^{-3}). By mapping the relevant genes of *ELF1* to the protein-protein interaction network, we found that ELF1 acted with RB1 and NFKB1, both of which were protein products of cancer-related genes. Moreover, *in vitro* experiments found that the overexpression of *ELF1* inhibited cell migration and increased apoptosis in the lung cancer relevant cell line A549. In summary, our integrative analyses suggest that *ELF1* could play an import role in lung cancer as a suppressor.

Copyright © 2018 American Society of Human Genetics. All rights reserved

308

Germline genetic variants associated with immune response in colorectal cancer and their contribution to survival. *J.R. Huyghe[1], T. Hamada[2], B. Banbury[1], T.A. Harrison[1], C.S. Grasso[3], K. Nosho[4], K. Inamura[4], M. Giannakis[4,5], R. Nishihara[2,6,7,8,9], L. Hsu[1], W. Sun[1], P.A. Newcomb[1], S. Ogino[6,2,9,5], A.T. Chan[10,11,12,5,9,13], U. Peters[1].* 1) Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA; 2) Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, MA; 3) Parker Institute for Cancer Immunotherapy, San Francisco, CA; 4) Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA; 5) Broad Institute of Harvard and MIT, Cambridge, MA; 6) Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 7) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA; 8) Department of Nutrition, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA; 9) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA; 10) Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA; 11) Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 12) Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA; 13) Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA.

Immunotherapies are transforming cancer care and hold great promise in delivering cure for many patients. Yet, only a proportion of patients responds to immunotherapy, and underlying factors are poorly understood. Treatment success has been associated with the presence of tumor-infiltrating lymphocytes (TILs) in the tumor microenvironment (TME). TILs lie at the heart of cancer immunotherapies, and their abundance has also been associated with a good clinical outcome regardless of treatment in many different cancer types, including colorectal cancer (CRC). Efforts to elucidate factors that contribute to the observed variability in TIL abundance have mostly focused on tumor characteristics. Little attention has been paid to the contribution of germline (inherited) genetic factors and, therefore, their role is poorly understood. We hypothesized that germline genetic variants contribute to the observed between-patient variability in immune response to CRC tumors. To investigate this hypothesis, we performed a genome-wide association study (GWAS) for abundance of TILs using mRNA-seq and germline SNP6 genotype array data from The Cancer Genome Atlas. Using MCP-counter, we estimated abundance of 10 immune and stromal cell populations in the TME of 438 primary tumors from European ancestry CRC cases. We imputed to the Haplotype Reference Consortium panel. We found one association signal (*P*=2e-11; minor allele frequency ~5%) for T cell abundance at chr11q13.5 that reached genome-wide significance after Bonferroni adjustment for the 10 cell populations analyzed (5e-8/10). Nearby gene *B3GNT6* is a strong candidate that encodes the enzyme creating the core 3 structure of O-glycans, which are important precursors of mucin-type glycoproteins. This enzyme is primarily expressed in gastrointestinal tissues and plays a critical role in the maintenance of the colonic mucus barrier. In mice, defective core 3-derived O-glycans have been shown to lead to severe spontaneous chronic colitis and colorectal tumors. At the time of submitting this abstract, we are working on a replication analysis using GWAS data and relevant immunopathologic lymphocytic reaction measures obtained from several hundred CRC cases from the Nurses' Health Study and the Health Professionals Follow-Up Study. Additionally, we are evaluating whether this locus is associated with CRC-specific and overall survival in ~20,000 CRC cases from the International Survival Analysis of Colorectal Cancer Consortium.

309

Germline variants associated with immune infiltration in solid tumors. *S. Shahamatdar[1], M.X. He[2,3,4], E. Van Allen[2,3], S. Ramachandran[1].* 1) Center for Computational Molecular Biology, Brown University, Providence, Rhode Island; 2) Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts; 3) Broad Institute of Harvard and MIT, Cambridge, Massachusetts; 4) Harvard Graduate Program in Biophysics, Boston, Massachusetts.

The advent of immunotherapies has changed the treatment landscape of multiple cancers, inducing durable responses in patients with advanced disease. While numerous studies have probed the tumor-immune system relationship and gained insight into somatic genomic correlates of response and resistance to immunotherapies, germline features that associate with specific immune infiltrates in cancers remain incompletely characterized. Here, we present a pan-cancer analysis of 700,374 autosomal germline variants in the TCGA cohort (5796 European-ancestry samples across 30 cancer types) and their effects on the tumor immune microenvironment (TME), whose composition has been shown to be linked to response to immunotherapy and patient outcome. We analyzed germline associations with 58 published immune-related phenotypes that describe the TME, e.g. relative fraction of various immune cells in bulk tumor samples. Our genome wide association (GWA) study identified two SNPs ($p < 4e-9$) associated with immune response in solid tumors: (1) rs738034 (3' UTR of gene *IL17RA*, whose product is the receptor for the IL-17 cytokine that is produced by Th17 cells) is associated with a decrease in fraction of Th17 cells; and (2) rs167723 (intronic variant in gene *PPP2R2B*, a regulatory subunit of PP2A, a regulator of cell division and growth) is associated with an increase in BCR richness. We then conducted gene-level tests using PEGASUS (Nakka et al. Genetics, 2016) to identify genes underlying immune response in solid tumors. Gene-level association tests using PEGASUS resulted in 44 gene hits across 23 phenotypes. *CNTF* ($p = 4.16e-6$), previously implicated in hematopoietic development through cytokine signaling, is associated with fraction of macrophages in tumor samples. *IL17RA* ($p = 3.04e-10$) was identified in both SNP-level GWA and by PEGASUS as associated with fraction of Th17 cells in the TME. Finally, pathway analysis using HotNet2 (Leiserson et al. Nat Genet, 2014) revealed gene subnetworks that are significantly perturbed in various phenotypes, including subnetworks that are part of immune-related pathways: notably, genes *IFNA2*, *IFNA6,* and *IRF9* in the cytokine signaling pathway. Broadly, these results reveal multiple germline genetic features that associate with specific tumor-immune phenotypes across cancer, and may indicate new modes of probing cancer immunotherapy efficacy.

Copyright © 2018 American Society of Human Genetics. All rights reserved

310

Integrative genomic analyses to identify susceptibility genes for somatic mutations in human cancers. *Z. Chen[1], J. Bao[2], W. Wen[1], J. Long[1], Q. Cai[1], X. Shu[1], W. Zheng[1], X. Guo[1].* 1) Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN 37203, USA; 2) College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China.

Both somatic mutations and germline genetic variants influence carcinogenesis. Recent studies have revealed that somatic mutations can be characterized by distinct patterns, termed signature mutations. Several signature mutations have been have been found to be driven by defective cancer genes, such as *BRCA1*, *TP53* and *APOBEC*. Genome wide association studies (GWAS) have identified numerous cancer related common germline variants. Most GWAS-identified variants or variants in strong linkage disequilibrium (LD) are believed to play regulatory roles in gene expression. However, it is unclear whether target genes regulated by these germline variants are associated with somatic mutations through which to contribute to carcinogenesis. Here, we conducted a cis-expression quantitative trait loci (cis-eQTL) analysis for GWAS-identified variants for nine common cancer types using transcriptome data from The Cancer Genome Atlas (TCGA) and the Genotype-Tissue Expression (GTEx) project. We identified approximately hundreds of target genes for these cancer types. We further evaluated associations between expression of these target genes and mutation signatures for each cancer type. We found that the expression of two genes, *GPR143* and *SHROOM2*, are positively associated with distinct mutation signatures in colorectal cancer, at Benjamini-Hochberg (BH)-adjusted $P < 0.05$. In breast cancer, a total of 16 genes were identified using the same cutoff, including *APOBEC3A* and *DCLRE1B*, known DNA repair genes. In addition, we also constructed a polygenic risk score (PRS) for each cancer type based on GWAS-identified variants. Our result showed that, at BH-adjusted $P < 0.05$, PRS was associated with one mutation signature in lung squamous cell carcinoma (LUSC) and multiple signatures in melanoma. In summary, our findings provide additional insights into the understanding of genetic risk variants contributing to carcinogenesis by regulating genes with affecting somatic mutations.

311

Pathogenic mutations in GABAergic transporters (GAT1, GAT3) and biosynthetic enzymes (GAD65, GAD67) cause a spectrum of genetic generalised epilepsy syndromes. *S.K. Chung[1], K. Everett[2], E. Dudley[1], J.G. Mullins[1], I. Scheffer[3], S. Berkovic[3], M.I. Rees[1].* 1) Swansea University Medical School, Swansea University, Swansea, Wales, United Kingdom; 2) St George's University of London, Cranmer Terrace, London, UK; 3) Epilepsy Research Centre, Department of Medicine, University of Melbourne, Austin Health, Heidelberg 3084, Victoria, Australia.

The link between inhibitory GABAergic neurotransmission and genetic generalized epilepsy (GGE) is a well-supported paradigm generated from molecular genetics of GABA receptor subunits, transgenic model systems, pharmacology, neuropathology, and electrophysiology. The focus on postsynaptic GABA channelopathies has historically attracted much attention, however, the research focus is evolving towards a proteomic synaptopathy approach. Consequently, we investigated the genetic variation within GABAergic transporters (*SLC6A1*; GAT1 & *SLC6A11*; GAT3) and biosynthetic proteins (GAD65 & GAD67) in 708 GGE cases from Australia, UK and New Zealand using a combination of light-scanner heteroduplex screening and direct Sanger sequencing. Eighteen novel or rare GABAergic variants were discovered in 34 GGE cases with a strong preference for absence seizure syndromes (CAE, MAE, JAE). All 18 GABAergic missense variants had deleterious outcomes in mutation-prediction software and protein-structure molecular modelling. FLAG-tagged and untagged mutation expression-constructs were prepared for all 18 GABAergic gene-variants in addition to a further 7 published *de novo* GAT1 mutations in myotonic-atonic epilepsy, but without functional validation. GAT1 and GAT3 variants were analysed using various *in vitro* functional assays including a bespoke non-radiolabelled GABA-activity assay based on gas-chromatography electron impact mass spectrometry (GCMS) and cell-surface trafficking status. Reduced GABA activity or cell-surface defects were detected for the majority of GAT1 and GAT3 gene-variants. Consistent with previous data, all GAT1 cases presented with severe learning disabilities whilst GAT3 cases did not present with this co-morbidity. In addition, the epilepsy syndrome presentation became more severe as a direct correlation with the degree of transporter knockdown. GAD65 and GAD67 mutation constructs were tested in a florescence-based GAD enzymatic assay using a resazurin-linked GABase microplate format. This revealed altered GAD activity and reduced binding properties to co-factor, pyridoxal 5′-phosphate (PLP). This study confirms the association between GABA transporters / GABA biosynthesis with GGE syndromes, which is supported by extensive *in vitro* evidence and has important patient / family-specific outcomes. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

312

Evaluation of copy number burden in specific epilepsy types from a genome-wide study of 18,329 subjects. *L.M. Niestroj[1,2] on behalf of the Epi25k Consortium.* 1) Cologne Center for Genomics, Cologne, NRW, Germany; 2) University of Cologne, Cologne, NRW, Germany.

Rare copy number variants (CNVs) are strongly implicated in the etiology of epilepsy. However, it is not clear to which degree CNVs contribute to specific epilepsy types. Here we present the largest epilepsy CNV burden analysis to date including >11k patients and >7k controls ascertained in the 'Epi25k' Consortium. Samples were genotyped using Illumina Global Screening Array-24 v1.0. After quality control, we analyzed a cohort of 11,051 European epilepsy cases and 7,278 ancestry-matched controls for burden of rare CNVs (< 1% frequency). Epilepsy type analyses were performed for: Non-acquired focal epilepsy (NAFE, n=4,597), genetic generalized epilepsy (GGE, n=3,538), epileptic encephalopathy (EE, n=1,278), lesional focal epilepsy (LFE, n=1,238) and unclassified epilepsy (UE, n=400). Significant enrichment of deletions was observed for all cases (odds ratio (OR) =1.93, 95%-CI=1.53-2.45, $P=3.1\times10^{-8}$), with the strongest signal coming from hotspot deletions in GGE (OR=2.69, 95%-CI=2.06-3.53, $P=5.8\times10^{-13}$) and NAFE (OR=1.63, 95%-CI=1.63-1.23, $P=7.1\times10^{-4}$). The hotspot signal in GGE was mainly driven by two CNVs, *15q13.3* (OR= 37.2, 95%-CI=5.87-1538.89, $P=2.4\times10^{-8}$) and *16p13.11* (OR=21.72, 95%-CI=5.3-190.7, $P=7.4\times10^{-9}$). Large deletions (>2Mb) were enriched in EE (OR=3.73, 95%-CI=2.1-6.07, $P=9.5\times10^{-7}$), GGE (OR=1.32, 95%-CI=1.11-1.55, $P=9.96\times10^{-4}$) and LFE patients (OR=3.13, 95%-CI=1.5-5.66, $P=6.5\times10^{-4}$). Patients with UE were the only group, which showed significant enrichment for gene covering duplications (OR=1.55, 95%-CI=1.27-1.89, $P=1.4\times10^{-5}$). No patient group showed genome-wide burden for noncoding CNVs. Exome-wide gene level analysis showed strongest enrichment for genes in hotspot regions at *15q13.2* and *15q13.3* in GGE cases (*FAN1*, *OTUD7A* both OR=23.8, 95%-CI=5.87-208, $P=9.3\times10^{-10}$; *KLF13,* OR=22.8, 95%-CI=5.59-199, $P=2.6\times10^{-9}$) and at *15q11*, *15q12* and *15q13* in EE cases (*GABRA5, GABRB3* both OR=inf, 95%-CI=12.84-inf, $P=5.4\times10^{-9}$; *ATP10A,* OR=inf, 95%-CI=11.3-inf, $P=3.6\times10^{-8}$). Even when excluding hotspots CNVs, significant enrichment remained for deletions covering genes under evolutionary selection (pLI>0.95) in GGE (OR=1.57, 95%-CI=1.28-1.9, $P=1.0\times10^{-5}$). In summary, using the largest patient cohort to date we are able to discover novel CNV associations with epilepsy types. We provide evidence for deletion burden outside of known hotspot regions for GGE affecting constrained genes and for the first time implicate large deletions in the genetic architecture of LFE.

313

***De novo* pathogenic variants in neuronal differentiation factor 2 (*NEUROD2*) cause a form of early infantile epileptic encephalopathy.** *A. Sega[1], E. Mis[1], K. Lindstrom[2], S. Andrews[3], W. Ji[1], M. Konstantino[1], M. Cho[4], J. Juusola[4], M. Khoka[1,5], S. Lakhani[1].* 1) Pediatric Genomics Discovery Program, Department of Pediatrics, Yale University School of Medicine, New Haven, Connecticut, USA; 2) Division of Genetics and Metabolism, Phoenix Children's Hospital, Phoenix, Arizona, USA; 3) Genetics and Genome Biology Program, Research Institute, The Hospital for Sick Children, Toronto, Ontario, Canada; 4) GeneDx, Gaithersburg, Maryland, USA; 5) Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, USA.

Early infantile epileptic encephalopathies are severe disorders consisting of early-onset refractory seizures accompanied often by significant developmental delay. The increasing availability of next generation sequencing has facilitated the recognition of single gene mutations as an underlying etiology of congential diseases, including forms of early infantile epileptic encephalopathies. This study was designed to identify candidate genes as a potential cause of early infantile epileptic encephalopathy, and then to provide genetic and functional evidence supporting patient variants as causative. We used whole exome sequencing to identify candidate genes. To model the disease and assess the functional effects of patient variants on candidate protein function, we used *in vivo* CRISPR/Cas9-mediated genome editing and protein overexpression in *Xenopus* tadpoles. We identified novel *de novo* variants in *NEUROD2* in two unrelated children with early infantile epileptic encephalopathy. Depleting *neurod2* with CRISPR/Cas9-mediated genome editing induced spontaneous seizures in tadpoles, mimicking the patients' condition. Overexpression of wildtype NEUROD2 induces ectopic neurons in tadpoles; however, the patient variants were markedly less effective, suggesting that both variants have loss of function and are likely pathogenic. This study provides clinical and functional support for NEUROD2 as a cause of early infantile epileptic encephalopathy, the first evidence of human disease caused by *NEUROD2* variants.

Copyright © 2018 American Society of Human Genetics. All rights reserved

314

The genetic landscape of the developmental and epileptic encephalopathies. *A.M. Muir[1], C.T. Myers[1], M.G. Mehaffey[1], K. Boysen[2], G. Hollingsworth[2], C. King[3], A. Schneider[2], A. Buttar[1], A. Chowdhary[1], A.B.I Rosen[1], M. Sud[1], N. Weed[1], J.X. Chong[4], M.J. Bamshad[4], D.A. Nickerson[4], UW Center for Mendelian Genomics[5], L.G. Sadleir[3], I.E. Scheffer[2,6,7], H.C. Mefford[1].* 1) Pediatrics, University of Washington, Seattle, WA. USA; 2) Epilepsy Research Centre, Department of Medicine, Austin Health, The University of Melbourne, Heidelberg, Victoria, 3084, Australia; 3) Department of Paediatrics and Child Health, University of Otago, Wellington, New Zealand; 4) Department of Genome Sciences, University of Washington, Seattle, WA, US; 5) University of Washington, Seattle, WA, US; 6) Florey Institute of Neuroscience and Mental Health, The University of Melbourne, VIC 3010, Australia; 7) Department of Paediatrics, Royal Children's Hospital, The University of Melbourne, Parkville, Victoria, 3050, Australia.

Developmental and epileptic encephalopathies (DEEs) are a group of severe, early onset epilepsies characterized by refractory seizures, frequent epileptiform activity, developmental delay and/or regression, and often a poor prognosis. DEEs are highly genetically heterogeneous with over 60 genes implicated. Despite significant progress, about half of DEEs remain unexplained with current clinical diagnostic testing. We sought to identify novel genetic causes of DEEs through exome sequencing of168 proband-parents trios, 21 proband-parent duos, and 25 proband only singletons. Prior to exome sequencing, all probands were negative on screening for pathogenic variants in 58 known DEE genes. We prioritized ultra-rare, nonsynonymous variants for review. We explored various modes of inheritance including *de novo* (constitutive and somatic mosaic), compound heterozygous, newly homozygous, and X-linked hemizygous. We identified pathogenic or likely pathogenic variants (according to American College of Medical Genetics and Genomics criteria) in 36/214 probands providing a molecular diagnosis for 16.8% of our unsolved cohort. Pathogenic or likely pathogenic *de novo* variants in known DEE genes were identified in 13/36 (36%) cases. These variants were missed in previous screening attempts due to insufficient coverage at the variant site or undetected mosaicism of the variant in the proband or his/her parent. Biallelic pathogenic or likely pathogenic variants were identified in 11/36 (31%) solved cases – a higher than expected occurrence of recessive inheritance in a cohort of apparently sporadic DEE. Finally, we identified pathogenic or likely pathogenic *de novo* variants in genes associated with related neurodevelopmental disorders including intellectual disability and autism, in 12/36 (33%) solved cases. In addition, we identified *de novo* variants of uncertain significance in 110 genes not previously linked to DEE or other neurological disorder. Multiple *de novo* variants in the same candidate gene were not identified. We are currently preforming targeted sequencing of 55 of these candidate genes in a cohort of ~750 unsolved DEE patients to identify additional cases. To date, we have identified second cases with *de novo* variants in 4 genes: *CMPK2*, *CPSF1*, *IRF2BPL* and *KCNV2*. Exome sequencing in 214 DEE cases identifies novel candidate DEE genes, implicates novel pathways, and confirms the genetic heterogeneity of this group of disorders.

315

Large-scale trans-ethnic genome-wide association study reveals novel loci, causal molecular mechanisms and effector genes for kidney function. *A.P. Morris[1], A. Akbarov[2], M. Tomaszewski[2], N. Franceschini[3], COGENT-Kidney Consortium.* 1) Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom; 2) Division of Cardiovascular Sciences, University of Manchester, Manchester, United Kingdom; 3) Department of Epidemiology, University of North Carolina, Chapel Hill, NC.

Chronic kidney disease (CKD) is a major public health burden and affects nearly 10% of the global population. We assembled published and de novo genome-wide association studies of estimated glomerular filtration rate (eGFR), a measure of kidney function used to define CKD, in up to 312,468 individuals of diverse ancestry. We identified 93 loci attaining genome-wide significant evidence of association with eGFR ($p<5x10^{-8}$), including 20 mapping outside those previously reported for kidney function, with the strongest novel signals mapping to/near *MYPN* (rs7475348, $p=8.6x10^{-19}$), *SHH* (rs6971211, $p=6.5x10^{-13}$), *XYLB* (rs36070911, $p=2.3x10^{-11}$) and *ORC4* (rs13026220, $p=3.1x10^{-11}$). Across loci, we identified 127 distinct association signals at locus-wide significance ($p<10^{-5}$), including four in each of the regions mapping to *SLC22A2* and *UMOD-PDILT*. Allelic effects on eGFR of index variants were consistent across populations, with no evidence of heterogeneity due to ancestry (Bonferroni correction, $p_{HET}<0.00039$), consistent with a model in which causal alleles are shared between ethnicities. Integration with functional and regulatory annotation revealed that eGFR association signals were jointly enriched in coding sequence, kidney-specific histone modifications and binding sites for HDAC2 and EZH2. Class I histone deacetylases (including HDAC2) are required for embryonic kidney gene expression, growth and differentiation. Annotation-informed fine-mapping, taking advantage of differential patterns of linkage disequilibrium across diverse populations, revealed 40 variants accounting for more than 50% of the posterior probability (π) of driving eGFR association signals. These included: coding alleles *GCKR* p.Leu446Pro (rs1260326, $p=2.0x10^{-35}$, π=86.1%), *CPS1* p.Thr1406Asn (rs1047891, $p=1.5x10^{-29}$, π=98.1%), *CERS2* p.Glu115Ala (rs267738, $p=1.7x10^{-10}$, π=55.3%) and *CACNA1S* p.Arg1539Cys (rs3850625, $p=2.5x10^{-29}$, π=99.0%); and expression quantitative trait loci in kidney tissue from the TRANSLATE Study for *FGF5* (rs12509595, $p=4.7x10^{-16}$, π=57.1%), *TBX2* (rs887258, $p=2.7x10^{-13}$, π=62.2%), and both *UMOD* and *GP2* for one signal at the *UMOD-PDILT* locus (rs77924615, $p=1.5x10^{-54}$, π=100.0%). These results define novel causal molecular mechanisms underlying kidney function association signals, and highlight genes through which their effects are mediated, offering a potential route to clinical translation and CKD treatment development.

Copyright © 2018 American Society of Human Genetics. All rights reserved

316

High resolution fine mapping of 406 smoking/drinking loci via a novel method that synthesizes the analysis of exome-wide and genome-wide association statistics. *Y. Jiang, the GWAS and Sequencing Consortium of Alcohol and Nicotine Use.* Penn State College of Medicine , PA.

Smoking and drinking are leading heritable and modifiable risk factors for many diseases. Recently, the GSCAN consortium meta-analysis identified 406 loci with $p<5x10^{-8}$, using >1 million individuals, of which 395 were significant for the first time. Functional dissection of these loci can lead to considerable advancement for addiction genetics. As a first step, we performed fine mapping of the 53 loci for the cigarettes per day (CPD) phenotype, while the fine-mapping of other traits are on-going. We aggregated summary statistics of exome-array data from 19 studies and augmented them with GWAS data from UK Biobank data. Together, the aggregated dataset consists of 14.8 million variants with a maximal sample size of 433,216. Not all contributing studies have both exome array and GWAS data. The association statistics for many variants were thus missing from some contributing studies. This missingness may skew the correlation between marginal score statistics, and lead to the incorrect determination of causal variants using existing fine mapping methods. We developed a novel method called partial correlation-based scores statistic (PCBS), which allows the correct estimation of joint effects when the contributed summary association statistics contain missing data. We further extended this method to incorporate functional genomic data from ENCODE, Roadmap and GTEx for fine mapping. To proceed, we defined a locus as the 500KB window surrounding the sentinel variant. In total, 69% of the loci contain multiple independently associated variants. Our fine-mapping results pinpointed causal variants with high resolution, with a median of 8 variants in 95%-credible sets. The 95%-credible sets were narrowed down to a single SNP for 25% of the CPD loci. Fine mapping elucidated the causal variants in many known addiction pathways and pointed to novel causal genes. For example, the causal variant for the loci surrounding rs75596189 is mapped to rs3025383 in *DBH*, a gene key to the dopamine system reward pathway. Causal variants in the nicotine receptor gene loci are fine-mapped to rs938682 (CHRNA3), rs4344703 (CHRNB4), rs8034191 (AGPHD1), all with probably damaging functional prediction. Together, our results significantly expanded our understanding of addiction genetics. The PCBS-based methods are particularly useful for the next phase of fine mapping study, where GWAS and sequence data (e.g. from TOPMed) that are not measured on all study subjects are synthesized.

