American Society of Human Genetics 67th Annual Meeting
October 17–21, 2017 in Orlando, Florida

PLATFORM ABSTRACTS

Tuesday, October 17, 5:30-7:00 pm:			**Abstract #'s**
4	Featured Plenary Abstract Session I	South Hall B	#1-#4
Wednesday, October 18, 9:00-10:30 am, Concurrent Platform Session A:			
6	Genetics of Vascular, Valvular, and Syndromic Disorders	Room 220B	#5-#10
7	Modeling Mega-cohorts: Insights & Innovation	Room 220F	#11-#16
8	Genetics and Epigenetics in Mental Illness	Room 230C	#17-#22
9	Genome Structure and Function: The Contribution of Mutations to Human Genetic Diversity, Disease, and Evolution	Room 230G	#23-#28
10	Disease Gene Discovery Strategies	Room 310A	#29-#34
11	Therapeutic Advances in Mendelian Disease	Room 310C	#35-#40
12	Detection and Impact of Mosaicism in Human Disease	Room 320	#41-#46
13	Pleiotropism and Penetrance in Cancer-causing Genes	Room 330A	#47-#52
14	Landscape of Cancer: Bioinformatic Analyses	Room 330C	#53-#58
Wednesday, October 18, 11:00 am-12:30 pm, Concurrent Platform Session B:			
15	Screening Cancer Cohorts for Novel Germline Cancer Genes	Room 220B	#59-#64
16	Cancer Genetic Testing: Approaches, Barriers, and Psychosocial Impact	Room 220F	#65-#70
17	Advances in Association Analysis	Room 230C	#71-#76
18	Strategies for Variant Interpretation	Room 230G	#77-#82
19	From Association to Function for Cardiometabolic Traits	Room 310A	#83-#88
20	Reproductive Genetics: Detection, Treatment, and Natural History of Errors	Room 310C	#89-#94
21	Leveraging Human Knockouts to Understand Biology	Room 320	#95-#100
22	Detection and Interpretation of Structural Variation	Room 330A	#101-#106
23	Neurodevelopmental Disorders: Causes and Mechanisms	Room 330C	#107-#112
Thursday, October 19, 9:00-10:30 am, Concurrent Platform Session C:			
29	Gene Discovery in Skeletal Phenotypes	Room 220B	#113-#118
30	Repeats and Rearrangements: New Methods, Genes, and Mechanisms in Neurological Disease	Room 220F	#119-#124
31	Secondary and Incidental Findings from WES/WGS	Room 230C	#125-#130
32	Computational Methods for Causal Inference in Complex Traits	Room 230G	#131-#136
33	Microbiome, Variation, and Disease	Room 310A	#137-#142
34	Genetic Architecture of Neurological Traits	Room 310C	#143-#148
35	High Throughput Functional Analysis of Enhancers and Variants	Room 320	#149-#154
36	Genetic Architecture of Rare Variants Across Diseases	Room 330A	#155-#160
37	Non-coding Variation and Epigenetic Effects in Cancer	Room 330C	#161-#166
Thursday, October 19, 11:00 am-12:30 pm, Concurrent Platform Session D:			
38	Neuromuscular Disease	Room 220B	#167-#172
39	Advances in the Genetics of Autoimmune Disease	Room 220F	#173-#178
40	Defining High Risk in Cancer	Room 230C	#179-#184
41	Natural Selection on Human Phenotypes	Room 230G	#185-#190
42	Consumers and Health Care Providers: Perspectives of Genetic Technology	Room 310A	#191-#196
43	Gene Discovery and Functional Models of Intellectual Disability	Room 310C	#197-#202
44	Polygenic Risk Scores and Genetic Correlation in Complex Disease	Room 320	#203-#208

45	Single Cell Omics Technologies	Room 330A	#209-#214
46	Sequencing in Neonatal and Pediatric Disorders	Room 330C	#215-#220
Friday, October 20, 9:00-10:00 am, Concurrent Platform Session E:			
56	Genomic Testing: A Focus on Results	Room 220B	#221-#224
57	Exploring the Impact of Archaic Ancestry	Room 220F	#225-#228
58	Data Sharing to Improve Genomic Variant Interpretation	Room 230C	#229-#232
59	Congenital and Pediatric Heart Diseases	Room 230G	#233-#236
60	Transcriptomics in Complex Neurological/Neuropsychiatric Disease	Room 310A	#237-#240
61	Context Matters: Genes, Environment, and Sex (Part 1)	Room 310C	#241-#244
62	New Paradigms for Regulatory Variant Contribution to Disease Risk	Room 320	#245-#248
63	Gene Expression Studies of T2D	Room 330A	#249-#252
64	Measuring Effects of Genetic Variants with High-Throughput Assays	Room 330C	#253-#256
Friday, October 20, 10:15-11:15 am, Concurrent Platform Session F:			
65	DNA Methylation	Room 220B	#257-#260
66	Splicing in Complex Traits	Room 220F	#261-#264
67	Ocular Development and Disease	Room 230C	#265-#268
68	Blood Omics in Large Cohorts	Room 230G	#269-#272
69	Genetics of Addictive Behaviors	Room 310A	#273-#276
70	Context Matters: Genes, Environment, and Sex (Part 2)	Room 310C	#277-#280
71	Autism	Room 320	#281-#284
72	Clinical Genomics in Cancer	Room 330A	#285-#288
73	Transcriptomic Analysis of Genetic Variation and Disease	Room 330C	#289-#292
Friday, October 20, 5:30-7:00 pm:			
87	Featured Plenary Abstract Session II	South Hall B	#293-#296
Saturday, October 21, 8:30-9:30 am, Concurrent Platform Session G:			
88	Functional Analyses of Cancer Genes	Room 220B	#297-#300
89	Modelling Candidate Disease Variants in Cellular and Animal Models	Room 220F	#301-#304
90	Cerebral Palsy and Epilepsy	Room 230C	#305-#308
91	Diverse Approaches to the Genetics of T2D	Room 230G	#309-#312
92	Population-based Diagnostic Sequencing	Room 310A	#313-#316
93	Identification and Function of Enhancers	Room 310C	#317-#320
94	Genetic Associations for Behavioral Phenotypes	Room 320	#321-#324
95	Host-pathogen Interactions in the Genetics of the Immune System	Room 330A	#325-#328
96	Investigating the Role of Non-coding Variants in Disease	Room 330C	#329-#332
Saturday, October 21, 9:45-10:45 am, Concurrent Platform Session H:			
97	Transcriptome-wide Association Studies	Room 220B	#333-#336
98	Improved Interpretation of Missense Variants	Room 220F	#337-#340
99	Big Data Approaches in Support of Population Studies	Room 230C	#341-#344
100	Novel Genetic and Environmental Contributions to Cancer Risk	Room 230G	#345-#348
101	Neurological Disorders: Chromatin in the Spotlight	Room 310A	#349-#352
102	Enhancers and Human Disease	Room 310C	#353-#356
103	The Genetics of Obesity	Room 320	#357-#360
104	Advancing Drug Discovery by Genetic Analysis in Large Cohorts	Room 330A	#361-#364
105	Regulation of Gene Expression in Metabolic and Vascular Tissues	Room 330C	#365-#368
Saturday, October 21, 11:00 am-12:30 pm:			
106	Featured Plenary Abstract Session III	South Hall B	#369-#372

1

Novel loci associated with skin pigmentation identified in African populations. *N. Crawford[1], D. Kelly[1,2], M. Hansen[1], M. Holsbach Beltrame[1], S. Fan[1], S. Bowman[3,4], E. Jewett[5,6], A. Ranciaro[1], S. Thompson[1], S. Pfeifer[7], J. Jensen[7], S. Wata Mpoloka[9], G. Mokone[10], T. Nyambo[11], D. Wolde Meskel[12], G. Belay[12], H. Rothschild[13], Y. Zhou[14,15], M. Kovacs[16], M. Xu[16], E. Oceana[20], Y. Song[5,6,21,22,23], E. Eskin[24], K. Brown[16], M. Marks[3,4], S. Loftus[17], W. Pavan[17], M. Yeager[18,19], S. Chanock[24], S. Tishkoff[1,25].* 1) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; 2) Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, PA, 19104, USA; 3) Department of Pathology & Laboratory Medicine, Children's Hospital of Philadelphia Research Institute, Philadelphia, PA 19104; 4) Department of Pathology & Laboratory Medicine and Department of Physiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; 5) Department of EECS, University of California, Berkeley, CA, 94704; 6) Department of Statistics, University of California, Berkeley, CA, 94704; 7) School of Life Sciences, Arizona State University, Tempe, AZ 85287; 8) Department of Biology, Howard University, Washington D.C; 9) Department of Biological Sciences, University of Botswana, Gaborone Botswana; 10) Dept. of Biomedical Sciences, University of Botswana School of Medicine, Gaborone, Botswana; 11) Department of Biochemistry, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania; 12) Department of Biology, Addis Ababa University, Addis Ababa, Ethiopia; 13) Stem Cell Program, Division of Hematology/Oncology, Pediatric Hematology Program, Boston Children's Hospital and Dana Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA; 14) Harvard Stem Cell Institute, Harvard University, Cambridge, MA 02138, USA; 15) Stem Cell Program, Division of Hematology/Oncology, Pediatric Hematology Program, Boston Children's Hospital and Dana Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA; 16) Laboratory of Translational Genomics (LTG), Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute, National Institutes of Health, Bethesda MD 20892; 17) Genetic Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892; 18) Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD; 19) Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Frederick, MD; 20) Department of Molecular Pharmacology, Physiology and Biotechnology, Brown University, Providence, RI 02912; 21) Chan Zuckerberg Biohub, San Francisco, CA 94158; 22) Department of Biology, University of Pennsylvania, Philadelphia, PA; 23) Department of Mathematics, University of Pennsylvania, Philadelphia, PA; 24) Department of Computer Science, Department of Human Genetics, University of California, Los Angeles, 90095; 25) Department of Biology, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA.

Despite the wide range of variation in skin pigmentation in Africans, little is known about its genetic basis. To investigate this question we performed a GWAS on pigmentation in 1,593 Africans from populations in Ethiopia, Tanzania, and Botswana. We identify significantly associated loci in or near *SLC24A5*, *MFSD12*, *TMEM138[DDB1],* and *OCA2* and *HERC2*. Allele frequencies at these loci in global populations are strongly correlated with UV exposure. At *SLC24A5* we find that a non-synonymous mutation associated with depigmentation in non-Africans was introduced into East Africa by gene flow, and subsequently rose to high frequency. At *MFSD12,* we identify novel variants that are strongly correlated with dark pigmentation in populations with Nilo-Saharan ancestry. Functional assays reveal that *MFSD12* codes for a lysosomal protein that influences pigmentation in cultured melanocytes, zebrafish and mice. CRISPR knockouts of murine *Mfsd12* display reduced pheomelanin pigmentation similar to the grizzled mouse mutant (gr/gr). Exome sequencing of gr/gr mice identified a 9 bp in-frame deletion in exon two of *Mfsd12*. Thus, using human GWAS data we were able to map a classic mouse pigmentation mutant. At *TMEM138[DDB1],* we identify mutations in melanocyte-specific regulatory regions associated with expression of UV response genes. Variants associated with light pigmentation at this locus show evidence of a selective sweep in Eurasians. At *OCA2* and *HERC2* we identify novel variants associated with pigmentation and at *OCA2,* the oculocutaneous albinism II gene, we find evidence for balancing selection maintaining alleles associated with both light and dark skin pigmentation. We observe at all loci that variants associated with dark pigmentation in African populations are identical by descent in southern Asian and Australo-Melanesian populations and did not arise due to convergent evolution. Further, the alleles associated with skin pigmentation at all loci but *SLC24A5* are ancient, predating the origin of modern humans. The ancestral alleles at the majority of predicted causal SNPs are associated with light skin, raising the possibility that the ancestors of modern humans could have had relatively light skin color, as is observed in the San population today. This study sheds new light on the evolutionary history of pigmentation in humans.

2

An atlas of 8,342 mosaic structural variants reveals genetic drivers of clonal hematopoiesis. *P. Loh[1,2], G. Genovese[2,3,4], R.E. Handsaker[2,3,4], H.K. Finucane[1,5], Y.A. Reshef[6], P. Palamara[1,2], B.M. Birmann[7], S.F. Bakhoum[8,9], S.A. McCarroll[2,3,4], A.L. Price[1,2,10].* 1) Department of Epidemiology, Harvard T.H. Chan School of Public Health, 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; 5) Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA; 6) Department of Computer Science, Harvard University, Cambridge, MA; 7) Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 8) Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY; 9) Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY; 10) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA.

Clonal expansions of blood cells harboring somatic mutations are often observed in elderly individuals not known to have cancer (Forsberg et al. 2017 *Nat Rev Genet*). The somatic mutations observed in clonal expansions are enriched at genes commonly mutated in cancer; accordingly, detectable clonal mosaicism is known to confer >10x increased risk of future hematological malignancy. However, the mechanisms that shape most clonal expansions in healthy individuals remain unknown, and the effects of specific somatic events on incident cancers have been unresolved by previous studies of up to ~1,000 detected mosaic events. Here we describe insights from an analysis of 8,342 mosaic structural variants (SVs) which we ascertained in genotyping intensity data from 151,202 UK Biobank participants. We ascertained these SVs using a novel detection algorithm that enables sensitive detection of clonal expansions at low-to-modest cell fractions by harnessing long-range haplotype phase, which we inferred using Eagle2 (Loh et al. 2016b *Nat Genet*). We observed that detectable mosaic SVs cluster non-randomly across the genome, with different chromosomes displaying different enrichments for deletion, duplication, and copy number-neutral loss of heterozygosity (CNN-LOH) events in clonal expansions affecting different blood cell lineages. Genome-wide association analyses of mosaic status revealed six novel loci at which inherited variation associates strongly with the presence of nearby acquired deletions or CNN-LOH. Inherited alleles at these loci appear either to affect the probability of mutation or to be objects of clonal selection themselves. At three loci (*MPL*, *TM2D3*, and *FRA10B*), we identified a likely causal variant that acted in a highly penetrant manner (odds ratios of 103, 698, and 18). A fourth locus implicated a rare *ATM* haplotype. We further identified two autosomal loci that affect mosaic loss of female chromosome X and two *cis*-acting loci on chrX that bias which X chromosome is lost. We observed that in individuals with no previous cancer diagnoses, several somatic SVs commonly seen in hematological cancers strongly associate with risk of future malignancies (OR>100), and we found that early clonal expansions of somatic SVs can be used to predict incident chronic lymphocytic leukemia (CLL) with high accuracy (AUC=0.92). Our results reveal a multitude of paths toward clonal expansion in blood, with a wide range of effects on human health.

3

An intronic *ABCA7* tandem repeat affects Alzheimer's disease, gene expression, and alternative splicing. *A. De Roeck[1,2], L. Duchateau[1,2], J. Van Dongen[1,2], C. Van Broeckhoven[1,2], K. Sleegers[1,2], BELNEU Consortium.* 1) Neurodegenerative Brain Diseases Group, Center for Molecular Neurology, VIB, Antwerp, Belgium; 2) Institute Born-Bunge, University of Antwerp, Antwerp, Belgium.

Genome-wide association studies (GWAS) revealed common SNPs in ATP-Binding Cassette Subfamily A Member 7 (*ABCA7* [MIM 605414]) as risk factor of Alzheimer's disease (AD [MIM 104300]). In addition, subsequent next-generation sequencing (NGS) studies identified strong enrichment of rare heterozygous *ABCA7* loss-of-function (LOF) mutations in AD patients. Nevertheless, LOF mutation carriers exhibit wide variability in disease penetrance, onset age, and ABCA7 expression. Additionally, the biological relevance of *ABCA7* GWAS signals remains elusive, and the effect of other (structural) mutations in *ABCA7* is poorly understood. To address this knowledge gap, we investigated structural variants in NGS data of a Belgian cohort of 1529 AD patients and control individuals, and discovered a previously undocumented intronic *ABCA7* tandem repeat (TR). TR sizes were determined with Southern blotting in a selection of 193 AD patients and 171 healthy control individuals. Lengths varied from 300bp up to 11kb. The largest TR alleles (> 5.5kb) were significantly enriched in AD patients (odds ratio = 3.29 [95% confidence interval = 1.13 - 11.67]). In addition, increasing TR size was associated with the risk-allele haplotype of GWAS SNPs rs3764650 (p-value = 6.83×10^{-8}) and rs78117248 (p-value = 2.66×10^{-4}). Due to the importance of ABCA7 dosage in AD, and the proximity of the TR to splice sites, we next investigated ABCA7 expression using third generation long-read MinION sequencing (Oxford Nanopore Technologies) and identified novel splicing isoforms. Some of these alternative splicing events have the potential to rescue LOF mutation carrying transcripts; while others - including prominent splicing events flanking the TR - lead to loss of functional ABCA7. We quantified overall and transcript specific *ABCA7* mRNA expression in lymphoblast cell lines (n = 48) of AD and control individuals with varying TR lengths, with and without cycloheximide treatment to account for nonsense-mediated mRNA decay. Increasing TR length was significantly correlated with reduction of overall *ABCA7* expression ($\rho = -0.38$, p-value = 0.017) and increased alternatively spliced *ABCA7* ($\rho = 0.71$, p-value = 4.07×10^{-7}). In conclusion, we used unconventional techniques to study complex genomic and transcriptomic rearrangements, and are the first to provide evidence for a TR underlying AD GWAS. In addition, our results have strong implications for AD due to ABCA7 dosage reduction and possible therapies thereof.

4

Scalable computational quantification of gender representation and behavior at ASHG. *N. Telis[1], E.C. Glassberg[2,3], C. Gunter[5,6], J.K. Pritchard[2,3,4].* 1) Biomedical Informatics Program, Stanford, Stanford, CA; 2) Department of Biology, Stanford, Stanford, CA; 3) Department of Genetics, Stanford, Stanford, CA; 4) Howard Hughes Medical Institute, Chevy Chase, MD; 5) Emory University School of Medicine, Departments of Pediatrics and Human Genetics; 6) Marcus Autism Center, Children's Healthcare of Atlanta.

Multiple studies have demonstrated that gender bias in science, along with other specialties, is persistent and has both social and economic consequences. The American Society of Human Genetics (ASHG) has a strong interest in these trends, and quantifying the participation of its own members would motivate continuing efforts to improve representation. Although it is challenging to quantitatively evaluate trends in behavior, a wealth of data about scientific fields is available through conference meetings, in the form of participant information freely available through online abstract portals. We present a quantitative, automated computational pipeline for extracting information about demographics at ASHG meetings from 2014-2016. We developed an automated pipeline and mined each successive meeting website and abstract booklet PDF for author name, affiliation, session, presentation type, and presentation content. We integrated across multiple sources of gender-name connection, using worldwide data sources through the genderizer.io database alongside detailed US Census information, to assign probable gender based on author first names. This provides us with likely gender alongside presenter type and coarse- and fine-grained information about presentation topic. In our analysis, we find that although these three ASHG meetings overall approach gender parity in abstract submissions, there is extreme variance in subfields (from nearly 20% female to over 70% female). This variation is persistent from year to year, in spite of significant changes in available categories for submission. There are also persistent differences in fine-grained word usage between abstracts submitted by men and women, including disparities of words like "data" (male-biased) and "patients" (female-biased). Our results are consistent with observations of continuing differences between male and female participation in science, and they suggest that finer-scale demographic differences and behavioral differences persist year-to-year at ASHG. This survey, paired with local data collection on question-asking behavior at the meeting (abstract by Glassberg et al), provides a scalable pipeline for quantifying differences in the gendered behavior of scientists.

5

Functional characterization of modifier loci for Marfan syndrome reveals novel therapeutic strategies. *R.D. Wardlow[1], J.J. Doyle[1,2], A.J. Doyle[3], N.K. Wilson[1], D. Bedja[4,5], H.C. Dietz[1,6,7].* 1) Howard Hughes Medical Institute and Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, United States; 2) Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, United States; 3) William Harvey Research Institute, Barts and The London School of Medicine, Queen Mary University of London, London, United Kingdom; 4) Department of Cardiology, Johns Hopkins University School of Medicine, Baltimore, United States; 5) Australian School of Advanced Medicine, Macquarie University, Sydney, Australia; 6) Division of Pediatric Cardiology, Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, United States; 7) Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, United States.

Aortic root aneurysm, the leading cause of death in Marfan syndrome (MFS), associates with excessive activation of the canonical (SMAD) and noncanonical (ERK) TGFβ signaling cascades. There is extensive crosstalk between the TGFβ, mitogen activated protein kinase (MAPK, e.g. ERK) and Angiotensin II (AngII) cascades in the aortic wall, with documented therapeutic potential of all respective antagonists in a mouse model of MFS. We identified strain-specific differences in aneurysm severity and survival among mice heterozygous for a MFS allele (*Fbn1*$^{C1039G/+}$), with overt protection on the C57BL/6J (BL6) background compared to 129S6 (129). We generated a large multi-generational pedigree of intercrossed BL6/MFS and 129/MFS mice. Genome mapping revealed 2 QTLs on chromosomes 5 and 11 that strongly linked with aortic size in MFS mice with genome-wide significance (p=0.008 for both) and evidence for epistasis between the loci (MfLOD=12.8; p=0.0006). Our functional analyses focused on strain-specific genetic variation at these loci, with prioritization of a stop-loss mutation in *Mmp17* (rs29636438; p.X579W) and a non-conservative missense mutation in *Map2k6* (rs51129320; p.G76E). Mixed background MFS mice selected for homozygosity for the 129 variants at both loci showed an aortic growth performance identical to that observed on the pure 129 background, while MFS mice homozygous for targeted null alleles at both loci behaved identically to pure BL6 animals. MMP17 is a GPI-anchored matrix metalloprotease that has been shown to act as an allosteric agonist of the epidermal growth factor receptor (EGFR) in a manner that is independent of protease activity. MAP2K6 is an activating kinase for multiple MAPKs. Aggressive aneurysm progression in 129MFS mice associates with overt accentuation of SMAD, ERK and p38 activation in the aortic wall, with complete normalization of all clinical and biochemical abnormalities upon treatment with an AngII receptor blocker or ERK inhibitor. 129MFS mice uniquely showed altered processing of membrane-bound MMP17 that associated with excessive phosphorylation of EGFR; we observed full aneurysm prevention in these animals upon treatment with the FDA-approved EGFR antagonist Erlotinib that also prevented ERK activation. In summary, we have identified a pathway for protective modification of MFS and propose that pharmacologic agents that leverage nature’s success represent potentially potent therapeutic options for the care of MFS patients.

6

Transcriptome analysis of miRNA and mRNA in the PL/J mouse model of hypoxia-induced pulmonary arterial hypertension. *K.T. Ikeda, P.T. Hale, M.W. Pauciulo, N. Dasgupta, M.K. Pandey, W.C. Nichols.* Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH.

Previous genetic studies have shown that some pulmonary arterial hypertension (PAH) patients carry mutations in several genes, most commonly *BMPR2*. Not all patients carry these PAH gene mutations, suggesting that there are other factors causing the PAH. To identify additional genetic factors contributing to the development of PAH, we have focused on a mouse strain which shows a PAH-like phenotype under hypoxic conditions as evidenced by the highest right ventricular systolic pressure (RVSP) in a survey of RVSP after chronic hypoxia exposure of 32 mouse strains. We identified the PL/J strain as a high responder and the MRL/MpJ strain as a low responder, suggesting that the PL/J strain carries genetic variants that are critical for PAH development. To elucidate genetic factors related to the development of the PAH in PL/J mice, we performed RNA-seq analysis of mRNA and miRNA using mouse lungs from PL/J, MRL/MpJ strains at various hypoxic time points. Our results show that among the many genes demonstrating differential expression (FC≥1.5), the expression of Rock2 is 1.6 fold upregulated specifically in PL/J mice at 3 wks hypoxia compared to normoxia. Because it is reported that Rock2 can directly contribute to vasoconstriction, Rock2 was inhibited using the Rock inhibitor Fasudil by IP injection in PL/J mice undergoing right heart catheterization post hypoxic exposure. A 33% drop in RVSP post Fasudil was measured, suggesting that some portion of increased RVSP in PL/J was due to vasoconstriction mediated via Rock2. Furthermore, Rock2 was predicted to have target sites of miR-150 whose expression was downregulated at 3wks hypoxia in PL/J, suggesting that miR-150 may have a role in regulating vasoconstriction. We also found that expression of 41 out of 53 miRNAs from the Dlk1-Dio3 imprinting region on chr12 is upregulated at 3wks hypoxia in PL/J but not in MRL/MpJ. Among these miRNAs, miR-434, miR-541, and miR-381 are predicted to target the development of T cell subsets (e.g. CD3, CD4, CD8). These targets were downregulated specifically in PL/J mice at 3 wks hypoxia. FACS measurement of CD3, CD4, and CD8 T cell subsets in lung and spleen of PL/J and MRL/MpJ mice confirmed the reduced number of different T cell subsets in hypoxic PL/J. In fact, previous studies have suggested T cell abnormalities/reduction may predispose to PAH. Our results suggest that hypoxia may trigger this T cell abnormality in PAH via miRNA regulation.

7

Mechanistic interrogation of a gene-by-environment interaction informs the pathogenesis and treatment of Mendelian aneurysm disorders. *N.K. Wilson[1], J.J. Doyle[1], E. Gallo MacFarlane[1], R. Bagirzadeh[1], G. Yazdanifar[1], D. Bedja[2], S.K. Cooke[1], H.C. Dietz[1], MIBAVA Leducq Consortium.* 1) Institute of Genetic Medicine, Johns Hopkins Medical Institutions, Baltimore, MD; 2) Department of Cardiology, Johns Hopkins Medical Institutions, Baltimore, MD.

Bicuspid aortic valve with distal ascending aortic aneurysm (BAV/DAscAA) is the most common inherited aneurysm condition, affecting ~1% of people with a strong male bias. Unlike less common Mendelian aortopathies characterized by aneurysm of the aortic root (AoRA) such as Marfan syndrome (MFS) and Loeys-Dietz syndrome (LDS), the genetic basis for BAV/DAscAA is poorly understood, with heterozygous LOF mutations in *NOTCH1* explaining less than 1% of cases. It is unclear whether these diseases share mechanistic similarities despite their anatomic differences. We showed that a mouse model of MFS (*FBN1*$^{C1039G/+}$), which develops AoRA driven by TGFβ- and angiotensin II type 1 receptor (AT1R)-linked ERK activation, develops hyperacute dilatation of the DAscA in response to calcium channel blockers (CCBs). This environmentally-induced aneurysm (spatially similar to that observed with BAV) remains AT1R- and ERK-dependent and is accentuated in male mice. In an unbiased screen to mechanistically characterize this phenotype, we carried out RNA sequencing on MFS mouse aortic specimens, applying strict *a priori* filters to select for transcripts that displayed an acute change in expression in response to CCBs that was abrogated upon ERK antagonism. The top pathways enriched in this dataset related to Notch and androgen receptor signaling. The Notch inhibitor dibenzazepine (DBZ) dramatically increased DAscAA progression in MFS mice treated with CCBs, and this effect was also accentuated in males. Rescue of the CCB/DBZ phenotype in MFS mice by treatment with an AT1R or ERK antagonist attests to the pathogenic relevance of this exacerbated disease state to the underlying condition. Treatment with an androgen receptor antagonist was overtly protective in the MFS/CCB/DBZ model, providing the first inkling of mechanism for the male gender bias inherent to many aneurysm conditions. Finally, we have developed and rigorously tested a pathogenic model for DAscAA that integrates underlying genetic predisposition, regional variation in gene expression, gender, the protective influence of Notch, and the exacerbating effects of CCBs that collapses on the expression and activity of the regulator of G-protein signaling (RGS) family of proteins – potent modulators of AT1R signaling. We have also demonstrated the relevance of our mechanism, environmental sensitivities and therapeutic strategies to other inherited aneurysm conditions related to altered TGFβ signaling including LDS.

8

Identification of a novel marker for valve maturation: Loss of *ADAMTS19* function causes progressive valve disease in mice and men. *F. Wünnemann[1,2], A. Ta-Shma[3], M-P. Tremblay[1], C. Preuss[4], P. van Vliet[1], S. Leclerc[1], E. Audain[5], S. Gerety[6], M. Hurles[6], W. Makalowski[2], O. Elpeleg[3], M-P. Hitz[5], G. Andelfinger[1], MIBAVA Leducq Consortium.* 1) Cardiovascular Genetics, CHU Sainte Justine Research Center, Montreal, Quebec, Canada; 2) Institute of Bioinformatics, University of Münster, Münster, Germany; 3) Department of Genetics, Hadassah Hebrew University Medical Center, Jerusalem; 4) The Jackson Laboratory, Bar Harbor, Maine, United States of America; 5) Department of Congenital Heart Disease and Pediatric Cardiology, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany; 6) Wellcome Trust Sanger Institute, Cambridge, United Kingdom.

Valve disease is one of the most common cardiovascular complications with an estimated 2% frequency for any form of valve disease. Despite the high incidence of valve disease, many of the molecular pathways underlying valve formation and maturation remain enigmatic and only a few functional models for valve disease currently exist. Here we describe the identification of two consanguineous families, both with two offspring affected by progressive, polyvalvular disease. Exome sequencing analysis in both pedigrees identified rare, homozygous loss of function alleles for the gene *ADAMTS19* for all four affected individuals. Further functional characterization of *Adamts19* during murine valve development was performed using a lacZ reporter mouse model. Expression analysis in the mouse highlighted Adamts19 as a specific marker for valvular interstitial cells (VICs) post-valve formation and throughout postnatal maturation. High-resolution digital echocardiography on homozygous *Adamts19*$^{-/-}$ mice revealed progressive aortic valve regurgitation and stenosis in these animals. Histology of dysfunctional valves showed thickening of aortic valve leaflets together with disorganization of collagen, reminiscent of a fused bicuspid aortic valve. Molecular analysis of the valves suggests that *Adamts19* is required for extracellular matrix remodeling after initial valve formation. Taken together, these data identify *Adamts19* as a key player in aortic valve homeostasis after birth. We show that the progressive phenotype, which replicates closely the course of aortic valve conditions in humans, is specifically triggered by VICs at an early developmental time point, with far reaching consequences into the maintenance of adult valves. Complete loss of *Adamts19* function in mice, therefore represents a valuable new model for studying the effects of progressive aortic valve disease.

9

***LTBP3* recessive mutations cause amelogenesis imperfecta as well as aortic diseases.** *D. Guo[1], E. Regalado[1], J. Chen[1], A. Pinard[1], C. Rigelsky[2], L. Zilberberg[3], E. Hostetler[1], S. Wallace[1], M. Bamshad[4], D. Nickerson[4], D. Rifkin[3], D. Milewicz[1], University of Washington Center for Mendelian Genomics, Seattle, WA.* 1) Univ Texas/Houston McGovern Med Sch, Houston, TX; 2) Cleveland Clinic Foundation, Cleveland, OH; 3) New York University School of Medicine, New York, NY; 4) University of Washington, Seattle, WA.

The major diseases affecting the thoracic aorta are aneurysms and acute dissections. Pathogenic variants in a number of genes lead to Heritable Thoracic Aortic Disease (HTAD) and these genes encode proteins involved in vascular smooth muscle cell contraction and adhesion to the extracellular matrix or the TGF-β signaling pathway. By exome sequencing DNA from HTAD families, we identified rare variants in *LTBP3* in two probands. First, homozygous p.del678-681NFPGinTC variant predicted to result in a truncated protein missing critical amino acids of the LTBP3 EGF-like calcium-binding domain was identified in the proband of TAA376, who had an ascending aortic dissection, enamel deficiency and short stature. This proband's similarly affected sister has the same homozygous indel variant. Their parents and brother are heterozygous carriers of the indel variant and did not have aortic, teeth enamel or bone abnormalities. Second, compound heterozygous LTBP3 variants (p.P45Rfs and p.E750*) predicted to lead to nonsense mediated decay were identified in the proband of Trio9, who had aneurysms of the aortic root, abdominal aorta, axillary, iliac, hepatic, and celiac arteries, amelogenesis imperfecta and short stature. This proband's sister has the same compound heterozygous variants and has amelogenesis imperfecta and bone abnormalities, but has not been evaluated for aortic or arterial aneurysms. Their parents are heterozygous carriers and do not have symptoms. Homozygous loss of function *LTBP3* mutations have been associated with amelogenesis imperfecta and short stature, primarily in children, and these findings were recapitulated in *Ltbp3* null mice. However, vascular disease has not been described in these mice followed to 9 weeks, and survival to 3 months was similar to wild type mice. Interestingly, absence of *Ltbp-3* attenuates noncanonical (ERK1/2) TGF-β signaling in the aortas of *Ltbp3*$^{-/-}$ mice. Additionally, *Ltbp3*$^{-/-}$ mice have an increased number of elastic lamellae in the ascending aorta (9.3 ± 0.9. p<0.01for *Ltbp3*$^{-/-}$ mice versus 7.4 ± 0.5 for WT mice) and lower blood pressure than wild type mice, which are vascular features in common with knockout of the *Acta2*, a known HTAD gene. Furthermore, zebrafish treated with *ltbp3* morphants failed to form the ventral aorta and pharyngeal arch arteries. These findings indicate that homozygous and compound heterozygous *LTBP3* loss of function variants also predispose to HTAD and other vasculopathies.

10

Identification of an autosomal recessive form of Noonan Syndrome. *J. Johnston[1], J.J. van der Smagt[2], J.A. Rosenfeld[3], A. Alswaid[4], E.H. Baker[5], G. Borck[6], J. Brinkmann[7], W. Craigen[3], V.C. Dung[8], L. Emrick[9], D.B. Everman[10], K.L. van Gassen[2], S. Gulsuner[11], M.H. Harr[12], M. Jain[3], K.A. Leppig[13], D.M. McDonald-McGinn[14], C.T.B. Ngoc[8], E.R. Roeder[15], R.C. Rogers[10], J.C. Sapp[1], A.A. Schäffer[16], D. Schanze[7], N.E. Verbeek[2], M.A. Walkiewicz[17], E.H. Zackai[14], M. Zenker[7], C. Zweier[18], B. Lee[3], L.G. Biesecker[1], Members of UDN.* 1) Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD, USA; 2) Department of Genetics University Medical Center Utrecht, Utrecht, The Netherlands; 3) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA; 4) King Abdulaziz Medical City, Riyadh, Saudi Arabia; 5) Department of Radiology and Imaging Services, Clinical Center; NIH, Bethesda, MD, USA; 6) Institute of Human Genetics, University of Ulm, Ulm, Germany; 7) Institute of Human Genetics, University Hospital, Magdeburg, Germany; 8) Rare Disease and Newborn Screening Service, The National Children's Hospital, Hanoi, Vietnam; 9) Division of Neurology and Developmental Neuroscience, Baylor College of Medicine, Houston, TX, USA; 10) Greenwood Genetic Center, Greenwood, SC, USA; 11) Division of Medical Genetics, University of Washington, Seattle, WA, USA; 12) Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA; 13) Genetic Services, Kaiser Permanente of Washington, Seattle, WA, USA; 14) Division of Human Genetics and Department of Pediatrics, Children's Hospital of Philadelphia and the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA; 15) Department of Pediatrics and Molecular and Human Genetics, Baylor College of Medicine, San Antonio, TX, USA; 16) Computational Biology Branch, National Center for Biotechnology Information, NIH, Bethesda, MD, USA; 17) Baylor Genetics Laboratories, Houston, TX, USA; 18) Institute of Human Genetics, Friedrich-Alexander-Universität Erlangen- Nürnberg (FAU), Erlangen, Germany.

Noonan syndrome is part of a spectrum of disorders with overlapping phenotypes that include craniofacial features such as widely spaced eyes, ptosis, downslanted palpebral fissures, lowset, posteriorly rotated ears, wide or webbed neck, curly hair, and cardiovascular abnormalities. Many genes, including *PTPN11*, *SOS1*, *SOS2*, *RAF1*, *KRAS*, *NRAS*, *BRAF*, *SHOC2*, *CBL*, *RIT1*, and *LZTR1,* have been implicated in autosomal dominant Noonan syndrome. While autosomal recessive inheritance of Noonan syndrome has been suspected in a few families, this genetic etiology has never been confirmed. Here we report both clinical and molecular evaluations of nine families with a total of 17 children who have Noonan syndrome compatible with an autosomal recessive mode of inheritance. The phenotype ranged from mildly affected patients to patients with fatal cardiac disease and leukemia. In all families, the parents were unaffected. Genome-wide genetic linkage analysis, using a recessive model, in a family with four affected siblings supported only a single candidate region in chromosome 22q11, which includes the candidate gene *LZTR1*, previously shown to harbor mutations in patients with Noonan syndrome inherited in a dominant pattern. A combination of exome, genome and panel sequencing identified biallelic pathogenic variants in *LZTR1*, including putative loss of function, missense, and canonical and non-canonical splicing variants in all affected children, with heterozygous, clinically unaffected parents. Splice variants were evaluated by reverse-transcription PCR and shown to result in aberrant splice products. These clinical and genetic data confirm the existence of a form of Noonan syndrome that is inherited in an autosomal recessive pattern and identify biallelic mutations in *LZTR1* as the cause of this disorder.

11

Novel insights into clinically relevant variation using the diverse sample populations of the PAGE study. *E.P. Sorokin[1], G.M. Belbin[2], G.L. Wojcik[1], N. Abul-Husn[2], S. Bien[3], N. Zubair[3], P. Norman[4], G. Nadkarni[5], C. Hodonsky[6], J. Odgis[2], C. Avery[6], S. Buyske[7], T. Matise[8], J. Kocarnik[3], L. Hindorff[9], R. James[10], K.E. North[6], R. Loos[2], C. Haiman[11], C. Kooperberg[3], C. Carlson[3], C.D. Bustamante[1], C.R. Gignoux[12], E.E. Kenny[2], the Population Architecture using Genomics and Epidemiology (PAGE) study.* 1) Department of Genetics, Stanford University School of Medicine, Stanford CA; 2) The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York NY; 3) Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle WA; 4) Department of Structural Biology, Stanford University School of Medicine, Stanford CA; 5) Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York NY; 6) Department of Epidemiology, University of North Carolina, Chapel Hill NC; 7) Department of Statistics & Biostatistics, Rutgers University, New Brunswick NJ; 8) Department of Genetics, Rutgers University, New Brunswick NJ; 9) Division of Genomic Medicine, NHGRI, NIH, Bethesda MD; 10) Division of Clinical Research & Data Management, NIMHD, NIH, Bethesda MD; 11) Department of Preventive Medicine, University of Southern California Keck School of Medicine, Los Angeles CA; 12) Colorado Center for Personalized Medicine, University of Colorado, Aurora, CO.

The profusion and population-specificity of rare variants in human populations make questions of causality, penetrance and expressivity of clinically relevant variants (CRVs) difficult to ascertain across diverse populations, even within well-characterized disease genes. Therefore, the characterization of CRVs in multi-ethnic cohorts is a critical step in advancing global precision medicine. To address this goal, we genotyped ~52K individuals from 126 global populations from the PAGE study representing broad global diversity (www.pagestudy.org) on the Multi-Ethnic Genotyping Array (MEGA). The CRV content selected for MEGA includes 34,453 variants from ClinVar, HGMD, PharmGKB, of which 25,119 are polymorphic within PAGE. We observed a median of 23 pathogenic alleles per individual, with a higher median of 26 within African descent populations consistent with elevated heterozygosity. To distinguish truly deleterious variation from ancestry-specific variation that may be rare in Europeans, we examined elevated allele frequencies (annotated Risk Allele Frequency > 0.05) within PAGE. Even within the twelve large populations within PAGE (N>500), variants with a risk allele frequency greater than 0.05 in at least one group include: 2,317 benign/ likely benign, 35 uncertain, 661 conflicted, and 38 pathogenic/likely pathogenic variants, demonstrating the challenges of hard frequency thresholds for clinical pathogenicity assertion. These numbers increase as these populations are sub-divided into country- or region-level designations, as unique population histories enable unique signatures of genetic drift. We also characterized differentiation within 2,600 pharmacovariants, and observed 1,018 with Fst>0.15 between 1000 Genomes Europeans and each non-European PAGE populations. Finally, we characterized the population health impact of segregating CRVs using electronic health record data from the Mount Sinai BioMe biobank in New York City, using Phenome-Wide Association Study (PheWAS) with ICD-9 medical billing codes for common and rare variants. Within common CRVs (MAF>0.01), we identified 132 associations (FDR<0.05), including 47 known associations, 16 associations with supporting evidence in the literature, and 69 other findings including potentially novel associations. Association tests with rare variants are ongoing, with results to report at the meeting. This data provides a unique opportunity to study the trans-ethnic impact of clinically relevant variants.

12

Common and rare variants associated with adult height: The Million Veteran Program. *T.L. Assimes[1,2], J. Huang[3,4], J. Li[1,2], K. Cho[3,4], Y. Ho[3], Y. Sun[5,15], N. Sun[6,14], J.M. Gaziano[3,4], J. Concato[6,7], S. Pyarajan[3,4], S. Muralidhar[8], H. Hunter-Zinck[3], H. Zhao[13,14], P. Wilson[9,10], P. Tsao[1,2], E.R. Hauser[11,12], C.J. O'Donnell[3,4] on behalf of the VA Million Veteran Program.* 1) Department of Medicine, Stanford University School of Medicine, Stanford, CA; 2) VA Palo Alto Health Care System; 3) Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA; 4) Harvard Medical School, Boston, MA; 5) Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA; 6) Clinical Epidemiology Research Center, VA Connecticut HealthCare System, West Haven, CT; 7) Internal Medicine: General Internal Medicine, Yale School of Medicine, New Haven, CT; 8) Office of Research and Development - Veterans Health Administration, Dept. of Veterans Affairs, Washington, DC; 9) Atlanta VA Medical Center, Decatur, GA; 10) Emory Clinical Cardiovascular Research Institute, Atlanta, GA; 11) Cooperative Studies Program Epidemiology Center-Durham, Veterans Affairs Medical Center, Durham, NC; 12) Duke Molecular Physiology Institute, Duke University School of Medicine, Durham, NC; 13) VA Cooperative Studies Program Coordinating Center, West Haven, CT; 14) Department of Biostatistics, Yale School of Public Health, New Haven, CT; 15) Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA.

Background: Adult height is a highly heritable trait. Recent genome-wide association studies have identified 697 common and 83 low frequency-rare variants associated with height, yet these variants explain only ~1/4 of the heritability. **Design and Methods**: We conducted a GWAS of adult height in 353,938 participants of the Million Veteran Program (MVP) receiving care within the US Veterans Health Administration including 225,299 Europeans (EUR), 56,264 African Americans (AFR), and 20,672 Hispanics (HIS). We genotyped participants with a customized Affymetrix Axiom biobank array that included >720,000 variants across a broad spectrum of frequencies. Genotype data was then used to impute ~50 million genotypes included in the 1000 genomes panel. We derived participants' height by taking the average of multiple measurements available in electronic health records (EHR) within three years of enrollment. We then performed single variant association score tests on the sex and race stratified residuals of height after adjusting for age and 10 principal components. We defined a novel locus as a locus with a genome-wide significant association ($P<5x10^{-8}$) at least 500kb away from previously reported variants and a novel variant as a variant with a $P<1x10^{-4}$ even after conditioning on previously reported lead variants. We conducted pathway analyses of all known and novel variants using Ingenuity (IPA). **Results**: We observed a very high rate of replication of previously established common and rare variants. Our association analyses yielded 170 novel loci among EUR and 60 among AFR. Our step-wise conditional analyses identified an additional 925, 234, and 11 novel variants among EUR, AFR, and HIS, respectively. We also identified 109 novel functionally significant low frequency-rare variant associations. Top novel loci mapped to *COQ8A*, *RAPGEF6*, and *NCK1* for common variants as well as *ACAN*, *ADAMTS10*, and *SERPINA1* for rare variants. The top canonical pathway, upstream regulator, molecular/cellular function, and physiological system identified by IPA were “axonal guidance signaling”, *ESR1*, “gene expression”, and “connective tissue development & function”, respectively. **Conclusion**: Using a large, ethnically diverse biobank connected to a rich EHR, we identified multiple novel variants associated with adult height. Our results confirm the highly polygenic nature of this trait, and highlight the value of examining non-European race/ethnic groups in large numbers.

13

Integrated inference that accurately identifies close relatives in > 1 million samples. *W.-M. Chen, A. Manichaikul, J. Nguyen, S. Onengut-Gumuscu, S.S. Rich.* Center for Public Health Genomics, University of Virginia, Charlottesville, VA.

Rapid growth of genomics data creates a critical need for tools that accurately infer cryptic relatedness. We propose an integrated inference procedure that combines an identical by state (IBS)-based method to estimate the kinship coefficient and an identical by descent (IBD) segment method that estimates the proportion of chromosomal segments with both alleles shared IBD between each pair of individuals. The backbone of this new implementation is built upon our KING software. This tool includes a number of KING's computational improvements, including a multi-stage procedure to eliminate unrelated or more distant pairs using the estimated kinship coefficients as the filter. The use of the homozygote concordance rate as an additional filter can increase speed of the initial screening process. The use of the heterozygote concordance rate is a key to rapid and accurate determination of the IBD2 segments without the need of time-consuming haplotyping, regardless of the type of variants (*e.g.* array and sequencing data). Our heterozygote-concordance-based IBD2 segment method distinguishes and unambiguously confirms the inference of full sibs, which further improves accuracy of inference especially under scenarios such as the use of SNP panels with sparse genome-wide coverage (*e.g.*, the ImmunoChip array in our real example data covers 186 autoimmune-related loci), and the presence of admixture and/or inbreeding. We apply this integrated tool (KING 2.1) to a dataset consisting of 30,375 samples, each genotyped at 170,646 SNPs from the ImmunoChip array. We are able to infer all 48 pairs of duplicates/MZ twins, 12,176 pairs of parent-offspring (PO), and 6,439 pairs of full sibs (FS) instantly (3 seconds) with 100% accuracy. In contrast, the accuracy to infer full sibs using the original KING is 99.52% (*i.e.*, 0.48% of full sibs have estimated kinship<0.177). In order to examine the computational performance and scalability of the new KING implementation, we created a large dataset of 2,004,750 samples (by duplicating the above ImmunoChip data 65 times). Analysis of cryptic relationships took 3 hours and 50 minutes on a single computer (with 40 CPU cores) to infer all 53 million pairs of PO, 28 million pairs of FS, and 65 million duplicate pairs. Our integrated tool not only provides rapid inference that applies to > 1 million samples, more importantly, it is more accurate in the presence of a wide range of complex but realistic genetic scenarios.

14

Using genotyped relatives of ungenotyped type 2 diabetes cases as proxy-cases in a cohort based genome-wide association study. *B.N. Wolford[1], S. Lee[2,3], W. Zhou[1], J.B. Nielsen[4], L.G. Fritsche[5,6], M. Lin[4], H.M. Kang[2,3], M. Gabrieldsen[5,6], O. Holmen[5,6,7], K. Hveem[5,6,8], G.R. Abecasis[2,3,6], M. Boehnke[2,3,6], C.J. Willer[1,4,6,9].* 1) Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI; 2) Department of Biostatistics, University of Michigan, Ann Arbor, MI; 3) Center for Statistical Genetics, University of Michigan, Ann Arbor, MI; 4) Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI; 5) HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway; 6) K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, Norwegian University of Science and Technology, Trondheim, Norway; 7) St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway; 8) The Mindich Child Health Development Institute, The Icahn School of Medicine at Mount Sinai, New York, NY; 9) Department of Human Genetics, University of Michigan, Ann Arbor, MI.

Type 2 diabetes (T2D) and myocardial infarction (MI) are complex diseases for which genome-wide association studies (GWAS) have identified dozens of associated risk loci. A recently published method by Liu et. al. introduces the concept of GWAS by proxy (GWAX): performing case-control genetic association studies using unaffected first-degree relatives of cases (proxy-cases) in the (near) absence of true cases. Here, we extend GWAX to model genetic liability in large cohorts for which cases, proxy-cases, and controls are available. We describe a score test without covariates where, in addition to the cases and controls, we include the probability of a genotype conditional on disease status for proxy-cases. We performed 1,000 simulation replicates to evaluate this statistical test in an idealized cohort (N=100,000) with disease prevalence of 0.1 and heritability of 0.5 (cases=10,000; proxy-cases=15,507; controls=74,493) where all proxy-cases have one affected relative and the sample-wide MAF is 0.1. Across a variety of common minor allele frequencies, odds ratios, and reasonable case/proxy-case/control ratios our type one error is well controlled. The standard GWAS model compares cases to controls and includes proxy-cases, typically unidentified in a cohort, as controls. By simply removing proxy-cases from controls, we increase power from 78.2% to 89.8% for an odds ratio of 1.175 and this trend holds across similar odds ratios. By modeling proxy-cases with cases and controls we increase power to 97.8%. Next, we used this statistical test to identify genetic variants associated with T2D and MI in the Norwegian Nord-Trøndelag Health Study (HUNT). Using self-reported family history, we partitioned our sample of 69,635 European individuals into cases, proxy-cases, and controls. We tested for association in ~10 million genotyped and imputed genetic variants (MAF > 0.5%). We used a linear mixed model due to relatedness within our sample and used relationship-to-case as a semi-continuous trait (F=1 for cases, F=0.5 for proxy-cases, and F=0 for controls). We replicate 5 known T2D loci at genome-wide significance compared to 3 loci found by simply modelling T2D cases and controls. We are evaluating this method in additional cohorts and continuing methods development to account for covariates and relatedness. With the increasing availability of biobank data, this work demonstrates the potential advantage of statistically modeling proxy-cases in a cohort based GWAS.

15

Sparse linear mixed models for pedigrees with millions of individuals. *T. Shor[1], D. Geiger[1], Y. Erlich[3,4], O. Weissbrod[2].* 1) Computer science, Technion, Haifa, Israel; 2) Computer science, Weizmann Institute, Rehovot, Israel; 3) Columbia University, New York City, New York; 4) New York Genome Center, New York City, New York.

The rapid digitization of genealogical and medical records enables the assembly of extremely large pedigree records spanning millions of individuals. Such pedigrees provide the opportunity to answer genetic and epidemiological questions in scales that are much larger than was previously possible. Linear mixed models (LMMs) are often used for analysis and prediction of phenotypes in pedigree data. However, LMMs cannot naturally scale to large pedigrees spanning millions of individuals, owing to their steep computational and storage requirements. Here we propose a novel modeling framework called Sparse-LMMs that alleviates these difficulties by exploiting the sparsity patterns found in large pedigree data. The proposed framework can construct a matrix of genetic relationships between trillions of pairs of individuals in several hours, and can perform a moment-based or a restricted maximum likelihood (REML) estimation with multiple matrices in several days. We demonstrate the capabilities of Sparse-LMM via simulation studies and by estimating the heritability of longevity in a very large pedigree spanning millions of individuals and over five centuries of human history. The Sparse-LMM framework enables the analysis of extremely large pedigrees that was not previously possible.

16

Personalized feedback on the genetic risk of common complex diseases: The potential of a large population-based biobank, and methodological challenges. *K. Fischer[1], K. Läll[1,2], R. Mägi[1], T. Esko[1], L. Leitsalu[1], N. Tõnisson[1,3], A. Metspalu[1].* 1) Estonian Genome Center, University of Tartu, Tartu, Estonia; 2) Institute of Mathematics and Statistics, University of Tartu, Estonia; 3) Dept. of Clinical Genetics in Tallinn, Tartu University Hospital, Estonia.

We will provide an overview of the development and validation of algorithms for personalized prediction of the risk of Type 2 Diabetes (T2D) and Coronary Artery Disease (CAD) in the Estonian Biobank cohort. We also discuss various methodological challenges at different steps of the process. To provide accurate estimates of the risk of a complex disease, a Genetic Risk Score (GRS) needs to be combined with known environmental and lifestyle-related risk factors. Usually the GRS is defined as a linear combination of effect allele counts of several Single Nucleotide Polymorphisms (SNPs), whereas the SNPs and their corresponding weights are based on results of a large-scale meta-analysis of Genome-Wide Association Study (GWAS). To develop a personalized risk prediction algorithm that could be used for health counselling in general practice, the following steps are needed: 1) selection of the appropriate genomic predictor (selection of SNPs and their weights to be combined in a GRS) based on an initial cohort; 2) validation of the genomic predictor in an independent cohort; 3) selection of non-genetic predictors; 4) development of a prediction model that combines genetic and non-genetic predictors; 5) estimation of baseline (age-specific) risk level and combining the results in a model for absolute risk; 6) validation of the final model in an independent cohort; 7) identification of the optimal ways to communicate the risk estimates to individuals. These steps were undertaken to develop risk prediction algorithms for Type 2 Diabetes and Coronary Heart Disease for the participants of the Estonian Biobank cohort. We will show important stages of the process as well as the resulting risk prediction tool. We also discuss statistical challenges that are related to specific features of the population-based biobank data (left-truncation for some outcomes, mix of retrospective and prospective data for some others, etc.). In addition, some common mistakes and their consequences are pointed out (such as partial overlap of discovery and validation cohorts).

17

Landscape of allele-specific open chromatin in human iPSC-differentiated neurons and its implication for mental disorders. *S. Zhang[1,2], W. Moy[1], H. Zhang[1], H. McGowan[3], J. Shi[4], C. Leites[1], A. Sanders[1,2], Z. Pang[2], P. Gejman[1,2], 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, IL 60637, USA; 3) Department of Neuroscience and Cell Biology and Child Health Institute of New Jersey, Rutgers University, New Brunswick, NJ 08901, USA; 4) Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA.

Allelic Imbalance of gene Expression (AIE) plays important role during neurodevelopment. However, the underlying mechanism of AIE has not been fully understood. Since chromatin accessibility (openness) has a strong influence on gene transcription, we hypothesised that allele-specific open chromatin (ASoC) might regulate AIE. Also, as open chromatin regions (OCRs) overlap with gene regulatory sequences and a genetic variant showing ASoC likely affects gene expression, we reasoned that mapping ASoC in neurons at different developmental stages would help functionally interpret the abundant noncoding risk variants of neurodevelopmental disorders. Using human neurons derived from induced pluripotent stem cells (iPSCs) as a cellular model, we carried out a global OCR profiling by Assay for Transposase-Accessible Chromatin through sequencing (ATAC-seq) and a transcriptomic profiling by RNA-seq in the same cell populations of different differentiation stages. We found that ASoC and AIE are both widespread and cell-type specific, and the differentiated neurons tend to have a higher level of ASoC and AIE than in iPSCs. In iPSCs and neurons from the same human subject, out of the 10,048 heterozygous SNPs in OCRs, 101, 404, and 250 showed ASoC (P<0.05; binomial test) in iPSCs, day-30 and day-41 neurons, respectively. SNPs showing ASoC had little overlap between iPSCs and neurons, whilst a substantial overlap (n=44) between day-30 and day-41 neurons was observed. Interestingly, we found that genes showing ASoC are more likely to exhibit AIE at various developmental stages, suggesting a functional link between ASoC at the DNA level and AIE at the RNA level. We further explored whether the neuronal ASoC can help identify functional risk variants implicated by schizophrenia genome-wide association studies (GWAS). Out of the 12 schizophrenia GWAS-implicated SNPs that we found in neuronal OCRs of this single individual, two SNPs showed ASoC and are thus putatively functional: one lies within the 5'-UTR of *CHRNA5* (*cholinergic receptor, nicotinic, alpha 5*) and the other is in the promoter region of *VPS45*, a Sec1 family gene involved in synaptic transmission. Our results not only provide novel insights into the epigenetic mechanisms of AIE during neural development but also suggest a new strategy for identifying functional noncoding disease risk variants that may influence chromatin accessibility and gene expression in schizophrenia and other neurodevelopmental disorders.

18

Alternative splicing of brain-expressed transcripts distinguishes major adult psychiatric disorders. *N. Akula[1], R. Kramer[2], Q. Xu[2], K. Johnson[3], S. Marenco[2], J. Apud[2], H. Rhodes[2], B. Harris[2], B.K. Lipska[2], F.J. McMahon[1].* 1) Human Genetics Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA; 2) Human Brain Collection Core, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA; 3) Bioinformatics Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA.

Gene expression studies in post-mortem brain have provided valuable clues to the biological basis of psychiatric disorders, but few such studies have taken into account the transcriptional complexity of the brain, where alternative splicing may generate multiple transcripts variants for each individual gene. We performed high-depth sequencing of RNA (RNA-seq) on ribosome-depleted libraries derived from subgenual anterior cingulate cortex, a region that has been implicated in psychiatric disorders. A total of 200 samples were studied, 39 from people with bipolar disorder (BD), 46 with schizophrenia (SCZ), 54 with major depression (MDD), and 61 without a known psychiatric disorder (controls). Stranded, paired-end sequencing of high-quality RNA (RIN ≥6) was performed on the Illumina HiSeq 2500. Of 54 billion 125 bp reads, 137M properly-paired reads/sample mapped to the reference genome (hg38) by HISAT2. StringTie identified 21K ENSEMBL genes with at least 10 reads each. These harbored over 85K transcripts, of which 44% were abundant (>100 reads). After quality control and quantile normalization, differential expression was estimated using DESeq2, with correction for RIN, race, and GC content. At FDR<10%, gene-level analysis identified 67, 53, and 11 genes that were differentially expressed in BD, SCZ, and MDD, respectively. Many of the same genes were differentially expressed in multiple disorders (mean overlap 42%). Compared to controls, the overlapping genes showed a consistent direction of differential expression and a significant positive correlation in fold-change values across all 3 disorders. Transcript-level analysis identified 336, 680, and 304 transcripts that were differentially expressed (FDR<10%) in BD, SCZ, and MDD, respectively. For some genes, multiple distinct transcripts were differentially expressed in the same disorder. Some of the same transcripts were differentially expressed in multiple disorders, but the mean overlap of differentially-expressed transcripts across disorders was only 18%. The likelihood of overlap in differential expression between disorders was 3.5 to 29 times greater at the gene level than at the transcript level. To our knowledge this is the deepest RNA-seq study in a large sample of human postmortem brain tissue. The results illustrate the great diversity of brain-expressed transcripts and suggest that alternative splicing is an important factor that distinguishes between major adult psychiatric disorders.

19

Sexually dimorphic DNA methylation in human brain and relevance to psychiatric disorders. *Y. Xia[1], R. Dai[1], K. Wang[1], Y. Xu[3], H. Li[3], J. Xi[3], C. Chen[1], C. Liu[1,2], the PsychENCODE Consortium.* 1) State Key Lab of Medical Genetics, Central South University, Changsha, Hunan, China; 2) Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, USA; 3) Xiangya school of medicine, Central South University, Changsha, Hunan, China.

Background: The prevalence of many psychiatric disorders, including autism spectrum disorders (ASDs), depression, intellectual disability (ID) and so on, has presented clear sex differences. Sex-associated biomarkers in the brain may help to reveal the mechanisms involved in biology and etiology of psychiatric disorders. DNA methylation is sexually dimorphic and thus is a candidate for such biomarkers. Methods: We systematically compared methylation in males and females in 697 human postmortem prefrontal cortex (PFC) samples, and another 210 samples to replicate the findings. We investigated sex-associated differential methylation at individual CpG loci (DMPs) and genomic regions (DMRs), their corresponding genetic regulators (methylation quantitative trait loci, meQTLs), and regulatory target genes. We also tested co-methylated CpG sites for association with sex. Lastly, we investigated whether SNPs and genes involved in sex-dimorphic methylation were implicated in psychiatric disorders. Results: We identified 13,352 DMPs and 4,490 DMRs. Those DMRs spanned genes enriched for calcium signaling pathway (adjusted p = $2.3e^{-10}$) and neuroactive ligand-receptor interaction (adjusted p = $1.5e^{-9}$), and enriched for ASDs (adjusted p = $2.4e^{-18}$) and depression (adjusted p value=$4.9e^{-10}$) and so on. There were 99,848 meQTLs pairs, 1,066 DMPs and gene expression pairs related to sexually dimorphic methylation. Co-methylation analysis identified a sex-associated module containing 4,825 DMPs. From candidate genes with de novo mutations which have been reported in schizophrenia, ASDs, epilepsy and ID, we found 112 DMR target genes, and eleven of them were reported in at least two disorders above. Among the previously reported 108 genome-wide significant schizophrenia-associated loci, we also found seven loci which could regulate the DMPs by meQTL. One SNP, rs7085104, could regulate the DMP and then affect the expression of *C10orf32*. Discussion: This is a comprehensive analysis of sexually dimorphic methylation in the largest number of human brain tissues up to now. Our results were more systematically investigated the DMPs which contained its upstream genetic regulators and downstream target genes' expression. Our findings of the association between the sexually dimorphic methylation and psychiatric disorders moved forward the understanding of the different prevalence of psychiatric disorders between males and females, and provided a window to explain the GWAS signals.

20

Genome-wide methylomic analysis of neonatal blood from Danish twins discordant for mental illness. *S. Weinsheimer[1,2], A. Starnawska[1,3,4], C.S. Hansen[1,5], A. Buil[1,2], J. Bybjerg-Grauholm[1,5], M. Bækvad-Hansen[1,5], D.M. Hougaard[1,5], T. Sparsø[1,2], M. Bertalan[1,2], P.B. Mortensen[1,6,7], C.B. Pedersen[1,6,7], T.M. Werge[1,2,8].* 1) iPSYCH - The Lundbeck Foundation's Initiative for Integrative Psychiatric Research, Denmark; 2) Institute of Biological Psychiatry, Mental Health Center, Sct. Hans, Mental Health Services, Copenhagen, Denmark; 3) Aarhus University, Aarhus Denmark; 4) iSEQ, Centre for Integrative Sequencing, Aarhus, Denmark; 5) Statens Serum Institute, Copenhagen, Denmark; 6) National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark; 7) Centre for Integrated Register-Based Research, Aarhus University, Aarhus, Denmark; 8) Institute of Clinical Sciences, Faculty of Medicine and Health Sciences, University of Copenhagen, Copenhagen, Denmark.

Purpose: Emerging evidence implicates altered DNA methylation in mental illness including autism, ADHD, bipolar disorder, major depressive disorder, anorexia and schizophrenia. However, it is unclear whether the DNA methylation changes observed to date are causative or reflect disease progression or treatment. The neonatal period is a time of rapid neurodevelopment during which alterations in DNA methylation may contribute to the risk of mental illness later in life. Hence, we explored whether differences in DNA methylation in neonatal blood taken at birth were associated with twin discordance for mental illnesses including autism, ADHD, affective disorder, anorexia, schizophrenia or bipolar disorder. Methods: A total of 597 pairs of twins (220 monozygotic) discordant for mental illness born between 1981 and 2005 were identified for methylomic comparison. Blood samples obtained from neonatal Guthrie cards were used for DNA extraction and genome-wide profiling of DNA methylation with the use of Infinium HumanMethylation450 BeadChip or EPIC array from Illumina. Quality control, data pre-processing and statistical analysis was performed using the *minfi* package in R. Data were normalized using the ssNoob method and adjusted for batch effects using the Combat tool. Blood cell composition was estimated using FlowSorted.CordBlood.450k. Using linear regression models and including potential confounders such as sex, blood cell composition and zygosity, we observed differentially methylated positions (DMPs) associated with mental illness. We used VarElect to identify which genes are known to directly associate with mentall illness. The GeneMania tool was used to visualize gene interactions in a network. Results: We observed significant DMPs (P<10-06) for ADHD (mapping to *TET2*, *HNRNPH2*, *HMGN5*), autism (mapping to *ATP1B4*), and anorexia (*ITGB4*, *GJA3*, *NXN*). Interestingly, there is an enrichment of DMPs mapping to genes in the dopaminergic and serotonergic synapse KEGG pathways including *KCNJ5*, *PRKCA*, *CACNA1D*, *CREB5*, and *ALOX12* (P<0.05). In addition, we identified 67 DMPs (P<10E-5) mapping to genes which have known direct association with at least one mental illness and are connected in a complex genetic network. Our data indicate that DNA methylation differences are quantifiable in neonatal blood from twins discordant for mental illness later in life and suggest that susceptibility to mental illness is conferred by dysregulated neurodevelopmental genes.