317

Fine-mapping of type 2 diabetes and glycemic traits with whole genome sequence data using 49,022 individuals from the NHBLI's TOPMed WGS Program. *A.K. Manning[1,2,3], D. Dicorpo[4], J. Wessel[5], TOPMed Diabetes Working Group.* 1) Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Boston, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Harvard Medical School, Boston, MA; 4) School of Public Health, Boston University, Boston, MA; 5) Indiana University, Indianapolis, IN.

Whole genome sequence (WGS) association studies afford the opportunity to perform trans-ancestry fine-mapping without depending on imputation. We have leveraged large, phenotypic-rich and ancestry-diverse cohorts from NHBLI's Trans-Omics for Precision Medicine (TOPMed) WGS Program to refine credible sets and discover novel distinct associations with type 2 diabetes (T2D), and fasting glucose (FG) and fasting insulin (FI) levels. We initially focussed on loci with known associations with T2D and glycemic traits or genes involved in monogenic diabetes and insulin resistance syndromes, and then expanded to describe novel associations for which we are seeking additional support. We performed ancestry-specific genetic association analysis using GENESIS mixed models with common (minor allele frequency [MAF]>1%), low-frequency (0.01%<MAF<1%) and rare (MAF<0.01%) variants, correcting for relatedness and population structure with a genetic relationship matrix derived from pruned common variants. We used PAINTOR for fine-mapping, and further refined our credible sets by leveraging ancestry-specific linkage disequilibrium (LD) and regulatory and chromatin accessibility annotations from tissues shown to be enriched in common variant association signals: pancreatic islets for T2D, liver, adipose and muscle for FG/FI. Our analysis included data from 16 TOPMed projects and 5 ancestries. For the T2D analysis: European N=4,781 with T2D/21,365 without T2D; African-American: 3,783/9,470, Hispanic: 612/1628, Asian: 427/1,973, Samoan: 185/922). For FG/FI: European N=13,749 individuals without T2D; African-American: 7,256, Hispanic: 2,005, Asian: 2,235, Samoan: 922. For T2D, in ancestry-combined meta-analyses, 90 variants met genome-wide significance (P<5e-8), all common: 4 variants in *SLC30A8*, 76 variants at *TCF7L2* of which 8 are multi-allelic variants, 2 variants at *KCNQ1* and 8 variants at *FTO* of which 1 is a short insertion/deletion. A potentially novel association at *MYO1F* shows a rare variant signal specific to African-ancestry individuals (MAF=0.002; P=2e-10). For FG, significant associations were seen at 6 loci with previously reported signals: *GCKR*, *G6PC2*, *GCK*, *SLC30A8*, *MTNR1B*, and *FOXA2*, all common variants. Nominally significant (P<5e-6) low-frequency variant associations were seen at 6 loci: *VPS13C*, *PRDM16*, *SLC2A1-AS1*, *INS*, *ACSL1*, and *CDKAL1*. We soon will extend our analysis into the next release of WGS data from TOPMed (N~100,000 individuals).

Copyright © 2018 American Society of Human Genetics. All rights reserved

318

Single-base resolution of autoimmune disease associations using molecular phenotypes. *K. Kundu[1,2], S. Watt[1], A. Mann[1], K. De Lange[1], L. Vasquez[1], BLUEPRINT Consortium[1], L. Chen[3], J. Barrett[1], C. Anderson[1], N. Soranzo[1,2].* 1) Department of Human Genetics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1HH, UK; 2) Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge, Cambridgeshire, CB2 0PT, UK; 3) West China Second University Hospital, State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610041, People's Republic of China.

In-depth understanding of molecular mechanisms of disease informs the development of new therapeutic approaches. Autoimmune diseases collectively affect almost 10% of the world's population. To date, close to a thousand genetic loci have been associated with the risk of autoimmune diseases through genome-wide association studies. Characterising the causal genetic variants, putative effector genes and molecular mechanisms underpinning these associations is the necessary next step to harness the power of these genetic discoveries. Here we extend the evaluation of molecular QTLs generated as part of the BLUEPRINT project to systematically map molecular mechanisms and causal genetic variants at 14 different autoimmune diseases with publicly-available summary statistics. We first recomputed molecular QTLs for high-resolution genetic, epigenetic, and transcriptomic profiling in three primary human immune cell types (i.e., monocyte, neutrophil, and T-cell) using a denser genotype map. We then used colocalisation analysis to identify shared genetic effects between each molecular and disease trait, and showed 176 unique, non-HLA disease loci achieved high posterior probability of colocalisation (PP≥0.99). We next sought to test the relative resolution of disease and molecular QTLs for defining credible sets of causal variants at colocalised loci. We optimized a Bayesian fine-mapping framework for analysis of disease and molecular QTL summary statistics and applied it to fine-map associations at 346 independent disease-QTL colocalised loci. We show that fine-mapping of molecular traits systematically improves resolution of causal variants compared to disease summary statistics alone. We were able to resolve causal credible sets (>95% posterior probability) to less than 20 variants for approximately 67% of loci using molecular QTLs, compared to 22% based on disease summary statistics alone. More importantly, molecular QTLs provide interpretable molecular mechanisms at a large set of putative causal variants. For example, fine mapping of the *ITGA4* locus associated with inflammatory bowel disease yields smaller credible sets for expression (n=2 variants), H3K4me1 (n=3) and H3K27ac (n=5) QTLs compared to use of disease summary statistics alone (n=11), and highlights a putative role for one promoter variant affecting CEBPB binding. Overall, our analysis clearly demonstrates how the use of molecular data empowers the interpretation of disease associations.

319

Fine-mapping causal regulatory variants using massively parallel reporter assays. *N.S. Abell[1], M.K. DeGorter[2], M.J. Gloudemans[3], B. Balliu[2], K.S. Smith[2], S.B. Montgomery[1,2].* 1) Department of Genetics, Stanford University, Stanford, CA; 2) Department of Pathology, Stanford University, Stanford, CA; 3) Biomedial Informatics Program, Stanford University, Stanford, CA.

Studies like the 1000 Genomes Project have produced large cohorts of individuals for which gene expression can be measured and associated with genetic variation. However, a frequent challenge in eQTL mapping is that associative signals may be statistically indistinguishable due to linkage disequilibrium (LD) blocks, which complicates their biological or clinical interpretation. In such cases, genetic associations require either additional information to break apart LD blocks, or some degree of experimental evaluation prior to interpretation. In this study, our goal was to resolve LD-linked eQTL regions derived from several 1000 Genomes populations using a combined analytical and experimental approach. First, we identify eQTLs consisting of statistically tied sets of variants across populations and perform a meta-analysis to prioritize candidate causal variants. We characterize the genomic properties of these loci, specifically their enrichments across transcription factor binding sites and histone marks, and use them as a platform for experimental follow-up. Second, we design and perform a massively parallel reporter assay (MPRA) targeting ~50,000 tied and candidate causal variants derived from ~750 LD-tied eQTL regions. To achieve this scale, we obtained an oligonucleotide library consisting of ~100,000 sequences and attached randomized 20-bp barcodes by PCR, using sequencing to associate barcode to genomic sequence. Following insertion of a GFP reporter gene coupled to a minimal promoter, we transfected our randomly barcoded reporter library into NA12878 cells, recovered GFP mRNA and sequenced the attached barcodes. We successfully recover >99% of sequences and identify significant regulatory and allele-specific effects. We also characterize the epigenetic determinants of activity and overall concordance with causal variant resolution based on multi-population meta-analysis, linking a generalized computational method to an experimental read-out. Finally, we present a simple set of software utilities, MPRAutils, for the design and analysis of MPRAs based on association studies. These include tools to design oligonucleotide libraries from summary statistics alone, recover barcode-oligonucleotide maps for random barcoding protocols, and extract barcode-level and/or sequence-level counts for downstream processing (including by existing tools which perform inference on MPRA count data and require count matrices as input).

Copyright © 2018 American Society of Human Genetics. All rights reserved

320

Inferring enhancer and noncoding RNA dysregulation underlying 2,419 UK Biobank phenotypes. *A. Amlie-Wolf[1,2], L. Qu[2], E.E. Mlynarski[2], P.P. Kuksa[2], Y.Y. Leung[2], C.D. Brown[2,3], G.D. Schellenberg[2,3], L.S. Wang[2,3].* 1) Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 2) Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 3) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

The majority of variants identified by genomewide association studies (GWAS) affect regulatory elements outside coding genes such as transcriptional enhancers and noncoding RNAs. We developed INFERNO (http://inferno.lisanwanglab.org), which integrates GWAS data with hundreds of functional genomics datasets to identify causal noncoding variants underlying association signals and their affected regulatory elements, tissue contexts, and target genes. INFERNO uses the COLOC method to identify colocalized GWAS and target gene eQTL signals overlapping enhancers in matching tissue classes and characterizes co-regulatory networks of targeted long noncoding RNAs (lncRNAs). Empirical enrichments of tissue-specific enhancer overlaps are quantified. We applied INFERNO to summary statistics for 191 case/control and 2,228 quantitative traits across anthropological and healthrelated phenotypes from the UK Biobank. We identified 1,389,198 significant variants ($p \leq 5 \times 10^{-8}$) in 2,298 phenotypes, 42.34% of which were found in multiple phenotypes. We pruned significant variants into a median of 93 independent signals by European linkage disequilibrium (LD) and expanded these into LD blocks with a median of 509 candidate causal variants. While only 1.04% of all candidate variants were in coding exons, 2.77% overlapped FANTOM5 enhancers, 50.60% overlapped Roadmap enhancers, and 46.61% overlapped transcription factor binding sites across phenotypes. On average, variants overlapping FANTOM5 or Roadmap enhancers were associated with 5.6 and 1.9 phenotypes, respectively. A subset of variants affected non-coding RNAs: 0.07% overlapped 12 classes of small RNA and 1.6% overlapped microRNA binding sites for 2,059 miRNAs, including 2,056 affected in multiple phenotypes. INFERNO identified strongly colocalized signals for 3,400 genes spanning all 44 GTEx tissues, 55% of which were supported by enhancer overlaps in the matching tissue categories. These included 522 lncRNAs coregulating an average of 247 genes. We also identified 2,078 significant enhancer overlap enrichments in 31 tissue classes including lung for asthma, eye for diabetes related eye disease, heart for atrial fibrillation and hypertension, adipose for waist circumference, and brain for age at completion of education and anxiety. These analyses support the utility of INFERNO for inferring the molecular mechanisms underlying noncoding GWAS signals in a huge range of both case/control and quantitative phenotypes.

321

The GTEx Consortium atlas of genetic regulatory effects across human tissues. *T. Lappalainen[1,2], GTEx Consortium.* 1) New York Genome Center; 2) Department of Systems Biology, Columbia University.

Understanding the molecular effects of genetic variants remains a challenge for the interpretation of genetic associations with complex disease. The Genotype Tissue Expression Consortium (GTEx) has built the most comprehensive catalog to date of genetic effects on transcriptome variation across tissues. Here, we describe the findings of the final consortium analysis of the v8 data release, with RNA-sequencing data from 17,382 samples across 49 tissues from 838 individuals with genome sequencing data. We discovered 26,499 genes with *cis*-eQTLs (5% FDR), 237 genes with trans-eQTLs (10% FDR), thousands of *cis* splicing QTLs, as well as data of haplotype-based allelic expression over 153 million genes and samples. As many as 96% of coding genes have at least one cis-eQTL, with widespread allelic heterogeneity and discovery of multiple independent eQTLs per gene increasing linearly with increasing sample size. Statistical fine-mapping with three methods provided an unprecedented catalogue of thousands of likely causal variants, elucidating genetic regulatory mechanisms. We demonstrate that co-localization of regulatory variants with tissue-specific regulatory elements is an important determinant of their tissue-specific effect size, which we further linked to tissue-specific expression of transcription factors that regulate eQTL activity. Computationally characterized cell type composition of each GTEx tissue sample was shown to be a key driver of inter-individual transcriptome variation. We mapped cell type interacting eQTLs for neutrophils in whole blood (1258 at 5% FDR) and neurons in brain (281 in meta-analysis with local false signal rate < 0.01) to demonstrate their enrichment among tissue-specific eQTLs and their value in providing additional resolution to cellular mechanisms of disease-associated loci. In order to characterize how regulatory genetic variation contributes to disease associations, we standardized GWAS summary statistics for 110 traits and applied multiple QTL colocalization methods and the PrediXcan method to identify regulatory changes across tissues underlying complex disease risk. Altogether, the GTEx consortium provides an unprecedented resource for biological discovery into molecular effects of genetic variants and their contribution to complex disease risk.

Copyright © 2018 American Society of Human Genetics. All rights reserved

322

Leveraging gene expression to understand the consequences of polygenic risk scores for disease in healthy individuals. *A. Claringbould[1], U. Võsa[1,2], H.-J. Westra[1], T. Esko[2], L. Franke[1], eQTLGen Consortium, BIOS Consortium.* 1) Department of Genetics, University Medical Centre Groningen, Groningen, Groningen, Netherlands; 2) Estonian Genome Center, University of Tartu, Tartu, Estonia.

Understanding the functional consequences of the >10,000 variants recently associated through genome-wide associations studies has proven challenging. Most efforts have focused on a single variant approach, like correlating single disease variants to various molecular data layers. However, single variant approaches do not account for the additional risk inferred by multiple independent disease risk alleles. Polygenic risk scores (PRS) can be instead used to represent the overall risk of an individual. While PRS is yet unable to efficiently predict disease status, we hypothesize that healthy individuals with high PRS for a disease may share molecular mechanisms with affected patients. To identify such mechanisms, we investigated the impact of the PRS for 1,263 complex diseases and traits on gene expression (ePRS), in a large meta-analysis of ~32,000 individuals with available genetics and blood transcriptomics data. In total, we identified 15,328 significant ePRS associations (false discovery rate of < 0.05), involving 1,973 genes. Aside from associations reflecting the known biology of a trait, our analysis also identified relevant candidate genes for traits that are less directly linked to blood. For example, the PRS for 'ever versus never smoking' was associated with increased expression of lymphocyte regulating gene *GPR15*, previously found to be overexpressed in smokers. We also identified ePRS associations for neuronal traits, such as educational attainment, which is positively associated with *STX1B*, a gene involved in synaptic transmission. Additionally, ePRS analyses highlighted genes and pathways known to be associated with monogenic diseases. The PRS for serine, glycine, its derivative n-acetylglycine and creatine were all negatively associated with the expression of *PHGDH*, *PSAT1*, and *AARS* genes. Mutations in those genes are known induce serine biosynthesis defects, causing low serine and glycine concentrations in blood. Finally, PRS for several auto-immune diseases such as rheumatoid arthritis, celiac disease, and primary biliary cirrhosis, were associated with genes primarily expressed in specific blood cell types. We hypothesize that cell type abundance plays a role in these diseases, and validated the patterns using single-cell RNA-sequencing data. ePRS analysis allows for new interpretations of both blood and non-blood trait associations and as such provides a novel avenue to investigate the molecular consequences with polygenic disease risk.

323

The transethnic portability of predictive models for gene expression. *K.L. Keys[1], A.C.Y. Mak[1], W. Eckalbar[1], M.J. White[1], C. Eng[1], D. Hu[1], S. Huntsman[1], J. Liberto[1], S. Oh[1], S. Salazar[1], J. Rodríguez-Santana[2], R. Hernandez[3], J. Ye[3], N. Zaitlen[1], E. Burchard[1,2], C.R. Gignoux[4].* 1) Department of Medicine, University of California, San Francisco, CA; 2) Centro de Neumología Pediátrica, San Juan, Puerto Rico; 3) Bioengineering and Therapeutic Biosciences, University of California, San Francisco, CA; 4) Colorado Center for Personalized Medicine, University of Colorado, Denver, CO.

Genetic variants can contribute to complex traits directly or through modulation of gene expression. While genome sequencing costs continually drop, transcriptome sequencing (RNA-Seq) remains costly. Recent interest in genetic variants affecting protein abundance has led to the development of public transcriptome repositories such as the Genotype-Tissue Expression (GTEx) repository. These invaluable repositories enable researchers to impute gene expression levels from genotype data using linear predictive models from PrediXcan and perform powerful gene-based transcriptome-wide association studies (TWAS). However, we note that 85% of the subjects in GTEx are of European (EUR) descent, in contrast to worldwide genetic diversity. The abundance of genomes from diverse populations with unpaired RNA-Seq data portends scenarios where transcriptome imputation into non-EUR populations uses EUR transcriptomes. Prediction quality in non-EUR or admixed populations remains underexplored. Furthermore, related initiatives on polygenic risk scores have revealed challenges with transethnic portability. Therefore, it is of great interest to quantify the extent to which predictive models trained on expression data from one population perform in another. To do so, we investigate the transethnic portability of transcriptome imputation by constructing predictive models using EUR and African (AFR) genome sequences and lymphoblastoid cell RNA-Seq data from the GEUVADIS project.We derive models for imputed gene expression using TWAS pipelines for eQTL-based predictive modeling. We find across genes that cross-population prediction accuracy remains low: models trained in EUR attain mean R2 of 2.1% in AFR, while models trained in AFR yield 0.8% mean R2 in EUR. Strikingly, 45% of genes have negative correlations between predicted and real gene expression, which will bias TWAS predictions in non-EUR populations. We discuss the consequences of using PrediXcan GTEx models on non-European study subjects, particularly in the context of locus-specific ancestry in African Americans.We stress that our results do not demonstrate a shortcoming of the TWAS approaches. These methods have demonstrated power and utility, albeit mostly in populations of EUR descent. Instead, our work highlights the need for continued transcriptome generation in populations from across the world to ensure research and medical benefits for all individuals.

Copyright © 2018 American Society of Human Genetics. All rights reserved

324

Identifying novel epigenetic allele-specific expression effects in GTEx. *S.N. Kravitz, W.C. Huang, E. Ferris, A.R. Quinlan, C. Gregg.* Human Genetics, University of Utah, Salt Lake City, UT., Select a Country.

Cells are generally thought to express both alleles of a gene equally. However, genetic and epigenetic mechanisms can drive the preferential expression of just one allele in a cell. A few well studied examples of epigenetic drivers of monoallelic expression include genomic imprinting and random X-chromosome inactivation in females. Monoallelic expression can have important consequences: if a pathogenic allele is the only one expressed due to epigenetic effects, the impact of a heterozygous variant may be severe. Recently, we uncovered thousands of genes in the mouse and macaque brain where populations of monoallelic cells preferentially express one allele in vivo in a developmentally regulated manner. These Differential Allele Expression Effects (DAEEs) act like X-linked genes; one allele is epigenetically silenced to create mosaic subpopulations of cells that preferentially express one gene copy. When the preferentially expressed allele is pathogenic, DAEEs could introduce dysfunctional cell populations within a tissue, contributing to phenotypic variation and disease etiology. In this study, we developed a novel statistical framework, "madRD," to identify human genes that exhibit similar epigenetic allelic effects. This approach leverages variation in allele-specific expression (ASE) across individuals and genetic eQTL information to distinguish genes with random epigenetic ASE, recapitulating a random X-inactivation model for autosomal genes. By examining RNA-seq data in over 50 tissues from healthy individuals in the Genotype-Tissue Expression (GTEx) project, we identify genes that exhibit DAEEs, tissue differences in the prevalence of these effects, and explore the impact of monallelic expression on genes linked to disease. By applying the madRD approach to the entire GTEx dataset, we have generated a preliminary atlas of epigenetic allelic effects for all genes across healthy human tissues. We show that we can separate genes with DAEEs due to genetic versus epigenetic regulatory effects in a healthy population. We also show epigenetic DAEEs cause monoallelic expression in post-mortem human brain tissue samples, revealing the potential to shape the impact of heterozygous mutations on disease risk in humans..

325

TIVAN: Tissue-specific cis-eQTL single nucleotide variant annotation and prediction. *L. Chen[1], B. Yao[2], A. Mitra[3].* 1) Health Outcomes Research and Policy, Auburn University - Harrison School of Pharmacy, Auburn, AL; 2) Department of Human Genetics, Emory University School of Medicine, Atlanta, GA; 3) Department of Drug Discovery and Development, Auburn University - Harrison School of Pharmacy, Auburn, AL.

Annotating and prioritizing genetic regulatory variants, most of which locate in non-coding regions of genome, still remains a challenge in genetic research. Among all non-coding regulatory variants, cis-eQTL single nucleotide variants (SNVs) are of particular interest for their crucial role in regulating the gene expression and aiding the interpretation of genome-wide association studies (GWAS) associations. Since different biological mechanisms are believed to contribute to the etiologies of different phenotypes, it is desirable to characterize the impact of cis-eQTL in a context-dependent manner (e.g. tissue, cell type). Though multiple computational methods developed to measure the potential of variants for being pathogenicity or deleteriousness are well established, methods for annotating and predicting tissue-specific eQTL SNVs are under-developed. To cater for this need, we develop a machine learning method TIVAN (TIssue-specific Variant Annotation and prediction) to predict cis-eQTL SNVs by considering tissue-specificity. For each tissue, TIVAN is trained in the framework of ensembled decision trees by using a collection of comprehensive feature set including genome-wide genomic and epigenomic profiling data. To evaluate the predictive performance, we first conduct comprehensive tests across 44 different tissues belonging to 27 different tissue classes. We observe that TIVAN could accurately discriminate cis-eQTL SNVs from non-eQTL SNVs and perform favorably to other competing methods by obtaining higher five-fold cross-validation AUC (CV-AUC) values. To further validate the advantage of TIVAN, we adopt a Leave-Chr-Out CV (LCO-CV) strategy, which test the prediction of cis-eQTL SNVs on one autosome by training a prediction model based on the cis-eQTL SNVs on the remainder of autosomes. Again, TIVAN outperforms other methods by obtaining higher LCO-CV-AUC values. Finally, TIVAN achieves an overall better predictive performance than other competing methods on an independent testing dataset, which include 7 tissues in 11 studies. These results suggest that the great potential of TIVAN in identifying novel cis-eQTL SNVs in a tissue-specific manner. We believe the findings of TIVAN will have profound implications for mechanistic interpretability how GWAS variants make their impacts on diseases/phenotypes through gene regulation by eQTL-GWAS comparisons. The software TIVAN and pre-computed scores of 44 tissue/cell types for hg19 are public available. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

326

Predicting the impact of cis-regulatory variation on alternative polyadenylation. *N. Bogard[1], J. Linder[2], A.B. Rosenberg[1], G. Seelig[1].* 1) Department of Electrical Engineering, University of Washington, Seattle, WA; 2) Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA.

Alternative Polyadenylation (APA) is a post-transcriptional regulatory process by which multiple RNA isoforms with distinct 3'-ends, and consequently differential stability, subcellular localization and translational efficiency, can be derived from a single gene. Genetic variants that interfere with proper regulation of APA have been implicated in disease. For example, a mutation of a PAS in the *FOXP3* gene can cause IPEX, a rare disorder of the immune system. Here, we use deep learning to interpret the sequence-regulatory code of polyadenylation sites and predict APA from DNA sequence alone. We trained our model (APARENT, APA REgression NeT) on cleavage and isoform expression data from over three million APA reporters, built by inserting random sequence into twelve distinct 3'UTR contexts. Predictions are highly accurate across both synthetic and genomic contexts; when tasked with inferring APA in human 3'UTRs, APARENT outperforms models trained exclusively on endogenous data. Visualizing features learned across all network layers reveals that APARENT recognizes sequence motifs known to recruit APA regulators, discovers previously unknown sequence determinants of cleavage site selection, and integrates these features into a comprehensive, interpretable cis-regulatory code. We use our model to screen disease-associated 3'UTRs for variants predicted to strongly impact APA isoform distribution and find that the model detects known pathogenic variants in a wide range of contexts from clinical variation datasets. Interestingly, the model predicts a large shift in APA regulation from a number of cryptic APA variants in the USE and DSE of polyadenylation signals which previously have been linked to disease through mutation of its core sequence element (CSE), a conserved hexamer motif. For example, SNP rs33978907 within the CSE of a polyadenylation signal in the *HBB* gene has been linked to beta-Thalassaemia, and our model recognizes several additional cryptic SNVs in its DSE which significantly alter the isoform distribution, probably by disruption or creation of downstream regulatory motifs. We suggest there are many potential mutations beyond that of CSE variation within these UTRs which could disrupt 3'-end processing and be equally deleterious. APARENT could prove a useful tool for narrowing down such unclassified genetic variants into candidates with high likelihood of affecting APA.

327

Genetically regulated gene expression in brain and peripheral tissues types associated with PTSD. *N.P. Daskalakis[1], M.S. Breen[2], S. van Rooij[3], J. Hartmann[1], K. Girdhar[2], CommonMind Consortium[4], PGC PTSD[5], T. Jovanovic[3], C. Nievergelt[6], H.K. Im[7], J.D. Buxbaum[2], P. Sklar[2], K.J. Ressler[1], E.A. Stahl[2], L.M. Huckins[2].* 1) Harvard Medical School/ McLean , Belmont, MA; 2) Icahn School of Medicine at Mount Sinai, New York, NY; 3) Emory University; 4) CommonMind Consortium; 5) PGC PTSD; 6) University of California San Deigo; 7) University of Chicago.

PTSD is a debilitating mental disorder occurring in some trauma-exposed individuals. The latest PGC-PTSD GWAS demonstrates genetic heritability and correlation with other psychiatric disorders. PTSD development involves multi-systemic dysregulation in multiple brain and peripheral tissues. Transcriptomic Imputation (TI) models leverage large expression quantitative trait loci reference panels to translate genome-wide genotype data into more biologically meaningful measures. Here, we apply CommonMind Consortium and Genotype-Tissue Expression derived TI-models to impute genetically regulated gene expression (GReX) in PGC-PTSD cases and controls (9,185/24,409). We performed transethnic meta-analysis and stratified analyses according to trauma type (civilian vs. combat trauma), and ancestry. We identified 10 significant associations, corresponding to 8 unique genes (CNOT1, GCM1, MARCH11, NEDD9, RPS3, SENP1, SLC9B2, SNRNP35, SYNGR2) at the level of 7 unique tissues; 6 brain regions , and 1 peripheral (heart). These results include three genome-wide significant associations with PTSD. Importantly, our results suggest substantial genetic heterogeneity between civilian and military cohorts. Our predicted PTSD effects in peripheral blood correlated with observed PTSD effects in blood gene expression (R=0.34, P=1e-12), and this correlation was much stronger compared to predicted effects in any other tissue. Predicted downregulation of *SNRNP35* in DLPFC showed the strongest association with PTSD (β=-1.39, p=1.23e-09). DLPFC is a brain region of interest in PTSD as it is involved in many stress-related neurobiological systems, and functional alterations in this region have been described in PTSD. SNRNP35 is part of the minor spliceosome, which catalyzes the removal/splicing of an atypical class of introns – U12-type, from mRNAs. We then showed that knockdown with specific *SNRNP35*-specific shRNAs has functional impacts on U12 splicing. Interestingly, the *SNRNP35* gene contains a high number of glucocorticoid-receptor (GR) binding sites indicating possible regulation by stress-hormones. We have shown that dexamethasone downregulates PFC *Snrnp35* in mice. Moreover, deployment-stress downregulated blood *SNRNP35* (N=164) in marines with post-deployment PTSD, together with increased GR-signaling (adj. p<5E-02, bias-corrected z-score<2.0). Gene discovery based on TI , followed by translational findings in mice and humans support a role for SNRNP35 in PTSD.

Copyright © 2018 American Society of Human Genetics. All rights reserved

328

Dissecting the genetic, transcriptomic, and phenotypic complexity of PTSD across 9400 individuals and 30 million phenotypic observations. *Y. Li[1], V. Michopoulos[2], A. Lori[2], A. Lott[2], B. Bradley[2], T. Jovanovic[2], K. Ressler[3], M. Kellis[1], N. Daskalakis[3].* 1) Massachusetts Institute of Technology, Cambridge, MA; 2) Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, Georgia; Yerkes National Primate Research Center, Atlanta, Georgia; 3) Division of Depression & Anxiety Disorders, McLean Hospital, Department of Psychiatry, Harvard Medical School.

Post traumatic stress disorder (PTSD) is a severe mental disorder occurring in some trauma-exposed individuals, but its molecular basis remains uncharacterized. Here, we study the relationship between genetically-regulated transcriptional variation, and phenotypic variation in PTSD, in a cohort of 9400 inner-city, low-socioeconomic-status, primarily-African-American patients of the Grady Memorial Hospital. Each individual was ascertained phenotypically using interview-based assessments and self-report questionnaires including childhood trauma history and PTSD diagnosis, and 12 lab tests including endocrine, metabolic and immune markers (e.g., cortisol, cholesterol, IL-6). In addition, we imputed expression for ~5000 genes across 48 tissues using GTEx eQTLs (v7), and selecting the top 200 most variable and tissue-specific genes. To study the comorbidity patterns of tissue-specific gene expression and phenotypic information across individuals, we developed MixEHR, a Bayesian method for phenotype correlation and imputation analysis, which explicitly deals with the non-missing-at-random nature of phenotypic information and lab tests. We applied MixEHR systematically, and found several noteworthy disease topics of high probability of PTSD diagnosis that were concordant with PTSD symptom severity, depression symptom severity and childhood trauma levels, thereby recapitulating the PTSD clinical presentation. We found several noteworthy enrichments for well-studied genes and tissues in specific disease topics across multiple tissues. For example, a human-lineage-specific gene, NBPF3, which is implicated in neurodevelopmental disorders, showed high probability for PTSD across multiple brain tissues. Similarly, the top tissues with the highest marginal PTSD associations are basal ganglia, amygdala, substantia nigra, and pituitary, which have been associated with emotion regulation and PTSD. We also constructed a ranked list of PTSD genes by calculating the marginal probabilities of each gene with respect to PTSD across multiple disease topics. The top 3 genes make biological sense, and also yield new biological insights on PTSD: (1) MXRA8, the top gene, is important in the maturation and maintenance of blood-brain barrier; (2) ATP6AP1L, ranked #2, is implicated in the glucocorticoid receptor pathway and neural responses to stress; and (3) AP3S2, ranked #3, is essential for vesicles delivery into neurites and nerve terminals.

329

Comprehensive functional genomic resource and integrative model for the adult brain. *D. Wang[1], S. Liu[2,3], J. Warrell[2,3], H. Won[4,5,6], X. Shi[2,3], F. Navarro[2,3], D. Clarke[2,3], M. Gu[3], P. Emani[2,3], M. Xu[2,3], Y. Yang[2,3], J. Park[2,3], S. Rhie[10], K. Manakongtreecheep[2,3], H. Zhou[2,3], A. Nathan[2,3], J. Zhang[2,3], M. Peters[11], E. Mattei[12], D. Fitzgerald[13], T. Brunetti[13], J. Moore[12], N. Sestan[14], A. Jaffe[15], K. White[13], Z. Weng[12], D. Geschwind[4,5,6,7], J. Knowles[8], M. Gerstein[2,3,9], PsychENCODE Consortium.* 1) Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY; 2) Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT; 3) Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT; 4) Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA; 5) Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA; 6) Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA; 7) Department of Psychiatry, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA; 8) SUNY Downstate Medical Center College of Medicine, Brooklyn, NY; 9) Department of Computer Science, Yale University, New Haven, CT; 10) Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA; 11) Sage Bionetworks, Seattle, WA; 12) Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA; 13) Institute for Genomics and Systems Biology, Department of Human Genetics, University of Chicago, IL; 14) Department of Neuroscience and Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT; 15) Lieber Institute for Brain Development, Johns Hopkins Medical Campus; Departments of Mental Health and Biostatistics, Johns Hopkins Bloomberg School of Public Health Baltimore, MD.

Disorders of the brain affect nearly one fifth of the world's population but our molecular-level understanding of how genomic variants relate to brain disorders is still limited. To address this challenge, the PsychENCODE consortium has generated ~5,500 genotype, transcriptome, chromatin, and single-cell datasets from 1,866 individuals. By uniformly processing and analyzing PsychENCODE data together with publically available data, we have developed a comprehensive functional genomic resource and integrative model for the adult brain (available via Adult.PsychENCODE.org). In particular, we deconvolved the gene expression of bulk tissue using single-cell data, finding that differences in the proportions of cell types explain >85% of the cross-population variation observed. Moreover, using chromatin and Hi-C data from reference prefrontal-cortex samples, we found ~79,000 brain-active enhancers and linked them to genes and transcription factors in an extended regulatory network. We identified ~2.5M eQTLs (comprising ~238K linkage-disequilibrium-independent SNPs) and many additional QTLs associated with chromatin, splicing and cell-type-proportion changes. We, also, leveraged our QTLs, Hi-C data and regulatory network to connect more genes to GWAS variants for psychiatric disorders than possible before (e.g., 304 for schizophrenia). Finally, we developed a deep-learning model embedding the regulatory network in a framework connecting genotype to observed traits. Our model achieves a ~6X improvement in disease prediction over an additive model, highlights key genes for disorders, and allows imputation of missing transcriptome information from genotype data alone.

Copyright © 2018 American Society of Human Genetics. All rights reserved

330

Single-cell transcriptomic catalog of mouse cortical development. *J.M. Simon, L. Loo, L. Xing, E.S. McCoy, J.K. Niehaus, J. Guo, E.S. Anton, M.J. Zylka.* UNC Chapel Hill, Chapel Hill, NC.

The development of the mammalian cerebral cortex depends on careful orchestration of proliferation, maturation, and migration events, ultimately giving rise to a wide variety of neuronal and non-neuronal cell types. To better understand cellular and molecular processes that unfold during late corticogenesis, we utilized Drop-seq to transcriptionally profile ~20,000 single cells from the developing cerebral cortex at a progenitor driven phase and at birth—after neurons from all six cortical layers were born. We developed a new computational procedure that iteratively identifies and refines cell-type classifications to maximize robustness of each intrinsic cell type. Using this approach, we identified 22 distinct cell types in both the embryonic and neonatal cortex, including numerous classes of neurons, progenitors, and glia, as well as their proliferative, migratory, and activation states, and their relatedness within and across age. Using a combination of pseudotiming and immunofluorescence, we identified and validated an apparent differentiation trajectory connecting Layer I Cajal-Retzius cells to their progenitors in the cortical hem. Further, using the cell-type-specific expression patterns of genes mutated in neurological and psychiatric diseases, we identified putative disease subtypes that were associated with clinical phenotypes. Our single-cell transcriptomic catalog therefore serves as a valuable resource to reconstruct complex neurodevelopmental processes, permitting a deeper understanding of cortical development and providing a window into the cellular origins of brain diseases.

331

Machine learning of ultra-deep whole-genome sequencing of human brain reveals somatic retrotransposition in both neurons and glia. *X. Zhu[1,2], B. Zhou[1,2], R. Pattni[1,2], K. Gleeson[3], C. Tan[3], S. Steven[4], A. Fiston-Lavier[5], J. Lennington[6], L. Tomasini[6], T. Bae[7], L. Faching[6], MP. Snyder[2], D. Petrov[8], A. Abyzov[7], FM. Vaccarino[6], BA. Barres[4], OH. Vogel[9], C. Tamminga[3], DF. Levinson[1], AE. Urban[1,2].* 1) Psychiatry and Behavioral Sciences, Stanford University, Palo Alto, CA; 2) Genetics, Stanford University, Palo Alto, CA; 3) University of Texas Southwestern Medical Center, Dallas, TX; 4) Neurobiology, Stanford University, Palo Alto, CA; 5) Computer Science, University of Montpellier 2, Montpellier, Hérault, France; 6) Child Study Center, Yale University, New Haven, CT; 7) Department of Health Sciences Research, Mayo Clinic, Rochester, MN; 8) Biology, Stanford University, Palo Alto, CA; 9) Pathology, Stanford University, Palo Alto, CA.

Somatic retrotransposition contributes to mosaic neuronal diversity during human neurogenesis and is speculated to contribute to neuropsychiatric conditions. Various methods have been developed to study brain somatic retrotranspositions, including retrotransposon capture sequencing (RC-seq), single neuron L1 insertion profiling (L1-IP) and single neuron whole genome sequencing. However, capture sequencing and whole genome amplification of single cells both create PCR chimeras. In addition, the single cell approach cannot readily be scaled up to studies of large cohorts, and so far has only been utilized to study neurons. Here we describe a novel approach for detecting somatic L1 and Alu retrotransposition in a direct and unbiased manner, by analyzing ultra-deep whole-genome sequencing, separately of the neuronal and glial cell fractions of the human brain. We developed a random forest classification algorithm, RetroSom, to extract features specific for novel retrotranspositions and to drastically improve the filtering of false positives. We tested RetroSom in three independent whole genome sequencing datasets: 1) multiple clones expanded from single neural cells from two subjects, 2) libraries generated with or without PCR, and 3) *in vitro* mixing of genomic DNA with frequency ranging from 0.04% to 25%. We verified that RetroSom is highly precise, performs well in libraries created with PCR, and detects retrotranspositions with frequencies of >1% for L1 and >0.2% for Alu. We applied RetroSom to ultra-deep sequencing data (>200x coverage) from neurons, glia and non-brain control specimens (heart or fibroblast) of six individuals, including a fetus, an elderly person, and two “blind” schizophrenia-control pairs. RetroSom identified two somatic brain-specific L1 insertions, found in both neurons and glial cells, as well as two heart-specific Alu insertions. We orthogonally validated all four insertions by digital droplet PCR to quantify their frequencies, and by nested PCR to characterize the insertion breakpoints and target site duplications. This study shows for the first time that genomic mosaicism by retrotransposons can manifest also in glia, i.e. in the non-neuronal cells that comprise approximately half of a human brain. The patterns of the detected somatic retrotranspositions in adult brain suggest that the events occurred in neural stem cells and were inherited by both the neurons and glia during differentiation.

Copyright © 2018 American Society of Human Genetics. All rights reserved

332

Using sequence graphs to fully characterize the variability of pathogenic repeat loci. *E. Dolzhenko[1], F. Schlesinger[1], V. Deshpande[1], J.J.F.A. van Vugt[2], G. Narzisi[3], S. Chen[1], C. Reeves[3], L. Winterkorn[3], N.S. Wexler[4,5], J.H. Veldink[2], R.J. Taft[1], D.R. Bentley[6], M.A. Eberle[1], The US-Venezuela Collaborative Research Group.* 1) Illumina Inc., 5200 Illumina Way, San Diego, CA, USA; 2) Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands; 3) New York Genome Center, 101 Avenue of the Americas, New York, NY, USA; 4) Columbia University, 1051 Riverside Drive, New York, NY USA; 5) The Hereditary Disease Foundation, 601 W 168th St, New York, NY, USA; 6) Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, UK.

Expansions of short tandem repeats (STRs) have been implicated in over 30 monogenic disorders. Typically, a repeat is defined as pathogenic if it exceeds a certain size threshold. It is critical to accurately determine the length of each repeat allele (genotype), as even a small error could lead to an incorrect diagnosis. However, repeat genotyping remains a challenging problem for both NGS- and PCR-based assays. In particular, sequence complexity and the presence of inserted or deleted sequences in the region flanking the repeat can confound the estimation of the repeat size. To investigate how the accuracy of repeat genotyping could be improved, we analyzed WGS data from 120 individuals with repeat expansion disorders and 150 controls. During this process, we identified and annotated many STRs that are flanked by other STRs. In particular, we observed that several pathogenic repeat loci harbor an STR a few base pairs away from the pathogenic repeat: the CAG repeat that causes Huntington disease (HD) is flanked by a CCG repeat; the GAA repeat that causes Friedreich ataxia (FRDA) is flanked by a homopolymer repeat; the CAG repeat that causes Spinocerebellar ataxia type 8 (SCA8) is flanked by an ACT repeat. These flanking repeats have high variability in our data with sizes ranging from 8 to 14 repeat units (HD), 14 to 24 repeat units (FRDA), and 5 to 15 repeat units (SCA8).Motivated by these observations, we developed a computational method which leverages sequence graphs to genotype tandem repeats. These graphs incorporate prior knowledge about the full complexity of each repeat region by directly encoding the repeat and the sequence variation surrounding the repeat. Aligning reads directly to sequence graphs makes it possible to unbiasedly genotype each constituent variant and to dramatically improve the genotyping accuracy of pathogenic repeats. For example, the number of Mendelian conflicts in 50 trios for pathogenic repeats in *NOP56* and *ATXN3* genes was reduced from over 80% to just one when these repeats were represented by graphs. Additionally, genotyping of variants flanking the pathogenic event may pave the way towards the discovery of new modifiers of disease onset and progression. We will present a detailed analysis of 24 known pathogenic repeat regions that demonstrates improvements brought about by our method. We have made our method publicly available by incorporating it into Expansion Hunter, an open-source repeat expansion detection software.

333

Genome-wide analyses identify *CTNNA2* and *SULT2A1* as candidate genes for cleft palate only in the African population. *A. Butali[1], P.A. Mossey[2], W.L. Adeyemo[3], M.A. Eshete[4], L.J.J. Gowans[5], M.L. Marazita[6], J.C. Murray[1], A.A. Adeyemo[7].* 1) University of Iowa, Iowa City, Nigeria; 2) University of Dundee; 3) University of Lagos; 4) Addis Ababa University; 5) Kwame Nkurumah University of Science and Technology; 6) University of Pittsburgh; 7) National Human Genome Research Institute.

Orofacial (OFCs) are the most visible common birth defects affecting one in 700 live births worldwide. We successfully conducted the first GWAS for OFC in Africa. A total of 3,353 participants were genotyped on the Illumina Multi Ethnic Genotyping Array (MEGA). IMPUTE2 was used for imputation into the 1000 Genomes Phase 3 reference panel. The final dataset that passed quality control consisted of 3,178 (1,133 male; 2,045 female) participants enrolled from Ethiopia (30%), Ghana (43%), and Nigeria (27%). The dataset included 814 cases of CLP, 205 cases of isolated CP, and 2,159 related and unrelated controls. The imputation yield was ~45 million SNPs and ~17 million that passed our quality control filter were included in the final analyses. We conducted two separate GWAS analyses (for CL/P and CP) due to the known differences in the developmental and genetic basis of isolated CL/P versus CPO. Single-variant association tests were done for imputed dosage data filtered for imputed allelic dosage frequency < 0.01 and info < 0.3 using logistic mixed models as implemented in the GMAAT package. After replication and meta-analyses, we identified novel loci for isolated cleft palate (CPO) at or near genome-wide significance on chromosomes 2 (near *CTNNA2* leading SNP rs140938806, $p = 2.76 \times 10^{-9}$) and 19 (*SULT2A1* leading SNP rs6252985, $p= 7\ 7.83 \times 10^{-8}$). Each of these loci is supported by functional data from model organisms. The previously-reported 8q24 locus (leading SNP, rs72728755, $p = 1.52 \times 10^{-6}$) for CL/P was the most significant in this study and we replicated several previously reported loci including *IRF6, PAX7, VAX1, PTCH1* and *COL8A1*. This first GWAS of OFC in Sub Saharan Africans identified novel loci for CPO and confirmed several findings previously reported from other ancestral populations.

Copyright © 2018 American Society of Human Genetics. All rights reserved

334

Quantifying genetic risk of prosthetic joint infection through biobank data and electronic health records. *D.J. McGuire, D. Liu.* Penn State College of Medicine, Hershey, PA.

Prosthetic joint infection (PJI) is a devastating outcome for those afflicted. Given many of the patients that receive joint or knee replacement are elderly, the complications from PJI can be fatal. Previously identified risk factors for PJI include obesity, high body mass index, diabetes, and rheumatoid arthritis. PJI displayed clear patterns of inter-individual differences with different individuals having different predisposition to PJI. While previous studies have investigated genetic susceptibility to PJI, the sample sizes have been in the hundreds, and to our knowledge systematic genome-wide association analysis has yet to be conducted. In this study, we utilize biobank data and patient electronic health records to identify patients with adverse outcomes after prosthetic joint replacement, and perform genome-wide association analysis to identify genetic loci associated with increased risk for PJI. Using ICD9 codes and CPT operation procedure codes, we defined individuals with prosthetic hip/knee revision CPT codes 27090, 27091, 27486, 27487, 27488 combined with ICD9 infection code 996.66 as cases, and patients with knee/hip replacement codes 27130 and 27447 who have no recorded history of joint revision or infection as controls. The phenotypes from the UK Biobank were defined similarly using the episodes data. A total of 440 probable cases and 11815 controls were phenotyped from the UK biobank and BioVU datasets. We used BOLT-LMM and BOLT-REML variance component analysis to conduct genome-wide association analysis for UK Biobank and BioVU data separately and then conducted meta-analysis. Preliminary results identified loci in or near genes associated with immune response, including HLA gene. In the BioVU dataset, exonic variants rs41553512 in HLA-DRB5 and rs17879702 in HLA-DRB1 region (p=1.29e-8 and p=2.12e-8), as well as intergenic variant rs10456411 in the HLA-DQA1:HLA-DQB1 region (p=4.43e-8) and rs143303636 HLA-DRB6:HLA-DRB1 region (p=9.47e-9) reached GW significance. HLA-DQA1,HLA-DQB1, and HLA-DRB1 are all class II genes in the MHC , which function to make proteins that are present almost exclusively on the surface of certain immune system cells. In summary, our study identified putative novel loci that affect the risk for PJI, a type of treatment failure due to infection. More importantly, our study showcased the novel use of EMRs in Biobanks to quantify the genetic risk for treatment failures, which is a key step for precision medicine.

335

Optimization of high diversity imputation in cohort of more than 400,000 US veterans in the VA's Million Veteran Program (MVP). *Y. Shi[1,3], S. Ji[1,3], S. Gallagher[1,3], B. Gorman[1,3], J. Prescott[1,3], J. Huang[1,3], C. Pan[1,5,6], T. Assimes[1,5,6], S. Muralidhar[1,8], C. O'Donnell[1,3], E. Hauser[1,7,9], H. Zhao[1,4,10], P. Tsao[1,5,6], S. Pyarajan[1,2,3].* 1) Veterans Administration (VA), Boston, MA; 2) Harvard Medical School, Boston, MA; 3) VA Boston Healthcare System, Boston, MA; 4) Yale University School of Medicine, West Haven, CT; 5) VA Palo Alto Health Care System, Palo Alto, CA; 6) Stanford University School of Medicine, Stanford, CA; 7) Duke University, Durham, NC; 8) Office of Research and Development (ORD), Veterans Health Administration, Washington, DC; 9) VA Health System, Durham, NC; 10) VA Connecticut Healthcare System, VA Cooperative Studies Program, West Haven, CT.

Over 500,000 US veterans have now been enrolled and genotyped in the Million Veterans Program (MVP). While recent advances in phasing and imputation methodologies allow for statistical inference of unobserved genotypes in large mega-cohorts, performance of these techniques in highly admixed and ancestrally diverse cohorts has yet to be evaluated empirically. Admixed populations with ancestry from two continental populations account for a sizeable portion (~22%) of the MVP cohort. The unique diversity of MVP makes it an optimal resource for empirically testing the performance of widely adopted imputation tools in mega-cohorts. Imputed genotypes for individuals with European ancestry were found to be significantly more consistent when compared to those for individuals with predominantly non-European ancestry. Two, sub-releases of the MVP dataset were separately imputed using Eagle2 + Minimac3 pipeline. A total of 1,000 GBR and 1,000 non-GBR individuals were duplicated in both sub-releases of the MVP dataset. ADMIXTURE analyses show similar ancestry composition of the two sub-releases. Concordance in the imputed genotype calls of the duplicated samples was consistently higher in the single ancestry (GBR) than in the multi-ethnic (non-GBR) population suggesting a need for improvement in the imputation of diverse, highly admixed cohorts. To improve the performance of imputation in the diverse MVP cohort, we devised a pipeline that optimizes the number of conditioning haplotypes for different ancestry proportion compositions.