21

GWAS of the PTSD "re-experiencing" symptom cluster in the MVP sample, N>150,000. *J. Gelernter[1,2,13], N. Sun[3,4], R. Pietrzak[1,2], Q. Lu[4], Y. Hu[4], B. Li[4], Q. Chen[3,4], K. Radhakrishnan[3], M. Aslan[3,5], K.H. Cheung[3,6], Y. Li[3,7], N. Rajeevan[3,7], F. Sayward[3,7], K. Harrington[8,9], K. Cho[8,10], J. Honerlaw[8], S. Pyarajan[8,10], R. Quaden[8], J.M. Gaziano[8,10], J. Concato[3,5], H. Zhao[3,4], M.B. Stein[11,12] on behalf of: Dept Veterans Affairs Cooperative Studies Program (#575B) & Million Veteran Program.* 1) Psychiatry Service, VA Connecticut Healthcare System, West Haven, CT; 2) Department of Psychiatry, Yale University School of Medicine, New Haven, CT; 3) VA Clinical Epidemiology Research Center (CERC), VA Connecticut Healthcare System, West Haven, CT; 4) Department of Biostatistics, Yale University School of Public Health, New Haven CT; 5) Department of Medicine, Yale University School of Medicine, New Haven, CT; 6) Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT; 7) Yale Center for Medical Informatics, Yale University School of Medicine, New Haven, CT; 8) Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA; 9) Department of Psychiatry, Boston University School of Medicine, Boston, MA; 10) Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 11) Psychiatry Service, VA San Diego Healthcare System, San Diego, CA; 12) Department of Psychiatry, University of California San Diego, San Diego, CA; 13) Departments of Genetics and Neuroscience, Yale University School of Medicine, New Haven, CT.

Posttraumatic stress disorder (PTSD) is a major problem among the veteran population and presents treatment challenges. The Veterans Affairs (VA) Million Veteran Program (MVP) is building one of the world's largest medical and genetic information databases, currently >570,000 consented participants. ~350,000 enrollees have genotype information available, linked to VA EHR data and questionnaire responses—the largest current sample for studying PTSD-relevant traits. We report here results from a project underway to identify genetic risk factors relevant to PTSD and related traits. PTSD symptoms are categorized into 4 major symptom clusters: intrusive re-experiencing of the trauma, avoidance of trauma-associated stimuli, negative trauma-associated cognitions and mood, and alterations in arousal or reactivity. We conducted a GWAS on the re-experiencing symptom cluster score based on a sum of 5 items from the PTSD Checklist (recurrent intrusive thoughts/dreams/flashbacks of trauma; emotional or physiological response to reminders of trauma), total score, 5-25. After data cleaning, 146,660 European-Americans (EAs) and 19,983 African-Americans (AAs) were retained. In the EAs, 8 distinct common-variant genomewide-significant (GWS) regions were identified—three with significance >5x10E-10. These latter regions map to chrom. 3 – lead SNP rs2777888 (2.1E-11), gene *CAMKV*, same SNP previously implicated in "age at first birth"; chrom. 17 – lead SNP rs2532252 (4.5E-10), closest to *KANSL1* but within a long high-LD region that also includes *CRHR1* (corticotropin releasing hormone receptor 1); and chrom. 18 – lead SNP rs2123392 (5.4E-11), at *TCF4,* previously GWS-associated to schizophrenia. Other significant associations were observed at *KCNIP4, HSD17B11, MAD1L1*, and *SRPK2*. There were no GWS associations in the smaller AA sample. PrediXcan was used to calculate gene-level test statistics across 44 tissues and identified 30 significant genes after Bonferroni correction. Using LD score regression and tissue and cell type-specific annotations (GenoSkyline-Plus), many tissues/cell types showed significant enrichments in functionally annotated regions with brain anterior caudate being most significant (7.4E-06). From comparison between summary statistics from this study and those from published GWAS of 55 traits, 16 traits showed statistically significant evidence of genetic correlations, including neuroticism (3.7E-50) and depression (3.8E-45).

22

Identification of genome-wide significant shared genomic segments in large extended Utah families at high risk for completed suicide. *H. Coon[1], T.M. Darlington[1], W.B. Callor[2], E. Ferris[3], A. Fraser[4], Z. Yu[4], N. Williams[1], S.E. Crowell[5], L. Jerominski[1], D. Cannon[1], K.R. Smith[4], B. Keeshin[6], A.V. Bakian[1], E. Christensen[2], N.J. Camp[7], D. Gray[1].* 1) Psychiatry Department, University of Utah School of Medicine, Salt Lake City, UT; 2) Utah State Office of the Medical Examiner, Utah Department of Health, Salt Lake City, UT; 3) Department of Neurobiology & Anatomy, University of Utah School of Medicine, Salt Lake City, UT; 4) Pedigree & Population Resource, Huntsman Cancer Center, Salt Lake City, UT; 5) Department of Psychology, University of Utah, Salt Lake City, UT; 6) Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, UT; 7) Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT.

Introduction. Suicide is the 10th leading cause of death in the US. While environmental variables have undeniable impact, evidence suggests that genetic factors play a major role in completed suicide, with estimates of heritability of ~45%. Genetic risk is unlikely to be consistent across previously studied demographic groups or across psychiatric conditions known to be associated with suicide. Studies have also focused on a variety of outcomes, from suicidal ideation/behaviors to completed suicide; genetic risks for these outcomes may not overlap. Finally, candidate gene and GWAS studies have statistical limitations. Our study addresses many of these difficult study design issues. **Methods.** We have >4,000 DNA samples from completed suicide cases through a long collaboration with the Utah Medical Examiner. We have linked the records from these samples to the Utah Population Database which includes genealogies of founding pioneers of Utah, demographic data, and medical information on over 8 million individuals. This linking has resulted in large extended families (7-9 generations) with significant familial risk of completed suicide. Familial aggregation across distant relatives minimizes effets of shared environment. The families provide more genetically homogeneous risk groups, and magnify genetic risks through familial repetition, increasing power. We have used the Shared Genomic Segments method to identify genomic regions that have high likelihood of harboring variants leading to risk. **Results**. We analyzed DNA from 215 suicide cases in 43 high-risk families, identifying 16 regions with genome-wide significance in 11 families. There are an additional 13 regions with genome-wide suggestive evidence where the region overlaps in at least 2 families, increasing significance (p-values: 4.63E-09 to <1E-16). Of the 101 genes in these overlapping regions, 6 have been previously associated with suicide risk (*RGS18, BRINP3, RHEB, CDK5, CTNNA3,* and *HTR2A*); only 1 with specific suicide risk was expected by chance. **Conclusion**. The regions found provide a new source of candidate genes and gene pathways. The genome-wide significant regions may reveal rare variants specific to a family, while regions that overlap across families may reveal more common risk variants. Follow-up of specific variants in these regions using WGS in 57 of these cases is underway. Knowledge of specific genetic vulnerabilities could result in better understanding of underlying biological mechanisms.

23

Genetic associations of DNA replication timing inferred from deep sequencing of 106 human embryonic stem cell lines. *Q. Ding[1], C. Charvet[1], C.J. Hsiao[2], X. Zhu[3], F.T. Merkle[4], R.E. Handsaker[5,6], S. Ghosh[5,6,7], K. Eggan[5,7], S.A. McCarroll[5,6], M. Stephens[2,3], Y. Gilad[2], A.G. Clark[1], A. Koren[1].* 1) Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA; 2) Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA; 3) Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA; 4) Wellcome Trust - Medical Research Council Institute of Metabolic Science, University of Cambridge, Cambridge, UK; 5) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; 6) Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; 7) Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

DNA replication is a fundamental biological process that is tightly regulated. The temporal dimension of replication, known as DNA replication timing, specifies when each genomic locus replicates during the S phase. We previously found that replication timing is variable among individuals and is at least partially determined by DNA sequence. To further understand the genetic causes and consequences of DNA replication timing, we deep-sequenced 106 proliferating cultures of human embryonic stem cell lines (hESCs) of European ancestry. Analysis of DNA copy number fluctuations along chromosomes resulted in high-resolution replication timing profiles that were highly consistent across individuals yet revealed clear variation at hundreds of genomic sites. By comparing DNA replication timing to sequence variation in these cell lines, we mapped 642 replication timing quantitative trait loci (rtQTLs) at a 10% false discovery rate. The majority of rtQTLs mapped in *cis* to the sites of DNA replication origins, suggesting that human replication origin activity is heavily influenced by sequence determinants. rtQTLs were enriched for DNase I hypersensitivity sites, active chromatin states including active transcription start sites and enhancers, and active chromatin marks (e.g. H4K12ac and H3K4me2). rtQTLs were further enriched for several transcription factor binding sites and motifs, notably POU5F1 (Oct4), a central pluripotency factor required for stem cell self-renewal. Finally, rtQTLs overlapped GWAS hits that suggested an influence of DNA replication timing on adult height, age at natural menopause, and the risk of Amyotrophic Lateral Sclerosis (ALS). This study marks the most comprehensive exploration to date into genetic associations of human DNA replication timing.

24

Mutational origins and pathogenic consequences of multinucleotide mutations in 6,868 trios with developmental disorders. *J. Kaplanis[1,2], M.E. Hurles[1] on behalf of the DDD study.* 1) Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom; 2) Cambridge University, Cambridge, United Kingdom.

In genomic analyses we often consider single nucleotide variants (SNVs) as independent mutational events however it has been estimated that ~2% of de novo SNVs appear as part of a clustered mutation that create multinucleotide variants (MNVs). MNVs are an important source of genomic variability as they are more likely to cause a greater consequence change than a SNV. This has important implications in disease as well as evolution. We identified 70,000 MNVs and ~300 de novo MNVs in 6,868 exome sequenced trios from the Deciphering Developmental Disorders Study which consists of families with severe developmental disorders. De novo mutations are a major cause of developmental disorders. We analysed the mutational spectra of these MNVs for patterns that might inform their mutational origin. We observed mutational signatures associated with DNA polymerase zeta, an error-prone translesion polymerase, and signatures associated with APOBEC, a family of DNA deaminases. To examine the consequences of these MNVs we looked at those where two mutations occurred within the same codon. We found 97% of MNVs in the same codon result in a missense mutation. Most of these missense MNVs create an amino acid that could not have been created by a single base pair change. These amino acids are on average more different physicochemically from the original amino acid compared to those that could have been created by a SNV. We found that missense MNVs are more depleted in highly constrained genes (pLI>=0.9) compared to missense SNVs and are under stronger purifying selection. We also observed that de novo MNVs are significantly enriched in genes previously associated with developmental disorders. This demonstrates that MNVs can be more damaging than SNVs even with an equivalent consequence annotation and are an important variant type to consider in relation to human disease.

25

Characterization of the noncoding regulatory landscape within human-specific duplicated regions. *P. Carmona Mora[1,2], C.J. Shew[1], E. Ha[1], M.Y. Dennis[1,2,3].* 1) Genome Center, University of California-Davis, Davis, CA; 2) MIND Institute, University of California-Davis, Sacramento, CA; 3) Department of Biochemistry & Molecular Medicine, University of California-Davis, Davis, CA.

Human-specific segmental duplications (HSDs, or regions >1 kbp at >98% sequence identity), which arose uniquely in the human lineage, are known drivers of evolution and diversity based on their ability to create novel genes with innovative functions. HSDs also contribute to 'genomic hotspots'—by sensitizing regions to deletions and duplications—which can cause neurodevelopmental disease, including autism spectrum disorder (ASD), epilepsy, schizophrenia, and intellectual disability. Until recently, many HSDs were erroneously assembled in human reference builds due to high sequence similarities. Considerable effort has gone into identifying novel genes created by these duplications, while noncoding elements, that can alter expression of multiple genes, have largely been ignored. To address this, we performed targeted assessment in HSD regions of the chromatin marks H3K27ac and H3K4me1 as well as enhancer-associated transcripts using publicly available ChIP-seq and RNA-seq datasets from human, chimpanzee and rhesus macaque lymphoblasts and three brain regions. By comparing with non-human primates we distinguished ancestral regulatory signals with novel ones from human duplications. We identified over 200 enhancers within HSDs originally dismissed in the public datasets, including cerebellum-specific enhancer of neuronally-implicated *SRGAP2*. We increased in nine-fold the overall enhancer discovery rate and confirmed that regulatory elements are duplicated along with HSD genes. Interestingly, duplication of such elements was seen in regions associated to neurodevelopmental disorders and ASD, including Williams-Beuren syndrome, Dup15q, 1q21.1, 2q21.1 and other loci. Using a capture Hi-C method in human and non-human primate cells, we identified differences of novel regulatory elements with genome-wide contacts between ape species, and are correlating results with public RNA-seq data. We premise that HSD regulatory elements have affected the expression of nearby genes contributing to novel human neurobehavioral traits when these duplications arose in our hominin ancestors. Such findings will allow us to understand how chromatin states were shaped upon duplications, and moreover, to identify novel candidate pathways altered in neurological disorders.

26

Single-molecule mapping of complex genomic regions across 26 human populations reveals population specific variation patterns. *P. Kwok[1,2,3], C. Chu[2], A. Hastie[4], E. Lam[4], A. Leung[5], L. Li[5], J. McCaffrey[6], M. Levy-Sakin[2], Y. Mostovoy[2], S. Pastor[6], A. Poon[1], R. Rajagopalan[6], J. Sibert[6], W. Wang[4], E. Young[6], H. Cao[4], T. Chan[5], K. Yip[5], M. Xiao[6].* 1) Institute for Human Genetics; 2) Cardiovascular Research Institute; 3) Department of Dermatology, University of California, San Francisco, CA; 4) Bionano Genomics, Inc., San Diego, CA; 5) Chinese University of Hong Kong, Shatin, Hong Kong, SAR; 6) Drexel University, Philadelphia, PA.

Whole-genome analysis of structural variations has been challenging because short-read sequencing contigs cannot span across repetitive elements and due to the diploid nature of the human genome. We have produced *de novo* genome maps[1] based on long DNA molecules for 156 individuals from all 26 human populations of the 1000 Genomes Project. These long molecules (all >150 kb and many >300 kb) allow us to identify all structural variations (SVs) >3 kb and characterize regions that are typically inaccessible with other methods. For example, by clustering genome map data that did not align to the reference genome hg38, we were able to identify and resolve multiple acrocentric chromosome arms. We also detected subtelomeric sequences that varied by population and resolved the haplotypes of a complex, hyper-variable segmental duplication region (22q21) that remains unresolved in the reference genome. By combining genome map data with barcoded "linked-read" whole genome sequencing data[2], we detected both the origins and the insertion sites of large insertions. Finally, by performing multiple alignments, we were able to visualize the structural patterns in different populations over very long genomic regions. In this presentation, we show the power of long single-molecule mapping in resolving complex SVs in the human genome and provide new human population based references for complex regions that are associated with important human diseases. The population specific SV patterns also shed light on the origins of the complex regions and provide another way to trace the spread of human populations across the world. References: 1. Mak AC et al. Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays. Genetics. 2016; 202:351-62. 2. Mostovoy Y et al. A hybrid approach for de novo human genome sequence assembly and phasing. Nat Methods. 2016;13:587-90.

27

Evolutionarily young LINE elements initiate recurrent DNA breaks forming different-sized CNVs via both NAHR and microhomology-mediated DNA replication mechanisms. *P. Szafranski[1], E. Kośmider[1,2], J. Wambach[3], L. Currie[4], S. Parkash[4], G.K. Suresh[5], M.T. Harting[6], M.D. Weaver[6], A.M. Khan[6], N. Tatevian[6], A.M. Breman[1,7], C.A. Shaw[1,7], E. Popek[8], C.R. Beck[1], A. Gambin[2], P. Stankiewicz[1,7].* 1) Dept. of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland; 3) Edward Mallinckrodt Department of Pediatrics, Washington University School of Medicine, St. Louis, MO; 4) Maritime Medical Genetics Service, IWK Health Centre, Halifax, Nova Scotia, Canada; 5) Department of Pediatrics, Section of Neonatology, Texas Children's Hospital, Baylor College of Medicine, Houston, TX; 6) UT Health McGovern Medical School, Houston, TX; 7) Baylor Genetics, Houston, TX; 8) Dept. of Pathology and Immunology, Baylor College of Medicine, Houston, TX. *Equal contribution.

Highly repetitive sequences such as LINE-1 and *Alu* elements contribute significantly to structural variation in human genomes. These transposable elements can alter the genome by insertion into new loci or by mediating non-allelic homologous recombination (NAHR), resulting in CNVs or balanced chromosomal translocations. Our genomic analyses of 49 CNV deletions in the *FOXF1* locus on 16q24.1, causative for a lethal neonatal lung disorder alveolar capillary dysplasia with misalignment of pulmonary veins, revealed that two-thirds of their breakpoints are located within *Alu* or LINE-1 elements. Interestingly, distal breakpoints of nine different-sized CNV deletions map within one of two neighboring and evolutionarily young LINE-1 elements, L1PA2 (5 cases) and L1PA3 (4 cases). Six of these deletions likely arose via NAHR and three may have resulted from microhomology-mediated replication-based repair. To better assess the role of LINE elements in genome-wide non-NAHR-mediated CNV formation and to identify other potential LINE genomic instability hotspots, we bioinformatically queried the Baylor Genetics (BG) database of 25,550 CNVs detected and reported in 19,537 patients referred for clinical chromosomal microarray analysis. Interestingly, we identified a three-fold LINE-1 enrichment for CNV deletions whose one breakpoint region harbor LINE-1 element and the other is devoid of any putative NAHR substrate. Moreover, intersection of 146 intact, full-length LINE elements (PMID: 27924012) with these CNVs revealed 12 and 16 LINE-1s involved in at least two different-sized non-homology driven CNV deletions or duplications, respectively. Importantly, 25 LINE-1 elements likely catalyzed both nonrecurrent CNV deletions and nonreciprocal duplications. For example, we observed breakpoint clustering of multiple CNV deletions and duplications from BG and DECIPHER databases around one L1HS element within the *IMMP2L* gene on 7q31.1. Using customized high-resolution array CGH and Sanger sequencing, we found that breakpoints of three different-sized CNV deletions map within or just adjacent to this L1HS. Our data demonstrate that in addition to transposition and NAHR, LINE-1 elements can substantially destabilize the human genome by initiating recurrent DNA breaks. We propose that intact full-length LINE-1 elements may propagate DNA breaks, leading to genomic instability hotspots and forming subsequent nonrecurrent rearrangements, including clinically relevant CNVs.

28

Human immune defense mechanisms drive rapid genome evolution in vaccinia virus. *T. Sasani, K. Rogers-Cone, R. Layer, N. Elde, A. Quinlan.* Department of Human Genetics, University of Utah, Salt Lake City, UT.

Viruses are locked in conflict with host organisms and rapidly adapt to combat antiviral host responses. For example, the vaccinia virus (VACV) genome encodes two proteins, E3L and K3L, that each disrupt key elements of the human immune response to viral replication and propagation. In prior work, populations of VACV that lacked the E3L gene were shown to rapidly adapt under selective pressure by duplicating K3L in tandem gene arrays. Additionally, these populations began to accumulate K3L^{H47R} single-nucleotide mutations, which appeared to confer a fitness benefit comparable to duplication of wild-type K3L. The interplay between these two genomic adaptations has remained mysterious, largely because short-read sequencing technologies are unable to sequence through tandem arrays of K3L duplications. In this study, we utilized the Oxford Nanopore (ONT) long-read platform to characterize K3L copy number and K3L^{H47R} accumulation in VACV populations under selective pressure during successive passages in HeLa cell lines. By sequencing the genomes of experimentally evolved VACV, we gained insight into two key mechanisms of vaccinia adaptation. Using long reads, we directly characterized viral genomes harboring up to 21 tandem copies of K3L, likely a result of recombination-driven gene expansion. Interestingly, population distributions of K3L copy number are nearly identical following 10, 15, and 20 passages (P10, P15, and P20), suggesting that viral recombination can generate stable copy-number increases at the population level. We also discovered that the K3L^{H47R} allele spreads rapidly over the course of vaccinia evolution, from a frequency of 12% at P10 to 90% by P20. Using long ONT sequencing reads, we were able to move beyond a population-level view of K3L^{H47R} accumulation, and tracked the spread of the variant *within* tandem K3L arrays in individual viral genomes. We determined that K3L^{H47R} rapidly homogenizes within these arrays during virus evolution. After 10 passages through HeLa cells, nearly all tandem arrays are composed of wild-type alleles, but by P20 these arrays contain almost entirely K3L^{H47R} alleles. Using long read sequencing methods, we demonstrate that human immune defense mechanisms can drive rapid genome evolution in vaccinia virus, in the form of both copy number and single nucleotide variation. These observations reveal a new and exciting facet of viral evolution during host-pathogen conflict.

29

Integrated sequence technology approaches to genomic diagnosis of birth defects. *K. Meltz Steinberg[1,2], J. Wambach[3], D. Wegner[3], D.E. Baldridge[3], D. Spencer[1,2], F.S. Cole[3].* 1) McDonnell Genome Institute, Washington University, St. Louis, MO; 2) Department of Medicine, Washington University, St. Louis, MO; 3) Department of Pediatrics, Washington University, St. Louis, MO.

Structural birth defects affect approximately 3% of US born infants and are a common cause of infant mortality. Although maternal medications, fetal infections, and uterine anatomy can cause birth defects, a majority is associated with inherited or *de novo* genetic variants that disrupt highly regulated developmental pathways. Associated pathogenic variants are frequently rare or private due to reduced reproductive fitness and are underrepresented in adult genomic databases. To improve diagnostic success through genotype-driven discovery of missing heritability in infants and children with structural birth defects, we used multi-algorithm functional prediction strategies and integration of whole genome and transcriptome sequencing data from 37 affected infant/parent trios recruited from St. Louis Children's Hospital/Washington University in St. Louis. We confirmed the sensitivity of our approach by identifying, in a blinded fashion, pathogenic variants in four trios previously diagnosed by clinical exome sequencing. The causal coding variants from all four families are either completely novel or extremely rare in the Exome Aggregation Consortium data. For example, in an infant with Ebstein's anomaly, clubfeet, craniofacial dysmorphism, single fused kidney and undescended testes, we identified a novel *de novo* structural nucleotide variant (SNV) in *CDK13* which expands the syndromic congenital heart disease phenotype recently associated with this gene locus. We then identified and characterized SNVs and small insertions/deletions, copy number (CNV) and structural variants (SV), and variants that disrupted transcription in the remaining 33 infant/parent trios. To investigate usefulness of 10X Genomics linked read technology, we also compared structural variant discovery with our whole genome sequencing approach in 16 affected infants. We found that 10X Genomics technology identified all CNVs observed in whole genome sequencing of trios without access to parental sequence and additionally discovered more complex SVs such as inversions that impact coding genes. We conclude that integrated sequencing technology improves diagnostic success in infants and children with structural birth defects.

30

New strategies for analyzing exomes from patients with rare and unknown disorders. *K. Schmitz-Abe[1,2,3], P. Agrawal[1,2,3].* 1) New Born Department, Children's Hospital Boston, Boston, MA; 2) Division of Newborn Medicine, Boston Children's Hospital, Boston, MA, USA; 3) Harvard Medical School, Boston, MA, USA.

The Gene Discovery Core (GDC) of The Manton Center for Orphan Disease Research at Boston Children's Hospital (BCH) has been enrolling families, mostly trios, with rare and unknown disorders for the last six years. Of those enrolled patients, whole exome sequencing (WES) has been performed on 1111 individuals from 428 families without an initial clinical diagnosis. We aim to determine the genetic and molecular basis on those by re-analyzing the WES data. Reportedly, WES is able to determine the cause of disease in a third to half of the patients with genetic rare diseases. One of the reasons for the relatively low yield may be that we are missing important variants in the non-coding regions or copy number variants not detected by WES. Another important reason is the high number of rare coding pathogenic variants in candidate genes per family with scant information about them in the literature. This makes linking the genetic variant with phenotype extremely challenging, especially when we try to determine the functional effects using cellular or animal modeling. To overcome this issue, we re-processed the entire WES data using our own in-house bioinformatics and analysis pipeline (Variant Explorer Pipeline), sub-grouped the families with similar phenotypes and interrogate the data for candidate genes. Genes carrying mutations in several non-overlapping phenotypes were likely to be non-pathogenic and were removed from the candidate gene list. Using this approach, we have reduced the number of potential candidate genes by 60%. We are currently reviewing the reanalyzed data to determine the true candidates that can go for the functional determination and Matchmaker stage. We highly recommended processing all WES data from patients with variable phenotype using the same pipeline.

31

The Human-Mouse Disease Connection (HMDC) Portal: Comparing mouse and human disease data to enable discovery. *C. Smith, S. Bello, J. Kadin, J. Richardon, Mouse Genome Informatics Team.* Mouse Genome Informatics, Jackson Laboratory, Bar Harbor, ME.

The now routine sequencing of exomes and genomes of patients provides the opportunity to identify potential genetic causes of hereditary human diseases. However, the large number of genetic variants uncovered from a patient presents challenges in identifying the causal gene or genomic regions. The laboratory mouse, as the most studied of mammalian model systems, provides unique insights into the dissection of genetic mechanisms of human disease. Comparative phenotyping and directed gene mutation can aid in identification of candidate gene mutations and development of mouse models that recapitulate specific human genetic mutations, especially for the study of rare disorders in which data for multiplex kindreds may be lacking, or for using mouse embryos to study human congenital diseases. The Human-Mouse Disease Connection is a translational tool (www.diseasemodel.org) designed for exploring and comparing human and mouse phenotypes and their associations with known human diseases, and to provide rapid access to mouse model resources and supporting references. Searches can be initiated based on human or mouse data using one or more parameters, including genes, genomic locations, or phenotypes or disease terms. Results are displayed initially as a visual color-coded grid comparing phenotypes and diseases associated with human and mouse orthologs; or users can choose to view data in tabular format based on genes or on diseases that are responsive to the search parameters. New features and data include searches by Disease Ontology terms to group and display disease classes, and the incorporation of human disease and phenotype relationships from Orphanet in addition to the existing OMIM disease-to-phenotype relationships from the Human Phenotype Ontology (HPO) project. Congenital diaphragmatic hernia (CDH), a protrusion of abdominal viscera into the thorax through an abnormal opening or defect in that is present at birth, is a common and life--threatening birth defect in humans. The molecular etiology of CDH is incompletely understood. We will show examples of comparisons of candidate genes from rare variants from CDH patients and mouse model data for these genes. By gathering additional functional and phenotype information from mouse models, high priority genes that contribute to the CDH phenotype can be identified and studied further.

32

Gene discovery via direct-to-family engagement using MyGene2. *J.X. Chong[1], R. Cornejo, Jr.[1], A.G. Shankar[1], A.E. Tattersall[1], D.A. Nickerson[2], M.J. Bamshad[1,2,3], University of Washington Center for Mendelian Genomics.* 1) Department of Pediatrics, University of Washington, Seattle, WA; 2) Department of Genome Sciences, University of Washington, Seattle, WA, USA; 3) Division of Genetic Medicine, Seattle Children's Hospital, Seattle, WA, USA.

Most rare diseases are Mendelian conditions (MC), which means that mutation(s) in a single gene can cause disease. Approximately 70% of families with a putative MC who undergo clinical exome sequencing/whole genome sequencing (ES/WGS) remain undiagnosed despite the increasing pace of gene discovery. Some families have used social media in an attempt to "match" with other families, but this process is highly inefficient and requires technical expertise. We recently launched a website, MyGene2 (http://www.mygene2.org), to streamline this process and provide equitable and timely access to genetic and phenotypic information for MCs to families, clinicians, and researchers. MyGene2 enables families with MC to publicly share health and genetic information and allows clinicians and researchers to publicly share de-identified genetic data on behalf of their patients and/or research participants. In the past year, >1,000 MyGene2 profiles have been created. Families contributed 25% of these, including >20 that describe a novel MC and/or candidate gene. Families in whom prior ES/WGS did not identify a causal variant can share their VCFs with other MyGene2 users and the University of Washington Center for Mendelian Genomics (UW-CMG) for re-analysis. Families can also apply for ES via MyGene2 and may be approved for either self-pay CLIA-certified research ES or for free research ES via the UW-CMG. All candidate genes, and optionally VCFs, are shared with families and the public through MyGene2. The impact of MyGene2 on the rare disease community is illustrated by several stories. For example, the family of a child with seizures and developmental delay elected to self-pay for ES because clinical ES was unavailable. Analysis of their data identified a pathogenic variant in *GNB1*, a known gene underlying epilepsy that had been reported only recently. Within one month of receiving this result, the family had identified and enrolled in ongoing genotype-phenotype studies of their condition and made contact with another family with the same causal variant. The family also used tools made available via MyGene2 to explore their VCFs and independently identified an actionable secondary result. Our early experience with engaging directly with families suggests that families who are highly motivated to obtain a diagnosis despite being unable to obtain ES/WGS through traditional routes are a rich source of new discoveries and understanding of genotype-phenotype relationships in MC.

33

Initiative on Rare and Undiagnosed Diseases in Pediatrics (IRUD-P) in Japan: Recent achievement and statistics. *Y. Matsubara[1], K. Hata[1], T. Kaname[1], K. Kosaki[2], IRUD-P consortium.* 1) Research Institute, National Center for Child Health and Development (Japan), Tokyo, Japan; 2) Center for Medical Genetics, Keio University School of Medicine, Tokyo, Japan; 3) IRUD, Japan.

The Initiative on Rare and Undiagnosed Diseases (IRUD) is a national consortium designed to help patients and their families suffering from rare and undiagnosed disease conditions in Japan. The project started in July 2015. The aims of the project are to make diagnosis on patients with rare and undiagnosed diseases, to construct their genome database with clinical information, and to make banking system of precious specimens. The pediatric version of IRUD (IRUD-P) is coordinated by dual centers, the National Center for Child Health and Development (NCCHD) and Keio University. The IRUD-P developed a nation-wide network for patient recruitment involving 17 regional core clinical centers, and mainly performed whole exome sequencing on children with undiagnosed diseases and their parents. Till now, more than 1,500 patients, who passed the first screening in the core clinical centers, were consulted to the IRUD-P centers. Specimens accompanied with medical information (n=4,600) were collected from patients and their families, mainly in trios, and sent to the centers. Of 827 patients analyzed, genetically confirmed diagnosis in 271 patients (diagnostic yield was 32.8%). Diagnostic yield of family-trio analysis and proband-only or proband-one parent analysis were 45.0% and 20.8%, respectively. Nine patients were found to have pathogenic variants in novel genes. Patient cohort represents children with a wide range of undiagnosed disorders (malformation syndromes 33.6%, neuromuscular disorders 26.2%, cardiovascular diseases 21.5%, skeletal and connective tissue diseases 3.6%, renal diseases 2.2%, liver diseases 1.7%, others 11.3%). Modes of inheritance in diagnosed patients were autosomal dominant 77.4%, autosomal recessive 9.1% (homozygote 2.3%, compound heterozygote 6.8%), X-linked 12.0% and others 1.5%. For the remaining undiagnosed patients, the IRUD-P consortium has started data-sharing network system, IRUD-Exchange. The IRUD-P also built international collaboration and data sharing on rare and undiagnosed diseases.