Copyright © 2018 American Society of Human Genetics. All rights reserved

336

Expanding the scope of the GWAS catalog to improve drug target identification and prioritisation. *A. Buniello[1,2], O. Vrousgou[1,2], E. Mountjoy[2,3], J. Hayhurst[1], L. Harris[1], C. Malangone[1], J. MacArthur[1], T. Burdett[1], M. Ghoussaini[2,3], J. Barrett[2,3], I. Durham[1,2], H. Parkinson[1], F. Cunningham[1].* 1) EMBL-EBI , Cambridge, UK, Cambridgeshire, United Kingdom; 2) Open Targets, an academic-industrial collaboration between the Wellcome Sanger Institute, European Bioinformatics Institute (EMBL-EBI), GlaxoSmithKline plc., Biogen Inc., Takeda Pharmaceuticals and Celgene Corp; 3) Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.

Open Targets provides robust integration of evidence from different public data sources to associate genes, proteins and pathways with diseases, with the aim of prioritising targets for drug discovery. A major source of genetic evidence is the NHGRI-EBI GWAS Catalog, a manually curated open resource repository of all published GWAS results. As of end of May 2018 the GWAS Catalog includes 5,168 GWAS and 62,156 unique SNP-trait associations from 3,395 publications. To ensure consistency across publications, the Catalog maintains strict eligibility criteria. Up to September 2017, inclusion in the Catalog was limited to array-based GWAS, including association analysis of >100,000 SNPs with genome-wide coverage and SNP-trait associations with a p-value <1x10-5. Targeted and exome array studies were considered outside the scientific scope, but these data are however valuable to Open Targets' scientific goals of target prioritisation. Open Targets have therefore collaborated with the GWAS Catalog to expand the scope of the Catalog to include specific prioritised association studies carried out using targeted arrays such as MetaboChip, ImmunoChip and Exome array.Broadening the type of studies hosted in the GWAS Catalog increases the number of variants known to be associated with immunologic, metabolic and oncologic phenotypes (key therapeutic areas for Open Targets). A total of 120 new independent association studies and approximately 800 SNP-trait associations have been already extracted from the first 55 studies prioritised for inclusion by Open Targets as part of this collaboration, and many more will be added with curation of additional targeted/exome array publications. Concomitantly, the GWAS Catalog user interface has been implemented to support searching, displaying, filtering and download of targeted and exome array studies. We are also combining our efforts to develop a comprehensive database of all available GWAS summary statistics (with no p-value cut-off), stored in a common format and harmonised across studies to enable easy comparison and downstream analysis. These data will be made publicly available via the GWAS Catalog website and, in the future, through a dedicated service (API). This new collaborative initiative aims to contribute to the "big data" revolution and to guide researchers through their drug discovery journey, ultimately leading to improved target identification and disease prognosis.

337

Joint analysis of phenotypes enables pathway detection. *H. Julienne[1,4], V. Laville[1], C. Lasry[1], A. Ziyatdinov[2], P. Kraft[2], P. Lechat[1], H. Ménager[1], B. J. Vilhjálmsson[3], V. Guillemot[1,4], H. Aschard[1,4].* 1) Institut Pasteur, Paris, France; 2) Harvard School of Public Health, Boston, Massachusetts; 3) QIAGEN; 4) Centre de Biologie et (C3BI), Paris, France.

The large number of publicly-available GWAS results has enabled a range of cross-phenotype analyses: assessing whether significant loci for a phenotype of interest are also associated with other traits (e.g. Locke 2015); computing pairwise genetic correlation between phenotypes (e.g. Pickrell 2016); or detecting new associations with multi-trait tests (e.g. Liu Z. 2018). We argue that, beyond mapping improvement, multivariate analysis also offer a powerful framework to better characterize variant effects and to decipher molecular pathways. We applied a three-step procedure on real data: 1) detect significant multivariate effects for a set of related traits, 2) cluster SNP z-scores to identify distinct multivariate signatures, and 3) determine if SNP clusters match specific biological pathways. We use JASS (Joint Analysis of Summary Statistics), a package we recently developed (http://jass.pasteur.fr/index.html), to analyze 45 traits split into clinically relevant subsets. We present an in-depth analysis of metabolism related traits (TG, HDL, LDL, total cholesterol, fasting-insulin, fasting-glucose, HOMA-beta, Homa-IR, T2D). Our analysis identified 39 new signals missed by univariate GWAS. We focus on a subset of loci better identified with a multivariate test than with univariate GWAS (i.e. showing a joint p-value 10 times lower than the minimum of univariate p-values). Clustering these loci yielded 6 signatures on which we assessed functional enrichment using FUMA (http://fuma.ctglab.nl/). Most clusters impact lipid traits and are enriched for genes associated with coronary heart disease. The clusters have unique signatures corresponding to specific pathways. The cluster that conjointly increase TG, total cholesterol and LDL is enriched for genes associated to PPAR-alpha. Another cluster, raising TG while lowering HDL was specifically enriched for the pathway 'positive regulation of lipid storage' and for genes related to body mass index. Intriguingly, the cluster that jointly raises TG and total cholesterol is associated with genes of the immune system but not coronary heart disease. These results illustrate how the proposed joint analysis of phenotypes can offer new insights about cross-phenotype pathways and propose gene candidates for drug targets.

Copyright © 2018 American Society of Human Genetics. All rights reserved

338

A systematic classification of heritable risk factors influencing common diseases. *A. Cortes[1,2], C. Dendrou[3], L. Fugger[2,4,5], G. McVean[1].* 1) Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7FZ, UK; 2) Oxford Centre for Neuroinflammation, Nuffield Department of Clinical Neurosciences, Division of Clinical Neurology, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DS, UK; 3) Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK; 4) MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DS, UK; 5) Danish National Research Foundation Centre PERSIMUNE, Rigshospitalet, University of Copenhagen DK 2100, Denmark.

Sharing of genetic risk across multiple common human diseases implies underlying molecular connectivity between clinical phenotypes. This connectivity is not captured well by current disease classifications but can have important medical implications if characterised systematically. Here, we use routine healthcare and genome-wide genetic data from the 500,000 participants in the UK Biobank to map cross-trait associations against 19,628 diagnostic terms with the TreeWAS analysis framework. We find that 14% of the assayed genetic variants show evidence of association with at least one diagnostic term and GWAS catalogue SNPs were found highly enriched within this group (OR = 5.62, $P = 1.37 \times 10^{-7}$). Cross-trait association analysis allowed for the identification of highly pleiotropic genetic variants across diagnostic terms and the quantification of sign-heterogeneity in genetic effects. For example, we found the SNP rs6025, known as the Leiden mutation in the *F5* gene, to be one of the most pleiotropic SNPs in the genome, associated with 260 ICD-10 diagnostic codes and the derived allele A (R506Q) increasing risk to 92% of these and decreasing risk to 8%. Overall, we found that 6.62% of the associated variants have evidence of sign heterogeneity. Utilising cross-trait analysis we cluster SNPs into groups with similar association profiles to the phenome and explore the different modalities of SNP associations with hypertension, the most prevalent diagnostic term in the UK Biobank cohort. We find evidence for 16 clusters with at least 10 independent variants across the genome with different patterns of risk. For example, we find evidence for clusters that affect hypertension only and others that impact distinct combinations of high cholesterol, heart disease, diabetes and thalassemia. Lastly, we explore the use of genome-wide SNP association profiles to redefine the classification of diseases. By focusing on genetic variants associated with immune-mediated diseases, we demonstrate that the likelihood of the data can be improved over the existing structure defined by the ICD-10, but that cluster profiles are inconsistent with any single hierarchical ontology. Our analyses show that the identification of pleiotropic associations can help to define the genetic architecture of complex traits, provide insight into evolutionary biology, and pave the way towards improved disease diagnosis, prognosis and therapy.

339

Measuring gene expression inequality in single cells using adjusted Gini index. *X. Zheng[1], C. Green[1], A. Shami[1], Q. Ma[1], S. Hammoud[1], J. Li[1,2].* 1) Department of Human Genetics, University of Michigan, Ann Arbor, MI; 2) Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI.

Single-cell RNA sequencing technologies have provided the potential for defining new cell types and detailed mapping of developmental trajectories. Recently, the Gini index, originally used in economics for measuring income inequality, has been adopted in single cell RNA-sequencing studies to find individual genes with highly uneven expression among cells, particularly those expressed in very rare cells. Here, we apply this measure to characterize transcriptional diversity of cells, and show how it represents a new, biologically meaningful attribute of distinct cell types. Generally, a larger Gini implies a higher level of uneven distribution, from the most abundant to the least abundant transcripts. Such a cell devotes most of its transcripts on a smaller spectrum of unique genes, and this in turn may reflect a more focused regulatory program, or a terminally differentiated state. We observed an important caveat that, in sparse counts data typically encountered in single cell RNA-seq, the Gini index has systematic dependencies with the cell size factor (the total number of transcripts of the cell) and the sparsity of the data (the proportion of zeros). By simulating RNA-seq data with known distribution properties we developed an adjusted Gini index (AGI) to correct for the count-based biases. Applying it to an in-house Drop-seq dataset of >33,000 single cells from the adult mouse testis we found that the somatic supporting cells have comparable AGI's as the spermatogonia cells, whereas the progression from spermatocytes to elongating spermatids is accompanied by progressively higher AGI, suggesting that more developed germ cells are increasingly focused on highly specialized functions. This is consistent with our prior reports that mature sperm cells are dominated by a small subset of spermatogenesis genes such as protamine. As one of the quantitative measures of transcriptomic diversity, AGI is related to other entropy-related metrics that correlate with cells' differentiation potency or plasticity (PM27345837, PM28569836), providing a new perspective of biological function of single cells. Further, our finding bears resemblance to the dependency of α- and β-diversity measures on species abundance in ecological studies. *Supported by NIH-R21HD090371 and the Michigan Center for Single-Cell Genomic Data Analytics.*.

Copyright © 2018 American Society of Human Genetics. All rights reserved

340

Decomposing cell identity for transfer learning across omics, tissues, and species. *G.L. Stein-O'Brien[1,2,3,4], B.S. Clark[2], T. Sherman[3], S. Blackshaw[2,5,6,7,8], E.J. Fertig[3,4], L.A. Goff[1,2].* 1) McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins School of Medicine, Baltimore, MD; 2) Department of Neuroscience, Johns Hopkins University, Baltimore, MD; 3) Department of Oncology, Division of Biostatistics and Bioinformatics,The Johns Hopkins School of Medicine, Baltimore, MD; 4) Institute for Data Intensive Engineering and Science, Johns Hopkins University, Baltimore, MD; 5) Department of Ophthalmology, The Johns Hopkins School of Medicine, Baltimore, MD; 6) Institute for Cell Engineering, Johns Hopkins University, Baltimore, MD; 7) Center for Humans Systems Biology, Johns Hopkins University, Baltimore, MD; 8) Department of Neurology, The Johns Hopkins School of Medicine, Baltimore, MD.

The rise of single cell RNA sequencing (scRNAseq) has enabled unprecedented examination of cell population heterogeneity, cell-cell variation, and has challenged traditional morphological based classification and signatures derived from bulk RNA sequencing (RNAseq). It is becoming increasingly important to develop methods to decompose the transcriptional signature of individual cells into consequential features. These features, represented as reduced dimension latent spaces, can be used to describe the contribution of discrete biological and technical effects within a dataset. Further, these feature dimension can be independently compared across diverse datasets. Here we propose that 1) cell identity should map to a reduced set of dimensions, the unique combination of which defines the cell in high-dimensional expression space and 2) these dimensions can be used to learn meaningful relationships across diverse data sets including different tissues, model systems, technologies, and even species. To this end, we have developed two software packages: scCoGAPS and projectR. scCoGAPS learns robust feature dimensions from scRNAseq that represent meaningful biological processes and the use of these processes across groups of cells. ProjectR is a general framework to relate continuous valued feature dimensions in biologically related data across diverse data types. Combining scCoGAPS with ProjectR enables transfer learning via dimension reduction, i.e. the learning of the biological processes in one training dataset and subsequent querying of its occurrence and significance in new test dataset. We apply these methods to a scRNAseq dataset of ~125k cells across mouse retina development. We demonstrate how dimensions learned with scCoGAPS can capture shared aspects of biology in 1) independent retina datasets from different sequencing technologies, 2) ATAC data, 3) a catalog of mouse tissue scRNAseq data, and 4) both mouse and human cortical development. This implementation of transfer learning via dimension reduction represents a platform for rapid exploration of discrete biological processes across diverse experimental conditions and hypothesis generation where knowledge from multiple data sets can be leveraged to inform the selection of meaningful feature dimensions for biological validation. Thus, these methods represent an important advance for the study of human diseases where information is gleaned across limited patient samples and multiple model systems..

341

Detecting differential splicing events from single-cell RNA sequencing data with or without unique molecular identifiers. *Y. Hu[1], N.R. Zhang[2], M. Li[1].* 1) Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA; 2) Department of Statistics, University of Pennsylvania, Philadelphia, PA.

The emergence of single-cell RNA-seq (scRNA-seq) technology has made it possible to measure gene expression variations at cellular level. This breakthrough enables the investigation of a wider range of problems including analysis of splicing heterogeneity among individual cells. However, compared to bulk RNA-seq, scRNA-seq data are much noisier due to high technical variability and low sequencing depth. Here we propose SCATS (Single-Cell Analysis of Transcript Splicing) for differential splicing analysis in scRNA-seq, which achieves high sensitivity at low coverage by accounting for technical noise. SCATS models scRNA-seq data either with or without Unique Molecular Identifiers (UMIs). For non-UMI data, SCATS explicitly models technical noise by accounting for capture efficiency and amplification bias through the use of external spike-ins; for UMI data, SCATS models capture efficiency and further accounts for transcriptional burstiness. A key aspect of SCATS lies in its ability to group “exons” that originate from the same isoform. Grouping exons is essential in splicing analysis of scRNA-seq data as it naturally aggregates spliced reads across different exons, making it possible to detect splicing events even when sequencing depth is low. To evaluate the performance of SCATS, we analyze both simulated and real scRNA-seq datasets, and compare with existing methods, including Census, DEXSeq and MISO. In simulation of non-UMI data, we weaken the true splicing difference between conditions by introducing technical noise through a generative model. We show that SCATS has well-controlled type I error rate, and is more powerful than existing methods, especially when splicing difference is small. In contrast, Census suffers from severe type I error inflation, whereas DEXSeq is substantially more conservative. In simulation of UMI data, we show that SCATS accurately estimates bursting frequency and is substantially more powerful than existing methods in detecting differential splicing events. We also apply SCATS to a mouse brain scRNA-seq dataset with ~1,600 single cells. SCATS identifies more differential splicing events with subtle difference across 49 cortical cell types compared to MISO, and these events were confirmed by qRT-PCR experiment. We are currently applying SCATS to multiple real scRNA-seq datasets that include both UMI and non-UMI data. With the increasing adoption of scRNA-seq, we believe SCATS will be well-suited for various splicing studies.

Copyright © 2018 American Society of Human Genetics. All rights reserved

342

Identification of meiotic recombination events through gamete genome reconstruction by linked-read sequencing technology. *Z. Chong, P. Xu.* Genetics and Informatics Institute, University of Alabama at Birmingham, Birmingham, AL.

Meiotic recombination (MR), which transmits exchanged genetic materials between homologous chromosomes to offspring, plays a crucial role in shaping genomic diversity in eukaryotic organisms. In humans, thousands of meiotic recombination hotspots have been mapped by population genetics approaches. However, direct identification of MR events for individuals is still challenging because it is difficult to resolve the haplotypes of homologous chromosomes and to reconstruct the gamete genome. Whole genome linked-read sequencing (lrWGS) can generate haplotype sequences of mega-base pairs (N50 of 0.8-2.8Mb) after computational phasing. However, the haplotype information is indistinguishable between a large number of fragmented genomic regions and is also limited by switch errors (as high as 1%), impeding its further application in the chromosome-scale analysis. In this study, we developed a tool MRLR (Meiotic Recombination identification by Linked-Read sequencing) for the analysis of individual MR events in trios. By leveraging pedigree information with lrWGS haplotypes, our pipeline is sufficient to reconstruct the whole human gamete genome with 99.8% haplotyping accuracy. By analyzing the haplotype exchange between homologous chromosomes, MRLR identified 462 high-resolution MR events in 6 human trio samples from Genome In A Bottle (GIAB) and Human Genome Structural Variation Consortium (HGSVC). In three datasets from the HGSVC, our results recapitulated 149 (92%) previously identified high-confident MR events and discovered 85 novel events. About half (40) of the new events are supported by single-cell template strand sequencing (Strand-seq) results. We found that the MR breakpoint regions are enriched in PRDM9 and DMC1 binding sites. Besides, 332 (71.9%) breakpoint regions co-localize with recombination hotspots (>10 cM/Mb) in human populations, suggesting a proportion of MR events may be private to an individual. In addition, 48% (221) breakpoint regions were detected inside a gene, suggesting these MRs may directly increase haplotype diversity of genic regions. Taken together, our approach provides new opportunities in individualized and haplotype-based genomic analysis of human inheritance. The MRLR software is implemented in Perl and is freely available at https://github.com/ChongLab/MRLR.

343

Determining KIR repertoires using single cell transcriptomics. *A. Goncalves[1], R. Vento-Tormo[2], M. Efremova[2], S.A. Teichmann[2].* 1) DKFZ, Heidelberg, Germany; 2) Wellcome Trust Sanger Institute, Cambridge, United Kingdom.

Natural killer (NK) cells are important regulators of immune responses, with functions that extend from killing infected or transformed cells, to interactions with dendritic cells, macrophages and fetal trophoblast cells. NK cells are regulated in part by a family of receptors that recognise MHC class 1 molecules, the killer-cell immunoglobulin-like receptor genes (KIRs). Particular combinations of HLA and killer cell immunoglobulin–like receptor (KIR) genes have been associated with autoimmunity, viral infections, reproductive failure and cancer. However, the functional basis of the observed associations is poorly understood. An analysis of receptor–ligand interactions is currently required to understand how HLA–KIR genotypes contribute to disease. Similar to the HLA region, the KIR family, is highly polymorphic and displays extensive homology between genes, making it challenging to detect and quantify the expression of its members accurately. Here we propose a new computational method to genotype and quantify the expression of KIR genes using full-length single cell RNA-seq data. Briefly, we first perform a genotyping step based on k-mer detection, which is used to build a custom reference for each individual, and subsequently quantify expression by using a method to disaggregate allele isoform specific expression from multi-mapping reads. Using the results from this method, we investigate how decidual NK cells interact and regulate the invasion of fetal-derived trophoblasts in the early stages of pregnancy. For this, we analysed ~50,000 single cells (1278 NK cells) from 5 donors, assayed with plated-based Smart-seq2 (SS2). Our findings reveal the interactions between fetal trophoblast cells, maternal immune cells and multiple decidual NK cell subsets, which orchestrate the support of early pregnancy.

Copyright © 2018 American Society of Human Genetics. All rights reserved

344

Single-cell co-expression network demonstrated superior biological signal in tissue-specific network analyses. *T. Li[1,2], A. Kim[2], A. Battle[3], L. Kasper[2,4].* 1) MD-PhD Program, Johns Hopkins School of Medicine, Baltimore, MD; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Department of Biomedical Engineering, Johns Hopkins School of Medicine, Baltimore, MD; 4) Massachusetts General Hospital, Harvard Medical School, Boston, MA.

Single-cell RNA sequencing (scRNA-seq) technologies have emerged as a powerful tool to dissect cellular heterogeneity in a variety of tissues. Despite recent interests in co-expression networks based on scRNA-seq data, there is limited evidence showing direct utility of scRNA-seq co-expression networks in understanding biological circuitry. Here we obtained thalamic reticular nucleus (TRN) tissues enriched for neurons from six adult mice (99-113 days), and sequenced single-nucleus full-length transcriptomes using optimized Smart-Seq2 and Nextera protocols (N = 694 cells). Using a modified WGCNA pipeline, we constructed a TRN co-expression network with 229,205 interactions spanning 11,934 genes (TRNNet). We then applied a random forest classifier ("Quack" algorithm) trained on 306 neurodevelopmental pathways curated from the Molecular Signature Database (MSigDB). TRNNet exhibited superior performance in recapitulating neuronal pathway architecture (AUC = 0.80) compared to generic co-expression network constructed using bulk RNA-Seq (from GEO Database, AUC = 0.78), and as negative controls, cancer dependency and cell-perturbation-based networks (AUC < 0.70). To further demonstrate the ability of TRNNet to perform tissue-specific network analyses, we applied TRNNet-specific Quack model to predict novel genes implicated in autism spectrum disorders (ASDs), using 65 well-established ASD risk genes. TRNNet recapitulated ASD candidate genes from recent sequencing studies including *FAM47A*, *DOCK8* and *SETBP1*. By integrating frontal-cortex-specific eQTL data based on GTEx V7, TRNNet also nominated *TDO2* (Quack probability = 0.90), suggesting thalamocortical regulation as a potential mechanism for its association with ASD as previously established. Overall, we have established a computational pipeline to leverage the unique potential of scRNA-seq data for high-resolution tissue-specific network analyses.

345

Systematic evaluation of yields from whole-genome sequencing of 8,954 individuals compared to conventional technologies for prenatal and pediatric diagnostics. *H. Brand[1,2,11], C. Lowther[1,2,11], B.B. Currall[1,2], J.L. Giordano[3], V. Aggarwal[4], H.Z. Whang[1], X. Zhao[1,2], D. Lucente[1], L. Margolin[2], J-Y. An[5], D.M. Werling[6], S. Dong[6], S.J. Sanders[5], B. Devlin[6], K. Gilmore[7], B. Powell[8], A. Brandt[8], A.H. O'Donnell-Luria[1,2,9], N.J. Lennon[2], D.B. Goldstein[10], H.L. Rehm[1,2], D. MacArthur[1,2], N.L. Vora[7], B. Levy[4,12], R. Wapner[3,12], M.E. Talkowski[1,2,12].* 1) Center for Genomic Medicine and Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA; 2) Program in Population and Medical Genetics and Genomics Platform, The Broad Institute of M.I.T. and Harvard, Cambridge, MA; 3) Department of Obstetrics & Gynecology, Columbia University Medical Center, New York, NY; 4) Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY; 5) Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA; 6) Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA; 7) Department of Obstetrics and Gynecology, Division of Maternal-Fetal Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC; 8) Department of Genetics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC; 9) Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA; 10) Institute for Genomic Medicine, Columbia University Medical Center, New York, NY; 11) These authors contributed equally; 12) Co-corresponding authors.

Clinical genetic screening in prenatal and pediatric cohorts have traditionally required a decision to test targeted genes or mutational classes, as evaluation of all variant classes has been intractable. In genome-wide screening, chromosomal microarray (CMA) can detect sub-microscopic copy number variants (CNVs) but misses balanced chromosomal anomalies (BCAs), which require karyotyping, and both methods are blind to coding single nucleotide variants (SNVs), accessible to sequencing technologies. Whole genome sequencing (WGS) has the potential to transform diagnostic testing by capturing variation from all three technologies; however, virtually all diagnostic WGS studies to date have only assessed coding SNV largely due to technical limitations for structural variation (SV) discovery and immature annotation of pathogenic noncoding variation. Here, we compared incremental yields from WGS to karyotype, CMA, and whole exome sequencing (WES) in a pediatric cohort of 2,100 quartet families with a proband diagnosed with autism spectrum disorder (ASD; n=8,400) and a prenatal cohort of 176 trios and 26 singleton cases with a structural defect detected on ultrasound, a subset of which were derived from a previous comparison of karyotyping to CMA in 4,340 fetuses (Wapner et al., 2012, NEJM). We first benchmarked our bioinformatic pipelines on 519 ASD quartets, discovering 3.4M SNVs, 0.3M indels, and 5,863 SVs per genome (CNVs, BCAs, and complex rearrangements), including 69 *de novo* SNVs/indels and 0.2 *de novo* SVs per individual. WGS recapitulated 99.6% of all CMA-predicted CNVs and >97% of all *de novo* coding variants from WES. Molecular validation of 171 *de novo* SVs revealed a 97% confirmation rate, but WGS provided only ~0.3% increased diagnostic yield over the combination of previous methods. Using ACMG interpretation criteria, we next evaluated WGS in the fetal samples. From Wapner et al., diagnostic rates of karyotype and CMA were 10.7% and 14.2%, respectively. Initial analyses of WGS discovered an additional pathogenic variant in 11.1% of cases, yielding an estimated 25% diagnostic rate across all variant classes. This study suggests a modest overall increased diagnostic yield of WGS compared to the combination of all conventional methods; however, WGS was superior to any individual method and provided a single test to displace all three conventional technologies with sufficient specificity to warrant evaluation as a first-tier screen in clinical diagnostics.

Copyright © 2018 American Society of Human Genetics. All rights reserved

346

Low-pass whole-genome sequencing: A study of 1,090 recurrent miscarriage couples. *Z. Dong[1,2,3], J. Yan[3,4,5], F. Xu[6,7], J. Yuan[6,7], H. Wang[1,2,8], T.Y. Leung[1,2,9,10], S.W. Cheung[9,11], C.C. Morton[12,13,14,15,16], H. Jiang[6,7], Z.J. Chen[3,4,5,17,18], K.W. Choy[1,2,9,10].* 1) Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong, Hong Kong, China; 2) Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China; 3) Centre for Reproductive Medicine, Shandong University, Jinan, China; 4) The Key laboratory of Reproductive Endocrinology (Shandong University), Ministry of Education, China; 5) National Research Center for Assisted Reproductive Technology and Reproductive Genetics, China; 6) BGI-Shenzhen, Shenzhen, China; 7) China National Genebank-Shenzhen, BGI-Shenzhen, Shenzhen, China; 8) Department of Central Laboratory, Bao'an Maternity and Child Healthcare Hospital Affiliated to Jinan University School of Medicine, Key Laboratory of Birth Defects Research, Birth Defects Prevention Research and Transformation Team, Shenzhen, China; 9) The Chinese University of Hong Kong-Baylor College of Medicine Joint Center For Medical Genetics, Hong Kong, China; 10) Hong Kong Branches of Chinese National Engineering Research Centers – Center for Assisted Reproductive Technology and Reproductive Genetics, Hong Kong, China; 11) Department of Molecular and Human Genetics, Baylor College of Medicine Houston, Texas, USA; 12) Department of Obstetrics and Gynecology, Brigham and Women's Hospital, Boston, Massachusetts, USA; 13) Harvard Medical School, Boston, Massachusetts, USA; 14) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 15) Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA; 16) Division of Evolution and Genomic Sciences, School of Biological Sciences, University of Manchester, Manchester Academic Health Science Center, Manchester, UK; 17) Center for Reproductive Medicine, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China; 18) Shanghai Key Laboratory for Assisted Reproduction and Reproductive Genetics, Shanghai, China.

Background: G-banded chromosome analysis remains the standard assay to detect chromosomal rearrangements in couples with recurrent miscarriage (RM) despite its limited resolution. Recently, whole-genome sequencing (WGS) shows its capability in detecting both balanced and unbalanced rearrangements with improved resolution. However, application of this state-of-the-art approach to make additional etiologic diagnoses for RM couples has yet to be assessed rigorously. **Methods**: We performed low-pass WGS retrospectively for 1,090 RM couples, all of whom had routine chromosome analysis. Chromosomal rearrangements and copy-number variants (CNVs) were identified blinded to previous karyotype results and the diagnoses were confirmed by fluorescence in-situ hybridization or by PCR studies. CNV were classified according to guidelines of the American College of Medical Genetics and Genomics. **Results**: Low-pass WGS yielded results in 1,077 of 1,090 couples (98·8%) with chromosomal abnormalities diagnosed in 9·5% (102/1,077), 25·5% (26/102) of whom were missed by previous cytogenetic testing. This method enabled making diagnoses in 26 (2·4%) additional couples, including eight with balanced translocations, six with pathogenic or likely pathogenic CNVs and 12 with potentially etiologic inversions in RM. **Interpretation**: In the context of chromosome analysis for RM couples, low-pass WGS identified additional etiologic chromosomal abnormalities, including balanced translocations, inversions and CNVs, compared to standard G-banded karyotype results. This utility makes clear its clinical value in comprehensive detection and precise delineation of chromosomal abnormalities, which are essential for prognosis and clinical management in RM couples. Our study establishes the potential application of low-pass WGS to facilitate diagnoses not only in RM couples, but with an extension to prenatal and postnatal testing. **Funding:** This project is supported by the National Key Research and Development Program of China (2016YFC1000202), the National Natural Science Foundation of China (8149073, 81370715 and 81300075), the Health and Medical Research Fund (04152666), and the National Institute of General Medical Sciences (GM061354).

347

Whole exome sequencing as a tool for the diagnosis of inborn errors of metabolism in cohort of 550 patients with ID and developmental defect. *J. Delanne[1,2,3], N. Houcinat[2], S. Moutton[1,2,3], A.L. Bruel[1,3], S. Nambot[1,2,3], P. Callier[1,3], A. Sorlin[1,2,3], P. Callier[3], N. Jean-Marçais[2], A.L. Mosca-Boidron[3], F. Tran Mau-Them[1,3], D. Lehalle[2], S. El Chehadeh[2], C. Francannet[4], M. Lebrun[5], L. Lambert[6], M.L. Jacquemont[7], M. Gérard-Blanluet[8], J.L. Alessandri[9], M. Willems[10], F. Feillet[11], T. O'Grady[12], Y. Duffourd[1,3], C. Philippe[1,3], L. Faivre[1,2,3], C. Thauvin-Robinet[1,2,3].* 1) Univ. Bourgogne Franche-Comté, UMR 1231 GAD team, Genetics of Developmental disorders, FHU TRANSLAD, CHU Dijon Bourgogne, France; 2) CHU Dijon, Centre de référence maladies rares Anomalies du Développement et Syndromes Malformatifs et Centre de référence maladies rares Déficiences Intellectuelles de causes rares, Centre de Génétique, FHU TRANSLAD, CHU Dijon Bourgogne, France; 3) Unité Fonctionnelle d'Innovation diagnostique dans les maladies rares; Laboratoire de Génétique chromosomique moléculaire, CHU Dijon Bourgogne, France; 4) Unité de Génétique Médicale, Hotel Dieu, CHU Clermont Ferrand, France; 5) Laboratoire de génétique, CHU de Saint-Etienne, Saint-Etienne, France; 6) Unité Fonctionnelle de Génétique Clinique, Service de Médecine Néonatale, Maternité Régionale Universitaire, CHU Nancy, France; 7) Unité de Génétique Médicale, Pole Femme-Mère-Enfant, Groupe Hospitalier Sud Réunion, CHU de La Réunion, La Réunion, France; 8) APHP, Department of Genetics, Robert Debré Hospital, Paris, France; 9) Service de Réanimation Néonatale, Pole Femme-Mère-Enfant, CH Felix Guyon, CHU de La Réunion, Saint-Denis, La Réunion, France; 10) Department of Medical Genetics, Reference Center for Rare Diseases, Developmental Disorders and Multiple Congenital Anomalies, Arnaud de Villeneuve Hospital, Montpellier, France; 11) Laboratory of Cellular and Molecular Pathology in Nutrition, EMI INSERM 00-14, Vandoeuvre-les-Nancy, France; 12) Royal College of Surgeons in Ireland Education and Research Centre, Beaumont Hospital, Dublin 9, Ireland.

Inborn errors of metabolism (IEM) are a group of rare genetic disorders resulting from a metabolic pathway disturbance. More than 500 IEMs have been described, with a wide phenotypical heterogeneity and a broad onset (from the embryonic development period to adulthood). Diagnosis approach in suspected IEM is mostly based on biochemical screening in first intention (blood and urine), +/- MRI and spectroRMN, and targeted biochemical or molecular testing in second intention. Considering that some IEM could sometimes be diagnosed in patients with no distinctive clinical features of IEM, we aimed at evaluating the power of whole exome sequencing (WES) to diagnose IEM within a cohort of 550 patients with ID. After uninformative clinical evaluation, WES (92% solo / 8% trio) was ordered for non-specific encephalopathy, intellectual disability, with or without congenital malformations. Causative variant was found in 117/550 cases (21%) including IEM in 20/117 cases (17%). An IEM was diagnosed in two fetuses: one with an *ALDH18A1* defect and one with a Nieman Pick C disease. Eighteen children [A1] cases were diagnosed with different IEM: mitochondropathies (3 cases) (1 *DRP1*, 2 *ADCK3*-linked Coenzyme Q10 deficiency); lysosomal storage diseases (9 cases) (2 mannosidosis, 4 Neuronal Ceroid Lipofucinoses (2 *PPT1* and 2 *CLN3*), 2 gangliosidoses (GM2-*HEXA*, GM3-*ST3GAL5*)); protein glycosylation disorders (2 cases) (*NGLY1, PIGN*); GLUT1 deficiency (1 case), SPR deficiency (1 case), GABA deficit (1 case) (*SLC6A1*), and a new *SLC13A5*-linked syndrome (1 case). Both mannosidosis diagnoses were early made in cases presenting with learning difficulties and hearing loss but without dysmorphism evocative of lysosomal storage disease. GLUT1 and SPR deficiencies led to specific treatment. For 9/20 cases, reverse biochemical testing remained normal or aspecific and would not have led to the right diagnosis in the absence of clinical orientation. With the diagnosis of IEM in 3,6% of the cohort (20/550 cases), WES appears a powerful tool to diagnose IEM, especially in our cohort of atypical phenotypes and/or when biochemical studies appear normal. It also permits to delineate new IEM with the discovery of new *SLC13A5*-linked syndrome. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

348

Frequency and features of incidental findings across 12,702 eMERGE network participants. *A.S. Gordon[1], H. Zouk[2], E. Venner[3], M.S. Leduc[3], L. Witkowski[2], C.M. Eng[3], B.H. Funke[2], L.M. Amendola[1], D.S. Carrell[4], R.L. Chisholm[5], W.K. Chung[6], R.C. Green[7], H. Hakonarson[8], I.J. Kullo[9], E.B. Larson[4], K.A. Leppig[10], D.M. Muzny[3], N.J. Lennon[11], C.A. Prows[12], L.J. Rasmussen-Torvik[13], M.E. Smith[5], I.B. Stanaway[14], S.L. Van Driest[15], K. Walker[3], G.L. Wiesner[16], M.S. Williams[17], D.R. Crosslin[14], R.A. Gibbs[3], H.L. Rehm[11], G.P. Jarvik[1], The Electronic Medical Records and Genomics (eMERGE) Network.* 1) Medical Genetics, University of Washington, Seattle, WA; 2) Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA; 3) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 4) Kaiser Permanente Washington Health Research Institute, Seattle, WA; 5) Center for Genetic Medicine, Northwestern University, Chicago, IL; 6) Department of Medicine, Columbia University Medical Center, New York, NY; 7) Genetics, Department of Medicine, Brigham and Women's Hospital, Broad Institute and Harvard Medical School, Boston, MA; 8) The Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA; 9) Department of Cardiovascular Diseases, Mayo Clinic, Rochester, MN; 10) Genetic Services, Kaiser Permanente of Washington, Seattle, WA; 11) The Broad Institute of MIT and Harvard, Cambridge, MA; 12) Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 13) Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL; 14) Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA; 15) Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, TN; 16) Department of Medicine, Division of Genomic Medicine, Vanderbilt University Medical Center, Nashville, TN; 17) Genomic Medicine Institute, Geisinger, Danville, PA.

Increasing numbers of patients are undergoing panel, exome, or genome sequencing tests in clinical care. Many of these tests will identify genetic variation that is medically actionable, but not related (termed incidental or secondary findings) to the indication for testing. Prior estimates of the rate of incidental findings are limited by cohort size and, critically, homogeneity of cohort age, ancestry, and indication for testing. Addressing these knowledge gaps is a major focus for Phase III of NHGRI's Electronic Medical Records and Genomics network (eMERGE), which developed a panel of 109 genes representing a variety of clinical and research questions. In addition to the "ACMG56," 12 genes had broad consensus within eMERGE for actionability for a total of 68 genes in which a Pathogenic or Likely Pathogenic (P/LP) finding is returned to participants. This "eMERGEseq" panel was deployed by 2 CLIA sequencing labs in over 25,000 participants with varying (or no) primary test indication, drawn from 9 different eMERGE sites across the US. As the rate of incidental findings may vary due to subtle differences between testing labs, variant interpretation was standardized between sequencing centers as best as possible through a pre-launch harmonization effort comparing variant repositories, followed by ongoing comparisons to identify potential discrepancies. As of submission, 13,953 samples have been sequenced and interpreted, each of 9 eMERGE sites equally represented; one site that selected participants based on previous research testing was not included in this analysis, leaving a cohort of 12,702 samples. Of these, 4812 (38%) had no indication for testing (unselected for trait), while the remaining 62% had a broad range of indications, the most frequent being hyperlipidemia, colorectal cancer/polyps, intellectual disability, and breast cancer. Among the 12,702, we identified a positive (P/LP) finding unrelated to test indication in 408 individuals for an overall incidental finding rate of 3.2%. Most common among these were variants associated with cardiomyopathy, cancer, and familial hypercholesterolemia. Ongoing analyses stratifying these results by patient ancestry, age, indication, and interpretation (P vs. LP) will provide valuable benchmarks for physicians seeking to quantify risks and benefits of genomic testing across diverse patient groups and clinical settings.

349

Microbial detection in rapid whole genome sequencing in an acute care setting. *M.N. Bainbridge, E. Sanford, L. Farnaes, M. Wright, D. Dimmock, S. Kingsmore.* Rady Children's Institute for Genomic Medicine, San Diego, CA.

Infectious agents are a major cause of morbidity in the acute setting (NICU/PICU) and can complicate the clinical presentation of an underlying genetic disease. To rapidly diagnose genetic disease we have implemented a rapid whole genome sequencing (rWGS) pipeline at Rady Children's Hospital which routinely returns results in less than 48 hours. In total, we have sequenced over 250 children from the NICU, PCIU and acute care settings and have an overall genetic diagnostic rate of ~35%. To test whether we could identify microbial agents we developed a pipeline that filters rWGS data for DNA sequence reads which are not of human origin and aligns them to a collection of microbial genomes In total 610 samples (255 probands, 355 parental) were analyzed. Microbial DNA was detected in 58 (22.7%) probands. We identified a range of human herpes viruses (HHV-4, -5 , -7) in our cohort as well as parvo and torque teno virus. Streptococcus, E. coli, staphylococcus and pseudomonas were the most common causes of disease (38 cases). In 3 cases of severe bacterial infection we could also identify DNA originating from bacteriophage. Interestingly, ~30% of cases positive by sequencing were blood culture negative. In 5 cases (8.6%) the infection was thought to be the primary cause of disease. In 20 cases the patient was immunocompromised or had a long term stay in the NICU which lead to infection and a complication in clinical presentation. In the remaining cases, the infectious agent was considered incidental to the disease presentation. In one case we were able to identify an unsuspected, subclinical sepsis in a child who later received chemotherapy which exasperated the infection and could have been prevented with timely care. Rapid diagnosis of infectious agents can drive therapeutic choices and make phenotypic presentation of genetic disease clearer. DNA from these agents can be detected in sequencing reads for little additional cost and aid in making a molecular diagnosis and should be made part of a primary analysis pipeline.

Copyright © 2018 American Society of Human Genetics. All rights reserved

350

Genetic testing for healthy individuals: A medically actionable panel finds a high positive rate for hereditary disease. *E. Haverfield[1], E.D. Esplin[1], S. Aguilar[1], K.E. Ormond[2], A. Hanson-Kahn[2], S. Macklin[3], C. Sak[4], S. Bleyl[5], P. Atwal[3,5], C. Fine[5], P.J. Hulick[6], O.K. Gordon[7], J. Gu[8], L. Velsher[8], M. Duquette[1], R.L. Nussbaum[1], S. Aradhya[1].* 1) Invitae Corporation, San Francisco, CA; 2) Stanford University, Palo Alto, CA; 3) Mayo Clinic, Jacksonville, FL; 4) Tucker Medical, Singapore; 5) Genome Medical, San Francisco, CA; 6) NorthShore University HealthSystem, Evanston, IL; 7) Providence Health & Services, Los Angeles, CA; 8) Medcan, Toronto, Canada.

Introduction Genetic information is of increasing interest to healthy individuals and their healthcare providers because of the potential to uncover unknown hereditary risks and allow early detection and/or prevention of actionable monogenic disorders. Educational support for clinicians and genetic counseling for patients are important components that enhance the value and utility of medically actionable genetic information. We report on the positive yield of a medically actionable multigene panel for healthy individuals and describe cases highlighting the clinical value and early outcomes of these results.**Methods** Under an IRB-approved protocol, we analyzed de-identified data from 1,828 individuals who underwent genetic screening with a panel of up to 139 genes for actionable Mendelian disorders. Clinician-documented health information, if provided, was also reviewed.**Results** Pathogenic/likely pathogenic (P/LP) variants were observed in 16.5% (301/1,828) of individuals. While not considered a primary positive result, carrier status for certain genes was reported in 24.8% (454/1,828) of cases. Findings were in genes related to cancer syndromes (43.5%), cardiovascular disorders (40.1%), and other types of medically actionable disorders (16.4%). Genes with P/LP results included *BRCA1*, *BRCA2*, *MSH6*, *PMS2*, *APOB*, *LDLR*, *MYH7*, and *MYBPC3*, among others. Reportable results included single nucleotide variants, indels, and exonic copy number variants. Evaluation of only the genes recommended by the American College of Genetics and Genomics (ACMG) for reporting secondary findings found a positive rate of 5.6%, or 3.4% if heterozygous variants in *MUTYH* and other moderate-risk alleles were excluded. Medical histories were submitted for ~30% of tested individuals, and although most lacked a personal or family history that would meet diagnostic testing criteria, over 14% were positive for a P/LP variant conferring an increased risk of disease.**Conclusions** Actionable, health-related genetic information is critical to the preventive implementation of precision health. We found that 16.5% of tested individuals had an increased risk for hereditary cancer and cardiovascular disease for which established management guidelines exist. This panel offers healthy individuals the opportunity to learn about clinically significant genetic risks for hereditary disorders, which otherwise would have gone undetected since these individuals do not meet diagnostic criteria for genetic testing.

351

A deep learning framework identifies a role for noncoding *de novo* variants in congenital heart disease. *F. Richter[1,2,3], K.M. Chen[4], J. Zhou[4], S. Morton[5], A. Kitaygorodsky[6], H. Qi[6], N. Patel[3], K.B. Manheimer[7], E.E. Schadt[3,7], J.W. Newburger[8], E. Goldmuntz[9], M. Brueckner[10], G.A. Porter[11], R.W. Kim[12], D. Srivastava[13], D. Bernstein[14], M. Tristani-Firouzi[15], J. Yost[15], M. Yandell[15], Y. Shen[6], W.K. Chung[16], J.G. Seidman[5], C.E. Seidman[5], O.G. Troyanskaya[4,17,18], B.D. Gelb[2,3,19].* 1) Graduate School of Biomedical Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY; 3) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 4) Flatiron Institute, Simons Foundation, New York, NY; 5) Department of Genetics, Harvard Medical School, Boston, MA; 6) Departments of Systems Biology and Biomedical Informatics, Columbia University Medical Center, New York, NY; 7) Sema4, a Mount Sinai venture, Stamford, CT; 8) Boston Children's Hospital, Boston, MA; 9) Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 10) Yale University School of Medicine, New Haven, CT; 11) University of Rochester Medical Center, Rochester, NY; 12) Children's Hospital Los Angeles, Los Angeles, CA; 13) Gladstone Institute of Cardiovascular Disease, San Francisco, CA; 14) Department of Pediatrics, Stanford University, Palo Alto, CA; 15) Nora Eccles Harrison Cardiovascular Research and Training Institute, University of Utah, Salt Lake City, UT; 16) Departments of Pediatrics and Medicine, Columbia University Medical Center, New York, NY; 17) Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ; 18) Department of Computer Science, Princeton University, Princeton, NJ; 19) Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY.

Congenital heart disease (CHD) is primarily genetic, but only 30% is explained even after whole exome sequencing. As coding *de novo* variants (DNVs) explain 8% of CHD, we hypothesized that noncoding DNVs are also contributory. We performed whole genome sequencing in 749 parent/affected child trios and compared these to 1616 Simons Simplex Collection control trios. We identified 74 and 71 DNVs/trio in cases and controls, respectively, with high PCR-confirmation rates (SNVs 98%, indels 85%). To predict impact at single-base resolution, we developed DeepHeart, which encompasses 202 cardiac noncoding regulatory features trained on 1000-bp genomic sequence context with a neural network. DeepHeart is an extension of the published algorithm DeepSEA; both predict molecular effect differences between any two alleles for every modeled regulatory feature. Per-feature DeepHeart scores were compared between cases and controls among DNVs associated with 4 gene sets: all genes, human CHD, mouse CHD, and genes highly expressed during heart development. Comparisons were made with a Wilcoxon rank-sum test, and features with FDR q-values <0.05 were deemed significant, accounting for P value correlations among regulatory features and gene sets. DeepHeart had receiver operator characteristic area under the curve values of 0.9 and 0.85 for mouse and human features, respectively, similar to DeepSEA. For DNVs in all genes and mouse CHD genes, we observed, respectively, 64 and 58 regulatory features with significantly higher functional impact scores among the case DNVs; ~1/2 of significant features overlapped the 2 gene sets. DNVs in embryonic mouse heart H3K27ac marks generated the lowest P value ($9x10^{-7}$, all genes). To assess whether the numbers of significant features were more than expected, DNVs were randomly re-assigned case/control status (10,000 permutations), showing our observed numbers exceeded expectations (P<0.007 for all and mouse CHD genes). Finally, we contrasted maximum functional scores between HGMD regulatory mutations (n=1564) and polymorphisms (n=642) to define biologically meaningful scores in DeepSEA and DeepHeart as >0.1. We observed a significant burden of DNVs with maximum functional scores >0.1 in cases compared to controls (OR=1.1, Fisher's exact test P=0.03). These results highlight the role of noncoding DNVs in CHD and the utility of predicting functional impact at single-base resolution using tissue-specific regulatory features.

Copyright © 2018 American Society of Human Genetics. All rights reserved

352

Biallelic truncating mutations in *TMEM94* are associated with neurodevelopmental delay, congenital heart defects and distinct facial dysmorphism. *J. Stephen[1], S. Maddirevula[2], S. Nampoothiri[3], M. Herzog[4], J.D. Burke[1], A. Shukla[5], K. Steindl[6], A. Eskin[4], S.J. Patil[7], P. Joset[6], H. Lee[8], L.J. Garrett[9], T. Yokoyama[1], N. Balanda[10], A. Elkahloun[11], A. Zheng[4], K.M. Girisha[5], C. Rivas[9], G. Ramantani[12], S.M. Wakil[2], L. Mahmoud[13], A.B. Kulkarni[14], T. Ben-Omran[15], D. Colak[15], H.D. Morris[16], A. Rauch[6], F.S. Alkuraya[2,17,18], J.A. Martinez-Agosto[4], W. Gahl[1,10,19], M.C.V. Malicdan[1,10,19], Undiagnosed Diseases Network members.* 1) Section of Human Biochemical Genetics, Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA; 2) Department of Genetics, King Faisal Specialist Hospital and Research Center, King Saudi University, Riyadh 11211, Saudi Arabia; 3) Department of Pediatric Genetics, Amrita Institute of Medical Sciences and Research Center, Kerala 682041, India; 4) Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, 90095, USA; 5) Department of Medical Genetics, Kasturba Medical College, Manipal Academy of Higher Education, Manipal, India; 6) Institute of Medical Genetics, University of Zurich, Schlieren-Zurich 8952, Switzerland and radiz – "Rare Disease Initiative Zurich, Clinical Research Priority Program for Rare Diseases University of Zurich", Zurich 8032, Switzerland; 7) Mazumdar Shaw Medical Center, Narayana Health City, Bangalore, India; 8) Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, 90095, USA; 9) Embryonic Stem Cell and Transgenic Mouse Core, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA; 10) NIH Undiagnosed Diseases Program, NHGRI, National Institutes of Health, Bethesda, Maryland 20892, USA; 11) NHGRI Microarray Core, National Institutes of Health, Bethesda, MD 20892 USA; 12) Neuropediatrics, University Children's Hospital, Zurich 8032, Switzerland; 13) Section of Clinical and Metabolic Genetics, Department of Pediatrics, Hamad Medical Corporation, Doha, P.O. Box 3050, Qatar; 14) Functional Genomics Section, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, Maryland 20892, USA; 15) Department of Bio statistics, Epidemiology, and Scientific Computing, King Faisal Specialist Hospital and Research Center, Riyadh 11211, Saudi Arabia; 16) Mouse Imaging Facility, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 21042, USA; 17) Human Genome Program, King Abdulaziz City of Science and Technology, Riyadh 11442, Saudi Arabia; 18) Department of Anatomy and Cell Biology, College of Medicine, Alfaisal University, Riyadh 11533, Saudi Arabia; 19) Office of the Clinical Director, NHGRI, and the NIH Undiagnosed Diseases Program, NHGRI, National Institutes of Health, Bethesda, Maryland 20892, USA.