34

Contribution of novel disease gene discovery to clinical diagnosis and management. *J.E. Posey[1], J.A. Rosenfeld[1], Z.H. Coban Akdemir[1], S. Jhangiani[2], T. Harel[1,3], M.K. Eldomery[1,4], A. Stray-Pedersen[1,5], I.K. Chinn[6,7], S. Lalani[1,8], P. Stankiewicz[1,8], M. Walkiewicz[1,8], P. Liu[1,8], M. Leduc[1,8], L. Meng[1,8], F. Xia[1,8], X. Wang[1,8], R. Xiao[1,8], C.A. Shaw[1,8], C.M. Eng[1,8], D. Muzny[2], R.A. Gibbs[1,2], E. Boerwinkle[2,9], V.R. Sutton[1], S.E. Plon[1,2,6,10,11], J.S. Orange[7,11,12,13], Y. Yang[1,8], J.R. Lupski[1,2,6,11].* 1) Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX, USA; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA; 3) Department of Genetic and Metabolic Diseases, Hadassah-Hebrew University Medical Center, Jerusalem, Israel; 4) Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA; 5) Norwegian National Unit for Newborn Screening, Women and Children's Division, Oslo University Hospital, 0424, Oslo, Norway; 6) Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA; 7) Center for Human Immunobiology, Texas Children's Hospital, Houston, TX, USA; 8) Baylor Genetics, Baylor College of Medicine, Houston, TX, USA; 9) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA; 10) Texas Children's Cancer Center, Texas Children's Hospital, Houston, TX, USA; 11) Department of Pediatrics, Texas Children's Hospital, Houston, TX, USA; 12) Department of Pathology and Immunology, Baylor College of Medicine, Houston, TX, USA; 13) Dan L Duncan Cancer Center, Baylor College of Medicine, Houston, TX, USA.

The broader availability of genomic sequencing technologies has fueled novel disease gene discovery. The Baylor Hopkins Center for Mendelian Genomics (BHCMG) is one of many large-scale sequencing programs funded by the National Institutes of Health to elucidate genotype-phenotype relationships in Mendelian disease. We explored the contribution of disease gene discoveries in the BHCMG to clinical diagnostics and medical management. To date, a total of 218 novel or candidate disease genes were identified at Baylor College of Medicine (BCM). Review of approximately 10,000 sequential cases referred to the Baylor Genetics diagnostic laboratory for non-cancer phenotypes yielded 3,214 cases for which a molecular diagnosis was achieved through WES. Of these, 3.4% (109/3214) of molecular diagnoses explaining part or all of the reported clinical phenotype involved novel or candidate disease genes identified at BCM. The 109 molecular diagnoses involved 29 (13.3%, 29/218) of the novel or candidate disease genes identified. *PURA* (mental retardation, autosomal dominant 31, OMIM 616158) and *TANGO2* (metabolic encephalomyopathic crises, recurrent, with rhabdomyolysis, cardiac arrhythmias, and neurodegeneration, OMIM 616878) contributed to over one-third (39/109; 24 and 15 cases respectively) of molecular diagnoses reported. Molecular diagnoses had significant implications for medical management of 21 unrelated individuals (19.3%, 21/109), including the potential for immunologic cure through hematopoietic stem cell transplant in 1 individual diagnosed with immunodeficiency due to phosphoglucomutase deficiency (*PGM3*, OMIM 615816), recommended surveillance for life-threatening arrhythmias for 15 individuals diagnosed with recurrent metabolic encephalomyopathic crises with rhabdomyolysis, cardiac arrhythmias, and neurodegeneration (*TANGO2*, OMIM 616878), and cancer surveillance for 5 individuals diagnosed with congenital heart defects and skeletal malformations due to *ABL1* variants. We find that 13.3% of novel and candidate disease gene discoveries through the BHCMG have already contributed to molecular diagnoses reported by a clinical diagnostic laboratory, impacting clinical management for at least 21 individuals. These findings suggest that the first 5 years of the BHCMG have already impacted both clinical diagnostics and medical management, demonstrating unequivocally a successful 'bench-to-bedside' approach.

35

Shedding light into voltage-gated sodium channel associated neurodevelopmental disorders. *D. Lal[1,2,3], J. Du[3], C. Dühring Fenger[4], E. Perez-Palma[3], A.J. Campbell[1], A. Allen[1], D. Baez-Nieto[1], H.R. Wang[1], J. Cottrell[1], F. Wagner[1], J.Q. Pen[1], H. Stammberger[5], I. Helbig[10], P. DeJonge[5], S. Weckhuysen[5], B. Sheidley[6], S. Zuberi[7], A. Poduri[8], S. McCarroll[8], A. Brunklaus[7], R.S. Møller[3], M. Daly[1,2], A. Palotie[1,2,9].* 1) Broad Institute of Harvard and M.I.T., Cambridge, MA, MA., USA; 2) Analytical Translational Genetics Unit,Massachusetts General Hospital focuses, Boston, MA, USA; 3) Cologne Center for Genomics, University of Cologne, Germany; 4) The Danish Epilepsy Centre, Dianalund, Denmark; 5) Neurogenetics Group, Center for Molecular Neurology, VIB, Antwerp, Belgium; 6) Boston Children's Hospital, Boston, MA, US; 7) The Paediatric Neurosciences Research Group, Royal Hospital for Sick Children, Glasgow, United Kingdom; 8) Department of Genetics, Harvard Medical School, Boston,US; 9) Institute for Molecular Medicine Finland (FIMM), University of Helsinki,Helsinki, Finland; 10) Division of Neurology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pa., USA.

Genetic variants affecting the four voltage-gated sodium channel encoding genes *SCN1A*, *SCN2A*, *SCN3A* and *SCN8A* are associated with 1-2% of patients with neurodevelopmental disorders (NDDs). The underlying biology how variants in these genes cause a heterogeneous group of NDDs is not understood limiting therapeutics strategies and drug development. We report on the largest cohort to date including >520 patients with neurodevelopmental disorders positive for either rare single nucleotide or large copy number variants in the four selected genes. We perform a range of novel genotype-phenotype analyses and correlate our observations with self-generated large scale single cell human cortex RNA sequencing data and public RNA tissue sequencing studies at different developmental stages. In addition, we *in silico* transferred patient and control variants (from the gnomAD database) across the different channels and performed local enrichment analyses to prediction gain and loss of function sensitive sites of the proteins. We subsequently validated the predictions using >40 patch clamp experiments. We show that for missense variant carrying patients the age of seizure onset follows gene expression order of *SCN2A*, *SCN8A* and *SCN1A*. Missense and truncating variants as well as full deletions all cause infantile onset epilepsy “Dravet syndrome” in >95% of *SCN1A* variant carriers. Missense variants in *SCN2A*, *SCN8A* near or in the pore are associated with early onset epilepsies whereas truncating variants and deletions are associated with developmental disorders with or without later onset seizures, predominantly autism. We bioinformatically predict and molecular validate across all four channels that gain-of-function (GoF) sensitive sites are near the pore region whereas loss-of-function sensitive regions are extracellular located. Full gene duplications of *SCN2A* and *SCN8A* mimic the early onset epilepsy phenotypes associated with GoF missense variants in the same genes. The contrasting disease biology of *SCN1A* vs. *SCN2A*, *SCN8A* can be explained by higher expression ratios at inhibitory vs. excitatory neurons. In summary, we present a novel model explaining the heterogeneity of sodium channel associated neurodevelopmental disorders. The novel developed model and ability to predict functional consequences of variants set the foundation of precision medicine in sodium disorder pathologies, have direct consequences in clinical practice and will facilitate drug development.

36

CFTR-targeted therapy for a subset of splice-site and nonsense variants that allow protein production. *M.J. Pellicore[1], T.A. Evans[1], E. Davis[1], S.T. Han[1], A. McCague[1], A. Joynt[1], Z. Lu[1], Y. Akhtar[1], N. West[1], C. Merlo[1], K.S. Raraigh[1], P.R. Sosnay[1], C. Cotton[2], G.R. Cutting[1], N. Sharma[1].* 1) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) Case Western Reserve University, Cleveland, OH.

CFTR-specific small-molecule therapies are not prescribed to CF individuals carrying nonsense or splice-site variants because such variants are not expected to produce CFTR protein. However, we speculated that variants which permit expression of CFTR, even aberrantly or in reduced quantity, may respond to targeted therapy. Consequently, we studied naturally-occurring disease-causing splice-site and nonsense variants to assess CFTR protein production and targeted drug response. To test our hypothesis, we selected 56 *CFTR* variants (50 intronic and 6 exonic) based on splicing predictions by CryptSplice, a variant annotation program. mRNA studies in HEK293 cells revealed that 50 of the 56 variants caused aberrant splicing. Missplicing introduced premature termination codons in 34 of the 50 variants, leading to nonsense-mediated decay (NMD), as in nasal cells of a c.3717+5G>A/F508del individual which showed 89% reduction in misspliced *CFTR* transcript relative to F508del transcript. Further, CFTR function did not improve upon addition of compounds that improve CFTR folding (correctors) or activate CFTR channels (potentiators). However, the remaining 16 variants generated a fraction of normally-spliced *CFTR* transcript. Functional testing completed on 4 of the 16 showed two to four-fold improvement upon corrector and potentiator treatment. RNA stability, protein processing, and protein function were assessed for 23 *CFTR* nonsense variants. We confirmed that transcript carrying a nonsense variant with multiple downstream exon junctional complexes undergoes NMD, as expected, with the exception of L88X, an N-terminus nonsense variant. RNA collected from the nasal cells of an L88X/F508del individual revealed that L88X mRNA abundance was no different than F508del. Of the 23, 12 produced no protein or core glycosylated protein. However, the remaining 11 variants produced complex glycosylated truncated protein that responded to corrector and potentiator treatments. Elucidating the molecular consequences of 79 nonsense and splice site variants in *CFTR* allowed classification of variants according to those that are candidates for FDA-approved CFTR-targeting drugs and those that will require alternative therapeutic approaches. This study demonstrates that variants predicted to disrupt gene function at the RNA level require functional assessment to optimally assign precision therapies.

37

Correlating CFTR function with key clinical features to inform targeted treatment of cystic fibrosis. *A. McCague, S.T. Han, M. Atalar, M.J. Pellicore, T.A. Evans, E.F. Davis, N. Sharma, K.S. Raraigh, P.R. Sosnay, G.R. Cutting.* Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD.

Treatment of cystic fibrosis (CF) has dramatically improved with the development of drugs that recover function of CFTR, a chloride channel expressed in epithelial tissues of the lung, pancreas and sweat gland. We addressed two key questions: 1) how does recovery of CFTR function correlate with clinical effectiveness and 2) which clinical measure has the greatest utility to gauge response to CFTR-targeted therapies? To analyze *CFTR* variants in a native cellular context, we introduced a single recombinase-targetable integration site into the genome of a human bronchial airway epithelial cell line lacking endogenous *CFTR* expression to create CF8 cells. Robust correlation was observed between mRNA expression (qRT-PCR) and function (I_{sc} measurements) (R^2=0.78) in 22 independent cell lines expressing wild-type CFTR. RNA level and chloride channel function of WT-CFTR expressed in heterologous CF8 cells was comparable to primary airway cells from healthy individuals. These results enabled normalization of variant CFTR function to WT-CFTR in a physiologically relevant range. CF8 cell lines were generated bearing 38 *CFTR* missense variants associated with a broad range of clinical features. The %WT-CFTR function was calculated for each variant and averaged across at least two independent clones (n≥3 mRNA and I_{sc} measurements for each clone). A logarithmic relationship was observed between %WT function and sweat chloride concentration (R^2 = 0.62). Inclusion of previously published data from 54 minimal to low function CFTR variants expressed in Fischer rat thyroid cells did not alter the correlation substantially (R^2=0.67, n=92). The relationship between %WT function and lung function is also logarithmic but less robust (R^2=0.11 in CF8s, R^2=0.29 with FRTs). The logarithmic relationship indicates that small functional improvements for CFTR variants with minimal residual function will result in greater clinical effect than for variants with higher levels of residual function. Furthermore, sweat chloride concentrations, as opposed to lung function measurements, provide the most useful *in vivo* measures of CFTR function across the full spectrum of CF-causing variants. Together, these results can be used to estimate clinical effect of CFTR-targeted treatments and to select which patients can benefit most from their use.

38

A CRISPR-C2c2 based therapy to target toxic RNA in microsatellite expansion diseases. *N. Zhang, T. Ashizawa.* Stanley H. Appel Department of Neurology- Neuroscience Research Program, The Methodist Hospital Research Institute, 6670 Bertner Avenue, Houston, Taxes, USA, TX77096.

Microsatellite expansion diseases encompass more than 30 incurable neurological and neuromuscular disorders, including Huntington's disease, various spinocerebellar ataxias, and myotonic dystrophy. They are characteristic of repeat expansions of 3-6 nucleotides in both coding and non-coding regions. When repeat expansion occurs in non-coding regions, the expanded RNA adopts unusual secondary structures, sequesters a myriad of RNA binding proteins, and forms insoluble nuclear RNA foci. This toxic RNA gain-of-function depletes RNA binding proteins from the cellular pool and causes disruption in essential biological processes such as transcription regulation, transcript processing, RNA translocation, protein quality control and apoptosis. As a result, microsatellite expansion diseases often manifest at a multisystem level, and elimination of toxic RNA and foci has become the main focus of many studies. One of the leading therapeutic approach to toxic RNA elimination is through the use of antisense oligonucleotides (ASOs). Often chemically modified for stability, ASOs hybridize to toxic RNA and elicit RNase H-dependent degradation. However, several drawbacks of this approach are evident. For instance, ASO "gapmers" have been shown to cause severe thrombocytopenia. Toxic doses (> 5 mg/kg) of ASO are often used for drug delivery and uptake, which can cause off-target effects. Lipid- or GalNAc-based packaging often leads to localization in the liver than the target tissue. Here, we report the development of a CRISPR-C2c2 based strategy to target and eliminate toxic RNA. The C2c2 protein from *Leptotrichia Shahii* is an RNA-guided RNase (Abudayyeh *et al*, Science, 2016). We show that the recombinant C2c2 protein is capable of degrading different RNA repeats *in vitro*: CUG repeat in myotonic dystrophy, Fuchs endothelial corneal dystrophy and Huntington disease-like 2, CAG repeat in polyglutamine diseases, GGGGCC repeat in C9orf72-amyotrophic lateral sclerosis/frontotemporal dementia, and AUUCU repeat in spinocerebellar ataxia 10, each with a specifically designed guide crRNA. We have constructed a vector encoding the entire C2c2-CRISPR-crRNA apparatus. By transiently transfecting myotonic dystrophy type 1 fibroblast cells, our preliminary data show a significant reduction in toxic RNA foci formation. By introducing tissue-specific promoters and different guide crRNAs, a wide range of neuromuscular diseases could potentially be treated by this approach.

39

The first viable mouse model of *cblC* deficiency displays growth failure and reduced survival which are rescued by hydroxocobalamin and AAV gene therapy. *J.L. Sloan[1], M.L. Arnold[1], N.P. Achilly[1], G. Elliot[2], P. Zerfas[3], V. Hoffman[3], C.P. Venditti[1].* 1) MGMGB, NHGRI, Bethesda, MD; 2) Mouse and ES Core Facility, NHGRI, Bethesda, MD; 3) Diagnostics and Research Services Branch, Office of the Director, NIH, Bethesda, MD.

Combined methylmalonic acidemia and homocysteinemia, *cblC* type (*cblC*), the most common inborn error of cobalamin metabolism, is caused by mutations in the *MMACHC* gene. MMACHC transports and processes intracellular cobalamin into the two active cofactors, 5'-adenosylcobalamin and methylcobalamin, necessary for the enzymatic reactions of methylmalonyl-CoA mutase and methionine synthase, respectively. Disease manifestations include growth failure, anemia, heart defects, neurocognitive impairment and a progressive maculopathy and pigmentary retinopathy. To explore disease pathophysiology and develop novel therapies, we created a mouse model of *cblC* using TALENs targeting exon 2 of *Mmachc* focusing our studies on two alleles: c.165_166del p.Pro56Cysfs*4 (Δ2) and c.162_164del p.Ser54_Thr55delinsArg (Δ3). We observed a decreased number of homozygous mutant pups at birth ($p<0.05$). Dissection of the pregnant dams at E18.5 showed that *Mmachc*$^{\Delta 3/\Delta 3}$ embryos were growth retarded. The median survival was less than 7 days with 90% of the mutant mice perishing before 3 weeks (Δ2 n=13; Δ3 n=42 $p<0.0001$). The weights of *Mmachc*$^{\Delta 3/\Delta 3}$ mice (n=5) were reduced relative to controls (n=23; $p<0.0001$) at 2 weeks. Mutant mice (Δ2 n=4, Δ3 n=6) recapitulated the biochemical features of *cblC*, with significantly elevated plasma methylmalonic acid, homocysteine, cystathionine and decreased methionine compared to wild type controls (n=7) ($p<0.05$). Tissue pathology revealed severe hepatic lipidosis and variable hydrocephalus in *Mmachc*$^{\Delta 3/\Delta 3}$ mutants. To explore systemic gene therapy as a treatment for *cblC*, we generated two AAV vectors: rAAVrh10-CBA-m*Mmachc* and rAAV9-CBA-h*MMACHC* that were delivered by a single intrahepatic injection (1×10^{11} GC/pup) and compared to treatment with weekly hydroxocobalamin (OHcbl) injections. *Mmachc*$^{\Delta 3/\Delta 3}$ mice treated with AAV vectors (AAVrh10 n=9, AAV9 n=11) or OHcbl (n=9) displayed dramatically improved clinical appearance with improved growth ($p<0.05$) and increased survival ($p<0.0001$), with the oldest treated mutants living beyond 1 year. In summary we developed the first viable animal model of *cblC* which recapitulates several disease manifestations including intrauterine growth retardation, decreased survival, poor growth and metabolic abnormalities. We describe successful treatment of this lethal model with both continual OHcbl injections and AAV gene transfer, a novel therapeutic approach for this disorder.

40

ZFN-mediated *in vivo* genome editing results in continuous high levels of GLA activity and effective substrate reduction in Fabry mice. *S. Pagant[1,3], M. Yasuda[1,3], M.W. Huston[2], T. Wechsler[2], S. St. Martin[2], M.C. Holmes[2], R.J. Desnick[1].* 1) Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY., USA; 2) Sangamo Therapeutics, USA; 3) co-first authors.

Fabry disease is an X-linked disorder due to the deficient activity of the lysosomal exoglycohydrolase, alpha-galactosidase A (GLA). The disease is characterized by progressive accumulation of the enzyme's substrates, globotriaosylceramide (Gb3) and globotriaosylsphinogosine (lyso-Gb3), leading to cardiac, renal, and cerebrovascular disease and early demise. Although enzyme replacement therapy is often effective for the treatment of Fabry disease, it necessitates a lifetime of biweekly infusions. Therefore, the development of a therapy with long-lasting effectiveness is desirable. Here, we report the preclinical evaluation of a liver-directed gene therapy that utilizes zinc finger nuclease (ZFN) technology to permanently insert a human *GLA* cDNA into the first intron of the murine albumin locus, thereby generating an albumin-GLA fusion protein that will be efficiently secreted into the plasma for uptake by various tissues via the mannose-6-phosphate receptor-mediated uptake system. The two obligate heterodimeric ZFNs and *GLA* donor were individually packaged into AAV vectors and intraveneously co-administered into Fabry knockout mice that have no endogenous GLA activity. Plasma GLA activities were markedly elevated by day 7 post-administration and were sustained for the two month study duration, reaching levels up to ~300-fold over those in normal mice. In the liver, heart, and kidney, GLA activities were increased up to ~145-, 32- and 2-fold over normal levels, respectively. Significant substrate reduction was achieved in all of these tissues, with Gb3 and lyso-Gb3 content averaging < 10% and < 20%, respectively, of those detected in untreated Fabry mice. Importantly, appropriate glycosylation of the liver-secreted GLA enzyme was confirmed and will facilitate efficient uptake into target tissues. These studies provide proof-of-concept for the development of ZFN-mediated *in vivo* genome editing as a therapy for Fabry disease.

41

Parental mosaicism for apparent *de novo* variants from exome sequencing of 10,000 trios. *R. Torene, K. Arvai, Z. Zhang, E. Butler, D. McKnight, J. Juusola, K. Retterer.* GeneDx, Gaithersburg, MD.

De novo variants (DNVs), while individually rare, collectively have a substantial contribution to genetic disease. True DNVs arise as post-zygotic events in the developing embryo. If a pathogenic variant occurs during gametogenesis in a parent, or a parent has somatic mosaicism including the gonads, the inherited variant may appear *de novo* when assuming Mendelian inheritance, but the recurrence risk for future pregnancies is significantly increased. We analyzed apparent DNV loci to improve parental mosaicism (PM) detection from NGS data, and, in turn, provide more accurate recurrence risks to families. We examined DNVs in nearly 10,000 exome sequenced trios (predominantly blood) with probands who have Human Phenotype Ontology terms associated with a neurodevelopmental disorder (NDD). After filtering for high confidence variants in genes with autosomal dominant and X-linked inheritance, 6,669 variants were examined for PM in 4,684 trios. Putative PM was defined by presence of ≥3 variant allele reads, with the fraction of variant reads being greater in the proband than in either parent. Additionally, the variant allele was observed in <3 probands in the cohort, and we used a binomial test to exclude variants that were likely heterozygous rather than mosaic in a parent. Of the trios examined, 124 putative PM variants were identified in 118 (2.5%) trios. This is consistent with a rate of 3.8% recently reported (PMID: 26656846). Putative PM was detected as low as 2% allelic fraction. The putative PM variants were split evenly among maternally and paternally inherited (112 variants each). While we observed no association between putative PM and parental age, there was an enrichment for C>T transitions ($p<0.05$) for putative PM variants over non-mosaic DNVs, indicating that CpG methylated sites may be prone to mosaicism. Putative PM variants were classified using ACMG guidelines. Half of the pathogenic/likely pathogenic PM variants were confirmed by Sanger sequencing, while the remainder were below the detection limit. Even for such low level mosaicism, a recent study showed a recurrence risk of >24% compared to <1.3% if the mutation was limited to the proband (PMID: 26656846). We conservatively estimate that 2% of apparent *de novo* variants in NDD are mosaic in parental blood. We expect that parental mosaicism would be observed in other disease areas as well. Identifying parental mosaicism from NGS will enable more accurate recurrence risk assessment and counseling.

42

Deep amplicon resequencing identified parental mosaicism for approximately 10% "*de novo*" *SCN1A* mutations in Dravet Syndrome families and was capable of multiple validation of mosaicism. *X. Yang[1], X. Xu[2], Q. Wu[3], Y. Dou[2], K. Wang[4], A.Y. Ye[1], A.U. Huang[1], Y. Zhang[2], L. Wei[1].* 1) Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing, China; 2) Department of Pediatrics, Peking University First Hospital, Beijing, China; 3) School of Life Sciences, Peking University, Beijing, China; 4) Key Laboratory of Genomics and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Chaoyang, Beijing, China.

Background: The accurate identification of the origin and transmission of mutations causing severe Mendelian disorders such as Dravet syndrome (DS, mainly caused by *de novo SCN1A* mutations) is critical in genetic counseling. In current practice, a mutation is considered "*de novo*" if Sanger sequencing detects the mutant allele in peripheral blood DNA of the patient but not in their parents. However, the "*de novo*" mutations determined by Sanger sequencing might in fact be inherited from undetected parental mosaicism, largely due to technical challenges to detect and quantify low-fraction mutant alleles. **Method**: To investigate the origin of novel mutations, we developed, benchmarked and applied a new method named PASM that combined PGM semiconductor deep amplicon resequencing with a Bayesian model for mosaicism to detect and quantify mosaic mutations with mutant alleleic fraction (MAF) as low as 0.5%. We validated PASM results with digital PCR as well as pyrosequencing. **Results and Conclusions:** Of 174 "*de novo*" *SCN1A* mutations in DS probands by Sanger sequencing, PASM identified 15 cases (8.6%) of parental mosaicism. PASM also validated another five mosaic detectable by Sanger sequencing. MAFs in the 20 cases of parental mosaicism ranged from 1.1% to 32.6%. Twelve (60% of 20) mosaic parents did not have any epileptic symptoms, and their MAFs were significantly lower than those mosaic parents with epileptic symptoms ($P = 0.016$). Varied MAFs were detected in multiple samples obtained from the same donor, demonstrating that postzygotic mutations could affect multiple somatic cells as well as germ cells. These results suggest that more sensitive tools for detecting low level mosaicism in parents of families with seemingly "*de novo*" mutations will allow for better informed genetic counseling. PASM were further applied for validation of mosaicism identified by a bioinformatics pipeline for mosaicism from whole-genome or whole-exome sequencing data named MosaicHunter and a DNA circulation based highly-sensitive detection framework for mosaicism named o2n-seq. In the whole-exomes of 2321 autistic families, 1248 and 285 putative child and parental mosaicisms were identified by MosaicHunter, and PASM validated 51% of them. Using Bayesian model coupled with Ampliseq sequencing, PASM validated 81.8% putative low-fraction mutations with MAF ranging between 0.1% and 10% identified by o2n-seq, demonstrating that PASM is a sensitive validation method for low-fraction mosaicism.

43

A clinical survey of mosaic variants in disease-causing genes detected by whole exome sequencing. *M.J. Tokita[1], W. He[2], F. Vetrini[2], A.V. Dharmadhikari[2], J. Zhang[2], T. Sim[1], X. Ge[1], P. Ward[2], A. Braxton[2], S. Narayanan[2], M. Leduc[1,2], X. Wang[1,2], L. Meng[1,2], R. Xiao[1,2], W. Bi[1,2], F. Xia[1,2], M. Walkiewicz[1,2], C. Shaw[1,2], C. Eng[1,2], P. Stankiewicz[1,2], Y. Yang[1,2], P. Liu[1,2].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas; 2) Baylor Genetics, Houston, Texas.

Mutations can occur at any time point in an individual's life cycle. Those that occur in germ cells or in early embryological development are likely to have clinical consequences as they affect most or all cells in the organism. Mosaic events arising later in development have been thought to contribute to a minority of genetic disorders because 1) the mutant allele fraction may not reach the deleterious threshold for the gene, 2) the mutation may be restricted to tissues not pertinent to the gene, or 3) the mutation may arise after the critical timeframe for gene function. To further investigate the frequency of clinically relevant mosaic variants and characteristics of the corresponding genes, we queried reported mosaic variants from 9,765 samples submitted for clinical whole exome sequencing (WES, mean coverage ~120X). We found 65 mosaic variants in proband samples and 42 in parental samples involving 96 genes. Of mosaic variants detected in probands, 36 (55%) were classified as pathogenic or likely pathogenic and 29 (45%) as variants of uncertain significance. Mosaic variants occurred in genes associated with AD or AD/AR (52%), X-linked (26%), AR (8%), and AD/somatic (8%) inheritance. Of note, 6% of variants were found in genes in which only somatic events have been reported. Fourteen mosaic variants (12 AD, 1 AR, 1 X-linked) were detected in paternal samples and 28 (8 AD, 1 AD/somatic, 7 AD/AR, 1 AR, 11 X-linked) in maternal samples. One mosaic variant in an AD gene was present in a sample from a paternal grandmother. Across the cohort, mean minor allele fraction of variants detected by WES (n=72; mean coverage 202x; range 24-854x) was 18.8% ± 10.3% for AD/AR/somatic and female X-linked variants and 36.3% ± 21.4% for male X-linked variants. Eight genes had mosaic events in two individuals, and mosaic variants in *CREBBP* and *MTOR* were present in 3 individuals each. Variants included 69% missense, 14% nonsense, 13% small indel, and 4% splice changes. In summary, we have reported proband or parental mosaicism in 108 samples submitted for clinical WES. For variants detected by exome, level of mosaicism ranged from 6% to 80%. Our estimate of the prevalence of mosaic variants is limited to those detectable by WES and likely represents an underestimate of the total mosaic cases in our WES cohort as exome by design favors breadth over depth of coverage. This analysis provides insights into genes in which mosaic variants are likely to be clinically relevant.

44

Mosaic *EFTUD1* mutation causes Shwachman-Diamond syndrome through dysregulating ribosome assembly. *S. Lee[1], C.H. Shin[2], C.R. Hong[3], J.-D. Kim[4], J.M. Ko[3], T.-J. Cho[5], S.-W. Jin[4,6], H.J. Kang[3], H.H. Kim[2], M. Choi[1].* 1) Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, South Korea; 2) Department of Health Sciences and Technology, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Korea; 3) Department of Pediatrics, Seoul National University Hospital, Seoul, Korea; 4) Yale Cardiovascular Research Center, Section of Cardiovascular Medicine, Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, USA; 5) Division of Pediatric Orthopaedics, Seoul National University Children's Hospital, Seoul, Korea; 6) School of Life Sciences, Gwangju Institute of Science and Technology, Gwangju, Korea.

Shwachman-Diamond syndrome (SDS) is a kind of ribosomopathies caused by variants in the *SBDS* gene, which encodes a protein that plays an important role in ribosome assembly. However, approximately 10% of the cases that were clinically diagnosed with SDS do not harbor pathogenic variants in the gene, suggesting the existence of additional genetic mechanisms that may lead to the disorder. EFTUD1 is a component of 60S pre-ribosome complex and a GTPase that couples hydrolysis of GTP and eviction of eukaryotic translation initiation factor 6 (eIF6), allowing 60S subunit to bind with 40S to form a mature 80S ribosome by directly collaborating with SBDS. Here we present two Korean SDS patients that carry incomplete but identical homozygous EFTUD1 p.Thr1069Ala variant due to a mosaic uniparental disomy in chromosome 15, localized to the hematopoietic system. The variant is asymptomatic when in a heterozygous status, and found in 0.035 % of the healthy East Asian population. To further elucidate molecular mechanism of the mutant protein function and pathogenesis process, we observed polysome profiles of CRISPR/Cas9-mediated ablation of *EFTUD1* (*EFTUD1*$^{-/-}$) and wild type cell lines and demonstrated that 80S ribosomal peak amplitude was significantly reduced in the mutant cells, suggesting that EFTUD1 is required for ribosomal assembly. Exogenous introduction of *EFTUD1* (*EFTUD1*$^{-/-}$-wt) completely rescued the 80S peak, while introduction of *EFTUD1*$^{p.Thr1069Ala}$ (*EFTUD1*$^{-/-}$-mut) did not. Compared to the *EFTUD1*$^{+/+}$ cells, ribosome profiling analysis (RPA) of *EFTUD1*$^{-/-}$ or *EFTUD1*$^{-/-}$-mut did not show significant difference of ribosome-protected fragment distribution along the transcripts. On the other hand, transcriptome profiles of *EFTUD1*$^{-/-}$ or *EFTUD1*$^{-/-}$-mut cells showed upregulation of genes involved in RNA processing or ribosome biogenesis which was similar to the previously reported result of *Sbds* knock-down in zebrafish. Indeed, morpholino-induced knock-down of *Eftud1* in zebrafish resulted in abnormal development of neutrophil and erythrocyte as well as increased apoptosis of somatic cells. Moreover, phenotypic analyses of CRISPR/Cas9-driven *Eftud1*$^{-/-}$ *and Eftud1*$^{p.Thr1076Ala}$ (equivalent to human *EFTUD1*$^{p.Thr1069Ala}$) mice confirmed that the mutation is hypomorphic and the gene is essential for the proper translational control. Our results strongly suggest the direct association between EFTUD1 dysfunction, translational control and human ribosomopathy pathogenesis mechanism.

45

Assessing the landscape of selfish *de novo FGFR2* mutations in human testes. *H.K. Ralph[1], G.J. Maher[1], Z. Ding[1], E. Giannoulatou[1], N. Koelling[1], S.J. McGowan[1], G. McVean[2], A.O.M. Wilkie[1], A. Goriely[1].* 1) MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, Oxfordshire, United Kingdom; 2) Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford.

We have been exploring a phenomenon taking place in the human testis termed “selfish spermatogonial selection” in which mutations that confer a selective advantage are preferentially enriched, resulting in a skewed mutational profile of sperm over time. Importantly, such mutations are potentially heritable thus offering a novel approach to study *de novo* mutations (DNM) directly in the tissue of origin, the male germ line. Studying this process is technically challenging and, to date, only a small number of assays have been developed to directly assess individual genomic positions in sperm or testis samples. These positions include three codons (251 – 253) of a known selfish gene, *FGFR2*, in which mutations are associated with craniosynostosis syndromes (Apert, Pfeiffer and Crouzon). Here, we describe an approach to systematically screen human testes for spontaneous mutations and present as an example *FGFR2*. RainDance, a massively parallel droplet PCR platform, was used to identify low level DNM (down to 0.1%) across a 58 kb gene panel. Primers were designed to cover hotspots in candidate genes, including the majority of the coding regions of *FGFR2*. DNA was obtained by dissecting 10 testicular slices from 5 men (age range: 34 to 90, mean: 73 years) into 288 biopsies. 16 libraries were constructed and sequenced on the Illumina HiSeq2000 platform to an average depth of 25,000x. A statistical prioritisation pipeline was developed to generate calls of variant sites with elevated mutation levels. Outliers were selected for further validation. Of the 288 testicular biopsies, more than 90% reached a coverage depth >10,000x across the entire targeted region of *FGFR2*. The allelic ratio of the variants called ranged from 0.1% to 2.95%. A number of genomic positions in *FGFR2* were identified as recurrently mutated – the majority of which have previously been reported either in germline disorders associated with craniosynostosis or in cancer (COSMIC). The relative geographic positions of testis biopsies can be visually reconstructed revealing localised clusters of mutations. We show that the high multiplex PCR capability of RainDance offers a viable option to directly detect mutations arising in the male germ line at levels as low as 0.1%. This has enabled a wider assessment of the landscape of selfish DNM directly in human testes than previously possible.

46

Genomewide association and inference of clonal mosaicism implicates germline variation in XPO1 as a driver of genome instability. *Y. Jakubek[1], S. Vattathil[2], P. Auer[3], P. Scheet[1].* 1) Epidemiology, UT MD Anderson Cancer Center, Houston, TX; 2) Department of Genome Sciences, University of Washington, Seattle, WA; 3) Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI.

Errors in DNA replication and cell division create daughter cells with copy number changes and copy-neutral loss of heterozygosity that, by selection or drift, propagate to establish genetically distinct cell populations within an individual. This phenomenon, known as clonal mosaicism, is positively correlated with age and has been established as an important factor in the development of cancer, neurodegenerative disease, miscarriage, birth defects, and developmental delay. Blood mosaicism is associated with higher risk of type 2 diabetes and is a strong prognostic marker for hematological cancers. Germline variation has been associated with a rare mosaic variegated aneuploidy syndrome characterized by high rates of somatic chromosomal abnormalities and cancer susceptibility. However, the role of less penetrant genetic variants in mosaicism of healthy tissue has not been established. This is due, in part, to critical difficulties in phenotyping. Here we sought to overcome this challenge by conducting a genome-wide association study of clonal mosaicism in samples that were characterized by a haplotype based method to obtain high-fidelity phenotyping of mosaicism (Vattathil & Scheet, 2016, AJHG). Our analysis included 5 studies from the GENEVA consortium comprised of 16,668 individuals, of which 748 exhibited detectable mosaicism. Meta-analysis following standard imputation-based techniques revealed a significant association with mosaicism at marker rs114740129 on chromosome 2 (OR 3.7, P < 2e-8). This SNP is 52kb upstream of XPO1 (“chromosomal maintenance 1”). This gene regulates nuclear export of RNA and proteins involved in cell cycle regulation, including P53. In addition, four loci exhibited suggestive evidence for association (P < 5e-7). One of these was an intronic variant in FAT1 (chr. 4, rs75128406). Somatic mutations in this gene have previously been documented in multiple cancers including lymphoblastic leukemia and shown to lead to aberrant Wnt activation. In summary, we have identified inherited forms of genetic variation that potentially explain the development of detectable clonal mosaicism, an established prognostic factor in hematological malignancy, which may offer insights into the etiology of diseases such as cancer.

47

Cancer risk in neurofibromatosis 1 (NF1): Nation-wide, population-based Danish cohorts followed up to 3/4 century. *J.J. Mulvihill[1], L. Kenborg[2], J.H. Olsen[2], J. Rosendahl-Østergaard[5], H. Hasle[3], A. Redzkina[3], S.A. Sørensen[4], J.F. Winther[2].* 1) Pediatrics, University of Oklahoma, Oklahoma City, OK; 2) Danish Cancer Society Research Center, Copenhagen, DK; 3) Aarhus University Hospital, Aarhus, DK; 4) University of Copenhagen, Copenhagen, DK.

Individuals with neurofibromatosis 1 (NF1) have an increased risk for cancer, especially in the nervous system. Our 4-decade follow-up (*N Engl J Med* 1986;314:1010) of a cohort of NF1 patients from 1924 to 1944 (Borberg A, *Acta Psychiatr Neurol Scand Suppl* 1951;71:1) left uncertainty about other cancer types. In his original work, Borberg reviewed records in all Danish hospitals (excepting a few psychiatric facilities) and visited the home of most probands, examining them and available relatives and taking family histories. Combining hospital and patient number as well as name and birthday provided in Borberg's doctoral thesis, it was possible to reconstruct each family through the Danish Civil Registration System and church records. We updated the original nation-wide Borberg cohort (n=198) and established a cohort of their unaffected descendants (n=1317) as well as a more recent large registry-based NF1 patient cohort (n=2417) by linking to the population-based Danish National Patient Register and the Danish Cause-of-Death Register. To assess their risk for cancer, all cohort members were linked to the Danish Cancer Register to compare observed numbers of cancers in NF1 patients and their unaffected relatives with the expected numbers derived from the general Danish population. Standardized incidence ratios (SIRs) and 95% confidence intervals (CIs) were calculated. The overall risk for cancer was 2.3 (95% CI 1.8–2.9) in the updated Borberg cohort and 2.7 (2.5–2.9) in the register-based cohort, with particularly high risks seen in children and adolescents (SIRs 26.1–51.1). Exceptionally high risks were found for cancers traditionally associated with NF1, i.e., tumors in the peripheral nerves and autonomic nervous system (SIR 667; 504–867) and spinal cord and cranial nerves (72.0; 56.8–90.2). Significant excess risks were also reported for tumors of the mediastinum (20.5), adrenal glands (18.2), and small intestine (15.1), bone tumors (12.3), melanomas (2.5), and prostate (2.1) and breast (1.7). Borberg relatives without NF1 experienced no excess risk for cancer. Our results suggest that patients with NF1 face not only an enormous risks for nervous system tumors, but also increased risks for various other cancers. .

48

Looking beyond the lamppost: Population-based exome sequencing to ascertain prevalence and penetrance in the *DICER1* syndrome, a rare tumor-predisposition disorder. *U.L. Mirshahi[1], J. Kim[2], K. Manickam[1], M.F. Murray[1], D.J. Carey[1], D.R. Stewart[2] on behalf of the Geisinger-Regeneron DiscovEHR Collaboration.* 1) Weis Center for Research, Geisinger Clinic, Danville, PA; 2) Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.

DICER1 syndrome arises from mutations in the miRNA-processing gene *DICER1,* and is associated with certain rare malignancies and multi-nodular goiter. The true prevalence, penetrance and phenotypic spectrum of *DICER1* variation is unknown. We identified pathogenic (P) and likely pathogenic (LP) *DICER1* variants in the DiscovEHR exome sequence database of individuals linked to comprehensive longitudinal electronic health records. P variants were defined as nonsense, frameshift, canonical splice sites, *DICER1* RNase IIIb mutation hotspot variants, or known P variants. LP variants are missense variants with MAF < 0.1% that were determined as deleterious by *in silico* prediction tools. EHR queries and chart reviews, focused on cancers and thyroid disease, were conducted for all P/LP carriers. We compared prevalence of conditions in P/LP carriers to non-carriers. Among 60,873 individuals there were 11 unique *DICER1* P variants in 21 carriers (1:2900) and 32 unique LP variants in 64 carriers (1:950). In P variant carriers, we observed 3 (14%) with thyroidectomy, 9 (43%) with hypothyroidism, 1 (5%) with thyroid cancer, and 4 (19%) with thyroid nodules. In LP variant carriers, we observed 2 (3%) with thyroidectomy, 7 (11%) with hypothyroidism, 1 (1.6%) with thyroid cancer and 3 (5%) with thyroid nodules. When compared with non-carriers, risks for thyroid related conditions reached statistical significance for P carriers. The odds ratios (95% confidence interval; p values) were: hypothyroidism 3.2 (1-8; p=0.01), thyroidectomy 7.6 (3-30; p=0.01), and thyroid nodules 5.0 (2-20; p=0.02). Among P carriers we observed 1 individual with thyroid cancer, 1 with renal cell carcinoma, and 1 with pineoblastoma and benign meningioma (14% malignancy), cancers previously observed in individuals with *DICER1* syndrome. Imputed pedigrees demonstrated clustering of *DICER1*-related conditions. Our data show increased prevalence of thyroid disease and rare cancers in *DICER1* P variant carriers. Ascertainment bias can influence estimates of prevalence and penetrance of rare diseases. Our analysis of *DICER1* P variation in a large clinical population is novel and produced higher than previously reported estimates. Identifying carriers of these variants could improve understanding of *DICER1* and related syndromes and lead to targeted medical screening and treatment and prevention. Our approach provides a model in large population cohorts to better understand genetic risk for rare disorders.

49

Higher-than-expected population prevalence of potentially pathogenic germline *TP53* variants in individuals unselected for cancer history. *K.C. De Andrade[1,2], L. Mirabello[1], D.R. Stewart[1], E. Karlins[3], R. Koster[4], M. Wang[3], S.M. Gapstur[5], M.M. Gaudet[5], N.D. Freedman[6], M.T. Landi[7], N. Lemonnier[8], P. Hainaut[8], S.A. Savage[1], M.I. Achatz[1].* 1) Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA; 2) International Research Center, A.C. Camargo Cancer Center, São Paulo, Brazil; 3) Cancer Genomics Research Laboratory, National Cancer Institute, Division of Cancer Epidemiology and Genetics, Leidos Biomedical Research Inc., Bethesda, MD, USA; 4) Laboratory of Genetic Susceptibility, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA; 5) Epidemiology Research Program, American Cancer Society, Atlanta, GA; 6) Metabolic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA; 7) Integrative Tumor Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA; 8) Institute for Advanced Biosciences, Inserm U 1209 CNRS UMR 5309 Université Grenoble Alpes, Site Santé, Allée des Alpes, 38700 La Tronche, France.

Li-Fraumeni syndrome (LFS) is an autosomal dominant cancer predisposition disorder caused by pathogenic germline variants in the *TP53* gene, with a near-complete lifetime penetrance of cancer. However, the actual population prevalence of pathogenic germline *TP53* mutations is still unclear. The aim of this study was to estimate the prevalence of potentially pathogenic *TP53* variants in databases of individuals unselected for cancer history. We investigated *TP53* variants in 63,989 unrelated individuals with no known history of cancer from three sequencing databases. Potential pathogenicity was defined using an original algorithm combining allele frequency, pathogenicity prediction tools, known clinical significance, and functional data. We identified a total of 35 different potentially pathogenic *TP53* variants in 132/63,989 individuals (0.2%). Most of these variants fell within the DNA-binding domain of *TP53* (29/35, 83%), with an enrichment for specific variants not previously identified as LFS mutation hotspots, such as the p.R290H and p.N235S variants identified in 58 of 132 individuals (44%) with potentially pathogenic variants. Sequencing databases reveal that the population prevalence of potentially pathogenic *TP53* variants may be 0.2% or up to 10 times higher than previously described. However, the current study is unable to address potential confounding factors such as clonal hematopoiesis. This study suggests that cancer penetrance in individuals with LFS due to pathogenic germline *TP53* variants may be influenced by additional genetic or epigenetic modifiers, whereas rare *TP53* variants in some populations may be associated with cancer family history patterns that do not meet clinical criteria for LFS.

50

Large-scale phenome-wide scan in twins helps identify candidate variants associated with seborrheic keratosis. *S. Hebbring[1], Z. Ye[1], J. Pathak[2,3], S. Kim[1], L. Bastarache[4], J. Mayer[1], J. Liu[1], Y. Cheng[3], S. Schrodi[1], J. Denny[3], M. Brilliant[1].* 1) Marshfield Clinic Research Institute, Marshfield, WI; 2) Weill Cornell Medicine, New York, NY; 3) Mayo Clinic, Rochester, MN; 4) Vanderbilt University, Nashville, TN.

Twin studies have long been a powerful study design when understanding the genetic and environmental contributions to human disease. However, twins are rare and collecting relevant phenotypic data can be challenging. Using twin populations linked to electronic health record (EHR) systems may alleviate such challenges. We identified 28,278 twins from Mayo and Marshfield Clinic patient populations and measured familial aggregation and sibling recurrence risk ratios (RR) in 5,671 diseases to identify phenotypes with unappreciated genetic etiologies. Results suggest that very few diseases are random in twins and that shared environmental and genetic exposures contribute significantly to disease risk. Furthermore, diseases with the most significant aggregation P-values and RRs are strongly enriched for genetic disorders that may include seborrheic keratosis; a benign skin lesion sometimes resembling melanoma. The first GWAS of seborrheic keratosis (2,977 and 8,555 unrelated cases and controls, respectively) identified multiple genome-wide significant signals that overlapped candidate cancer genes/SNPs; most were independently replicated (1,362 cases and 17,389 controls). This included multiple independent variants in *TERT* (e.g., rs10069690, P=1.1E-11, OR=0.75; rs62332591, P=2.9E-11, OR=1.3). TERT is the rate limiting factor in telomere elongation and plays an important role in tumorigenesis and cellular senescence. When measuring lymphocytic DNA telomere length (L-TL) in 1,909 unrelated adults, including 431 diagnosed with seborrheic keratosis, rs62332591 was associated with L-TL (P=0.003, β=-0.025 units). Conversely, L-TL by itself or other SNPs known to be associated with L-TL (e.g., *TERC*, rs10936599, P=0.004, β=-0.029 units) were not associated with seborrheic keratosis. When measuring TL in 20 seborrheic keratosis tissues, 80% had reduced TL compared to adjacent normal tissues (P=0.0092). The totality of these functional studies suggests *TERT* and TL may influence the pathophysiology of seborrheic keratosis through tissues specific mechanisms. In conclusion, this study represents the first comprehensive assessment of thousands of clinical phenotypes measured in families in a single experiment to identify and map diseases with genetic predispositions. Importantly, this study may provide insights into the future of epidemiologic research when extensive clinical data is combined with family and genomic data in an EHR to predict, prevent, and treat disease.

51

Oncogenic potential of germline mutations in the lysosomal storage disease-associated genes. *J. Shin[1,2], D. Kim[2], Y. Koh[1,2], S. Yoon[1,2].* 1) Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea; 2) Cancer Research Institute, Seoul National University, Seoul, South Korea.

Introduction: Decades of clinical observations have shown consistent association of LSDs with cancer. However, the extent of contribution of each LSD-associated genetic variant to specific subtypes of cancer and the molecular pathogenic mechanism accountable for the tumorigenesis are largely unclear. Methods: Matched tumor-normal WGS, tumor RNA-Seq, and clinical information data of 2582 cancer patients constituting the PCAWG project were analyzed. For control, Variant call sets from the 1000 Genomes project phase 3 and ExAC release 0.3.1 were used. Among the germline single nucleotide variants and micro-insertions and deletions within the loci of the 42 LSD-associated genes, 434 putative pathogenic variants (PPVs) were selected based on the functional annotation of variants, curated databases, and manual review of medical literatures. We investigated the association of selected PPVs with cancer and its biological and clinical implications. Results: PPV prevalence was 20.7% in the Pan-Cancer cohort, which was significantly higher than the 13.5% of the 1000 Genomes cohort (OR, 1.67; 95% CI, 1.43-1.94; $p = 1.2 \times 10^{-11}$). This association was significant after adjustment for the population structure with 3 major principal components. SKAT-O tests revealed 44 significant associations between specific histologic subtypes of cancer and genes, with pancreatic adenocarcinoma being one of the most strongly associated. The cancer-PPV associations were well validated with the use of the ExAC cohort as control. Pancreatic cancer patients with germline PPVs developed cancer at earlier age than the wild-type patients (onset age, 61 versus 68.5; $p < 0.001$). Generally applicable gene set enrichment analysis revealed significant bidirectional difference in expression of pathways involved in complement and coagulation cascades, calcium signaling, Rap1 signaling, PPAR signaling, Ras signaling, and MAPK signaling. *KRAS* was the most commonly mutated gene both in the PPV-positive and PPV-negative pancreatic cancer. Conclusion: LSD-associated genetic variants are strongly associated with multiple histologic subtypes of cancer, revealing previously unknown connection between the lysosomal machinery and carcinogenesis. Patients with these variants may develop early-onset cancer via alteration of a wide spectrum of intracellular signaling cascades, revealing the intricate biology of LSD-related cancers.

52

Coverage matters: High rate of promoter 1B deletions in a large *APC* testing cohort. *A.J. Stuenkel, K. Jasperson, H. LaDuca, M. Richardson, S. Gutierrez, K. Blanco, L. Hoang, C. Espenschied.* Ambry Genetics, Aliso Viejo, CA.

Germline analysis of the *APC* gene is a long-standing first-tier test for patients with multiple colorectal adenomas. Improvements in testing methodologies (i.e. gross deletion/duplication (del/dup) analysis), increased gene coverage (i.e. promoter 1B), and the emergence of next-generation sequencing (NGS) multi-gene panel tests (MGPTs) have optimized detection rates of *APC*-associated polyposis conditions in recent years. Here we present results from >5 years of high volume clinical *APC* testing. Data from *APC* single gene tests (SGTs) and MGPTs containing *APC* ordered 10/01/2011-12/31/2016 were retrospectively reviewed. In all cases, *APC* analyses included sequencing and gross del/dup coverage of all coding regions and gross del/dup coverage of promoters 1A and 1B. Cases with an alteration classified as “pathogenic” or “likely pathogenic” (P/LP) were considered positive. The common p.I1307K moderate-risk mutation was excluded from the analysis. The final dataset contained 2,501 SGTs and 62,868 MGPTs. A total of 355 unique P/LP alterations were identified among 716 positive cases. *APC* positive rate varied drastically by test ordered: 17.8% (446/2,501) for SGTs and 0.4% (270/62,868) for MGPTs. Nonsense and small insertions/deletions were the most commonly identified mutation types (75.1%, 538/716), followed by gross deletions (12.3%, 88/716), splice junction (9.9%, 71/716), missense (2.5%, 18/716), and gross duplications (0.1%, 1/716). Promoter 1B deletions accounted for nearly 5% of positive cases (4.6%, 33/716) and represented the most common *APC* deletion (37.5%, 33/88). No gross deletions isolated to promoter 1A were identified, suggesting these alterations may not be responsible for *APC*-associated polyposis. Notably, 5 positive probands had reported previous negative *APC* testing. In 4/5 of these cases, a gross deletion had escaped detection on previous testing due to sequence analysis only (2 cases) or lack of promoter 1B gross deletion coverage (2 cases). In the final case, a point mutation was not detected previously by Sanger sequencing at an external laboratory, but was detected on re-analysis by NGS. In summary, promoter 1B gross deletions are common *APC* mutations. For patients suspected of having an *APC*-associated polyposis condition who previously tested negative, our data supports re-testing for promoter 1B deletions if not covered in the initial testing. Phenotype analysis is in progress to better understand possible genotype-phenotype correlations.

53

Passenger mutations in 2500 cancer genomes: Overall molecular functional impact and consequences. *M. Gerstein[1,2,3], S. Kumar[1,2], P. McGillivray[1,2], W. Myerson[2], L. Salichos[1,2], S. Li[2], A. Harmanci[1,2], J. Warrell[1,2], E. Khurana[4], A. Fundichely[4], C. Chan[5], C. Herrmann[5], M. Nielsen[6], X. Li[2], Y. Zhang[7].* 1) Molecular Biophysics and Biochemsitry, Yale Unviersity, New Haven, CT; 2) Program in computational biology, Yale Unviersity, New Haven, CT; 3) Department of computer science, Yale Unviersity, New Haven, CT; 4) Weill Medical College of Cornell University, New York, NY; 5) Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany; 6) Department of Molecular Medicine (MOMA), Aarhus University Hospital, Aarhus, Denmark; 7) Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH.

The Pan-cancer Analysis of Whole Genomes (PCAWG) project provides an unprecedented opportunity to comprehensively characterize a vast set of uniformly annotated coding and non-coding mutations present in thousands of cancer genomes. Classical models of cancer progression posit that only a small number of these mutations strongly drive tumor progression and that the remaining ones (termed “nominal passengers”) are considered inconsequential for tumorigenesis. In this study, we leverage the comprehensive variant data from PCAWG to ascertain the molecular functional impact of each variant, including nominal passengers. This allows us to decipher their overall impact uniformly over different genomic elements, both coding and non-coding. The molecular impact distribution of PCAWG mutations shows that, in addition to high-impact drivers and low-impact passengers, there is a group of medium-impact passenger variants predicted to influence gene expression or activity. Furthermore, we find that functional impact relates to the underlying mutational signature: different signatures confer contrasting molecular impact, differentially affecting distinct regulatory subsystems and different categories of genes. Also, we find that molecular functional impact varies based on subclonal architecture (i.e. early vs. late mutations) and can be related to patient survival. Finally, we speculate on how the overall burdening of cancer mutations might be related to the existence of both weak positive and negative selection during tumor evolution.

54

The impact of *PRDM9* expression on the cancer genomic rearrangement landscape. *A. Ang Houle[1,2], M. Agbessi[1], V. Bruat[1], F. Lamaze[1], L. Stein[1,2], P. Awadalla[1,2], Pan-Cancer Analysis of Whole Genome Consortium.* 1) Informatics and Bio-Computing, Ontario Institute for Cancer Research, Toronto, Ontario, Canada; 2) Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

Homologous recombination is a process allowing for the exchange of genetic information between homologous chromosomes, and when impaired, has been linked to multiple genomic alterations ranging from indels to larger scale aneuploidies. During meiosis, PRDM9 binding sites determine positions of double strand breaks as SPO11 is recruited, a topoisomerase-like protein that catalyzes double strand breaks leading to the initiation of meiotic recombination. Previously, allelic variation at the *PRDM9* locus was associated with paediatric acute lymphoblastic leukaemia, suggesting a role for PRDM9 in cancer. Although PRDM9 has a meiosis-specific function, and therefore is normally expressed solely in testes and in foetal ovaries, we found *PRDM9* expression in over 260 tumours across several cancer types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Project (n=1256), even after stringent homology correction. *PRDM9* expression levels are significantly different from those found in healthy tissues, implicating cancer-specific expression of *PRDM9* in somatic cells. To investigate the downstream effects of *PRDM9* expression in cancer, we focused on its putative association with somatic genomic rearrangements, reminiscent of PRDM9’s function in meiotic cells. Amongst all cancers expressing *PRDM9,* the most abundantly recognized motif neighbouring genomic rearrangements matches PRDM9’s binding motif, hinting at an association between the location of PRDM9 binding sites and the initiation of double strand breaks. Furthermore, PCAWG samples expressing *PRDM9* from several cancer types showed an enrichment of genomic rearrangements in recombination hotspots. Overall, these results suggest an association between the positions of PRDM9 activity and binding with the occurrence of genomic rearrangements in cancers where *PRDM9* is expressed. We employ two strategies to characterize mechanisms through which *PRDM9* might be expressed. First, using a differential gene expression analysis, we observe meiosis-specific genes consistently overexpressed in cancers expressing *PRDM9*. Second, recurrently mutated binding domains are associated with increased *PRDM9* expression. This study is the first to highlight the role of the meiosis-specific gene *PRDM9* on the transcriptomic and genomic landscape of tumours, and suggests a mechanism for aberrant homologous recombination in cancers.

55

Unlocking the genetic and molecular regulation of APOBEC mutagenesis in human cancers. *A.R. Banday, A. Bayanjargal, K.I. Udquim, O.O. Onabajo, A. Obajemu, L. Prokunina-Olsson.* Laboratory Of Translational Genomics, Division Of Cancer Epidemiology And Genetics, National Cancer Institute, NIH, Bethesda, MD.

APOBEC mutagenesis has been identified as one of the key mechanisms contributing to the generation of somatic mutations in tumors, fueling tumor evolution and resistance to therapies. Unlocking the genetic and functional regulation of APOBEC mutagenesis in human cancers is a new frontier in cancer biology that holds promise for new cancer therapeutic strategies. Here, we explored factors affecting APOBEC mutagenesis using data from public resources such as for 11,820 samples from 33 cancer types of TCGA, including germline and somatic variation, mRNA expression, CpG methylation, copy number variations and clinical data, as well as data for normal tissue from GTEx; protein structure data for APOBEC3 enzymes from Protein Data Bank, and lab-generated data. We then integrated statistical analysis of high dimensional data, protein structural modeling and functional *in vitro* analysis. We discovered that alternative splicing generates non-mutagenic and mutagenic isoforms of APOBEC3A and APOBEC3B, and the presence of these non-mutagenic isoforms is significantly associated with reduced burden of APOBEC mutagenesis. For example, cancers with high APOBEC mutagenesis, such as bladder, cervical, skin, head and neck, have low expression levels of non-mutagenic isoforms, while cancers with low APOBEC mutagenesis, such as hepatocellular, thymic and pancreatic, have high levels of non-mutagenic isoforms. We also found a significant interactions between APOBEC mutagenesis and SNP rs1014971, expression of specific *APOBEC3* splicing isoforms, environmental exposures (HPV infection) and DNA methylation patterns of *APOBEC3* genes across human cancers. We propose that sufficient expression of non-mutagenic APOBEC3 isoforms reduces the risk of APOBEC mutagenesis. Modulating the APOBEC3 isoform ratios can provide a therapeutic strategy for human cancers.

56

Accelerating pharmacogenomics discovery by imputing drug response in The Cancer Genome Atlas and beyond. *P. Geeleher[1], Z. Zhang[2], F. Wang[1], R.F. Gruener[1], A. Nath[1], G. Morrison[1], R.L. Grossman[2], R.S. Huang[1].* 1) Department of Medicine, University of Chicago, Chicago, IL; 2) Center for Data Intensive Science, University of Chicago, Chicago, IL.

Genomics-based studies such as The Cancer Genome Atlas (TCGA) have accelerated our understanding of the molecular basis of cancer. However, because it is difficult to collect accurate drug response information in large cohorts of cancer patients, these studies have not been effectively used for finding new biomarkers of drug response, which is crucially important if patient survival is to be improved. Thus, most cancer pharmacogenomics research is limited to pre-clinical disease models (e.g. the Genomics of Drug Sensitivity in Cancer (GDSC) project, drug screens in mouse patient-derived xenografts). However, these studies—among several limitations—are restricted by small sample sizes. Here, we present a machine learning based approach, which integrates data from TCGA with data from pre-clinical disease models and overcomes these critical obstacles—allowing studies such as TCGA to now be directly used for pharmacogenomics discovery. Our method works by learning models relating gene expression and drug response in pre-clinical data; we used the GDSC cancer cell lines but many suitable pre-clinical drug screening datasets now exist. Next, we used this model to impute drug response from tumor gene expression data in the TCGA clinical cohort. We then associated these imputed drug response data with measured variants (e.g. somatic mutations, copy number changes) in TCGA. Using this approach, we could recapitulate known clinically effective biomarkers and we have discovered several promising new indications for existing drugs. For example, we have conducted cell line and mouse based functional work to validate novel associations suggesting efficacy for Wee1 inhibitor AZD1775 in triple-negative breast cancer. Other functional validation has shown that amplification of ERLIN2 causes resistance to the cytotoxic chemotherapeutic vinorelbine, a finding that is clinically relevant in late-stage breast cancer. Our approach can be applied to any of the vast numbers of clinical cancer sequencing studies recently undertaken (TCGA, ICGC, TARGET, METABRIC etc.), meaning that it will now be possible to directly study pharmacogenomics using clinical sequencing data collected in all of these disease cohorts. To facilitate this, we have also developed a set of freely available computational tools. The TCGA imputed drug response data will be accessible on Genomic Data Commons (https://gdc.cancer.gov/)—the NCI's access portal for TCGA data, which we host at the University of Chicago.

57

Identifying and characterizing novel virus integrations in hepatocellular carcinoma genomes by virome-wide analysis of whole-genome sequencing data. *X. Chen[1], A. Sulovari[1], C. Jian[2], D. Li[1,3,4].* 1) Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT; 2) Department of Pathology, Yale School of Medicine, New Haven, CT; 3) Department of Computer Science, University of Vermont, Burlington, VT; 4) Neuroscience, Behavior, and Health Initiative, University of Vermont, Burlington, VT.

Hepatitis B virus (HBV) is a main cause of hepatocellular carcinoma (HCC). Using our newly developed bioinformatics tool, VIcaller, we analyzed the whole-genome sequencing data of 99 HCC tumor and matched adjacent normal tissues, 88 of which had been analyzed by others using a candidate virus approach. Using our virome-wide approach, we detected 388 HBV integrations and two adeno-associated virus (AAV) integrations. Compared to the candidate virus approach, 245 HBV integrations were consistently detected, and 143 HBV integrations were newly detected. Of the newly detected integrations, 23 were from the 11 HCC samples that had not been analyzed previously. In addition to HBV, we detected two AAV integrations, one of which had integration length of 212 base pairs with two base pairs deletion in the human genome at the integration site. To characterize the HBV integrations, we analyzed the junction sites in the HBV genome. We observed an enrichment at ~1.8 kilo base pair (where the x, core genes and enhancer 2 located) in the tumors based on the upstream junction sites, compared with the normal tissues. Particularly, the HBV integrations (with both upstream and downstream junctions) in two known oncogenes, *TERT* and *MLL4*, were found to harbor either viral enhancers or promoters. We further calculated the cellular proportion of each integration, and found the cellular proportions were significantly higher in tumors compared to matched normal tissues (t-test, $P = 2\times10^{-16}$). The average cellular proportions of the HBV integrations in *TERT* (N = 26, mean = 29%) and *MLL4* (N = 11, mean = 22%) were significantly higher than integrations located elsewhere in the human genome (N = 232, mean = 10%) (t-test, $P = 4\times10^{-10}$ and $P = 1\times10^{-3}$, respectively). At a single tumor tissue level, 24 among the 37 HBV integrations in *TERT* and *MLL4* ranked as the top among all HBV integrations in each tumor regarding cellular proportion. In addition, we found that HBV integrations in four other oncogenes (*GAS7*, *PRDM16*, *ARID1B*, and *AFF1*) also had the top cellular proportion ranks. To conclude, we conducted a virome-wide search for viral integrations, and our analysis demonstrated that our approach accurately identified the verified HBV integrations, and more importantly, we also identified novel HBV and AAV integrations. In addition, for the first time, we characterized both upstream and downstream junctions, integrated viral elements, and cellular proportion of each identified integration.

58

SiFit: A method for inferring tumor trees from single-cell sequencing data under finite-site models. *H. Zafar[1,2], A. Tzen[1], N. Navin[3], K. Chen[2], L. Nakhleh[1].* 1) Department of Computer Science, Rice University, Houston, Texas, USA; 2) Department of Bioinformatics and Computational Biology, the University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA; 3) Department of Genetics, the University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA.