Human brain development is precisely orchestrated by various genes essential for neuronal cell proliferation, maturation, migration and vascularization; hence, genetic etiologies and phenotype associated with neurodevelopmental disorders are highly heterogenous. Here we describe nine similarly affected individuals from five unrelated families of different ethnic origins presented with motor and speech delay, social, cognitive and learning disabilities, facial dysmorphism and congenital heart defects. Using SNP microarray, homozygosity mapping and whole exome/genome sequencing we identified biallelic truncating mutations in *TMEM94*, as a shared candidate gene in all five families. *TMEM94* encodes for an uncharacterized transmembrane nuclear protein that is highly conserved across mammals. Gene expression analysis in lyphoblastoid cell lines and skin derived fibroblasts from the affected individuals showed reduced abundance of TMEM94 mRNA. Global gene expression analysis on *TMEM94* mutant cells using gene expression microarray and RNA sequencing revealed transcriptional dysregulation of several genes essential for cell cycle growth and proliferation that predicted to have an impact on cardiotoxicity, hematological system and neurodevelopment. To further understand the role of *TMEM94* in development and because of high protein sequence similarity between human and mice, we targeted *Tmem94* gene in mice using CRISPR/Cas9 gene editing. Homozygous null mice were embryonic lethal and embryos harvested at different embryonic stages of E12.5, E15.5 and E18.5 displayed decreased body size, craniofacial abnormalities and areas of superficial hemorrhages. *Tmem94* expression analysis in the mutants showed 80-90% reduction of mRNA, supporting the knock out of this gene. Whole body MRI, CT and X-ray of embryos at E18.5 revealed cardiac and craniofacial abnormalities. Histopathological examination of brain and heart showed neuronal migration defects and disarray of heart muscle fibers suggesting a defect in neuro and cardiac development, which recapitulate the phenotype observed in our probands. In conclusion, our study defines a novel genetic etiology of a recognizable dysmorphic syndrome with neurological and cardiac defects due to truncating mutations in *TMEM94*. In the context of the loss-of-function mutations in the affected individuals and knockout mice phenotypes, we demonstrate the role for this gene in mammalian embryonic development.

353

Variants in *MAP4K4* cause a novel and potentially treatable form of neurologic dysfunction with cardiac anomalies. *E.J. Bhoj[1], D. Li[1], M.H. Harr[1], R. Smith[2], S. Ellingwood[2], M. Cho[3], J. Keller-Ramey[3], R. Person[3], A. Sidhu[4], S. Saliganan[5], S.L. Cassisi[5], D.K. Grange[6], X. Hu[7], Y. Shen[8], M. Maimaiti[9], Y. Luo[9], H.H. Hakonarson[1].* 1) Children's Hospital of Philadelphia, Philadelphia, PA, PA; 2) Maine Medical Partners Pediatric Specialty Care; 3) GeneDx, Gaithersburg, MD; 4) University of Iowa Stead Family Children's Hospital; 5) Michigan State University; 6) Washington University School of Medicine, St. Louis Children's Hospital; 7) Beijing Children's Hospital; 8) Boston Children's Hospital; 9) Capital Medical University Xinjiang Medical University.

Mitogen-Activated Protein Kinase Kinase Kinase Kinase 4 (MAP4K4) is an activator of the JNK pathway, which controls many vital cellular processes, including proliferation, embryonic development, and apoptosis. The *MAP4K4* gene is extremely intolerant to missense variation, with a Z score is 4.01 (>2 is significantly constrained), and a maximum Probability of Loss of function Intolerance (pLI) with a score of 1. In 2017 a *de novo* nonsense variant in *MAP4K4* was identified in a fetus with HLHS and fused kidneys, but these prenatal ultrasound diagnoses are the only reported phenotype data (Vora, 2017). Here we report seven individuals from four families with variants in *MAP4K4* and neurologic dysfunction and cardiac anomalies. Three variants are expected to lead to early truncation, and one is a missense variant in a highly conserved region. These individuals share a phenotype that resembles a rasopathy, including one family with a clinical diagnosis of "Noonan-like syndrome." Cardiac defects include pulmonary artery stenosis, tricuspid and pulmonary valve insufficiency, patent foramen ovale, and atrial and ventricular septal defects. Development ranges from moderate to severe developmental delay. MAP4K4 has already been identified as a highly impactful target for pharmacologic manipulation for non-Mendelian disorders. It has been the focus of therapeutic studies in metabolic dysregulation (Tang et al. PNAS 2006), inflammation (Aouadi et al. Nature 2009), and pathologic angiogenesis (Vitorino et al. Nature 2015). The loss of MAP4K4 can be mitigated by inhibition of Integrin α5β1, which has been shown to be vital to the neurite outgrowth and axonal regeneration of adult brain neurons (Vitorino et al. Nature 2015). Small molecule inhibitors of Integrin α5β1 have already been tested in animal models. Therefore, this may represent a targeted therapy for the refractory neurologic and psychiatric morbidities in these patients. Further studies on an animal model and patient cells are currently underway.

Copyright © 2018 American Society of Human Genetics. All rights reserved

354

De novo variants in congenital diaphragmatic hernia identify _MYRF_ as a new syndrome and reveal genetic overlaps with other developmental disorders. *X. Zhou[1,2], H. Qi[2,3], L. Yu[1], J. Wynn[1], H. Zhao[2,4], Y. Guo[2], N. Zhu[1,2], A. Kitaygorodsky[2,4], R. Hernan[1], G. Aspelund[5], V. Duron[5], F.Y. Lim[6], T. Crombleholme[6], R. Cusick[7], K. Azarow[8], M.E. Danko[9], D. Chung[9], B.W. Warner[10], G.B. Mychaliska[11], D. Potoka[12], A.J. Wagner[13], M. ElFiky[14], J.M. Wilson[15,16], D. Nickerson[17], M. Bamshad[17], F.A. High[15,16,18], M. Longoni[16], P.K. Donahoe[16,18], W.K. Chung[1,19,20], Y. Shen[2,4,21].* 1) Department of Pediatrics, Columbia University Medical Center, New York, NY; 2) Department of Systems Biology, Columbia University, New York, NY, USA; 3) Department of Applied Mathematics and Applied Physics, Columbia University, New York, NY, USA; 4) Department of Biomedical Informatics, Columbia University, New York, NY; 5) Department of Pediatric Surgery, Columbia University Medical Center, New York, NY, USA; 6) Cincinnati Children's Hospital, Cincinnati, OH, USA; 7) Children's Hospital & Medical Center of Omaha, University of Nebraska College of Medicine, Omaha, NE, USA; 8) Department of Surgery, Oregon Health & Science University, Portland, OR, USA; 9) Monroe Carell Jr. Children's Hospital, Vanderbilt University Medical Center, Nashville, TN, USA; 10) Washington University, St. Louis Children's Hospital, St. Louis, MO, USA; 11) University of Michigan, CS Mott Children's Hospital, Ann Arbor, MI, USA; 12) Children's Hospital of Pittsburgh, Pittsburgh, PA, USA; 13) Medical College of Wisconsin, Milwaukee, WI, USA; 14) Department of Pediatric Surgery, Faculty of Medicine, Cairo University, Cairo, Egypt; 15) 15.Department of Surgery, Boston Children’s Hospital, Boston, MA, USA; 16) Department of Surgery, Harvard Medical School, MA, USA; 17) University of Washington, Seattle, WA, USA; 18) Pediatric Surgical Research Laboratories, Department of Surgery, Massachusetts General Hospital, Boston, MA, USA; 19) Department of Medicine, Columbia University, New York, NY, USA; 20) Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA; 21) JP Sulzberger Columbia Genome Center, Columbia University, New York, NY, USA.

Congenital diaphragmatic hernia (CDH) is a severe birth defect that is often accompanied by other congenital anomalies. Previous exome sequencing studies for CDH have supported a role of *de novo* damaging variants but did not identify any recurrently mutated genes. To investigate further the genetics of CDH, we analyzed *de novo* coding variants in 362 proband-parent trios including 271 new trios reported in this study. We identified four unrelated individuals with damaging *de novo* variants in *MYRF* ($P=5.3x10^{-8}$), including one likely gene-disrupting (LGD) and three deleterious missense (D-mis) variants. Seven additional individuals with *de novo* LGD or missense variants were identified from our other genetic studies or from the literature. Common phenotypes of *MYRF de novo* variant carriers include CDH, congenital heart disease and genitourinary abnormalities, suggesting that it represents a novel syndrome. *MYRF* is a membrane associated transcriptional factor highly expressed in developing diaphragm and is depleted of LGD variants in the general population. All *de novo* missense variants aggregated in two functional protein domains. Analyzing the transcriptome of patient-derived diaphragm fibroblast cells suggest that disease associated variants abolish the transcription factor activity. Furthermore, we showed that the remaining genes with damaging variants in CDH significantly overlap with genes implicated in other developmental disorders. Gene expression patterns and patient phenotypes support pleiotropic effects of damaging variants in these genes on CDH and other developmental disorders. Finally, functional enrichment analysis implicates the disruption of regulation of gene expression, kinase activities, intra-cellular signaling, and cytoskeleton organization as pathogenic mechanisms in CDH.

355

Hypomorphic mutations in the deubiquitilation enzyme _OTUD5_ lead to multiple congenital defects. *D.B. Beck[1], H. Oda[1], A. Asmar[2], N. Sampaio Moura[1], E. Macnamara[1], P. D'Souza[1], J. Bodurtha[3], M. Walkiewicz[4], R. Wang[5], C.J. Tifft[1], I. Aksentijevich[1], D. Kastner[1].* 1) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA; 2) National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD 20892, USA; 3) Institute of Genomic Medicine, Johns Hopkins Hospital, Baltimore, MD 21287, USA; 4) National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD 20892, USA; 5) Children's Hospital of Orange County, University of California Irvine School of Medicine, Orange, CA 92868, USA.

Ubiquitylation, the covalent attachment of ubiquitin to proteins, is an essential post-translational modification that orchestrates many aspects of human development. Through attachment of either one ubiquitin molecule or chains of ubiquitin, typically linked through different lysine residues (Lys), ubiquitylation is able to regulate various substrate fates via degradation and altered intracellular signaling pathways. Enzymes that catalyze the deposition and removal of ubiquitin are critical for many cellular processes and their mutations underlie disorders of human development such as Angelman Syndrome, caused by loss of UBE3A expression, and a complex congenital anomaly disorder caused by mutations in *OTUD6B*. Here we report six male patients with novel hemizygous missense mutations in an X-linked gene *OTUD5,* which encodes for a Lys48/Lys63-chain-specific deubiquitylation enzyme previously not linked to human disease. Affected individuals have clinical manifestations including structural brain malformations, congenital heart disease, ambiguous genitalia, post-axial polydactyly, arthrogryposis, and craniofacial defects. We find that disease-causing variants result in decreased OTUD5 function through distinct mechanisms including reduced overall protein expression due to aberrant splicing and altered enzymatic activity through catalytic reprogramming from a dual Lys48/Lys63- to a single Lys63-chain-specific enzyme. Consistent with a loss of function phenotype, decreasing OTUD5 levels causes aberrant cellular differentiation *in vitro* and *in vivo* and leads to altered RNA expression in patient-derived cells. Furthermore, we have identified and characterized linkage specific substrates of OTUD5 and their role during differentiation. Taken together, identification and mechanistic dissection of this novel genetic syndrome provides fundamental insights into how ubiquitin and specific ubiquitin linkages regulate important steps of human development.

Copyright © 2018 American Society of Human Genetics. All rights reserved

356

Heterozygous missense mutations in CDK8, a regulator of the Mediator complex, cause a syndromic developmental disorder. *E. Calpena[1], S.M.A. Swagemakers[2], J.A.C. Goos[2], T. Kaserer[3], O. Popoola[3], M.J. Ortiz Ruiz[3], T. Barbaro Dieber[4], L. Bownass[5], E. Brilstra[6], E. Brimble[7], N. Foulds[8], T.A. Grebe[9], A.V.E. Harder[6], M.M. Lees[10], K. McWalter[11], R.A. Newbury-Ecob[5], K.R. Ong[12], D. Osio[12], F.J. Reynoso Santos[13], M.R.Z. Ruzhnikov[7], E. Torti[11], E. van Binsbergen[6], M.F. van Dooren[14], D.D.D. Study[15], P.J. van der Spek[2], J. Blagg[3], S.R.F. Twigg[1], I.M.J. Mathijssen[2], P. Clarke[3], A.O.M. Wilkie[1,16].* 1) Clinical Genetics Group, MRC-Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK; 2) Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands; 3) Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, London, United Kingdom; 4) Cook Children's Genetics, Fort Worth, TX, USA; 5) University Hospitals Bristol, Department of Clinical Genetics, St Michael's Hospital, Bristol, UK; 6) Department of Genetics, University Medical Center Utrecht, Utrecht University, The Netherlands; 7) Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA, USA; 8) University Hospital Southampton, Southampton, UK; 9) Department of Child Health, University of Arizona College of Medicine, Division of Genetics and Metabolism, Phoenix Children's Hospital, Phoenix, AZ, USA; 10) North Thames Regional Genetics service, Great Ormond street Hospital NHS Trust, London, UK; 11) GeneDx, Gaithersburg, MD, USA; 12) Department of Clinical Genetics, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK; 13) Joe DiMaggio Children's Hospital, Hollywood, FL, USA; 14) Department of Clinical Genetics, Erasmus MC, Rotterdam, The Netherlands; 15) Deciphering Developmental Disorders Study, Wellcome Trust Sanger Institute, Cambridge, UK; 16) Craniofacial Unit, Oxford University Hospitals NHS Foundation Trust, Oxford, UK.

The Mediator complex is a global regulator of RNA polymerase II transcription in eukaryotes, required for expression of all protein-coding and most non-coding RNA genes. Mediator activity is regulated by the reversible association of a four-subunit kinase module (CDK8 module) comprising CDK8, CCNC, MED12 and MED13, or paralogs CDK19, MED12L and MED13L. Mutations in MED12, MED13 and MED13L were previously identified in syndromic developmental disorders with overlapping phenotypes, and amplification of *CDK8* is frequent in colon cancer. Here we report mutations in *CDK8* causing a novel congenital disorder in humans. Using whole exome/genome sequencing, and by international collaboration and GeneMatcher exchange, we identified 8 different heterozygous missense mutations of *CDK8* (pLI= 0.95 and z(missense)= 4.12, constraint scores from ExAC database) in 12 unrelated patients. In 10 cases we showed that the mutation arose *de novo* in the proband; one recurrent mutation, c.185C>T (p.S62L) was present in 5 different subjects. All the encoded substitutions localize to the kinase domain and, by structural modelling, we demonstrate that they cluster around the ATP-binding pocket, suggesting a shared pathogenic mechanism. Affected individuals have overlapping phenotypes characterised by mild-moderate intellectual disability and dysmorphic facial features. Additional clinical problems present in at least half of cases were behavioural disorders (notably autism spectrum disorder/ADHD), hypotonia and congenital heart disease. Agenesis of the corpus callosum, anorectal malformations, seizures and visual impairment occurred in a minority of individuals. To evaluate the functional impact of the mutations, we measured phosphorylation at STAT1-Ser727, a known CDK8 target, in a *CDK8/19* CRISPR double-knockout cell line overexpressed with wild-type or mutant *CDK8* constructs. This demonstrated a reduction of the CDK8 kinase activity of the mutants, to a similar extent as a CDK8 kinase-dead positive control. The reduced activity associated with the mutations, together with their clustered distribution and function of CDK8 within the Mediator complex, suggest dominant-negative action as the most likely pathogenic mechanism. Our findings demonstrate that missense mutations in CDK8 cause a developmental disorder with phenotypic similarities to syndromes produced by mutations in other subunits of the CDK8 module. This adds a further gene to the list of "Mediatorpathy" syndromes.

357

Uncovering novel cytogenetic and molecular etiologies for male infertility. *S.L.P. Schilit[1,2], S. Menon[3], T. Kammin[4], E. Wilch[4], C. Redin[2,5,6], S. Jiang[7], A.J. MacQueen[8], J.F. Gusella[2,5,6,9,10], M.E. Talkowski[2,5,6,10,11,12], C.C. Morton[2,4,10,13,14].* 1) Program in Genetics and Genomics and Certificate Program in Leder Human Biology and Translational Medicine, Biological and Biomedical Sciences Program, Graduate School of Arts and Sciences, Harvard University, Cambridge, MA, USA; 2) Harvard Medical School, Boston, MA, USA; 3) Harvard College, Cambridge, MA, USA; 4) Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, Boston, MA, USA; 5) Molecular Neurogenetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; 6) Department of Neurology, Massachusetts General Hospital, Boston, MA, USA; 7) Baxter Laboratory for Stem Cell Biology, Stanford University School of Medicine, Stanford, CA, USA; 8) Department of Molecular Biology and Biochemistry, Wesleyan University, Middletown, CT, USA; 9) Department of Genetics, Harvard Medical School, Boston, MA, USA; 10) Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 11) Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA; 12) Department of Pathology, Massachusetts General Hospital, Boston, MA, USA; 13) Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA; 14) Manchester Academic Health Science Centre, University of Manchester, Manchester, UK.

Unexplained infertility affects 2-3% of reproductive aged couples. One approach to identifying genes involved in infertility is to study subjects with this clinical phenotype and a *de novo* balanced chromosomal aberration (BCA). While BCAs may reduce fertility by production of unbalanced gametes, a chromosomal rearrangement may also disrupt or dysregulate genes important in fertility. One such subject, DGAP230, has severe oligospermia and 46,XY,t(20;22)(q13.3;q11.2). By large-insert whole genome sequencing, chromosomal breakpoints were determined with nucleotide-level precision. Investigation of genes in the topologically associated domains at the sites of the rearrangement revealed exclusive dysregulation of *SYCP2* (overexpressed 20-fold by RNA [p<0.02] from the der(20) allele and five-fold by protein [p<0.0032]), which resides 1.5 Mb centromeric to the der(20) breakpoint. 4C-seq from the *SYCP2* promoter revealed interactions 8 Mb downstream of the der(20) breakpoint in chr22 (210-fold increased interaction [p<0.0042]). CRISPR/Cas9-mediated deletions of putative enhancers from this region are being used to validate a role in *SYCP2* dysregulation, supporting a model of enhancer adoption as the etiology for *SYCP2* overexpression. *SYCP2* encodes synaptonemal complex protein 2 (MIM 604105), a member of the synaptonemal complex (SC) involved in homologous chromosome synapsis in male meiosis I. Misexpression of *SYCP2* may impair spermatogenesis by disrupting meiosis. To assess the impact on meiosis, we misexpressed *RED1*, the functional axial element homolog of *SYCP2* in *S. cerevisiae*, using a β-estradiol-induced promoter system (overexpressed 19-fold by RNA relative to meiotic levels [p<0.026]). By performing Red1 immunolocalization on surface-spread meiotic nuclei, we discovered that excess Red1 forms polycomplexes, disrupting the structural integrity of the SC by preventing incorporation of the transverse filament Zip1 (p<0.0001). The resulting asynapsis of homologous chromosomes can explain the molecular etiology in DGAP230's oligospermia, because asynapsis in spermatocytes induces checkpoint-mediated apoptosis and subsequently decreases sperm count. In sum, this investigation illustrates the power of precision cytogenetics for annotation of the infertile genome and suggests that these mechanisms should be considered as an alternative etiology to that of segregation of unbalanced gametes in infertile men harboring a BCA.

Copyright © 2018 American Society of Human Genetics. All rights reserved

358

Ectopically expressed CGG repeats lead to ovarian dysfunction in a mouse model of the *FMR1* premutation. *K.E. Shelly[1], N.R. Candelaria[2], D.L. Nelson[1].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX.

Women heterozygous for an expansion of CGG repeats in the 5'UTR of *FMR1* are at risk to develop Fragile X-associated Primary Ovarian Insufficiency (FXPOI) and/or Tremor and Ataxia Syndrome (FXTAS). We investigated whether these expanded CGGs, independent of *FMR1*, are sufficient to drive ovarian insufficiency. A second aim was to elucidate whether dysfunction arises via expression of CGG-containing mRNAs or a peptide product translated from these RNAs. Heterozygous females from two mouse lines expressing either CGG RNA-only (RNA-only) or CGG RNA and its translated polyglycine product FMRpolyG (FMRpolyG+RNA) were used to assess fertility by continuous breeding and superovulation studies. Other phenotypic data were collected longitudinally and histology examined in aging mice. Morphology was correlated with gene expression changes, assayed by qRT-PCR on whole ovary tissue. Our data suggest that CGG RNA and FMRpolyG+RNA both contribute to ovarian dysfunction, albeit differently. Immunostaining shows FMRpolyG is present in oocytes and granulosa in early postnatal life and gonadotropin stimulation reveals that young RNA-only and FMRpolyG+RNA mice ovulate fewer oocytes/female. Continuously breeding FMRpolyG+RNA mice exhibit declining fertility with age, but this reduction is not seen in RNA-only mice. Cessation of breeding in FMRpolyG+RNA females is preceded by significant weight gain compared to control mice, and histology shows a lack of ovulation as well as hyperplastic stroma and disorganized theca in aged ovaries. qRT-PCR from whole ovary tissue collected at 8 months of age reveals gene expression changes consistent with the morphological findings. Unexpectedly, cysts lined with Cytokeratin 8+ epithelial cells were noted in a subset of both FMRpolyG+RNA and RNA-only ovaries aged beyond 6 months. The origin of these lesions is not yet known. Together, our data show globally expressed FMRpolyG+CGG RNA leads to anovulation, altered steroidogenic profile, and fertility decline with age. CGG RNA-only mice do not exhibit a decline in fertility or increased weight. Diminished response to superovulation suggests that CGG RNA-alone may affect ovarian function, although less robustly than when expressed in conjunction with FMRpolyG. This highlights potential differences in pathological mechanism between FXPOI and FXTAS as published studies of CGG RNA-only mice show normal motor phenotypes, whereas FMRpolyG+RNA mice have reduced motor function.

359

Largest genome-wide association meta-analysis of endometriosis and its subphenotypes including 21K cases and 482K controls reveals 21 loci for the condition and genetically based comorbidity with various pain conditions. *N. Rahmioglu[1,2], S. Mortlock[3], K. Banasik[4,5], R. Mägi[6], L. Stefansdottir[7], C. Turman[8], A. Giri[9], O. Uimari[10], Y. Sapkota[11], S. Macgregor[3], B. Marciniak[12], S.K. Low[13], D. Strapagiel[12], A. Campbell[14], C. Hayward[15], R. Danning[16,17], T. D'hooghe[18], D. Nyholt[19], P. Rogers[20], C. Becker[2], D. Chasman[16,17], P. Kraft[8], D. Whiteman[21], D.V. Edwards[9], V. Steinthorsdottir[7], M. Nyegaard[4,5], G. Montgomery[22], S. Missmer[8,23,24], A.P. Morris[25], K.T. Zondervan[1,2], The International Endometriosis Genomics Consortium (IEGC).* 1) Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK; 2) Endometriosis CaRe Centre Oxford, Nuffield Department of Gynaecology and Obstetrics, University of Oxford, Oxford, UK; 3) Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Australia; 4) Department of Biomedicine, Aarhus University, Aarhus, Denmark; 5) The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Denmark; 6) Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia; 7) deCODE genetics/Amgen, Reykjavik, Iceland; 8) Department of Epidemiology, Harvard University, Boston, USA; 9) Department of Obstetrics and Gynecology, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA,; 10) Department of Obstetrics and Gynecology, Oulu University Hospital, Oulu, Finland; 11) Epidemiology and Cancer Control Department, St. Jude Children's Research Hospital, Memphis, TN, USA; 12) Biobank Lab, Faculty of Environmental Protection, University of Lodz, Lodz, Poland; 13) Center for Integrative Medical Sciences, RIKEN, Yokohama, Japan; 14) Centre for Genomic and Experimental Medicine, Institute of Genetics & Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK; 15) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK; 16) Harvard Medical School, Boston, MA, USA; 17) Division of Preventive Medicine, Brigham and Women's Hospital, Boston MA, USA; 18) KULeuven, Department of Development and Regeneration, Organ systems, Leuven, Belgium; 19) Institute of Health and Biomedical Innovation, Queensland University of Technology, Queensland, Australia; 20) Obstetric and Gynaecology Royal Women's Hospital, University of Melbourne, Melbourne, Australia; 21) Cancer Control Group, QIMR Berghofer Medical Research Institute, Brisbane, Australia; 22) Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia; 23) Department of Obstetrics, Gynecology and reproductive biology, Michigan State University, Michigan, USA; 24) Boston Center for Endometriosis, Boston, MA, USA; 25) Department of Biostatistics, University of Liverpool, UK.

Endometriosis is a complex condition, which causes infertility and chronic pelvic pain. The heritability has been estimated to be ~50% with 26% due to common genetic variants. We have conducted the largest meta-analysis of endometriosis to date, including 20,933 cases and 482,225 controls, across 14 genome-wide association studies of European ancestry and one of Japanese ancestry. We have also undertaken sub-phenotype analyses for stage III/IV (3,711 cases), stage I/II (3,160 cases) and infertility (2,843 cases). Each GWAS was imputed up to reference panels from the 1000 Genomes Project (Phase 3) or Haplotype Reference Consortium. We identified 21 loci associated with endometriosis at genome-wide significance ($p<5x10^{-8}$), of which 7 mapped outside previously reported regions for the condition, represented by common (minor allele frequency >5%) lead SNPs: rs495590 in *DNM3* ($p=6.73x10^{-10}$, OR=1.07 95%CI= 1.05-1.10); rs62468795 near *IGF2BP3* ($p=8.05x10^{-9}$, OR=1.10 95%CI=1.07-1.14); rs10090060 near *GDAP1* ($p=5.72x10^{-11}$, OR=1.08 95%CI=1.06-1.11); rs1802669 in *MLLT10* ($p=5.52x10^{-9}$, OR=1.07 95%CI=1.05-1.10); rs796945 in *RNLS* ($p=1.78x10^{-9}$, OR=1.07 95%CI=1.05-1.10); rs7151531 in *RIN3* ($p=3.80x10^{-8}$, OR=1.07 95%CI=1.04-1.10); and rs66683298 in *SKAP1* ($p=1.73x10^{-10}$, OR=1.08 95%CI=1.06-1.11). Gene-based analysis revealed 34 genes passing genome-wide significance threshold ($p<2.67x10^{-6}$), of which 13 map to 10 genomic regions mapping outside of known endometriosis loci: *ATG7, HMGA1, CD109, RSPO3, ADAM22, TRPS1, SKIDA1, HOXC6, RP11-834C11.12, IGF1, NUP37, PARBP, BMF*. We investigated the shared genetic contribution of endometriosis with a range of co-morbid autoimmune, metabolic, reproductive and pain-related conditions and traits via LD-score regression. We identified significant ($p<1.28x10^{-3}$), positive genetic correlations with excessive/irregular menstruation, uterine fibroids, diabetes, osteoarthritis, dorsalgia, back pain in last 3 months, back pain, headache, hip, knee, neck, abdominal, pain all over the body in the last month, and significant negative correlations with age at menarche, menstrual cycle length, and age of first birth. On-going expanded meta-analyses, including 62K cases, and interrogation in RNAseq/microarray data from 344/63 eutopic/ectopic endometrium samples, will provide the most comprehensive view of the genetic contribution to endometriosis, and the causal genes and molecular mechanisms through which their effects on the condition are mediated.

Copyright © 2018 American Society of Human Genetics. All rights reserved

360

Endometriosis genome-wide association study in >288,000 women of European ancestry. *G. Galarneau[1], P. Fontanillas[2], T. Hu-Seliger[1], C. Clementi[1], U. Schick[1], D. Colaci[1], D.E. Parfitt[1], J.Y. Tung[2], P. Yurttas Beim[1], the 23andMe Research Team, the Celmatix Research Team.* 1) Celmatix Inc., New York, NY; 2) 23andMe, Inc., Mountain View, CA.

Endometriosis is characterized by the growth of endometrial-like tissue outside the uterus. The disorder, which affects ~10% of women, is frequently associated with severe dysmenorrhea and infertility. Despite its high prevalence, many aspects of the pathophysiology of endometriosis remain to be elucidated.We performed the largest endometriosis genome-wide association study (GWAS) to date, using data from research participants of the personal genetics company 23andMe, Inc. The study included 37,183 women who reported being diagnosed with or treated for endometriosis (cases) and 251,258 women who reported not being diagnosed with or treated for endometriosis (controls).Samples were genotyped on custom genome-wide genotyping arrays targeting 556-955k single nucleotide polymorphisms (SNPs). Up to 15M additional SNPs were imputed using phase 1 of the 1000 Genomes Project as a reference. Assuming an additive model, we tested association using logistic regression including age, first 5 principal components, and genotyping array version as covariates.Fourteen loci were associated with endometriosis at genome-wide significance ($p< 5\times10^{-8}$), including signals in 8 loci not previously associated with endometriosis through GWAS before: *NGF*, *CD109*, *CEP112*, *ATP1B1-F5*, *STK3-VPS13B*, *HDDC2-HEY2*, *RCN1-WT1*, and *TEX11*.The *NGF* locus overlaps with a locus identified in a GWAS on pain severity in dysmenorrhea. High NGF levels have been observed in women with endometriosis undergoing *in vitro* fertilization. A study in rats showed that silencing *Ngf* gene expression suppresses the growth of ectopic endometrial implants.The *STK3-VPS13B* locus overlaps with *OSR2*, which is regulated by progesterone receptor in human endometrial stromal cells. Studies have shown that *OSR2* transcription is downregulated and its promoter is significantly hypermethylated in endometriotic stromal cells compared to normal endometrial stromal cells.The *HDDC2-HEY2* locus overlaps with loci identified in breast and endometrial cancer GWASs. While the nearby gene *NCOA7* modulates the activity of the estrogen receptor, *HEY2* contains estrogen receptor binding sites and is estrogen responsive.*WT1* is involved in the development of the urogenital system and is down-regulated in endometriotic stromal cells.Our work adds strength to previously associated loci and sheds light on the potential role of novel factors, including genes involved in endometrial biology, in the pathophysiology of endometriosis.

361

Genome-wide association analysis identifies 27 novel loci associated with uterine leiomyomas revealing common genetic origins with endometriosis. *N. Mäkinen[1], C.S. Gallagher[2], H.R. Harris[3], O. Uimari[4,5], K.L. Terry[6], D.I. Chasman[7], S. Missmer[7,8,9], K.T. Zondervan[4,10], C.C. Morton[1,11,12,13], the 23andMe Research team, the FibroGENE consortium.* 1) Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 2) Department of Genetics, Harvard Medical School, Boston, MA, USA; 3) Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; 4) Oxford Endometriosis CaRe Centre, Nuffield Department of Women's and Reproductive Health, University of Oxford, John Radcliffe Hospital, Oxford, UK; 5) Department of Obstetrics and Gynecology, Oulu University Hospital, Oulu, Finland, PEDEGO Research Unit, University of Oulu and Oulu University Hospital, Oulu, Finland, Medical Research Center Oulu, University of Oulu and Oulu University Hospital, Oulu, Finland; 6) Obstetrics and Gynecology Epidemiology Center, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 7) Division of Preventative Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 8) Division of Adolescent and Young Adult Medicine, Department of Medicine, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA; 9) Department of Obstetrics, Gynecology, and Reproductive Biology, College of Human Medicine, Michigan State University, Grand Rapids, MI, USA; 10) Big Data Institute, Li Ka Shing Center for Health for Health Information and Discovery, Oxford University, Oxford, UK; 11) Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 12) Broad Institute of MIT and Harvard, Cambridge, MA, USA; 13) School of Psychological Sciences, University of Manchester, Manchester, UK.

Uterine leiomyomas (UL), also known as uterine fibroids, are the most common tumors of the female reproductive system and primary cause for hysterectomy, leading to notable morbidity and high economic burden. Genetic epidemiologic studies indicate heritable factors influence the risk for developing UL. Previous genome-wide association studies (GWAS) have identified five loci associated with UL at genome-wide significance ($P < 5 \times 10^{-8}$). To expand considerably upon the existing data for UL, we performed GWAS meta-analysis of four population-based cohorts (Women's Genome Health Study, Northern Finland Birth Cohort, QIMR Berghofer Medical Research Institute, UK Biobank) and one direct-to-consumer cohort (23andMe) of white European ancestry, totaling 35,474 UL cases and 267,505 controls. Along with replicating three of the five previously reported loci, we identify genome-wide significant associations at 24 novel independent loci. Various identified loci harbor genes previously implicated in cell growth including well-known oncogenes and tumor suppressor genes: *PDGFRA*, *TERT*, *ESR1*, *WT1*, *ATM*, *FOXO1*, and *TP53*. Furthermore, four loci identified in the analysis have been associated with risk for endometriosis – another common gynecologic disorder: 1p36.12 (rs7412010, OR [95% CI] = 1.13 [1.11 – 1.16], $P = 2.43 \times 10^{-29}$), 2p25.1 (rs35417544, OR [95% CI] = 1.09 [1.07 – 1.10], $P = 2.32 \times 10^{-19}$), 6q25.2 (rs58415480, OR [95% CI] = 1.19 [1.17 – 1.22], $P = 1.86 \times 10^{-54}$), and 11p14.1 (rs11031006, OR [95% CI] = 1.10 [1.07 – 1.12], $P = 5.65 \times 10^{-15}$). Linkage disequilibrium is strong between UL and previously reported endometriosis lead SNPs at three of the loci. Each of the overlapping genomic loci contains a gene(s) known to be involved in progesterone or estrogen signaling: *WNT4* at 1p36.12, *GREB1* at 2p25.1, *ESR1* at 6q25.2, and *FSHB* at 11p14.1. To characterize further the potential overlap between UL and endometriosis, we performed a large-scale epidemiologic meta-analysis across 112,406 women from three population-based cohorts (Women's Health Study, Nurse's Health Study II, UK Biobank). Results implicate history of laparoscopically confirmed endometriosis to be associated with elevated risk for developing UL (age-adjusted, OR [95% CI] = 3.34 [1.01 – 10.79]; multivariate-adjusted, OR [95% CI] = 3.48 [1.43 – 8.46]). These findings increase our understanding of the biological mechanisms underlying UL development, and suggest overlapping genetic origins with endometriosis.

Copyright © 2018 American Society of Human Genetics. All rights reserved

362

Exome sequencing of a Mayer-Rokitansky-Kuster-Hauser cohort reveals novel candidate genes and significant mutational burden. *A. Jolly[1], Z. Coban Akdemir[1], S. N. Jhangiani[2], D. M. Muzny[2], A. Koch[3], J. E. Dietrich[4,5], I. B. Van den Veyver[1,4], S. Brucker[3], R. A. Gibbs[1,2], J. R. Lupski[1,2,5], J. E. Posey[1].*
1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 3) University of Tübingen, Department of Obstetrics and Gynecology, Tübingen, Germany; 4) Department of Obstetrics and Gynecology, Baylor College of Medicine and Texas Children's Hospital, Houston, Texas; 5) Department of Pediatrics, Baylor College of Medicine, Houston, Texas.

Background: Mayer-Rokitansky-Kuster-Hauser (MRKH) syndrome is a developmental disorder of the female reproductive tract characterized by a spectrum of hypoplasia or agenesis of the uterus, cervix, and upper two-thirds of the vagina. Clustering of MRKH within families suggests heritability of this sex limited-trait, with an autosomal dominant mode of inheritance, albeit with variable penetrance and expressivity. Methods: Peripheral blood samples were collected from 71 probands with apparently sporadic MRKH type I or II, and extracted DNA sequenced by whole exome sequencing (WES). Exome variant analysis of both single nucleotide variants (SNVs) and copy number variants (CNVs) was performed to identify rare, potentially damaging variants in novel candidate MRKH genes, as well as previously-identified MRKH candidate and known disease genes. Results: Potential molecular diagnoses were identified in 57/71 (80%) of probands, and included 15 novel candidate MRKH genes, 11 previously identified MRKH candidate genes, and 13 known uterine malformation genes. In 35/57 (61%) of probands with a potential molecular diagnosis, multilocus variation was identified. Candidacy of each novel candidate MRKH gene was supported by the identification of rare variants in at least four unrelated probands, and includes 4 genes (*APC*, *SCRIB*, *TLN2,* and *HIVEP3*) associated with Wnt signaling in the urogenital and skeletal systems, and 3 genes (*SBF1, RGS22,* and *TPST2*) previously implicated in male fertility, spermiogenesis, or aberrant fertilization. Discussion: In the present cohort of 71 unrelated probands, we demonstrate a potential molecular diagnostic rate of 80%. Only one gene has previously been firmly associated with MRKH, *WNT4*, and earlier studies targeting a small number of gene candidates failed to show causative variation in a large number of patients. These findings support the hypothesis of genetic heterogeneity in MRKH and lend to the identification of a large number of previous candidate genes and novel gene candidates in our cohort, which represents the largest MRKH cohort sequenced by WES to our knowledge. Candidate novel MRKH genes underscore the importance of Wnt signaling in urogenital development and also suggest a possible role for paternal effect genes in the pathogenesis of MRKH. The high rate of multiple molecular diagnoses in this cohort supports multilocus variation in the etiology of apparently sporadic MRKH.

363

Genetic study of longitudinal pubertal height growth describes links with adult health. *D. Cousminer, Early Growth Genetics (EGG) Consortium.*
Children's Hospital of Philadelphia, Philadelphia, PA.

Distinct growth patterns during puberty correlate with adverse health outcomes such as poor cardiometabolic health; however, the genetic mechanisms mediating differences in growth trajectories remain largely unknown. We leveraged ~39,300 trans-ethnic samples from 20 cohorts with repeated height measurements across childhood. Longitudinal height was modeled using Super-Imposition by Translation And Rotation (SITAR) growth curve analysis for three parameters determined for each individual's growth curve: ***a-size***, representing taller or shorter than the mean; ***b-timing***, representing timing of the growth spurt earlier or later than the mean; and ***c-velocity***, which is the tempo of the pubertal growth spurt, a property of pubertal growth that has not previously been subjected to genetic analyses. We performed Haplotype Reference Consortium imputation followed by genome-wide association meta-analysis with GWAMA. Independent loci detected with GCTA-COJO were fine-mapped using trans-ethnic data with MR-MEGA, followed by credible set analysis. We observed six genome-wide significant loci. All were previously reported for related traits, including body size from infancy to adulthood, adiposity, and age at menarche. We also found additional, novel independent signals at these same loci. Next, we used LD score regression to investigate genetic correlations with traits and diseases. ***a-size*** strongly correlated with body size traits across the life-course and health traits such as coronary artery disease. ***b-timing*** was highly correlated with pubertal timing and body fat/BMI traits, plus adult fasting insulin and 2hr glucose adjusted for BMI. Interestingly, ***c-velocity*** correlated with puberty timing, body size (height more than adiposity) throughout life and measures of adult health, including glycemic traits (insulin, HOMA-IR), metabolites (HDL, VLDL concentrations), bone (femoral neck bone mineral density), lung function and lung cancer. Using UK Biobank data on >300,000 individuals, two-sample Mendelian randomization analyses inferred causal effects of pubertal growth on some of these outcomes. Combined, these findings suggest that the tempo of pubertal development (***c-velocity***), often challenging to assess and regularly overlooked in epidemiological studies, may provide insight into adult health outcomes. Additionally, our results should help in the identification of specific growth trajectories impacting lifelong health.

Copyright © 2018 American Society of Human Genetics. All rights reserved

364

Identifying novel longevity-associated variants from >90,000 whole-exome sequences of the DiscovEHR cohort. *P. Sin-Chan[1], A.H. Li[1], C. Gao[1], C. O'Dushlaine[1], J.G. Reid[1], J.D. Overton[1], D.H. Ledbetter[2], D.J. Carey[2], A. Baras[1], A.N. Economides[1], A.R. Shuldiner[1].* 1) Regeneron Genetics Center, Regeneron Pharmaceuticals, Inc, Tarrytown, NY; 2) Geisinger Health System, Danville, PA.

Aging is a complex process characterized by the degeneration in cellular functions leading to organismal dysfunction and increased susceptibility to disease and death. While our understanding of aging mechanisms has improved substantially, most studies have focused on animal models with extremely short lifespans. Large-scale, high-throughput sequencing studies of human aging are still at their infancy. The Regeneron Genetics Center and Geisinger Health System established the DiscovEHR study, which has since completed the sequencing of >90,000 whole exomes from patients with linked electronic health records [1-2]. We performed exome-wide association studies using age as the trait of interest (Age-ExWAS) and adjusted for Sex and principal components to identify variants related to human lifespan. Linear regression analysis (n= 87,400 individuals of European descent) using an additive model, in which 1,087,242 exonic variants were analyzed, revealed 30 SNPs that passed genome wide significance (p<5E-8). Intriguingly, Age-ExWAS analysis identified an enrichment of genes involved in immune response and lipid metabolism, such as the rs429358 *APOE* missense mutation. Notably, clonal hematopoiesis of indeterminate potential (CHIP) variants, which are associated with myelodysplastic syndromes (MDS), were highly enriched in our analysis and suggestive of somatic mutation accumulation in hematopoietic cells as a consequence of aging despite patients lacking MDS clinical symptoms. Binary trait association analysis in cases (>85 years, n=4,268) versus controls (40-70 years, n=47,927) indicate CHIP variants are more common at older ages, in contrast to known longevity-associated variants. Finally, sex-stratified analysis in 52,871 females and 34,519 males revealed a number of genetic variants that appear sex-specific, including *APOE*, *HLA-DQA1*, *HLA-DRB5*, *TP53* and *POT1*. Taken together, our analysis identified potential age-associated exonic variants in DiscovEHR, which will inform follow-up functional experiments and may yield insights into the biology of aging, with the goal of promoting healthy human aging. [1] Dewey et al. N Engl J Med. 2016 [2] Dewey et al. Science. 2016.

365

Large-scale genome-wide discovery and phenome-wide association analyses of genetic differences in leukocyte telomere length. *C. Li[1], L.A. Lotta[1], T. Loe[2], V. Codd[3], J. Tao[4], R.A. Scott[1], I.D. Stewart[1], N.D. Kerrison[1], F.R. Day[1], J. Luan[1], J.H. Zhao[1], C.P. Nelson[3,5], K. Ong[1], G. Matullo[6], J. Danesh[4,7], A. Butterworth[4], N. Samani[3], E.L. Denchi[2], N.J. Wareham[1], C. Langenberg[1], ENGAGE Consortium Telomere Group, EPIC-CVD Consortium, EPIC-InterAct Consortium.* 1) MRC Epidemiology Unit, University of Cambridge, Cambridge, UK; 2) Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA; 3) Department of Cardiovascular Sciences, University of Leicester, Leicester, UK; 4) MRC/BHF Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge, UK; 5) NIHR Leicester Cardiovascular Biomedical Research Unit, Glenfield Hospital, Leicester, UK; 6) Italian Institute for Genomic Medicine (IIGM), Turin, Italy; 7) Wellcome Trust Sanger institute, Hinxton, UK.

Introduction: Leukocyte telomere length (LTL) is a heritable biomarker of genomic ageing. Genetic determinants of LTL in the general population and phenotypic consequences of genetically-determined differences in LTL remain incompletely understood. *Methods*: We performed a genome-wide meta-analysis of LTL in up to 69,378 individuals by pooling densely-genotyped and imputed association results from the InterAct and EPIC-CVD studies (N=31,694) with genome-wide summary statistics from Codd et al. Conditional analyses, fine-mapping, multi-omic data integration and *in vitro* assays were used to prioritise likely-causal genes at identified regions. Associations of genetically-longer leukocyte telomeres with a broad spectrum of clinical diagnoses were estimated in UK Biobank and other large-scale cohorts with genetic and clinical outcome data. *Results*: We identified five genomic regions not previously known to be associated with LTL in or near *PARP1*, *POT1*, *TERF2*, *MPHOSPH6* and *PRRC2A*/*HLA*, and replicated 6 out of 7 previously reported regions for a total of 11 loci at p-value<$5x10^{-8}$. Independent lead variants explained over a fifth of the chip-based heritability (h^2=4.6%) of LTL. Causal-gene prioritisation analyses suggested 28 candidate genes at the 10 non-HLA loci, including 11 additional candidate genes at existing loci. Genetically-longer leukocyte telomeres were associated with lower risk for coronary artery disease (OR [95%CI] = 0.84 [0.77-0.91]), as previously suggested, but higher levels of established cardiovascular risk factors. Phenome-wide analyses in >350,000 UK Biobank participants linked genetically-longer LTL to significantly greater predisposition for a range of proliferative conditions at Bonferroni corrected p-values, specifically malignant but also non-malignant neoplasms. *Conclusions*: Our findings substantially expand current knowledge on the genetic determinants of LTL and their impact on human diseases and cancer development.

Copyright © 2018 American Society of Human Genetics. All rights reserved

366

Genomic underpinnings of lifespan allow prediction and reveal basis in modern risks. *P.R.H.J. Timmers[1], N. Mounier[2,3], K. Lall[4,5], K. Fischer[4], Z. Ning[6], X. Feng[7], A. Bretherick[8], D.W. Clark[1], X. Shen[1,6], T. Esko[4,9], Z. Kutalik[2,3], J.F. Wilson[1,8], P.K. Joshi[1,2].* 1) Centre for Global Health Research, Usher Institute of Population Health Sciences, Edinburgh, United Kingdom; 2) Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne, Switzerland; 3) Swiss Institute of Bioinformatics, Lausanne, Switzerland; 4) Estonian Genome Center, University of Tartu, Tartu, Estonia; 5) Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia; 6) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; 7) State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources, Key Laboratory of Biodiversity Dynamics and Conservation of Guangdong Higher Education Institutes, School of Life Sciences, Sun Yat-sen University, Guangzhou, China; 8) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, United Kingdom; 9) Broad Institute of Harvard and MIT, Cambridge, MA, USA.

Living long and healthy lives is of great interest to us all, yet investigation into the genomic basis of lifespan has been hampered by limited sample sizes, both in terms of gene discovery and identification of longevity pathways. Applying univariate, multivariate, and risk factor-informed genome-wide association to 1,012,240 parental lifespans from European subjects in UK Biobank and an independent replication cohort, we validate previous associations near *CDKN2B-AS1*, *ATXN2/BRAP*, *FURIN/FES*, *FOXO3A*, 5q33.3/*EBF1*, *ZW10*, *PSORS1C3*, 13q21.31, and provide evidence against associations near *CLU*, *CHRNA4*, *PROX2*, and *d3-GHR*. Our combined dataset reveals 21 further loci and shows, using gene set and tissue-specific analyses, that genes expressed in foetal brain cells and adult dorsolateral prefrontal cortex are enriched for genetic variation affecting lifespan, as are gene pathways involving lipoproteins, lipid homeostasis, vesicle-mediated transport, and synaptic function. We next perform a lookup of disease SNPs and find variants linked to dementia, smoking/lung cancer, and cardiovascular risk explain the largest amount of variation in lifespan. This, and the notable absence of cancer susceptibility SNPs (other than lung cancer) among the top lifespan variants, suggests larger, more common genetic effects on lifespan reflect modern lifestyle-based susceptibilities. Finally, we create polygenic scores for survival in independent sub-cohorts and partition populations, using DNA information alone, into deciles of expectation of life with a difference of more than five years from top to bottom decile.