Intra-tumor heterogeneity (ITH), as caused by a combination of mutation and selection, poses significant challenges to the diagnosis and clinical therapy of cancer. This heterogeneity can be readily elucidated and understood if the evolutionary history of the tumor cells was known. This knowledge, alas, is not available, since genomic data is most often collected from one snapshot during the evolution of the tumor's constituent cells. Consequently, using computational methods that reconstruct the tumor phylogeny from sequence data is the approach of choice. However, while intra-tumor heterogeneity has been widely studied, the inference of a tumor's evolutionary history remains a daunting task. Recently emerged single-cell DNA sequencing (SCS) technologies promise to resolve ITH to a single-cell level. However, inherent technical errors in SCS datasets, including false-positive errors, allelic dropout events, cell doublets and coverage non-uniformity significantly complicate this task. Moreover, single-cell-based phylogeny inference methods such as SCITE and OncoNEM operate under infinite sites model, which posits that each site in the dataset mutates at most once during the evolutionary history and the taxa form a perfect phylogeny. This assumption often gets violated in human tumors due to events such as chromosomal deletions, loss of heterozygosity and convergent evolution, necessitating the development of statistical inference methods that utilize finite-site models. Here we propose SiFit, a likelihood-based approach for inferring tumor trees from imperfect SCS genotype data with potentially missing entries, under a finite-sites model of evolution. SiFit employs an error model to account for technical errors and extends the Jukes-Cantor model of evolution to adopt it for cancer phylogeny. SiFit also employs a heuristic algorithm for exploring the joint space of trees and error rates in search of optimal parameters. We evaluate SiFit on a comprehensive set of simulated data, where it performs superior to the existing methods in terms of tree reconstruction and error rate estimation. Finally, we applied SiFit to experimental SCS datasets from a non-hereditary colorectal cancer patient and a metastatic colorectal cancer patient and it resulted in improved inference of clonal lineages and chronological order of mutations. SiFit is widely applicable to current and future SCS datasets and is a major step forward in understanding the tumor phylogeny.

59

Exome sequencing identifies *de novo* germline mutations in patients with early-onset cancer. *Z.K. Stadler[1], J. Vijai[1], M. Ronemus[2], S. Topka[1], T. Thomas[1], B. Yamrom[2], I. Iossifov[2], D. Villano[1], D. Levy[2], J. Kendall[2], C. Tran[1], S. Mukherjee[1], A. Maria[1], M. Robson[1], D. Bajorin[1], R. Kobos[1], B. Kushner[1], M. Walsh[1], L. Saltz[1], D. Feldman[1], G. Bosl[1], L. Norton[1], S. Modak[1], M. Seandel[3], M. Wigler[2], K. Offit[1].* 1) Memorial Sloan Kettering Cancer Center, NEW YORK, NY; 2) Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; 3) Weill Cornell Medical College, New York NY.

Genetic susceptibility plays an important role in the etiology of early-onset cancers. However, a genetic predisposition is identified in only a small fraction of individuals with early-adulthood or pediatric cancers and in the majority of cases, there is an absence of a cancer family history. To explore the extent to which *de novo* germline mutations contribute to cancer susceptibility, we performed exome-sequencing of early-onset cancer case-parent trios. Germline DNA from cancer affected proband and unaffected parents was ascertained under a prospective IRB approved protocol and included early-onset testicular, *BRCA1/2* deleterious mutation-negative breast, colorectal, bladder, and pediatric cancer patients. We report on the exome sequencing data from 122 case-parent trios in whom joint coverage per trio was >30% of the target covered at 20x. *De novo* events were called using GATK and custom designed algorithms. Likely gene-disrupting (nonsense, splice site, and frameshift) *de novo* mutations were identified in 9% (11/122) of cancer probands and were confirmed by Sanger sequencing. These included mutations in known cancer susceptibility genes including a *de novo* nonsense *WT1* mutation in a Wilms' tumor patient with no other features of *WT1*-related cancer susceptibility, and a nonsense *ERCC3* (C342X) mutation in a 27 year-old bladder cancer patient. On *ERCC3* functional assessment, in the UV repair-deficient XPCS2BA cell line viability after UVC radiation and repair of an UV-damaged reporter plasmid was significantly increased by overexpression of wild-type *ERCC3*. However, overexpression of the *ERCC3* C342X mutant could not rescue this phenotype. *De novo* germline truncating mutations in novel cancer-associated genes, such as *SIN3A*, a component of the histone deacetylase-dependent co-repressor complex and a regulator of MYC and E2F transcription, and *ACACA*, a translocation partner gene for *MLL* rearrangements associated with acute leukemias, in a neuroblastoma and a leukemia patient, respectively, were also identified. Additional rare *de novo* missense variants will be presented. Identification of *de novo* likely-gene disrupting mutations in nearly 10% of our early-onset cancer patients suggests that expansion of our study may identify either additional genes of interest or recurrent events in target genes. Our data support the model proposed in other human diseases that in seemingly 'sporadic' cases, *de novo* genetic mutations contribute to disease susceptibility.

60

Discovery and prevalence of germline and somatic mutations in patients with advanced renal cell carcinoma in MSK-IMPACT cancer genes. *S. Mukherjee[1], M.I. Carlos[1], Z. Stadler[1], J. Vijai[1], M. Walsh[1], A.G. Arnold[1], M. Sheehan[1], Y. Kemel[1], V. Ravichandran[1], Z. Shameer[1], D. Coskey[1], N. Pradhan[1], C. Stewart[1], A. Victor[1], A. Zehir[3], A.A Hakimi[2], J.A. Coleman[2], C.H. Lee[1], D.R. Feldman[1], M.H. Voss[1], D.B. Solit[1], M.F. Berger[3], M. Ladanyi[3], D. Mandelker[3], L. Zhang[3], M. Robson[1], R.J. Motzer[1], K. Offit[1].* 1) Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY; 2) Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY; 3) Department of Pathology, Memorial Sloan Kettering Cancer Center, NY.

Historically, 5% of RCC cases are reported to be associated with an inherited cancer syndrome, but more recent studies suggest this may be an underestimate. We studied the prevalence of germline cancer-susceptibility mutations in patients with advanced Renal Cell Carcinoma **(**aRCC). 213 patients with aRCC (stage III or IV), unselected for suspicion of an inherited cancer syndrome, were offered germline testing between October 2015 and December 2016. Next-generation sequencing of both somatic and germline DNA was performed using MSK-IMPACT, a platform that analyzes ≥400 cancer-associated genes. The results for 76 genes known to predispose to cancer using the ACMG guidelines were reported back to the patients. To investigate variants of uncertain significance (VUS), we assessed the multifactorial probability of VUS using combined likelihoods of family history, the clinical phenotypes and pathologic profile of the cancer, and co-occurrence with pathogenic germline and somatic mutations. 203/213 patients accepted testing (median age 55, range 13-55) of whom 73% had clear cell RCC (ccRCC), 92% had metastases, 20% were early onset (≤46 yrs at diagnosis), 9% had a family history of RCC, 6% multifocal RCC at diagnosis, and 15% ≥2 primary malignancies. Pathogenic/likely pathogenic mutations in genes: *CHEK2, FH, BAP1, MET, SDHA, SDHB, VHL, MSH6. BRCA2, PALB2, RAD51C,* were found in 35 patients (17%): 12 (6%) with mutations in genes associated with familial RCC; 10 (5%) mutations in high/moderate penetrance genes not linked to RCC.13 (6%) had mutations in genes of low/uncertain penetrance or for autosomal recessive disease. Germline mutations were present in 15% of ccRCC and 19% of non-ccRCC. *BAP1* germline mutant cases have a somatic hit, and IHC results showed loss of the protein in ccRCC and non-ccRCC tumors. Notably, 4/12 pts with mutations in familial RCC genes did not meet the American College of Medical Genetics (ACMG) criteria for testing (1 each *VHL*, *BAP1*, *SDHA*, *FH*). Prevalence of *CHEK2* mutations compared to population databases (ExAC) conferred a relative risk of 4.1 (CI=1.4-8.6) for RCC. Integrating the somatic and germline mutations provided an insight into the interplay between them. The VUS results will be presented . In conclusion, 17% of aRCC Patients had a germline mutation in a cancer-associated gene of which 33% of the high penetrance RCC germline mutations were not identified using standard clinical criteria, providing rationale for broad testing.

61

Rates and qualities of actionable and uncertain findings detected by eMERGE panel sequencing in 1155 colorectal cancer patients. *A.S. Gordon[1], H. Zouk[2], K.A. Leppig[3], D. Carrell[5], J. Ralston[5], A. Scrol[5], L. Witkowski[2], H.L. Rehm[2,4], E.A. Rosenthal[1], D.R. Crosslin[6], E. Larson[5], G.P. Jarvik[1].* 1) Medical Genetics, University of Washington School of Medicine, Seattle, WA; 2) Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA; 3) Genetic Services, Kaiser Permanente of Washington, Seattle, WA; 4) The Broad Institute of MIT and Harvard, Cambridge, MA; 5) Kaiser Permanente Washington Health Research Institute, Seattle, WA; 6) Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA.

Identifying pathogenic genetic variants in patients at risk for familial Colorectal Cancer or Polyposis (CRC/P) can change medical management and may prevent morbidity and mortality for the patient and their family. Although high penetrance variants for CRC/P have been identified in ~16 genes, the pathogenicity for most rare variation in these genes remains unclear, as does the penetrance of pathogenic variants. Although genome, exome, and panel-sequencing based tests including these genes are increasingly deployed in the clinic to inform cancer treatment and risk, there is little available data indicating the rates and types of variation likely to be detected by these tests in large patient cohorts unselected for age and family history. As part of the electronic Medical Records and Genomics (eMERGE) Network, we deployed a panel sequencing test centered on 109 genes with known or putative clinical associations (including the ACMG 56) in 1155 local Kaiser Permanente Washington patients diagnosed with CRC/P. The cohort was 55% female and had a mean age of 65 years (range: 29-98). In total, 224 patients (19.4%) received a non-negative report. Of these, we identified 55 patients (4.8%) with a positive (pathogenic or likely pathogenic) finding; 31 (2.7%) of these findings were associated with CRC/P, while 24 (2.1%) patients had additional findings not related to CRC/P. The remaining 169 patients (14.6%) carry at least one variant of uncertain significance (VUS) in a gene associated with CRC/P. Among these VUSs are novel predicted loss-of-function alleles and uncommon or novel missense alleles in dominant genes or as second alleles with a single heterozygous pathogenic recessive variant. Overall, our results indicate that nearly 1 in 20 CRC/P patients will receive a medically actionable result from a large panel test; if including VUS findings, this rate increases to nearly 1 in 5 CRC/P patients, highlighting the need to better understand how such variation affects disease etiology as these tests become the standard of care.

62

Cancer risks associated with known and putative predisposition gene mutations in a 341 gene panel sequenced on 10,000 individuals with advanced cancers. *J. Vijai[1], P. Sreenivasan[2], S. Mukherjee[1], C. Bandlamudi[2], Y. Kemel[1], V. Ravichandran[1], Z. Shameer[1], S. Topka[1], N. Bense[1], D. Mandelkar[3], A. Zehir[3], L. Zhang[3], M. Walsh[1], K. Cadoo[1], Z. Stadler[1], B. Taylor[2], D. Solit[2], M. Robson[1], M. Berger[2], K. Offit[1].* 1) Niehaus Center for Inherited Cancer Genomics, Memorial Sloan Kettering Cancer Center, New York, NY; 2) Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY; 3) Diagnostic Molecular Pathology Lab, Memorial Sloan Kettering Cancer Center, New York, NY.

Clinical genetic testing of individuals with family history of certain cancers is now routinely performed using multiplex targeted panels for known DNA repair gene mutations. In a series of 10,000 individuals with advanced malignancies whose tumors and matched normal samples were analyzed with MSK-IMPACT, we identified known correlations of *BRCA1/2* as well as other high risk or moderate risk genes in breast, pancreas, colorectal and prostate cancers. Here, we also report an expansion of the phenotype in rarer subtypes such as bowel , bladder, esophagogastric and lung cancer to known predisposition genes such as *BRCA1/2, ATM, CHEK2* and *CDH1*, including copy number alterations in addition to mutation calls. We describe the contribution of known genes in major cancer types and novel correlations such as germline *NF1* mutations in soft tissue malignancies and *CHEK2* in bladder, kidney, skin and lung cancers and associated population-specific founder mutations of *BRCA1/2, CHEK2, APC, NBN* and *ERCC3*. To discover putative predisposition genes, we re-classified missense variants using an ensemble of *in-silico* predictors to discover potentially damaging missense variants. We then estimated gene specific and functional domain specific risks for all pathogenic/likely pathogenic variants as suggested by the ACMG classification schema using an automated in-house variant curation program: *PathoMAN*. We compared the frequencies of pathogenic/likely pathogenic mutations across specific subsets of cancer types to the gNOMAD consortium data, comparing burden of founder mutations with matched population controls. Both single variant, gene and functional domain level risks vary across cancer types. We find a strong association of *ERCC3* R109X to colorectal, bladder, non-small cell lung cancer and glioma in Ashkenazi individuals. We show novel association tests for kinase and DNA binding domains across genes . Finally, for a subset of DNA repair pathway members, we analyzed mutations in functional domains that are known to physically interact to identify a polygenic model of cancer predisposition.

63

Prevalence of mutations in a large series of clinically ascertained ovarian cancer cases tested on multi-gene panels compared to reference controls. *J. Lilyquist[1,2], H. LaDuca[3], E. Polley[1], B. Tippin Davis[3], H. Shimelis[2], C. Hu[2], S.N. Hart[1], J.S. Dolinsky[3], F.J. Couch[1,2], D.E. Goldgar[4].* 1) Dept of Health Sciences Research, Mayo Clinic, Rochester, MN; 2) Dept of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN; 3) Ambry Genetics, Aliso Viejo, CA; 4) Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT.

Given the lack of adequate screening modalities for ovarian cancer (OC), knowledge of ovarian cancer risks for carriers of pathogenic alterations in predisposition genes is important for decisions about risk-reduction by salpingo-oophorectomy. The knowledge that some genes are not associated with OC also can reduce concerns of women found to carry pathogenic alterations in those genes. We sought to determine which of the genes assayed on multi-gene panels are associated with OC and to estimate the magnitude of the associations. 7768 adult ovarian cancer cases of European ancestry referred to a single clinical testing laboratory underwent multi-gene panel testing for detection of pathogenic alterations in known or suspected ovarian cancer susceptibility genes. A targeted capture approach was employed to assay each of 19 genes for the presence of pathogenic or likely pathogenic alterations. Mutation frequencies in Caucasian ovarian cancer cases were compared to mutation frequencies in Non-Finnish European individuals from the Exome Aggregation Consortium (ExAC). Analyses accounting for family and personal history of other cancers and by age at diagnosis were also performed. Significant associations (P<0.001) were identified between alterations in 11 genes and ovarian cancer, with seven of these displaying at least 5-fold increased risk (*BRCA1, BRCA2, MSH2, MSH6, PTEN, RAD51C, RAD51D*). Relative risks of ovarian cancer greater than two-fold could reliably be ruled out for *RAD50* and *CHEK2*. Further study of the remaining genes (*BARD1, APC, MLH1, MRE11A, NBN, and PMS2*) is needed to assess associations with OC. This is the largest study of germline multi-gene panel testing for ovarian cancers reported to date. These results will help inform clinical management of women found to carry pathogenic alterations in ovarian cancer susceptibility genes and may reassure women with alterations in genes not associated with ovarian cancer susceptibility.

64

Genome-wide association study (GWAS) identifies 9 novel breast cancer loci from analyses accounting for subtype heterogeneity. *H. Zhang[1], J. Lecarpentier[2], T.U. Ahearn[3], K. Michailidou[2,4], R.L. Milne[5,6], P. Kraft[7,8], J. Simard[9], P.D.P Pharoah[2,10], M. Schmidt[11,12], D. Easton[2,10], N. Chatterjee[1], M. Garcia-Closas[3] on behalf of the Breast Cancer Association Consortium.* 1) Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD; 2) Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 3) Division of Cancer Epidemiology and Genetics, National Cancer Institute , Rockville, MD; 4) Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus; 5) Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, Australia; 6) Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global health, The University of Melbourne, Melbourne, Australia; 7) Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA; 8) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA; 9) Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Québec City, QC, Canada; 10) Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK; 11) Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands; 12) Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands.

Background: Most common breast cancer risk loci have been identified in GWAS using standard tests for association with overall or ER-negative/triple negative (TN) disease. This approach is less well powered to identify SNPs when tumor heterogeneity is present. We conducted a GWAS to identify risk loci using subtype-specific case-control and case-case analyses by individual tumor characteristics, and a 2-stage regression that simultaneously accounts for multiple correlated tumor characteristics. **Method**: We included 108,946 cases and 96,201controls of European descent participating in 81 studies of the Breast Cancer Association Consortium using genotyped data from two genome-wide chips imputed to the 1000 Genomes. Chip data were analyzed separately and then meta-analyzed. We excluded loci with a minor allele frequency < 0.01 and those within 500Kb of or correlated at r^2>0.1 with known risk SNPs. We used standard logistic regression to identify SNPs associated with subtypes defined by ER, PR, HER2 or TN status. This included 9 case-control and 4 case-case analyses. We also performed a global test for association using a parsimonious 2-stage polychotomous logistic regression model that simultaneously test a SNP for an association with overall disease and subtype heterogeneity defined by ER, PR, and HER2 marker combinations. This method efficiently accounts for multiple testing, correlation between markers and missing tumor marker data. **Results**: We identified 4 loci associated ($P<3.85\times10^{-9}$; 5×10^{-8}/13 comparisons) with disease in at least one standard logistic regression model. The 2-stage model global test for association was significant ($P<5\times10^{-8}$) for 3 of these loci and identified 5 additional loci. Of the 9 novel loci identified, specific marker heterogeneity tests showed 4 loci had heterogeneity only by ER. The other 5 loci were associated with overall disease or were heterogenous for more than one marker.The strongest evidence was for an uncommon allele (MAF=0.01) in the 3'UTR of *TP53* (global 2-stage $P=2.3\times10^{-10}$), which was associated with increased ER-positive risk and decreased ER-negative risk. Analyses conditional on previously known risk SNPs are underway to confirm the independence of these signals. Additional analyses will incorporate data on tumor grade and histology to evaluate further heterogeneity. **Conclusion**: Using methods that account for tumor heterogeneity, we identified 9 novel breast cancer risk loci.

65

Using cancer status is better, simpler and more cost-effective than family history in determining breast cancer genetic testing eligibility. *N. Rahman[1,2,3], A. Turnbull[1], A. George[1,2], A. Strydom[1,3], Z. Kemp[1,2].* 1) Division of Genetics and Epidemiology, Institute of Cancer Research, London, United Kingdom; 2) Cancer Genetics Unit, Royal Marsden NHS Foundation Trust, London, United Kingdom; 3) TGLclinical, Institute of Cancer Research, London, United Kingdom.

Background: *BRCA1* and *BRCA2* (BRCA) gene testing is of proven value in cancer treatment optimisation and cancer prevention. Most countries recommend that individuals with >10% chance of having a BRCA mutation should have testing. Currently, complex family history (FH) criteria are used to identify these individuals. This system performs sub-optimally, is time intensive for patients and clinicians, is costly and has limited capacity. Our aim was to develop a simpler, more effective eligibility and testing process for cancer predisposition gene testing in breast cancer (BC) patients. **Methods:** We first defined five BRCA testing criteria based on cancer status: BC <40 years, Bilateral BC both <60 years, Triple-negative BC, BC + ovarian cancer, Male BC. We performed testing using the TruSight Cancer Panel (TSCP). We evaluated BRCA mutation detection rate overall and for each criterion separately. We also evaluated the mutation detection rate in eight other BC predisposition genes included on TSCP. **Results:** 1020 patients were tested, 110 had a BRCA mutation, giving an overall BRCA mutation detection rate of 10.8%. Furthermore, the mutation detection rate for each of the five criteria was also >10%. If FH eligibility criteria had been used instead, 56/110 (50.9%) BRCA mutations would have been missed. We also evaluated the mutation detection rate in 368 probands with FH but who did not fulfil the cancer criteria. The mutation rate was only 5.4% and 75% of the mutation carriers had a relative that fulfilled the cancer eligibility criteria. Using cancer-based criteria allows better assessment of genetic testing requirements, for example in the UK we estimate that 12,000 BC patients (22% of BC) would be eligible for testing per year. Health economic analyses showed that implementation would be highly cost-effective, primarily through the prevention of cancers in relatives in whom cascade testing and cancer prevention strategies are undertaken. **Conclusions:** Our data show that BRCA testing eligibility should be determined by cancer status rather than family history. Cancer-based criteria to determine individuals at >10% chance of having a mutation perform better and are simpler and more cost-efficient. The mutation rate in genes other than BRCA is very low. In many health systems panel testing to cover other genes may only be appropriate if it doesn't add to the cost and time of BRCA testing.

66

Discovery of germline pathogenic mutations in hereditary cancer syndromes with whole genome, low-coverage variant imputation. *H.P. Ji[2], S.U. Greer[1], O. Barad[2], I. Kela[2], Y. Waldman[1].* 1) Medicine/Oncology, Stanford University, Stanford, CA; 2) NRGene Ltd., Ness Ziona, Israel.

Increasingly, whole genome studies are being applied to large populations. Whole genome variant discovery has numerous advantages compared to exome analysis - non-coding variants are increasingly recognized as being important causes of human disease. Moreover, whole genome analysis has the potential to provide extended haplotype information that dramatically improves variant analysis and calling. However, whole genome sequencing approaches are associated with high cost for large-scale population studies. Even with recent drops in whole genome sequencing costs, large number of samples (in the tens of thousands if not more) and the associated computational analysis burden pose a significant challenge regarding resources and expense. Thus, achieving adequate statistical power using large number of samples remains a significant challenge in the use of population-based studies with whole genomes. As a solution to these issues, we have developed an imputation approach that enables genome-wide haplotype variants to be provided from low-coverage sequencing. This approach dramatically reduces the overall cost of whole genome studies in a variety of settings. Herein, we present initial results using this approach based on low coverage whole genome sequencing to impute whole genome variants from families and discover the pathogenic mutations of hereditary gastrointestinal cancers. We leverage hereditary family structure to effectively increase coverage, as many sequences are shared among family members. Each family member, either affected or not affected, is sequenced in relatively low coverage. Analyzing sequencing data together with pedigree structure and Mendelian inheritance patterns results in effective high coverage sequencing of variants observed in the family. Also, this enables one to phase variants into highly informative haplotypes. Segregating variants that may be associated with the disease can be readily identified. The performance is influenced by family size and structure, as well as the number of affected individuals. We demonstrate the power of this approach using the whole genome sequencing data from a series of highly characterized family trios. Subsequently, we apply this approach to a number of hereditary cancer pedigrees. Overall, this method allows sequencing larger number of individuals to identify disease-related variants and has immediate application in identifying pathogenic variants in hereditary disorders.

67

Population genetic testing for breast and ovarian cancer susceptibility. *I. Campbell[1], S. Rowley[1], L. Devereux[2], D. Goode[1], S. McInerny[3], N. Grewal[3], A. Lee[1], A. Trainer[3], M-A. Young[4], N. Li[1], P. James[3].* 1) Research Div, Peter MacCallum Cancer Ctr, Melbourne, Australia; 2) Lifepool, Research Div. Peter MacCallum Cancer Ctr, Melbourne, Australia; 3) Parkville Familial Cancer Ctr., Peter MacCallum Cancer Ctr., Melbourne, Australia; 4) Garvan Institute, Sydney, Australia.

Background. Germline mutations in certain genes account for a large proportion of inherited risk for breast and ovarian cancer. The identification of asymptomatic mutation carriers could significantly reduce the incidence of these diseases as active risk management can dramatically reduce the risk of developing cancer. In most countries, identifying high-risk individuals is based on their family history. In general, a family is first identified because one family member develops cancer and, because of high-risk indicators is referred to a familial cancer centre (FCC). However, current data suggests that many *BRCA1* or *BRCA2* mutation carriers do not have a remarkable history of cancer in a close relative. Population-based genetic testing would be a far more effective strategy for identification of at-risk individuals. To test the feasibility of such a strategy we are conducting a population genetic testing trial for actionable mutations in 11 breast/ovarian cancer predisposition genes among 15,000 healthy women from the Australian population. Methods. All subjects are female participants in the *LifePool* cohort (www.lifepool.org) who had no personal history of breast or ovarian cancer at the time of DNA collection. Participants found to carry an actionable germline mutation were notified by letter with an invitation to contact the PeterMac telephone genetic counselling service for further information and/or also invited for counselling at an FCC. Only participants with an actionable mutation were notified of their genetic testing result. Results. Of the 5,557 women tested to date, 40 (0.72%) were carriers of mutations that are currently actionable in the Australian context (*BRCA1* n=7, *BRCA2* n=15, *PALB2* n=15, *ATM* n=3). All 40 women accepted the invitation to attend a familial cancer centre for formal predictive testing. Less than 20% of the women would have met the minimum threshold for clinical genetic testing under current guidelines. A further 16 participants (0.29%) carried mutations in *BRIP1*, *RAD51C* and *RAD51D* but were not notified of the result as these genes are not currently actionable in Australia. Conclusions. A relatively large proportion of cancer free-women from Australia carry high-risk mutations in BRCA genes and subsequent uptake of clinical genetic testing was very high. Population-based genetic testing is well accepted and can identify a much larger proportion of the at risk-population than contemporary family history based approaches.

68

BRCA population screening in unaffected Ashkenazi Jewish women: A randomized controlled trial of different pre-test strategies. *E. Levy-Lahad[1,2], S. Lieberman[1,2], A. Tomer[1], A. Ben-Chetrit[3,4], O. Olsha[5], S. Levin[4], R. Beeri[1], A. Raz[6], A. Lahad[2,7].* 1) Medical Genetics Institute, Shaare Zedek Medical Center, Jerusalem, Israel; 2) Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel; 3) Department of Obstetrics&Gynecology, Shaare Zedek Medical Center, Jerusalem, Israel; 4) Clalit Health Services, Israel; 5) Breast Unit, Dept. of Surgery, Shaare Zedek Medical Center., Jerusalem, Israel; 6) Department of Sociology, Ben-Gurion University of the Negev, Beer Sheva, Israel; 7) Department of Family Medicine, Clalit Health Services, Jerusalem, Israel.

Background: About half of *BRCA1/BRCA2* carriers lack significant family history, and would only be identified through general testing. The Ashkenazi Jewish (AJ) population is a model for *BRCA* screening, given high mutation prevalence (1/40) and >95% sensitivity and specificity of testing for three common mutations. Towards screening implementation, we examine the impact of excluding pre-test in-person genetic counseling (GC) in the population screening setting. **Methods:** Healthy AJ women age > 25 years are randomized to two pre-test arms: written information only (WI) vs. GC. Post-testing, GC is provided to non-carriers indicating significant family history and to all carriers. Psychosocial outcomes (satisfaction with health decision, stress, anxiety, personal perceived control (PPC), knowledge) are assessed post-testing, at one week (before results-Q1) and at 6 months (post results-Q2). **Results:** Among the first 824 participants (mean age 46 years), we identified 12 carriers (1.4%). Only 4/12 carriers had significant family history. Post-testing, 96% of participants both of GC and WI were satisfied/very satisfied with testing. Stress (Impact of Events) scores were also similar in both groups. Knowledge and PPC scores were higher in GC vs. WI at both Q1 and Q2, but absolute differences were small. PPC scores were 62% and 69% in GC vs. 56% and 66% in WI, at Q1 (p=.0001) and Q2 (p=.04) respectively. The difference in knowledge was 1/10 points at Q1 (p=.0001) and 0.8/10 points at Q2 (p=.001). Carriers had higher PPC and knowledge than non-carriers. At Q2, carriers' stress was higher (14 vs. 5.7, p=.0004), as expected. Within GC, PPC increased and stress decreased over time (from Q1 to Q2, p=.003, 0.05), and within WI *all* outcomes improved over time, with greater satisfaction, lower IES and increased knowledge and PPC (p<.004). This may reflect the impact of post-test counseling in participants with suggestive family history. Overall, >85% at Q1 and > 90% at Q2 would recommend population screening. **Conclusions:** A streamlined screening process would identify substantially more carriers (regardless of family history) while addressing logistic and cost limitations. Our results suggest that compared to WI, pre-test GC is associated with a mild increase in knowledge, and a somewhat greater sense of control. Forgoing pre-test GC may therefore be a legitimate alternative in large scale screening, particularly if alternative methods for imparting knowledge are explored.

69

Barriers and facilitators to genetic testing among a population-based sample of young Hispanic and non-Hispanic White breast cancer survivors. *D. Cragun[1], A. Weidner[2], T. Pal[2].* 1) Global Health, University of South Florida, Tampa, FL; 2) Vanderbilt University Medical Center, Nashville, TN.

Background: Women with hereditary breast cancer (HBC) face high cancer risks; however, cancer screening and prevention options can reduce risks to near that for the general population. Given concerns about growing disparities, it is important to gain a better understanding of factors associated with access to and uptake of genetic testing for HBC. **Purpose:** The purpose of this study was to compare self-reported barriers and facilitators associated with access to and uptake of genetic testing between Hispanic and non-Hispanic White women (NHW). **Methods:** Following state mandated recruitment methods, living women diagnosed with breast cancer < age 50 in 2009-2012 were recruited through the Florida State Cancer Registry with oversampling of Hispanics compared to NHW. All participants were asked to complete a questionnaire and medical records release for verification of genetic test results. Using descriptive statistics and Chi-square tests, we compared barriers and facilitators to genetic testing for HBC. **Results:** Of the 1,182 participants who completed the questionnaire, 61% (174/285) Hispanic versus 65% (580/897) NHW had HBC testing at the time of enrollment. Untested Hispanics were more likely than NHW to report having never heard of genetic testing before the survey (28% Hispanic, 11% NHW; p<.0001). The two most commonly cited barriers were: lack of testing recommendation (44% Hispanics, 32% NHW; p=.021) and cost-related concerns (41% Hispanics, 40% NHW; p=.834). Over 70% of untested participants reported additional barrier(s) to testing besides or in addition to being unaware of testing or lacking a testing recommendation. Among tested participants, the top three facilitators were similar across ethnic groups: 1) “To benefit my family’s future” (70% Hispanic, 68% NHW); 2) “My doctor recommended I have testing” (60% Hispanic, 54% NHW); and 3) “Minimal cost to me” (59% Hispanic, 72% NHW). **Conclusions:** Rates of genetic testing awareness and testing recommendations are lower among untested Hispanics compared to untested NHW in our high-risk population. Although economic concerns should become less of a barrier now that testing costs have plummeted, other barriers to testing remain. A multi-pronged, personalized approach that goes beyond raising awareness among patients and healthcare providers may be needed in order to overcome other reported barriers and increase testing uptake.

70

Psychosocial outcomes of genetic counseling in a population based sample of Black breast cancer survivors. *S.T. Vadaparampil[1], M.L Kasting[1], D. Cragun[2], J.P. Kim[1], B. Augusto[1], J. Garcia[1], C.L. Holt[3], K. Ashing[4], C. Hughes-Halbert[5], T. Pal[6].* 1) Health Outcomes and Behavior, Moffitt Cancer Center, Tampa, FL; 2) University of South Florida, Tampa, FL; 3) University of Maryland, College Park, MD; 4) City of Hope, Duarte, CA; 5) Medical University of South Carolina, Charleston, SC; 6) Vanderbilt University Medical Center; Vanderbilt-Ingram Cancer Center, Nashville, TN.

Purpose: Black women are less likely to access genetic counseling (GC) and genetic testing (GT) for hereditary breast cancer (BC), compared to women from other racial/ethnic groups. Prior studies demonstrate minimal adverse psychological consequences among women participating in GC/GT. However, findings are based on predominantly Caucasian women of high socioeconomic status. To address this gap, we conducted a prospective follow up of a subset of participants from a population based study of Black BC survivors receiving GC/GT. Methods: Black women with invasive BC at age ≤50 years diagnosed between 2009 -2012 were recruited through the Florida Cancer Registry as part of a population-based study to investigate etiology and outcomes of early-onset invasive BC (n=456). Participants were offered telephone pre- and post-test GC; a subset completed questionnaires assessing sociodemographic, clinical, and psychosocial variables (n=354). After excluding those with prior GT, we examined psychosocial outcomes by *BRCA* status (positive, negative, variant of uncertain significance [VUS]) at 1-month (n=166), and 1-year (n=151) following results disclosure. Psychosocial outcomes include BC related distress (Impact of Events Scale [IES; range=0-75]), and anxiety and depression (Hospital Anxiety and Depression Scale [HADS; range=0-21 for both anxiety and depression]). Changes from baseline to 1- month and 1-year post results disclosure for BC related distress, anxiety, and depression, were compared using Wilcoxon signed-rank tests. All tests were two-sided, with alpha set at $p < 0.05$. Statistical analyses were conducted using SPSS (v.24). Results: Participants were 44.7 (+6.2) years of age, 31.3% graduated college, and half had private insurance (48.2 %). Cancer related distress did not significantly change for any of the *BRCA* groups at 1-month. At 1-year, distress decreased for *BRCA* negative (p=0.04) and VUS (p=0.03) women but did not change for *BRCA* positive. Anxiety significantly increased from baseline to 1 month for the VUS group (p=0.01) but did not change for *BRCA* positive or negative participants. There were no differences in anxiety at one year for any group. Depression scores did not differ from baseline to either time for any *BRCA* group. Conclusions: This study is among the first to examine GC/GT outcomes in Black women. Black women demonstrate minimal negative psychosocial outcomes, irrespective of test results.