367

Mitochondrial variants influence human complex diseases and molecular endophenotypes. *E. Yonova-Doing[1], C. Calabrese[2], N. Cai[3,4], I.D. Stewart[5], W. Wei[2], S. Karthikeyan[1], W.J. Astle[1], B. Prins[1], J. Peters[1,6], T. Jiang[1], P. Surendran[1], O. Stegle[3], T. Bolton[1,7], C. Langenberg[5], A. Wood[1], A.S. Butterworth[1,7], J. Danesh[1,4,6,7], N. Soranzo[4,6,8], P.F. Chinnery[2,9], J.M.M. Howson[1].* 1) British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; 2) MRC Mitochondrial Biology Unit, Cambridge Biomedical Campus, Cambridge, United Kingdom; 3) European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom; 4) Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 5) MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom; 6) Cambridge Substantive Site, Health Data Research UK, Wellcome Genome Campus, Hinxton, United Kingdom; 7) National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, United Kingdom; 8) Department of Haematology, University of Cambridge, Cambridge, United Kingdom; 9) Department of Clinical Neurosciences, Cambridge Biomedical Campus, University of Cambridge, Cambridge, United Kingdom.

Mitochondria are responsible for cellular energy production via the oxidative phosphorylation cycle. Perturbations in this cycle can lead to mitochondrial dysfunction, oxidative stress, apoptosis and various age-related diseases. Despite the pivotal role of mitochondria in human health and disease, the role of mitochondrial DNA (mtDNA) variants is largely unexplored in well-powered studies. We have addressed this using the UK Biobank (up to N=358,196) and INTERVAL (up to N=40,520) studies to test associations between mtDNA variants and over 4,000 phenotypes: diseases, anthropometric traits, physical fitness, longevity, metabolites, proteins and blood cell phenotypes. We performed stringent data quality control including manual genotype re-calling (as algorithms are not optimised for mtDNA variants). mtDNA variants were imputed to the NCBI reference panel. We performed single-variant analyses using linear or linear mixed models (EPACTS or RVTESTS), aggregate ("gene-based") tests and Bayesian joint analyses of genetic variants against hierarchical ICD10 diagnosis data (TreeWAS). Variants were considered statistically significant if they had $P<5\times10^{-8}$ (mitochondrial genome-wide significance). Significant associations between mtDNA variants and plasma metabolites were replicated in an independent cohort (up to N~12,000). We found more than 100 trait-variant associations, including three (rs28660704, $P=2\times10^{-10}$; rs879051705, $P=4\times10^{-11}$; and rs9743, $P=2\times10^{-12}$) independent associations with N-formylmethionine (fMet), an amino acid with a key role in initiation of translation in mitochondria. These associations were independent of the effect of fMet-associated variants in the nuclear genome. We found associations with sphingomyelins and accumulation of this class of lipid has been shown to cause mitochondrial dysregulation and cell death. As well as confirming previous associations between D-loop variants and longevity, we also found novel associations with blood cell traits, airway function, and nuclear-encoded mitochondrial proteins. Moreover, the same mtDNA variants were associated with multiple traits, which is suggestive of pleiotropic effects of mtDNA variants. The results suggest important synergistic roles for the nuclear and mitochondrial genomes beyond cellular energy production that could underlie some of the biological mechanisms responsible for age-related disorders.

Copyright © 2018 American Society of Human Genetics. All rights reserved

368

Mitochondrial DNA heteroplasmy is associated with overall mortality. *R.J. Longchamps[1], Y.S. Hong[2], C.E. Newcomb[1], J.A. Sumpter[1], M.L. Grove[3], J.D. Walston[4], B.G. Windham[5], J. Coresh[4], E. Boerwinkle[3], E. Guallar[2], D.E. Arking[1].* 1) Mckusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) Departments of Epidemiology and Medicine, and Welch Center for Prevention, Epidemiology, and Clinical Research, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD; 3) School of Public Health, Human Genetics Center, The University of Texas Health Science Center at Houston, Houston, TX; 4) Department of Epidemiology, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD; 5) University of Mississippi Medical Center Department of Medicine and Center of Biostatistics.

Several biological processes have been hypothesized to explain the critical role of mitochondrial dysfunction in disease, such as declines in energy production, altered rates of apoptosis, and elevated free radical production. These processes may be exacerbated by the accumulation of mitochondrial DNA (mtDNA) mutations leading to increased levels of heteroplasmy - the presence of multiple distinct mtDNA genomes within an individual. While previous reports have shown heteroplasmy levels increase with age, little is known about the specific role of heteroplasmy in human disease and mortality. We hypothesized higher levels of heteroplasmy would be associated with overall mortality. Heteroplasmy was measured from whole genome sequence data from 3,658 individuals of the Atherosclerosis Risk in Communities cohort using the mtDNA-server analysis pipeline. Individuals were removed if < 10,000 of the 16,569 mitochondrial bases achieved 250X coverage and haplogroup analysis indicated contamination. Heteroplasmies were identified with the following criteria: 1) Coverage ≥ 250X; 2) Minor allele frequency (MAF) ≥ 5%; and 3) MAF call differs < 4% between strands. Sites near known indel, transition, and transversion artifacts were excluded. After QC filtering we observed 1,219 heteroplasmies in 3,192 individuals. During 61,520 person-years of follow up we observed 1,341 deaths. To assess the association of heteroplasmic mtDNA with overall mortality, we performed a Cox proportional-hazards model adjusting for age, sex, DNA collection site, smoking status, body mass index, systolic blood pressure, low-density lipoprotein, history of myocardial infarction and type 2 diabetes status. We observed a hazard ratio (HR) of 1.25 (95% CI 1.11 - 1.40; P = 0.0002) for heteroplasmic mtDNA carriers. To enrich for biologically relevant variants, we investigated the added effect of variants which cause deleterious mutations as defined by scaled CADD scores > 15. Although not statistically significant, the 145 Individuals with these variants showed greater mortality with a HR of 1.14 (95% CI 0.89 - 1.45; P = 0.30). Furthermore, the 466 individuals with recurrent variants had similar rates of mortality compared to individuals with singletons indicating recurrent sites are similarly tolerated (HR = 1.07; 95% CI 0.88 - 1.30; P = 0.50). Together, our findings highlight heteroplasmy as a strong predictor of overall mortality which warrants further investigation into disease-specific mortality.

369

Directly measuring the dynamics of the human mutation rate by sequencing large, multi-generational pedigrees. *T. Sasani, B. Pedersen, A. Quinlan, M. Leppert, L. Baird, L. Jorde.* University of Utah, Salt Lake City, UT.

Developing an accurate estimate of the human germline mutation rate is critical to our understanding of evolution, demography, and genetic disease. Early phylogenetic analyses inferred mutation rates from the observed sequence divergence between humans and related primate species at particular genes and pseudogenes. However, as whole genome sequencing has become ubiquitous, these estimates have been refined using pedigree-based approaches. By identifying mutations present in offspring that are absent from their parents (*de novo* mutations), it is possible to more accurately approximate the human germline mutation rate.To obtain a precise, unbiased estimate of the mutation rate in humans, we performed deep whole-genome sequencing on blood-derived DNA from 34 of the original three-generation CEPH families from Utah, comprising a total of 603 individuals. These families, which each contain grandparents (P0 generation), parents (F1), and their children (F2), are considerably larger than any used in prior estimates of the human mutation rate, and offer unique power to detect and validate *de novo* mutation. With a median of 8 F2 individuals per pedigree, we were able to biologically validate putative *de novo* mutations in the F1 generation by assessing their transmission to a third generation. Using this dataset, we have generated a high-confidence estimate of the human mutation rate, observe significant parental age effects on the rate of *de novo* mutation, and identify wide variability in family-specific age effects across CEPH pedigrees. To our knowledge, this study represents the first example of a longitudinal analysis of the effect of parental age within individual families. Additionally, we have identified recurrent *de novo* variants present in multiple F2 offspring, which are likely the result of mosaicism in the parental germline.Finally, we have trained a classification model on the high-quality, transmitted *de novo* variants in our dataset, and used this model to identify *de novo* mutations in a large cohort of children from the Simons Foundation for Autism Research Initiative (SFARI). Combining the de novo mutations observed in 34 Utah families with the SFARI callset, we have generated a dense genomic map of spontaneous human mutation. We observe regional enrichment of *de novo* variation in the human genome, and explore the role of sequence context, as well as molecular processes like recombination and gene conversion, on the rate of human mutation.

Copyright © 2018 American Society of Human Genetics. All rights reserved

370

Genome-scale capture C promoter interaction analysis implicates novel effector genes at GWAS loci for bone mineral density. *A. Chesi[1], Y. Wagley[2], M.E. Johnson[1], M. Manduchi[1,3], C. Su[1], S. Lu[1], M.E. Leonard[1], K.M. Hodge[1], J.A. Pippin[1], K.D. Hankenson[2], A.D. Wells[1,4], S.F.A. Grant[1,5,6,7].* 1) Center for Spatial and Functional Genomics, Children's Hospital of Philadelphia, Philadelphia, PA; 2) Department of Orthopaedic Surgery, University of Michigan Medical School, Ann Arbor, MI; 3) Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 4) Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 5) Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 6) Divisions of Genetics and Endocrinology, Children's Hospital of Philadelphia, Philadelphia, United States; 7) Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA.

Osteoporosis is a devastating disease with an essential genetic component. Genome wide association studies (GWAS) in both children and adults have discovered genetic variants robustly associated with bone mineral density (BMD); however they only report genomic signals and not necessarily the precise localization of culprit effector genes. Therefore, we sought to carry out physical and direct ‘variant to gene mapping’ in a relevant primary human cell type. Given the notable paucity of genomic data available on bone in the public domain, and to improve upon the low resolution of typical Hi-C approaches, we developed a massively parallel, high-resolution Capture-C based method, SPATIaL-seq (genome-Scale, Promoter-focused Analysis of chromaTIn Looping), to characterize the genome-wide interactions of all human promoters in human mesenchymal progenitor cell-derived osteoblasts - facilitated by a custom Agilent SureSelect RNA library targeting *DpnII* restriction fragments overlapping 36,691 promoters of protein-coding, noncoding, antisense, snRNA, miRNA, snoRNA and lincRNA genes. We also generated ATAC-seq open chromatin maps from the same MSC-derived osteoblast samples to determine informative proxy SNPs residing in open chromatin for each of the 110 independent BMD GWAS signals at 107 candidate loci. By intersecting our SPATIaL-seq and ATAC-seq data, we observed consistent contacts between candidate causal variants and putative target gene promoters in open chromatin for ~30% of the loci investigated. In order to validate our findings, we targeted the expression of implicated genes using siRNA at two key loci in primary human MSC-derived osteoblasts from multiple donors: *CPED1* and *ING3* at the ‘*WNT16-CPED1*’ locus, and *EPDR1* and *SFRP4* at the ‘*STARD3NL*’ locus. Knockdown of either *ING3* or *EPDR1* yielded pronounced inhibitory effects on osteoblastogenesis, as assessed by staining for alkaline phosphatase and Alizarin red S. In summary, knockdown of two novel genes not previously associated with BMD but implicated by our combined ATAC-seq and SPATIaL-seq approach revealed strong effects on osteoblast differentiation (decreased ALP expression and absence of calcium phosphate mineral deposition), suggesting an important role for *ING3* and *EPDR1* in bone biology. Our approach therefore aids target discovery in osteoporosis and can be applied to other common genetic diseases.

371

Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *K. Lehmann[1,2,12,14], A. Kahles[1,2,12,14], N. Toussaint[9,14], M. Hüser[1,12,14], S. Stark[1,2,12,14], T. Sachsenberg[5], O. Stegle[4], O. Kohlbacher[5,6,7,8,9], C. Sander[10,11], G. Rätsch[1,2,12,13,14], The Cancer Genome Atlas Research Network.* 1) ETH Zurich, Zurich, Zurich, Switzerland; 2) Memorial Sloan Kettering Cancer Center, Computational Biology Department, New York, USA; 3) ETH Zurich, NEXUS Personalized Health Technologies, Zurich, Switzerland; 4) European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK; 5) University of Tübingen, Department of Computer Science, Tübingen, Germany; 6) Center for Bioinformatics, University of Tübingen, Tübingen, Germany; 7) Quantitative Biology Center, University of Tübingen, Germany; 8) Biomolecular Interactions, Max Planck Institute for Developmental Biology, Tübingen, Germany; 9) Institute for Translational Bioinformatics, University Medical Center, Tübingen, Germany; 10) Dana-Farber Cancer Institute, cBio Center, Department of Biostatistics and Computational Biology, Boston, MA, USA; 11) Harvard Medical School, CompBio Collaboratory, Department of Cell Biology, Boston, USA; 12) University Hospital Zurich, Biomedical Informatics Research, Zurich, Switzerland; 13) ETH Zurich, Department of Biology, Zurich, Switzerland; 14) SIB Swiss Institute of Bioinformatics, Zurich, Switzerland.

We analyze RNA and whole exome sequencing data of tumors from 8,705 donors spanning a range of 32 cancer types. Our study focuses on the analysis of specific alternative splicing (AS) events involving a small number of exons from RNA-seq data. We report results from comprehensive analyses of a) the underlying genetic changes leading to splicing variability in tumors, b) quantitative and qualitative changes of AS in tumors, and c) the extent to which splicing aberrations can be exploited for immunotherapy. To our knowledge, this study presents the first comprehensive analysis of known as well as novel alternative splicing events across all suitable samples of The Cancer Genome Atlas (TCGA). Depending on cancer type, we find an up to 40% increase in AS in tumor samples relative to normal and find on average more than 900 tumor-specific exon-exon-junction.We uncover strong splicing signatures for individual cancer types and subtypes, often overpowering the respective tissue-specific effects. Further, we combine the splicing phenotypes with variants obtained from exome sequencing data for a genome-wide splicing association analysis. This is the largest reported splicing quantitative trait loci (sQTL) study with respect to number of donors thus far. Here, we focus on variants that have been shown to occur as somatic in some individuals but may also occur in the germline genome in others. For the first time, the available data provides sufficient statistical power to detect trans-sQTL that were difficult to detect before. Aside from confirming known sQTL-variants in splicing factors U2AF1 and SF3B1, we also detect novel trans-sQTL in IDH1, TADA1 and PPP2R1A.Finally, our study is the first to comprehensively estimate to which extent AS in tumors leads to new RNA transcripts that are translated into tumor-specific peptides, that are potential targets for immunotherapy. Integrating data from TCGA and GTEx, we identify tumor-specific events. We use CPTAC protein mass spectra of two tumor types to show that the resulting mRNAs are indeed translated into tumor-specific peptides. From all peptides that are predicted to be MHC-I binders, we are able to confirm on average 1.7 splicing derived neo-epitopes per sample – an almost 3-fold increase over the number of neo-epitopes predicted with classic approaches taking into account only SNVs (≈0.6). To our knowledge this is the first comprehensive analysis of this type. .

Copyright © 2018 American Society of Human Genetics. All rights reserved

372

Uganda genomes resource enables inferences into population history and genomic discovery in Africa. *D. Gurdasani[1,2], T. Carstensen[1,2], S. Fatumo[1,2,3], G. Chen[4], CS. Franklin[1], J. Prado-Martinez[1], H. Bouman[1], Uganda Genomes Resource Investigators.* 1) Human Genetics, Wellcome Sanger Institute, Cambridge, Cambridgeshire, United Kingdom; 2) Department of Medicine, University of Cambridge, Cambridge, UK; 3) H3Africa Bioinformatics Network (H3ABioNet) Node, National Biotechnology Development Agency, Federal Ministry of Science and Technology, Abuja, Nigeria; 4) Center for Research on Genomics and Global Health, National Institute of Health, USA.

Genomic studies in African populations provide unique opportunities to understand disease aetiology, human genetic diversity and population history in a regional and a global context. In the largest study of its kind to date, comprising genome-wide data from 6,400 individuals from rural Uganda, and including whole-genome sequence from 1,978 individuals, we find evidence of geographically correlated fine-scale population substructure, as well as complex admixture from eastern African hunter-gatherer and Eurasian populations. We highlight the value of the largest sequence panel from Africa to date as a global resource for population genetics, imputation and understanding the mutational spectrum and its clinical relevance in African populations. Examining 34 cardiometabolic traits, we demonstrate systematic differences in trait heritability between European and African populations, probably reflecting the differential impact of genetic and environmental factors on traits. In the first multi-trait pan-African GWAS of up to 14,126 individuals, we identify 10 novel loci associated with anthropometric, haematological, lipid and glycemic traits. Our findings suggest that several functionally important signals at known and novel loci may be driven by differentiated variants within and specific to Africa, highlighting the utility of inclusion of diverse study populations in African GWAS. We provide a rich new genomic and phenotypic resource for researchers in Africa and globally.

3568

An atlas of polygenic risk score associations to uncover putative causal relationships across the human phenome. *S. Harrison, T.G. Richardson, G. Davey Smith.* Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, Bristol, United Kingdom.

The age of large-scale genome-wide association studies (GWAS) has provided us with an unprecedented opportunity to evaluate genetic predisposition to complex disease using polygenic risk scores (PRS). Along with helping predict lifelong risk of disease, PRS can be harnessed as unconfounded instrumental variables within a Mendelian randomization (MR) framework to help unravel causal relationships between modifiable risk factors and complex traits. In this study we have analysed 165 PRS derived from GWAS and 553 heritable traits in up to 334,970 individuals from the UK Biobank study. For selected PRS, we observed a several fold increase in the odds of developing disease when comparing the highest and lowest deciles. A web application (available at http://mrcieu.mrsoftware.org/PRS/) has been developed to query and visualise findings, which we envisage will help confirm known associations between risk factors and complex traits, as well as elucidate novel findings. We demonstrate the value of this resource by undertaking several in-depth evaluations of associations using MR methodology. Firstly, we investigate observed effects from a hypothesis-free scan of type 2 diabetes genetic liability on all 553 complex traits. Along with expected findings, we identified associations with reduced lung function (Beta=-0.011, 95% CI=-0.014 to -0.009) and birth weight (Beta=-0.016, 95% CI=-0.020 to -0.012) which were robust to follow-up analyses and validated using independent populations. We have also undertaken analyses to investigate bi-directional associations. For example, we observed much stronger evidence that hypertension influences chronic kidney disease risk (OR=2.724, 95% CI=2.516 to 2.932), rather than the converse relationship (OR=1.014, 95% CI=1.005 to 1.024). Finally, we apply mediation and multivariable MR frameworks to evaluate the effect of multiple risk factors on outcomes. For instance, the body mass index and triglyceride PRS were associated with gout risk (OR=1.091, 95% CI=1.062 to 1.120 & OR=1.135, 95% CI=1.106 to 1.164 respectively). However, our analyses indicate that these effects are mediated by increased uric acid levels along the causal pathway to gout risk. Our atlas should prove valuable for future studies which aim to unravel causal relationships between complex traits. Findings from these endeavours will help improve our capability to prevent and treat disease.

Copyright © 2018 American Society of Human Genetics. All rights reserved

3569

Re-identification of genomic datasets using long-range familial searches. *Y. Erlich[1], T. Shor[1], I. Pe'er[2], S. Carmi[3].* 1) Columbia University/MyHeritage, Tel Aviv, Israel; 2) Columbia University; 3) Hebrew Univresity.

Consumer genomics databases have reached the scale of millions of individuals. Recently, law enforcement authorities have exploited some of these open databases to identify suspects via distant familial relatives. The most notable case is the Golden State Killer and since April 2018 the media reported another 12 cases that were solved using these open databases. Here, we wondered about the power of this technique and its implications to genetic privacy. To this end, we examined genomic data of 1.28 million individuals tested with DTC providers. Our empirical results show that over 60% of the searches after individuals of European-descent US adults will return a third cousin or a closer relative, which can permit their identification using simple demographic identifiers. We also developed a theoretical model that returns the probability of a match given the size of the database. After validation of the model with empirical data, we found that a genetic database of 3 million European-descent US adults will return a 3rd cousin match to virtually anyone in this population. These sizes are within reach to some of the open consumer genomics databases in the near future. To better understand the risk to human subjects, we conducted a long-range familial search on a 1000 Genomes Project sample. We selected a female sample from the CEU cohort in Utah, whose husband was identified by us in the past using surname inference. We extracted her genome from the (publicly available) 1000Genomes data repository, re-formatted her genotypes to resemble a file released by DTC providers, and uploaded the genotypes to one of these open databases. After a day of genealogical work, we were able to trace her identity, which was the same person we have previously re-identified based on surname inference. We posit that long range familial searches require a reevaluation of the status quo regarding the identifiability of DNA data. The Revised Common Rule, which will regulate federally funded human subject research from January 2019, does not define genome-wide genetic datasets as identifiable information. Given the growing success of long-range familial searches, we encourage HHS to use their authority and consider genomic information as identifiable in order to further protect participants of genomic research. We also propose a technical measure using cryptographic signatures that can mitigate some of the risks and restore control to data custodians. .

3570

Patient-customized oligonucleotide therapy for an ultra-rare genetic disease. *T.W. Yu[1], C. Hu[1], J. Kim[1,2], A. Larson[5], A. Lee[1,2,3,4], L. Black[6], C.M. El Achkar[1,2], A. Soucy[1], J. Vaze[1], M. Armant[1], N.R. Belur[7], A. Kuniholm[1], J. Douville[6], E. Augustine[8], M. Pendergast[9], S. Goldkind[10], K. Tyndall[11], B. Goodlett[1,2], S. Waisbren[1,2], B. Riley[1,2], L. Cornelissen[1,2], L. Pereira[1,2], C. Reed[12], R. Snyder[13], A. Patterson[1], A. Poduri[1,2,3], J. Mazzulli[6], A. Biffi[1,2], O. Bodamer[1,2,4], C. Berde[1,2].* 1) Boston Children's Hospital, Boston, MA; 2) Harvard Medical School, Boston, MA; 3) Broad Institute of MIT and Harvard, Cambridge, MA; 4) Manton Center for Orphan Disease, Boston Children's Hospital, Boston; 5) University of Colorado School of Medicine, Aurora, CO; 6) Charles River Laboratories, Wilmington, MA and Montreal, Canada; 7) Northwestern University Feinberg School of Medicine, Chicago, IL; 8) University of Rochester Medical Center, Rochester, NY; 9) Pendergast Consulting, Washington, DC; 10) Goldkind Consulting, Potomac, DC; 11) Tyndall Consulting, Raleigh, NC; 12) Brain Hz Consulting, Del Mar, CA; 13) Brammer Bio, Alachua, FL.

Next generation sequencing has revolutionized the diagnosis of rare genetic diseases. However, many patients still suffer from a lack of therapeutic options for most of these conditions, which in aggregate impact tens of millions of individuals globally. Here, we demonstrate a dramatically new pathway for the treatment of even ultra-rare genetic diseases.A six year old girl developed progressive blindness, epilepsy, and neurocognitive regression. Whole genome sequencing and RNA-seq revealed a maternally inherited retrotransposon inserted into an intron of *MFSD8*/*CLN7*, a key lysosomal gene. The insertion was found to cause exon trapping, leading to gene inactivation. This mutation, in combination with a paternal missense mutation in the same gene, caused Batten Disease, a rare, recessive disorder of neuronal lysosomal storage.No treatments exist for *CLN7* Batten disease. Unchecked, it is rapidly progressive and ultimately fatal. We therefore set out to rapidly develop a novel antisense oligonucleotide drug, customized to her mutation. *In vitro* administration of this oligonucleotide to patient-derived cells relieved the exon trap, and reversed lysosomal dysfunction as well.In communication with the FDA, we arranged the manufacturing and formulation of our oligonucleotide drug, which we named milasen. In January of this year – only one year from first patient contact – we launched a single patient treatment trial of milasen, under an Individual Patient Expanded Access IND. In this ongoing trial, milasen treatment appears to have arrested further clinical deterioration, and has been associated with striking reductions in seizure intensity and frequency. No safety or tolerability issues have arisen.This study illustrates an end-to-end pathway from genomic diagnosis to precision therapy, and offers a possible template for future treatments of even ultra-rare genetic diseases, in a safe and timely fashion.

Copyright © 2018 American Society of Human Genetics. All rights reserved