71

Methods for meta-analysis of multiple traits using GWAS summary statistics with an application to lipid traits. *D. Ray, M. Boehnke.* Department of Biostatistics, University of Michigan, Ann Arbor, MI.

Genome-wide association studies (GWAS) for complex diseases have focused primarily on single trait analyses for disease status and disease-related quantitative traits. For example, GWAS on risk factors for coronary artery disease analyze genetic associations of plasma lipids such as total cholesterol, LDL-cholesterol, HDL-cholesterol, and triglycerides separately. However, traits are often correlated and a joint analysis may yield increased statistical power for association over multiple univariate analyses. Recently several multivariate methods have been proposed which require individual-level data. Here, we develop metaUSAT, a novel unified association test of a single genetic variant with multiple traits that uses only summary statistics from existing GWAS. This novel method does not require individual-level data and can test genetic associations of binary and/or continuous traits. One can also use metaUSAT to analyze a single trait over multiple studies, appropriately accounting for overlapping samples, if any. metaUSAT provides an approximate asymptotic p-value for association and is computationally efficient for implementation at a genome-wide level. Simulation experiments show that metaUSAT maintains proper type-I error at low error levels. It has similar and sometimes greater power to detect association across a wide array of scenarios compared to existing methods, which are usually powerful for some specific association scenarios only. When applied to plasma lipids summary data from the METSIM and the T2D-GENES studies (with few individuals that are common to both), metaUSAT detected genome-wide significant loci beyond the ones identified by univariate analyses. Evidence from larger studies suggest that the variants additionally detected by metaUSAT are, indeed, associated with lipid levels in humans. In summary, metaUSAT can provide novel insights into the genetic architecture of a common disease or traits.

72

Quantifying directional effects of transcription factor binding on polygenic disease risk using GWAS summary statistics. *Y. Reshef[1], H. Finucane[2], D. Kelley[3], A. Gusev[4], J. Ulirsch[2], L. O'Connor[5], B. van de Geijn[5], P. Loh[5], S. Grossman[2], G. Bhatia[5], S. Gazal[5], P. Palamara[5], L. Pinello[6], N. Patterson[5], R. Adams[1], A. Price[5].* 1) Computer Science, Harvard University, Cambridge, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Calico Labs, South San Francisco, CA; 4) Dana-Farber Cancer Institute, Boston, MA; 5) Epidemiology, Harvard School of Public Health, Boston, MA; 6) Massachusetts General Hospital, Boston, MA.

Connecting GWAS data to a biological process frequently involves an unsigned genomic annotation describing the importance of each SNP to that process (Finucane et al. 2015 Nat Genet). However, it is often possible to generate a signed annotation quantifying whether each SNP allele promotes or hinders a biological process, e.g. binding of a transcription factor (TF) in some cell type. We introduce a method, signed LD profile regression, that evaluates the directional genetic effect of a signed functional annotation on a disease or trait by regressing the vector of marginal GWAS z-scores on the product of the signed annotation and a reference LD matrix. We show via simulations using real genotypes that our method is well-calibrated under a variety of null models, including models confounded by unsigned enrichments, and that it is well-powered to detect true signals. We also show that, by modeling MAF-dependent directional effects, it is robust to confounding resulting from the genome-wide action of negative selection on both disease and TF binding. We applied our method to 42 diseases and traits (avg N=99,535) and 427 signed annotations, each quantifying the effects of SNP alleles on binding of one TF in one cell line and constructed using a neural network predictor of ENCODE ChIP-seq read density from genomic sequence (Kelley et al. 2016 Genome Res). We identified 57 significant associations at per-trait FDR<5%, representing 10 independent signals after accounting for linked annotations. Our results include a positive association between Crohn's disease (CD) and binding in myeloid cells of IRF1, a TF that lies in a CD GWAS locus, is differentially expressed in CD intestinal tissue, and has eQTLs with significant association to CD risk in the direction predicted by our finding; a positive association between CD and binding in LCLs of ELF1, a TF that lies in a CD GWAS locus and whose binding sites are over-represented among promoters of genes that are differentially expressed in CD intestinal tissue; and a negative association between systemic lupus erythematosus (SLE) and binding in myeloid cells of CTCF, consistent with experimental evidence that CTCF modulates the rate of myeloid differentiation as well as evidence of CTCF-mediated decrease in promoter methylation in CD4+ T cells of SLE cases vs. controls. Our method thus represents a promising new way to leverage signed functional annotations to interrogate the biology of diseases and complex traits.

73

Tissue specific transcriptome prediction leveraging GTEx datasets and gene-level association mapping and fine-mapping. *Z. Qi[1,2], Y. Guan[1,2,3].* 1) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 2) USDA/ARS Children's Nutrition Research Center, Houston, TX; 3) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.

Tissue-specific gene expressions have direct relevance to disease phenotypes, and knowing gene expression in disease-relevant tissues is advantageous in both detecting novel associations and fine-mapping known genetic associations. Unfortunately, many genome-wide association study(GWAS) datasets have no companion gene expression essay. This project aims to develop novel methods that can predict tissue-specific gene expression into GWAS data by leveraging genotype-tissue expression(GTEx) datasets. The idea of predicting gene expression to perform gene-level association mapping was pioneered by others. Their software PrediXcan uses penalized regression to predict gene expression. Here we present a novel Bayesian approach to impute tissue-specific gene expression using Bayesian variable selection regression(BVSR) implemented in a software package fastBVSR. Compared to the penalized regression, BVSR not only enjoys the benefit of shrinkage and sparse priors, but also has inherited advantage from Bayesian model averaging. We first compared predictive performance between fastBVSR and Predixcan. We used two datasets that have both dense genotypes and RNA-seq data, GTEx whole blood(n=338) and DGN whole blood(n=922). When using GTEx as training and DGN as testing, fastBVSR performs slightly better than PrediXcan overall, and a set of genes can be predicted better using PrediXcan and another set of genes can be predicted better using fastBVSR. When using DGN as training and GTEx as testing, however, fastBVSR significantly out-performed PrediXcan(Wilcox rank test P=2e-31) in 99% of genes. We applied fastBVSR to analyze a glaucoma GWAS dataset using GTEx as training datasets. We chose five tissues in GTEx after considering the representativeness of different germ layer, sample size, and tissue relevance to glaucoma. Using the predicted gene expression in specific tissues as surrogate genotype, we tested gene-level associations, and also performed fine mapping to dissect newly detected and known associations. The set of tissues for significant association hits are largely different. Among all significant associations that can be replicated in at least two tissues, three are known glaucoma GWAS hits and one is a novel candidate. The novel candidate is a lncRNA gene significant associated with glaucoma in both lung and thyroid. Functional annotation implied that a previously reported GWAS hit is likely associated with this lncRNA instead of the originally reported gene.

74

New IDEAS for GWAS loci: Using genome segmentation to identify causal variants and tissues driving disease associations. *Y. Zhang[1], K. Sieber[2], M.R. Nelson[2], C. Guo[2].* 1) Statistics, Pennsylvania State Univ, State College, PA; 2) GlaxoSmithKline, Philadelphia, PA.

Identifying causal genetic variants and relevant tissues underlying a genetic association greatly assists in pinpointing effector genes driving disease susceptibility and progression. Doing so can be challenging given that thousands of loci associate with hundreds of complex traits and diseases. In this study, we predicted disease-specific tissues and cell types for ~80,000 loci from the Systematic Target OPportunity assessment by Genetic Association Predictions (STOPGAP) database by applying genome segmentation results in 127 Roadmap and ENCODE human cell types. The genome segmentation was conducted using Integrative and Discriminative Epigenome Annotation System (IDEAS), a 2D genome segmentation method that accurately annotates regions of the genome with gene regulatory potential and highlights tissue specificity developed by Zhang and colleagues. By integrating the posterior probabilities from statistical fine mapping using PICS with the posterior probabilities for regulatory states identified by IDEAS, we increased the power to predict causal variants by an average of 31%, estimated by simulation assuming 1000 cases 1000 controls and effect size 0.1. Our approach is particularly useful for identifying functional variants within loci carrying >10 candidate variants in strong linkage disequilibrium. We were also able to identify cell type enrichments for putative functional variants across ~1,900 traits. Unlike existing enrichment analysis methods, we account for the correlation of regulatory events across cell types and estimate the cell type in which causal variants are most likely to be functional. For example, we found strong enrichment of IBD associated SNPs within predicted regulatory elements of CD4+ and macrophage cells. Overall, we saw evidence for enrichment in one or more cell types in 86% of traits included in this analysis. Interestingly, we found putative causal SNPs for obesity to be enriched in not only fat tissues, but also in mesenchymal stem cell types. In summary, we present a new integrative approach to identify causal variants and cell type specific enrichments, which will increase our understanding of the various pathways and tissues involved in multifactorial diseases. We expect this approach to substantially improve our ability to take advantage of GWAS to identify new prospective therapeutic targets.

75

Rare variant association in non-coding sequence: An analysis of deep coverage whole genome sequences and blood lipids in 16,324 individuals. *P. Natarajan[1,2,3], G.M. Peloso[4,5], S.M. Zekavat[3], M. Montasser[5], A. Ganna[3,7], M. Chaffin[3], W. Zhao[8], J. Bloom[1,3,7], J.R. O'Connell[6], S.E. Ruotsalainen[9], M. Alver[10], J.A. Perry[6], I.L. Surakka[9], T. Esko[10], S. Ripatti[9], A. Correa[11], B. Neale[1,2,3,7], G. Abecasis[12], B. Mitchell[6], S.S. Rich[13], J.G. Wilson[11], L.A. Cupples[4,5], J.I. Rotter[14,15], C.J. Willer[16], S. Kathiresan[1,2,3], NHLBI TOPMed Lipids Working Group.* 1) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114; 2) Department of Medicine, Harvard Medical School, Boston, MA 02115; 3) Program in Medical and Population Genetics, Broad Institute of Harvard & MIT, Cambridge, MA 02142; 4) Department of Biostatistics, Boston University, Boston, MA 02118; 5) Framingham Heart Study, Framingham, MA 01702; 6) School of Medicine, University of Maryland, Baltimore, MD 21201; 7) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114; 8) Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI 48109; 9) Institute for Molecular Medicine Finland, Helsinki Finland; 10) Estonian Genome Center, University of Tartu, Tartu, Estonia; 11) Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216; 12) Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109; 13) School of Medicine, University of Virginia, Charlottesville, VA 22908; 14) Institute for Translational Genomics and Population Sciences, LA Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA 90502; 15) David Geffen School of Medicine, UCLA, Los Angeles, CA 90095; 16) Departments of Human Genetics, Internal Medicine, and Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109.

Plasma lipids are heritable risk factors for cardiometabolic diseases. Prior studies using genotyping arrays and exome sequencing have identified hundreds of associations involving common, non-coding variants or a burden of rare, coding variants. Whole genome sequencing (WGS) allows for testing across the full spectrum of allele frequency and variant type in both non-coding and coding regions. We performed deep WGS (>20X) in 16,324 individuals: Old Order Amish, Framingham Heart Study, Jackson Heart Study, Multi-Ethnic Study of Atherosclerosis, FINRISK, and Estonian Biobank. 189M unique variants were identified. We individually tested 32M variants (MAF>0.1%) for association with total cholesterol, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides. Common variants in 31 known loci demonstrated association ($P<5x10^{-8}$). We observed an African American-specific haplotype (1% frequency) including a variant at the *LDLR* promoter associated with 28 mg/dl lower LDL-C ($P=2x10^{-11}$) and a 1-bp intronic deletion at 9p24.1 (MAF 2%) with HDL-C ($P=1x10^{-8}$). We confirmed that rare (MAF<1%) disruptive coding mutations in known Mendelian lipid genes (*LDLR*, *PCSK9*, *APOE*, *LCAT*, *APOC3*) associated with blood lipids ($P<2.5x10^{-6}$). We found that both rare disruptive coding mutations and an expanded polygenic risk score of 2M variants independently predicted extreme lipid phenotypes, but the effects differed by ancestry. Finally, we tested the hypothesis that rare, variants in regulatory (non-coding) sequence contribute to plasma lipid variation. We aggregated rare variants in non-coding sequence using a 3kb sliding window approach as well as three other methods which leverage ENCODE and Roadmap annotations for hepatocytes and adipose tissue: 1) regions overlapping enhancers within 20kb and promoters within 5kb of transcription start sites, 2) overlapping enhancers co-occurring with gene expression, and 3) overlapping enhancer/Hi-C contacts to transcription start sites. No burden-of-rare-variant association signals were detected in non-coding sequence with any of these approaches. In summary, we present a large-scale whole genome sequence analysis of plasma lipids in 16,324 ethnically diverse participants. Common variants in non-coding sequence as well as rare variants in coding sequence contribute to plasma lipid variation; however, association signals for rare mutations in non-coding sequence were not detectable.

76

Imaging-wide association study: Integrating imaging endophenotypes in GWAS. *Z. Xu, C. Wu, .W. Pan.* University of Minnesota, Minneapolis, MN.

A new and powerful approach, called imaging-wide association study (IWAS), is proposed to integrate imaging endophenotypes with GWAS to boost statistical power and enhance biological interpretation for GWAS discoveries. IWAS extends the promising transcriptome-wide association study (TWAS) from using gene expression endophenotypes to using imaging and other endophenotypes with a much wider range of possible applications. As illustration, we use gray-matter volumes of several brain regions of interest (ROIs) drawn from the ADNI-1 structural MRI data as imaging endophenotypes, which are then applied to the individual-level GWAS data of ADNI-GO/2 and a large meta-analyzed GWAS summary statistics dataset (based on about 74000 individuals), uncovering some novel genes significantly associated with Alzheimer's disease (AD). We also compare the performance of IWAS with TWAS, showing much larger numbers of significant AD-associated genes discovered by IWAS, presumably due to the stronger link between brain atrophy and AD than that between gene expression of normal individuals and the risk for AD. The proposed IWAS is general and can be applied to other imaging endophenotypes, and GWAS individual-level or summary association data.

77

Integrated analysis of exome sequencing and metabolomic profiling improves sequence variant interpretation, classification and diagnosis. *J.T. Alaimo[1,2], L. Hubert[1,2], M. Miller[1,2], H. Dai[1,2], R. Xiao[1,2], F. Xia[1,2], W. Bi[1,2], M. Leduc[1,2], M. Walkiewicz[1,2], V.R. Sutton[1,2], C.M. Eng[1,2], Q. Sin[1,2], S.H. Elsea[1,2], Y. Yang[1,2].* 1) Baylor Genetics Laboratories, Baylor College of Medicine, Houston, TX; 2) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.

The utilization of exome sequencing has substantially increased our capacity to identify disease-causing genes across a variety of clinical indications. Interestingly, clinical exome has a variable diagnostic rate ranging from 17-40%, leaving many individuals undiagnosed or with uncertain diagnoses. Methods to make a definitive diagnosis remain a cardinal challenge for clinical exome sequencing and a significant limitation is observed at the level of variant interpretation and classification largely due to the lack of functional information and clinical impact. One approach to curb this limitation is to integrate other genetic testing modalities in order to improve variant interpretation and classification. Metabolomic profiling is a low-cost, untargeted assessment of multiple substrates, intermediates and biological pathways often lending a useful source of functional information. Here, we describe a cohort of individuals from a single genetic diagnostic laboratory who received combinatorial genetic testing consisting of exome sequencing and metabolomic profiling. The metabolomic profiling data were treated as either strong pathogenic or benign functional evidence (PS3 or BS3) according to the ACMG/AMP guidelines for the interpretation of sequence variants. Our cohort consisted of 262 individuals that were 59% male and 41% female with a predominant (86%) neurological indication. We were successful in integrating metabolic profiling to a subset of variant classification scenarios in 8.4% of cases. First, metabolomic profiling successfully ruled out the contribution of either a single pathogenic variant or a bi-allelic pathogenic variant with a variant of uncertain significance (VUS) in *CPT1AD*, *MMACHC*, *HIBCH, MCCC2* and *UROC1*. Second, profiling results upgraded VUS or bi-allelic pathogenic variant with a VUS in *NDUFA4*, *PAH*, *ACADS*, *HIBCH, PYCR2*, *HSD17B4*, *TRMU*, *MTR*, *HPRT*, *ABAT* and *DDC* to either likely pathogenic or pathogenic variants aiding in subsequent diagnoses. Lastly, profiling confirmed exome pathogenic or bi-allelic pathogenic variants in *DDC*, *ALDH7A1*, *PEX6*, *SLC13A5*, *ABHD5* and *UROC1*. Taken together, our combinatorial genetic testing approach of metabolomic profiling and exome analysis considerably enhanced the interpretation and classification of sequence variants. This powerful approach may serve as a diagnostic tool for individuals with neurological indications of an unclear genetic etiology and contribute to medical management recommendations.

78

A CRISPR/Cas9 pipeline for functionally characterizing variants of uncertain significance in very early onset psychosis. *C.F. Mavros[1], A.H.M. Ng[2], C.A. Brownstein[1], K.G.C. Leeper[2], P.F. Chen[1], E.D. Buttermore[1], R.J. Kleiman[1], J.P. Rodriguez[1], K. Graber[1], S.K. Tembulkar[1], C. Genetti[1], P.B. Agrawal[1], A.H. Beggs[1], G.M. Church[2], J. Gonzalez-Heydrich[1].* 1) Genetics, Boston Children's Hospital, Boston, MA; 2) Department of Genetics & Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Boston, MA.

The advent of whole exome sequencing has led to an explosion in the identification of genetic variants of uncertain significance (VUS). In a cohort of 46 probands with very early onset psychosis (VEOP), we have identified 18 rare and predicted damaging single gene VUS in genes previously linked to developmental or degenerative brain disorders. We have developed a pipeline to confirm these VUS using CRISPR/Cas9 technology to induce mutations in human induced pluripotent stem cells (iPSCs) and rapidly inducing neuronal cultures. iPSCs are used to look for cell autonomous phenotypes that signal pathogenicity for the VUS and that can be used to screen for compounds reversing that phenotype. The screening process consists of parallel experiments with rapidly differentiated excitatory neurons from the following iPSCs: those engineered with CRISPR/Cas9 technology to contain the VUS on an isogenic background, the patient's iPSCs containing the VUS, and his/her parent's iPSCs without the VUS. These are used to look for cell autonomous phenotypes present in the two neuronal cultures carrying the VUS but not in the two that do not. Depending on the gene carrying the VUS, we triage them for developmental defects, electrophysiological properties or transcriptomic alterations. The cells are assayed for specific biomarkers that can identify a loss or gain of function through the induced mutation and differentiation process. Additionally, the resulting phenotypes are studied using electrophysiological techniques. Preliminary data indicates reduced survival of neurons from patients with a mutation in the ATP1A3 gene.

79

Analysis of variants of uncertain significance: Application of neoteric protein structural dynamics. *P.S. Atwal[1], P.R. Blackburn[1,2], M.T. Zimmerman[2], T. Caulfield[1].* 1) Mayo Clinic, Jacksonville, FL; 2) Mayo Clinic, Rochester, MN.

Introduction The issue of how to interpret variants of uncertain significance (VUS) pervades human genetics from basic science labs to clinical offices. Common strategies include analyzing population frequencies, the type of variant (e.g. missense vs nonsense), phenotypes in other patients with the same variant, simple prediction software (e.g. SIFT), and functional assays such as enzymatic analysis, histological staining or model-organism analysis. Some VUSs however cannot be analyzed using these strategies; these can include variants with no functional test available or novel variants. In addition, cost or expertise maybe prohibitive. We report a novel method of variant analysis using cutting-edge computational analysis to run complex neoteric protein molecular structural dynamic simulations (MDS) that represents a novel tool in the interpretation of VUSs. Methods Briefly, modeling of structure, refinement and molecular dynamic simulations of mutant and wild-type DNA2, SIM1, CYP11A1, and TGFBR2 were performed. Models were generated from a variety of sources including X-ray data sources, ab-initio calculations and homology models. We then performed MDS of both wild-type and mutant proteins was using state-of-the-art simulations. Using statistical mechanics approaches, deductions to the structural basis for the individual mutants is addressed in terms of the structural reorganization that alter function. Both local and global deviations in geometry were analyzed and correlated with their effect or protein interactions such as dimerization. Results We provide the first full-length human atomic model for DNA2 and SIM1. Dynamic simulations of wild-type and mutant protein from 100-150 nanoseconds were compared. We demonstrate: unique local fluctuations including alteration of hydrogen-bonding networks in DNA2, helix binding site disruption in SIM1, loss of cholesterol binding due to residue orientation in CYP11A1, and root mean square fluctuations destabilizing helix conformation in mutant TGFBR2, all showing patho-mechanism for loss of function. These data have been correlated in basic science assays such as nuclease activity (DNA2), reporter assay (SIM1), biochemical studies (CYP11A1), or were highly correlated to clinical phenotype (TGFBR2). Conclusions We demonstrate a novel method of variant interrogation using MDS. This method is a powerful new tool for researchers and clinicians alike in human genetics to analyze VUSs to help define pathogenicity. .

80

Next-Generation Mapping (NGM): A novel approach for genetic diagnosis of structural variants. *H. Barseghyan[1,4], W. Tang[1], R. Wang[1], M. Almalvez[1,4], E. Segura[1], M. Bramble[1,4], A. Lipson[1], E. Douine[1], H. Lee[3], E. Delot[1,2], S. Nelson[1], E. Vilain[1,2,4].* 1) Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA; 2) Department of Pediatrics, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA; 3) Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA; 4) Center for Genetic Medicine Research, Children's National Health System, Washington D.C.

Undiagnosed disorders are individually rare; however, their combined incidence is high. The associated diagnostic odyssey for rare disorders accounts for a significant portion of the resources of the healthcare system and is extremely stressful for families and patients due to delays in treatment. Many of these diseases remain a medical mystery without the identification of an underlying genetic cause and a clear basis for treatment. Exome and genome sequencing (ES/GS) solve approximately 25-35% of these rare diseases. They can reliably identify inherited and *de novo* single-nucleotide variants (SNVs), small insertions and deletions (INDELs). However, next-generation sequencing (NGS) relies on sequencing of 100-300bp DNA fragments that are mapped to a reference genome, failing to identify large structural variants (SVs), such as insertions, deletions, inversions, translocations and copy number variants, especially in repetitive regions of the genome. To overcome these limitations we performed next-generation mapping (NGM) using Bionano nanochannel arrays to produce high-resolution images of fluorescently labeled native-state DNA molecules (mega base size) for *de novo* genome assembly and SV detection. First, we investigated NGM's capacity to identify previously known pathogenic SVs in a series of 10 patients with Duchenne muscular dystrophy. We successfully identified all major SV types (insertion, deletion, inversion, translocation) ranging from 13Kb to 5.1Mb in size within or involving *DMD* gene. Second, we performed NGM on 50 patient trios with a variety of phenotypes, including autism, developmental delay, and disorders of sex development. On average we identified 3,300 insertions, 1300 deletions, 47 inversions and 1 translocation per individual, showing that, while less common than SNVs and INDELs, SVs account for a substantial fraction of genetic variation. SV data from144 healthy individuals and Database of Genomic Variants (DGV) was effectively used to filter out common variants. In addition, we identified 3-15 de novo SVs in each trio. We show the ability of NGM to detect large pathogenic structural variants otherwise missed by exome sequencing and chromosomal microarrays. In addition, we present results on NGM's ability to identify potential pathogenic SVs in patients without causative SNV identified by either ES or GS. We believe that NGM is poised to become a new tool in both the clinical diagnostic strategy and research.

81

A high frequency of previously reported pathogenic variants in nephropathy genes among healthy controls suggests potential for erroneous clinical interpretation of sequence variants for kidney disorders. *H. Milo Rasouly, D.A. Fasel, R. Heyman-Kantor, A. Mitrotti, R. Westland, E. Groopman, S. Sanna-Cherchi, D. Goldstein, A. Gharavi.* Department of Medicine, Columbia University Medical Center, New York, NY.

BACKGROUND: The American College of Medical Genetics guidelines for clinical interpretation of sequence variants is predicated on accurate prior reports of pathogenicity and appropriate filters for population minor allele frequency (MAF). We studied the prevalence of previously reported pathogenic and rare loss-of-function (LoF) variants in nephropathy genes among healthy individuals to assess the potential for false positive findings during diagnostic whole exome sequencing (WES) for kidney disorders. **METHODS**: We curated a list of OMIM and Orpha.net genes associated with genitourinary diseases and examined the number of pathogenic variants (PV) reported in ClinVar and HGMD in 7974 unrelated healthy individuals who had undergone WES at the Institute of Genomic Medicine at Columbia University. The exomes were sequenced using a variety of exome capture kits and analyzed using an in-house software, ATAV. **RESULTS**: We identified 172 dominant genes and 453 recessive genes associated with genitourinary disorders in OMIM and Orpha.net. With the application of a modestly restrictive ExAC popmax MAF <1%, up to 20% of healthy controls carried at least one PV annotated by ClinVar or HGMD for a dominant disorder (287 distinct variants detected in 80/172 genes). In addition, 2% were homozygotes or compound heterozygotes for PV's for one of the 453 recessive disorders. Furthermore, application of a more stringent MAF (ExAC popmax MAF<0.1%) reduced but did not eliminate the number of mutation carriers (5% for dominant disorders and 0.2% for recessive disorders). Finally, with ExAC popmax MAF<0.1%, up to 2% of the cohort carried at least one LoF variant in 107 genes for dominant disorders caused by haploinsufficiency (97 distinct LoFs detected in 40/107 genes) and 0.1% were homozygotes for LoF variants in the genes associated with recessive disorders. **CONCLUSIONS**: A large number of healthy individuals carry ClinVar and HGMD reported PVs for nephropathy, exceeding the known prevalence of genetic genitourinary disorders or expectations for incomplete penetrance. These findings predict that automated annotation of WES will generate many false positives PVs, increasing the burden of downstream clinical interpretation. These data indicate the need for systematic reassessment of pathogenicity for kidney-associated PVs in clinical databases and analysis of phenotypes associated with these PVs among healthy individuals to clarify their penetrance and clinical relevance.

82

Gene-specific allele frequency thresholds for benign evidence to empower variant interpretation. *D. Qian, S. Li, B.A.J. Sarver, Y. Tian, A. Elliott, H.M. Lu, M.H. Black.* Bioinformatics and Computational Biology, Ambry Genetics, Aliso Viejo, CA.

Allele frequency is often used as evidence of whether a variant is likely to be causative for a rare disease. However, current assessments of allele frequency for variant classification rely on either fixed or prevalence-adjusted thresholds and make no use of gene-specific information on variant pathogenicity. The publicly available Genome Aggregation Database provides an unprecedented spectrum of population-based human genetic variation that can be leveraged to examine the relationship between allele frequency and variant pathogenicity in classified variants on a gene-by-gene basis. Using a cohort of 176,280 patients who underwent genetic testing at a single diagnostic laboratory in 2012-2016, we assembled a training dataset of 1,388 classified missense variants in 10 genes (*BRCA1*, *BRCA2*, *CDH1*, *PALB2*, *PTEN*, *TP53*, *MLH1*, *MSH2*, *MSH6* and *PMS2)* included on hereditary cancer panels. The number of variants per gene ranged 44 to 438. We developed a constrained distribution fitting (CDF) approach to quantify gene-specific allele frequency thresholds (AFT) using data mining techniques of bounded constraints, monotonic distribution fitting and bootstrap sampling, each targeted to conservative estimates in case of uncertainty. Across variants in all 10 genes, positive predictive values (PPVs) for benign evidence were 0.40 ± 0.33, 0.43 ± 0.33 and 0.47 ± 0.33 using AFTs designated by a fixed 1% cutoff, prevalence-adjusted method, and CDF method, respectively. Negative predictive values (NPVs) were 1.00 ± 0.00 for fixed 1% cutoff and CDF methods, and 0.88 ± 0.31 for prevalence-adjusted method. Thus, the AFTs estimated by CDF showed 9% to 17% higher PPV than the other two methods and 12% higher NPV than the prevalence-adjusted method. Notably, differences between gene-specific AFTs estimated by CDF vs. other methods were striking for several genes. For example, using the CDF method, the AFT for benign evidence was as low as 0.0019% for *CDH1*, due to extremely rare pathogenic/likely pathogenic variants and as high as 0.43% for *PMS2* due to pathogenic/likely pathogenic variants having a wide range of frequencies. In contrast, AFTs estimated by the prevalence-adjusted method were nearly identical from 0.060% to 0.063% for both genes. Our results underscore the tremendous need for and practical usefulness of gene-specific allele frequency thresholds for benign evidence to empower variant interpretation.

83

Redrawing the map of blood pressure genes in a transcriptome-wide association study of over 301,000 participants in the Million Veterans Program and 145,000 from UK Biobank. *D.R. Velez Edwards[1], T.L. Edwards[2], J. Hellwege[2], A. Giri[2], E. Torstenson[2], Y.V. Sun[3], P. Elliot[4], E. Evangelou[4], M. Caulfied[5], P.W.F. Wilson[6], P.S. Tsao[7], C.P. Kovesdy[8], K.A. Birdwell[9], C. O'Donnell[10], A. Hung[11] on behalf of the VA Million Veteran Program and The UK Biobank.* 1) Vanderbilt Genetics Institute, Vanderbilt Epidemiology Center, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center; Tennessee Valley Health Systems VA, Nashville, TN; 2) Division of Epidemiology, Department of Medicine, Institute for Medicine and Public Health, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Tennessee Valley Healthcare System (626)/Vanderbilt University, Nashville, TN; 3) Department of Epidemiology, Emory University Rollins School of Public Health; Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA; 4) Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, UK; 5) William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, UK; 6) Atlanta VAMC and Emory Clinical Cardiovascular Research Institute, Atlanta, GA, 30233; 7) VA Palo Alto Health Care System; Division of Cardiovascular Medicine, Stanford University School of Medicine; 8) Nephrology Section, Memphis VA Medical Center, Memphis TN; 9) Division of Nephrology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; 10) VA Boston Healthcare, Section of Cardiology and Department of Medicine, Brigham and Women's Hospital, Harvard Medical School; 11) VA TVHS Nashville, Division of Nephrology & Hypertension, Vanderbilt University Medical Center, Nashville, TN.

High blood pressure (BP) is heritable and a risk factor for stroke, cardiovascular, and kidney diseases. Many genome-wide association studies (GWAS) of BP traits have been performed to date; however, identification of causal gene targets from GWAS loci remains a challenge. Utilization of genetically predicted gene expression (GPGE) can refine this search by integrating genomic and transcriptomic data to inform gene targets and potential mechanisms of action in a tissue-specific manner. To refine the search for gene targets associated with BP, we evaluated the association between GPGE and BP traits in various tissues using the MetaXcan method. Summary statistics from the largest GWAS of systolic and diastolic blood pressure (SBP and DBP, respectively) to date, using samples from the Million Veterans Program (MVP) and UK Biobank (UKB) cohorts in a total of over 440,000 individuals, were incorporated with gene expression prediction models for 44 tissues obtained from the Genome-Tissue Expression (GTEx) project. GPGE at 337 and 267 loci were associated with SBP and DBP respectively at p-value < $5x10^{-8}$, across all tissues. Of these, there were 133 unique genes represented for SBP and 85 for DBP. The strongest association was shared by both phenotypes with decreasing BP with increasing expression of *MTHFR* in skeletal muscle (p-value = $7.66x10^{-26}$ in DBP and p-value = $3.74x10^{-36}$ in SBP). *MTHFR* GPGE was also associated in aorta and whole blood, among many significant tissues, with both DBP and SBP. Several additional genes were associated with SBP in cardiovascular tissues, including increases by higher *CCDC71L* and *PRKAR2B* in aorta, and decreases with increasing *CLCN6* and *ATP2B1* in tibial artery. Overall, there were 56 significant GPGE associations with SBP that were in heart or arterial tissues. For the 133 genes unique to SBP, just 31 of them were the reported or mapped locus from a GWAS hit. Increasing GPGE of *CLCN6* and *ATP2B1* in tibial artery and *DNAJC5G* in the left ventricle of the heart were also associated with decreased DBP. There were 43 significant associations between GPGE and DBP in tissues of the heart and arteries. For the 85 genes identified as unique to DBP, 23 of them were identified as the reported/mapped locus from a previous BP GWAS. As the largest study to evaluate GPGE with BP traits, our study identifies novel genes in relevant tissues and implicates different genes as potentially causal loci than previously mapped genes. .

84

Insights to the population structure and genetic architecture of cardiometabolic traits in 20,029 Finnish exomes. *C.W.K. Chiang[1], A.E. Locke[2,3], K.M. Steinberg[3], S. Service[1], H. Abel[2], A.S. Havulinna[4,5], C. Chiang[2], L. Stell[6], H.M. Stringham[3], A.U. Jackson[3], M. Pirinen[4], D. Ray[3], D.E. Larson[2], D.C. Koboldt[2], L.J. Scott[3], R.S. Fulton[2], J. Nelson[2], T.J. Nicholas[2], P. Yajnik[3], V. Ramensky[1], N.O. Stitziel[2], I.M. Hall[2], C. Sabatti[6,7], S. Ripatti[4,8,9], V. Salomaa[5], A. Palotie[4,8,10], M. Laakso[11], M. Boehnke[3], R.K. Wilson[2], N.B. Freimer[1].* 1) Semel Institute for Neuroscience & Human Behavior, UCLA, Los Angeles, CA; 2) McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO; 3) Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI; 4) Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland; 5) National Institute for Health and Welfare, Helsinki, Finland; 6) Department of Biomedical Data Science, Stanford University, Stanford, CA; 7) Department of Statistics, Stanford University, Stanford, CA; 8) Department of Public Health, Hjelt Institute, University of Helsinki, Helsinki, Finland; 9) Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; 10) Broad Institute of MIT and Harvard, Cambridge, MA; 11) Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio, Finland.

Low frequency functionally deleterious variants are enriched in the Finnish population compared to non-Finnish European populations, due to a bottleneck when the population was founded. Further bottlenecks occurred when the Northern and Eastern parts of Finland were settled over the last 400 years. We sequenced the exomes of 20,029 individuals from these late settlement parts of Finland drawn from two large population-based longitudinal studies of metabolic and cardiovascular risk phenotypes (METSIM and FINRISK) and identified > 1.3M single nucleotide and indel variants. Combining the exome sequences with genome-wide array data, we surveyed fine-scale population structure in Northern/Eastern Finland using haplotype-based methods. We found enrichment of functional variants among the most bottlenecked subpopulations of Finland, and geographical clustering of variant origins even over the last 100 years. We then tested 64 quantitative cardiovascular and metabolic measures for genetic associations, after conditioning on known associated signals in the literature, and identified at exome-wide significance level ($5x10^{-7}$) 135 novel loci associated with at least one trait, and 52 traits with at least one novel association. These include novel loci for kidney function, HDL cholesterol, and anthropometric measures such as body weight and hip circumference. We observed a general tendency of geographical clustering as well as enrichment in frequency in Finland among the novel associations, again attesting to the power of leveraging population demographic history for gene discovery and the generalization of the Finnish Disease Heritage to quantitative complex phenotypes. Together, these results suggest that further expansion in sample size and detailed phenotyping in cohorts of special population history will be fruitful for future gene discovery experiments.

85

The genetic etiology of metabolic traits in people of Hispanic/Latino ancestry: Large-scale meta-analysis of single variant effects and gene-based functionally oriented analyses in 35,000 Hispanic/Latino individuals. *J.E. Below[1], L.E. Petty[1], M. Graff[2], X. Guo[3], Y. Hai[3], J. Yao[3], A. Manichaikul[4], C. Schurmann[5], C. Gao[6], D. Nousome[7], J.M. Mercader[8], X.Q. Wang[4], L.S. Emery[9], T. Sofer[9], C.L. Hanis[10], R. Loos[5], N.D. Palmer[6], J. McCormick[11], S. Fisher-Hoch[11], J.C. Florez[8,12,13], R. McKean-Cowdin[7], E.J. Parra[14], J.I. Rotter[3], K.E. North[2].* 1) Vanderbilt University Medical Center, Nashville, TN; 2) Department of Epidemiology, University of North Carolina, Chapel Hill, NC; 3) HARBOR-UCLA Medical Center, Torrance, CA; 4) Center for Public Health Genomics, University of Virginia, Charlottesville, VA; 5) Icahn School of Medicine at Mount Sinai, New York City, NY; 6) Wake Forest School of Medicine, Winston-Salem, NC; 7) Keck School of Medicine, University of Southern California, Los Angeles, CA; 8) Programs in Metabolism and Medical & Population Genetics, Broad Institute, MA; 9) Department of Biostatistics, University of Washington, Seattle, WA; 10) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX; 11) University of Texas Health Science Center at Houston, Division of Epidemiology, School of Public Health, Brownsville Campus, Brownsville, TX; 12) Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, MA; 13) Department of Medicine, Harvard Medical School, MA; 14) Department of Anthropology, University of Toronto at Mississauga, Mississauga, Ontario, Canada.

The genetics of type 2 diabetes (T2D) and dyslipidemia remain understudied in Hispanic/Latino populations despite worsening health disparities. Specifically, large-scale meta-analyses and functional interpretation have not previously been undertaken and represent an untapped resource for further discovery of metabolic disease etiology. In the most powerful genetic analysis of T2D and lipid traits in Hispanics/Latinos to date, we present both single variant (~20M SNPs) and functionally oriented, gene-based (~7200 genes) findings in 13,152 T2D cases and 21,511 controls, and ~23,400 samples with HDL, LDL, triglycerides, and total cholesterol data. Using MetaXcan, our approach leverages tissue-specific GTEx expression quantitative trait loci data in metabolically relevant tissues to give biological context to results and highlight genes whose genetically regulated expression levels may impact risk of metabolic disease in a heterogeneous and underrepresented minority population. Expression model validation for whole blood was performed using RNAseq data from an additional 75 Hispanic/Latino samples. Our findings identified 37 novel gene or locus associations and provide additional support for 72 genes previously implicated in T2D and lipids traits, giving new perspective on the effects of specific regulatory variants that may underlie broad GWAS peaks in Hispanics/Latinos. Single variant lipid findings of interest include a novel locus rs10416306 (p=1.65E-8) with HDL, which mapped to *TFPT*, rs174530 (p=1.33E-14) with LDL, intronic to *MYRF*, which is a known gene for other lipid traits and in the *FADS1* region, and rs1265044 (p=2.05E-8) with LDL, which maps to *PSORS1C1*, a gene linked to height and several autoimmune disorders including type 1 diabetes. T2D single variant results highlighted several novel associations, including rs72683520 (p=3.30E-8), downstream of *SLC45A4*. Gene-based analyses revealed 34 novel genes for lipid traits and one novel T2D gene, *SLC22A18,* located distally to known T2D gene *KCNQ1*. Gene ontology enrichment analysis of MetaXcan results identified very-low density lipoprotein remodeling, cholesterol homeostasis, and triglyceride homeostasis as the most significantly enriched biological pathways, lending further support to these novel association findings. Via large-scale meta-analysis and analysis based on functional annotation, results highlight potential regulatory signatures that are predictive of metabolic health in Hispanics/Latinos.

86

Identification and validation of novel regulatory genes for simultaneous lipid and blood glucose control in a large coronary artery disease (CAD) cohort using integrative RNA, DNA, metabolomics and clinical trait causal networks. *A. Cohain[1], N.D. Beckmann[1], D. Jordan[1], G.M. Belbin[1], A. Ruusalepp[2], R. Do[1], J.LM. Bjorkegren[1,2,3,4], E.E. Schadt[1,4].* 1) Genetics & Genomic Sciences, Icahn Institute of Genomics and Multiscale, New York, NY; 2) Department of Cardiac Surgery, Tartu University Hospital, Tartu, Estonia; 3) Cardiovascular Institute, Icahn School of Medicine at Mount Sinai, New York, NY; 4) Clinical Gene Networks AB, Stockholm, Sweden.

Coronary Artery Disease (CAD) is the leading cause of mortality and morbidity in the United States (US). Additionally, CAD is significantly co-morbid with Type II Diabetes (T2D), another major contributor to the US population health burden. While the molecular pathology of CAD has been extensively characterized, the mechanistic relationship between CAD and T2D remains poorly understood. T2D patients are observed to be at increased risk for cardiovascular diseases, while conversely patients taking cholesterol-lowering drugs often exhibit elevated blood glucose levels, and thus are at higher risk for T2D. As of 2012, 27.9% of all adults over 40 years old were prescribed statins to lower their cholesterol. Furthermore, recent research has shown through Mendelian randomization approaches that alleles that lower LDL increase T2D risk. For these reasons, we sought to use a data-driven approach to investigate and ultimately elucidate the mechanisms underlying lipid and blood glucose metabolism in CAD. We utilized the STARNET study, with over 600 patients undergoing coronary artery bypass graft surgery and 7 tissues collected (containing aortic root, mammary artery, skeletal muscle, liver, visceral fat, subcutaneous fat and whole blood). From these tissues, RNA was sequenced, DNA genotyped, metabolomics profiled and clinical phenotypes recorded. This is among the richest datasets to date for studying CAD and other metabolic disorders in a human disease cohort. Differential gene expression analysis between lipid traits (LDL, HDL, triglycerides, plasma cholesterol, cholesterol medication, hyperlipidemia) and blood glucose traits (HbA1c, blood glucose, T2D, insulin medication, oral anti diabetics medication) highlighted liver as the most informative tissue. Using coexpression networks analysis on liver, we found a group of 60 genes (containing HMGCR, PCSK9, NPC1L1) that were co-expressed only in the liver, and that correlated with decreased low-density lipoprotein cholesterol and increased blood glucose levels. This signal recapitulates the phenotypic effect of statins. We validated these findings in two external datasets – 1 human and 1 mouse. By incorporating the expression of these genes, lipid metabolomics, and clinical traits in multiscale causal networks, we were able to detect key novel gene regulators of this set of 60 genes. This study helps untangle the biology of the relationship between lipids and blood glucose and their co-regulatory mechanisms.

87

Functional annotation of common noncoding and rare coding variants in *ANGPTL3*. *X. Wang[1], A. Raghavan[2], A.C. Vourakis[3], A.E. Sperry[3], W. Li[1], W. Lv[1], A.C. Chadwick[1], K. Musunuru[1].* 1) Department of medicine, University of Pennsylvania, Philadelphia, PA; 2) Harvard Medical School, Boston, MA; 3) Harvard University, 7 Divinity Ave, Cambridge, MA.

Angiopoietin-like 3 (*ANGPTL3*) associates strongly with blood lipid phenotypes in human genetics studies and has thus emerged as promising therapeutic target for plasma lipids. Exome sequencing of patients with familial combined hypolipidemia identified rare nonsense *ANGPTL3* mutations, and genome-wide association studies (GWAS) identified common variants near the *ANGPTL3* locus that associate with decreased triglycerides (TG) and low-density lipoprotein cholesterol. In light of the seemingly favorable clinical consequences of *ANGPTL3* deficiency, we established an experimental framework to (1) identify causal common variants that regulate *ANGPTL3* expression and (2) determine the functional consequences of rare *ANGPTL3* missense mutations. All the common variants linked to the lead GWAS SNP ($r^2 \geq 0.5$) in the *ANGPTL3* locus were profiled by massively parallel reporter assays, and rs10889356 demonstrated significant allele-specific enhancer activity. To validate this, we introduced the minor allele into the H7 human pluripotent stem cell line (major/major at rs10889356) using the PiggyBac transposon system. We next differentiated edited H7 cells into hepatocyte-like cells (HLCs) and observed a 60% increase in *ANGPTL3* expression (P = 0.0004). We also used CRISPR-Cas9 to delete 36-39 base pairs flanking the SNP in the H7 line. When differentiated into HLCs, homozygous deleted H7 cells had a 67% increase in *ANGPTL3* expression ($P = 0.007$). These findings support rs10889356-*ANGPTL3* as a causal SNP-gene set. Next, we examined the coding regions of *ANGPTL3* for missense variants identified through exome sequencing and found 77 rare variants in individuals from separate human cohorts. We sought to experimentally define these variants *in vivo*. Using CRISPR-Cas9, we generated an *Angptl3* knockout mouse, which exhibited decreased TG (61%, P < 0.001) and decreased total cholesterol (31%, P < 0.002). We attempted to rescue this phenotype using adenoviruses expressing either wild-type or missense variant *ANGPTL3*. To date, 52 rare missense variants have been assessed, of which 19 were validated in terms of loss-of-function as severe (conferring <25% of wild-type activity as assessed by either TG or cholesterol levels), 15 were moderate (25-50%), 10 were mild (50-75%), and 7 were benign (75-125%) while 1 was gain-of-function (>125%), underscoring the need for functional characterization of variants of uncertain significance.

88

Zebrafish larvae as a model system for high-throughput, image-based genetic screens in cardiometabolic diseases. *M. den Hoed[1,2], A. Emmanouilidou[1,2], M. Bandaru[1,2], B. von der Heyde[1,2], T. Klingström[1,2], C. Wählby[2,3], P. Ranefall[2,3], A. Allalou[2,3], A. Larsson[4], E. Ingelsson[5].* 1) Department of Immunology, Genetics and Pathology, Uppsala University, Sweden; 2) Science for Life Laboratory, Uppsala University, Sweden; 3) Department of Information Technology, Division of Visual Information and Interaction, Uppsala University, Sweden; 4) Department of Medical Sciences, Clinical Chemistry, Uppsala University, Sweden; 5) Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, USA.

Background: Genome-wide association, exome array and whole-exome sequencing efforts have identified hundreds of loci that are robustly associated with the risk of cardiometabolic diseases and risk factors. With few exceptions, the causal genes through which these loci influence the risk of disease remain uncharacterized. While *in silico* genomic annotation using results from e.g. the ENCODE and RoadMap Epigenomics projects provides valuable insights, novel model systems that enable systematic, *in vivo* characterization of candidate genes are desirable. **Methods:** My group has developed and validated zebrafish model systems that make optimal use of: 1) the zebrafish' well-annotated genome, with orthologues of ≥71.4% of human genes; 2) recent developments in multiplex CRISPR-Cas9-based mutagenesis; 3) advances in automated positioning of non-embedded zebrafish larvae; 4) fluorescent transgenes and dyes; and 5) custom-written image-quantification pipelines. **Results:** Five days of overfeeding, dietary cholesterol supplementation and/or exposure to glucose induce atherogenic and insulin resistant phenotypes (higher/more whole-body LDL cholesterol (LDLc) levels; vascular foam cell formation and inflammation; β-cell number and volume; subcutaneous and hepatic accumulation of fat) that can largely be prevented by concomitant treatment with lipid-lowering or diabetes medication (N>4000). Proof-of-principle studies show that each additional mutated allele in the zebrafish' orthologues of *APOE* (apoea, apoeb) results in higher LDLc levels and more vascular foam cell formation and inflammation (N~384). In line with recent results in humans, treatment with LDLc-lowering drugs (N~400) and mutations in pcsk9 (N~384) both result in higher whole-body glucose levels. Finally, characterization of candidate genes in loci identified in a recent GWAS for heart rate variability helped identify genes that influence early-stage cardiac development, cardiac rate, and/or cardiac rhythm. **Conclusions:** Systematic, largely image-based characterization of candidate genes for cardiometabolic traits in zebrafish model systems will increase our understanding of human disease, and will likely identify novel targets that can be translated into efficient therapeutics. In addition, undesirable side effects can be quantified *in vivo* at an early stage, thereby preventing costly and time-consuming experiments for targets that would otherwise likely fail on the road towards clinical trials.

89

The landscape of chromosomal aberrations in *in vitro* fertilized preimplantation human embryos. *S. Madjunkova[1], R. Abramov[1], R. Antes[1], V. Kuznyetsov[1], C. Librach[1,2,3,4,5].* 1) Preimplantation Genetic Department, Create Fertility Centre, Toronto, Ontario, Canada; 2) Department of Obstetrics and Gynecology, University of Toronto, Toronto, Canada; 3) Institute of Medical Sciences, University of Toronto, Toronto, Canada; 4) Department of Physiology, University of Toronto, Toronto, Canada; 5) Department of Gynecology, Women's College Hospital, Toronto, Canada.

Introduction: Preimplantation genetic screening (PGS) of human embryos offers unprecedented information on chromosomal copy number enabling selection of the fittest embryo for transfer. The underlining molecular mechanisms for the high frequency and diversity of aberrations seen in human embryos is still understudied. Both meiotic and mitotic errors in chromosome (chr) segregation contribute to these abnormalities. Mitotic errors may lead to mosaic embryos with several cell lineages coexisting in one embryo. Next generation sequencing (NGS) has increased the detection resolution for mosaicism and chromosomal aberrations overall. The aim of this study was to evaluate the type and frequency of chromosomal aberrations detected by NGS screening in embryos derived from in vitro fertilization. **Methods:** PGS using lowpass whole genome NGS of trophectoderm biopsies was performed at the CReATe Fertility Centre Genetics laboratory. The resolution of chr aberration detection is ≥ 10Mb with mosaicism detection of ≥ 30%. **Results:** We screened 2064 embryos from 574 patients (median maternal age of 36.1). Egg donation was utilized for 1019 embryos (median age of 26.6). Overall whole chr aneuploidy was found in 20.5%, segmental aneuploidies in 3.1%, and diploid/aneuploid (D/A) mosaicism ≥30% in 20% of embryos, with 43.7% of them harbouring complex (≥2) aberrations. Maternal age was strongly associated with aneuploidy ($p<0.05$). In contrast, mosaicism and segmental aberrations showed no age association. Trisomies (T) were more prevalent than monosomies (M) in aneuploid embryos than in mosaic D/A embryos (54% vs 44.3%, $p=0.0004$), while segmental gains (52.5%) and loses (47.5%) were not significantly different. The rate of D/A mosaicism was higher than aneuploidy mosaicism (34.7% vs 22%, OR 1.879, 95%CI:1.47-2.404). The most frequent full chr aneuploidies were T: 16,22,15,21,9,13 and M: 16,22,21,X,15,13. In mosaic embryos the most frequent aberrations were M: 21,X,22,2,Y and T: 19,12,16, and 6. **Conclusion:** The overall rate of chr aberrations was 43.6% with high frequency of complex aberrations involving all chromosomes, and was significantly associated with maternal age. Mitotic errors were detected in 25% of embryos with a D/A mosaicism rate of 20% and this was not associated with maternal age. The finding of more common specific aberrations in aneuploid and mosaic embryos (M: 21, X and 22 and T 16) may support the hypothesis of self-correction and survival of the fittest cells.

90

Experience from the first live-birth derived from oocyte nuclear transfer as a treatment strategy for mitochondrial diseases. *T. Huang[1], H. Liu[2,3], S. Luo[1], Z. Lu[3], A. Chávez-Badiolaa[2], Z. Liu[3], M. Yang[3], Z. Merhi[4], S. Silber[5], S. Munne[6], M. Konstantinidis[6], D. Wells[6], J. Tang[7], J. Zhang[2,3].* 1) Department of Pediatrics/Division Human Gen, Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 2) New Hope Fertility Center, Punto Sao Paulo, Lobby Corporativo, Américas 1545 Providencia, Guadalajara, Mexico; 3) New Hope Fertility Center, 4 Columbus Circle, New York, NY 10019, USA; 4) Department of Obstetrics and Gynecology, Division of Reproductive Biology, NYU School of Medicine, 180 Varick Street, New York, NY 10014, USA; 5) Infertility Center of St Louis, St Luke's Hospital, St Louis, MO 63017, USA; 6) Reprogenetics, 3 Regent Street, Livingston, NJ 07078, USA; 7) Department of Obstetrics and Gynecology, The Mount Sinai Hospital, E 101st St, New York, NY 10029, USA.

Mutations in mitochondrial DNA (mtDNA) are maternally inherited and can cause fatal or debilitating mitochondrial disorders. The severity of clinical symptoms is often associated with the level of mtDNA mutation load or degree of heteroplasmy. Current clinical options to prevent transmission of mtDNA mutations to offspring are limited. Experimental spindle transfer (ST)[CF1] in metaphase II oocytes, also called mitochondrial replacement therapy, is a novel technology for preventing mtDNA transmission from oocytes to pre-implantation embryos. Here we report a female carrier of Leigh Syndrome (mtDNA mutation 8993T>G), with a long history of multiple undiagnosed pregnancy losses and deaths of offspring due to this disease, who underwent in-vitro fertilization following reconstitution of her oocytes by spindle transfer into the cytoplasm of enucleated donor oocytes. A male euploid blastocyst was obtained from the reconstituted oocytes, which had only a 5.7% mtDNA mutation load. Transfer of the embryo resulted in a pregnancy with delivery of a boy with neonatal mtDNA mutation load of 2.36-9.23% in his tested tissues. The boy is currently healthy at 12 months of age although long-term follow-up of the child's longitudinal development remains crucial. In this presentation, we will aslo discuss the safety of mitochondrial replacement therapy by nuclear transfer between oocytes of two different women, due to of the possibility of mitochondrial DNA (mtDNA) heteroplasmy drift from small amounts of mtDNA carryover due to the nuclear-mitochondrial incompatibility. It has been reported in several recent *in vitro* studies that, even though the low levels of heteroplasmy introduced into human oocytes often vanish, they can sometimes result in mtDNA genotypic drift and reversion to the original genotype in some of the reconstituted human embryonic-derived stem cell (hESC) lines. We will aslo discuss the rigor of these studies and whether these *in vitro* experiments reflect the *in vivo* applications. .

91

Non-invasive prenatal screening for single gene disorders in pregnancies with abnormal ultrasound findings or advanced paternal age.

J. Li[1], Y. Feng[1], J. Sinson[1], H. Dai[2], X. Ge[2], G. Wang[1], H. Mei[2], A. Breman[2], A. Purgason[1], A. Pourpak[1], X. Wang[2], I. Van den Veyver[2,3], A. Beaudet[2], L. Wong[2], C. Eng[2], J. Zhang[2]. 1) Baylor Genetics, Houston, TX; 2) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 3) Departments of Obstetrics and Gynecology, Baylor College of Medicine, Houston, TX.

Background and Purpose: The incidence of single gene disorders in liveborn is ~0.36%, while the aggregated incidence of chromosomal anomalies is 0.18%. Yet, current non-invasive prenatal testing (NIPT) is targeted towards detection of chromosomal abnormalities in the fetus, while a prenatal screening test for pathogenic variants in multiple single genes is not available. Here we report on the development, validation and our experiences in the initial clinical offering of the first non-invasive prenatal screening test on circulating cell-free DNA (cfDNA) in maternal blood for *de novo* or paternally inherited pathogenic variants in 30 genes associated with dominant monogenic diseases. **Methods:** After tagging cfDNA with unique molecular index by adaptor ligation and hybridization-based target enrichment followed by next-generation sequencing, the target region was analyzed with average read-depth of >1,000X. A set of regions containing highly polymorphic SNPs were used to determine fetal fraction. This test was validated using cfDNA from 43 pregnant women and 47 spike-in samples with known pathogenic variants. All positive results were confirmed by a secondary assay and/or Sanger sequencing on DNA from invasive or postnatal specimens. **Results:** The SNP-based fetal fraction calculation yielded consistent results with Y-chromosome method. Analytical sensitivity and specificity were >99% in regions with sufficient coverage (>200X) and the cut-off fetal fraction is 4.5%. In more than 100 consecutive samples tested, pathogenic variants including those in *COL1A1*, *FGFR3*, *NIPBL* and *RIT1* were successfully identified and confirmed in pregnancies with abnormal ultrasound findings or known paternal history of conditions included in the screening panel. **Conclusion:** We developed a highly sensitive and specific non-invasive prenatal screening method for *de novo* or paternally inherited pathogenic variants in maternal blood. Initial offering of this test demonstrated its usefulness for pregnancies with abnormal ultrasound findings, or positive paternal history in the related genes. Clinical studies on larger numbers of samples from pregnant women are in progress to evaluate the clinical performance of this new test which has a potential to be offered as a population-based non-invasive prenatal screening for single gene Mendelian disorders caused by *de novo* mutations, and in the setting of advanced paternal age as an extension of NIPT for aneuploidy.

92

Cell-based noninvasive prenatal testing enables detection of benign and pathogenic copy number variants at much higher sensitivity than cell-free NIPT methods.

L. Vossaert[1], A. Breman[1], J. Chow[2], L. U'Ren[2], Q. Wang[1], R. Salman[1], S. Qdaisat[1], A. Kim[1], X. Zhuo[1], E. Normand[1], C. Shaw[1], D. Henke[1], E. Chang[2], R. Seubert[2], J. Stilwell[2], E. Kaldjian[2], Y. Yang[1], I. Van den Veyver[1,3], A. Beaudet[1]. 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) RareCyte Inc., Seattle, WA; 3) Obstetrics and Gynecology, Baylor College of Medicine, Houston, TX.

Background: Cell-based noninvasive prenatal testing (CB-NIPT) is a promising alternative for the detection of fetal copy number abnormalities. The isolation of pure fetal genomic DNA from circulating single trophoblastic cells offers significant advantages over the currently available cell-free NIPT methodologies, for which the fetal fraction comprises approximately 10% of the total plasma DNA. We previously showed the clinical potential of our CB-NIPT assay to detect fetal microdeletions as small as 2.7 Mb in size (Breman *et al.,* 2016, Prenatal Diagnosis). To date, we also have repeatedly demonstrated the capacity of our assay to detect common benign copy number variants (CNVs) in the 1-2 Mb size range, thus further establishing the sensitivity of our assay. Methods: Maternal blood samples were collected from pregnant women, who in most cases also underwent an invasive procedure. Fetal trophoblastic cells were obtained after initial maternal white blood cell depletion, density-based enrichment, immunofluorescence staining and high-resolution scanning and analysis. Individual fetal cells were picked and underwent whole genome amplification. Single nucleotide polymorphism-based genotyping was performed to confirm their fetal origin. Genome-wide copy number data were collected through low-pass paired-end next-generation sequencing on an Illumina platform aiming for 5 to 10 million reads per sample. Concurrent clinical data were generated by array comparative genomic hybridization (arrayCGH) analysis of amniotic fluid and chorionic villi samples. Results: Concordance between our own CB-NIPT data and available clinical data was high, and consisted of multiple autosomal trisomies, sex chromosome aneuploidies and several pathogenic CNVs of 2 Mb or larger. For the latter, we confirmed cases with an 18.9 Mb deletion of chromosome 4p, a complex 15 Mb duplication of chromosome 16p, a 6 Mb deletion of chromosome 1p and a 2.7 Mb deletion of chromosome 15q. Additionally, multiple cases demonstrated benign polymorphic CNVs in the 1-2 Mb size range that were also concordant with the clinical arrayCGH data. Conclusions: We show here that our CB-NIPT test yields results with a higher sensitivity for smaller CNVs than the currently available cell-free testing, which is limited to events >7 Mb. We are currently in the process of optimizing our method further with the goal of its clinical application in the near future.

93

Individuals with monosomy X mosaicism (XO/XX) detected among 63,350 UK Biobank females appear asymptomatic of Turner syndrome. *M.A. Tuke[1], K.S. Ruth[1], R.N. Beaumont[1], J. Tyrrell[1], S.E. Jones[1], H. Yaghootkar[1], C.L. Turner[2], M. Donohoe[3], A. Brooke[3], M. Collinson[4], R.M. Freathy[1], A.R. Wood[1], M.N. Weedon[1], T.M. Frayling[1], A. Murray[1].* 1) Genetics of Complex Traits, University of Exeter Medical School, Exeter, Devon, United Kingdom; 2) Peninsula Clinical Genetics Service, Royal Devon and Exeter Hospital, Exeter, UK; 3) Macleod Diabetes & Endocrine Centre, Royal Devon and Exeter Hospital, Exeter, UK; 4) Wessex Regional Genetics Laboratory, Salisbury District Hospital, Salisbury, UK.

Introduction Women with X chromosome aneuploidy such as XO (Turner syndrome) or XXX (Triple X syndrome) present with a range of characteristics relating to stature, an increased risk of cardiac problems and premature ovarian failure. Women with Turner syndrome should receive regular medical screening for complications. Studies assessing X chromosome aneuploidy to date have likely been subject to ascertainment bias. We aimed to characterise the prevalence and phenotypic consequences of X chromosome aneuploidy in the general population using UK Biobank. *Methods* We quantile normalised SNP chip data from all UK Biobank females and processed them using PennCNV-Affy. Individuals were identified as having X aneuploidy if both their mean Log R Ratio and count of strict heterozygous B Allele Frequencies were outliers from the sample distribution. XO, XO/XX mosaics and XXX groups were identified and phenotypic variables were compared with 63,247 XX controls. *Results* We detected 8 "full" XO samples (estimated >80% XO cells), 55 mosaic XO/XX and 40 with XXX. X chromosome dosage was validated in a sample of known cases of X chromosome aneuploidy. All 63 samples with X chromosome loss were on average 14cm shorter than XX females in UK Biobank ($p=1x10^{-8}$), driven largely by the 8 full XO samples. The XO mosaic samples were not substantially shorter than XX females (2.2cm shorter on average, p=0.007). Conversely, samples with an additional X chromosome were on average 7cm taller ($p=6.5x10^{-11}$). Most XO cases did not report going through menarche, but mean menarcheal age in the 55 mosaic XO cases was not significantly different to controls (p=0.2). Only 8/55 reported never being pregnant and there was no increased incidence of miscarriage, termination or stillbirth with an average of 2.3 pregnancies. XXX cases had a natural menopause on average 7 years, 2 months earlier than controls ($p=3.9x10^{-8}$). *Conclusions* We identified 103 women in the UK Biobank with X chromosome aneuploidy. Females with XXX were taller than average and reported earlier menopause, whereas women with full XO were shorter and did not reproduce. In contrast, XO/XX mosaic females had a normal reproductive lifespan and birth rate, with no reported cardiovascular complication. This study characterises X chromosome aneuploidy phenotypes in a population-based sample of older individuals and presents implications for future management of women with an XO/XX mosaic karyotype, particularly those identified incidentally.

94

Discernment in early childhood: Neurodevelopmental outcome and early hormonal therapy (EHT) in a large, prenatally diagnosed population of 47, XXY Boys. *C. Samango-Sprouse[1,2,3], C. Keen[3], F. Mitchell[3], D. Sargsyan[4,5,6], L. Petrosyan[7], T. Sadeghin[3], A. Gropman[8,9].* 1) Deparment of Pediatrics, George Washington University 2121 I St. NW Washington, DC 20052; 2) Department of Human and Molecular Genetics, Florida International University 11200 SW 8th St. Miami, FL 33199; 3) The Focus Foundation 820 W Central Ave. #190 Davidsonville, MD 21035; 4) Janssen Research and Development 920 Route 202 Raritan, NJ 08869; 5) Cardiovascular Institute, Robert Wood Johnson Medical School 675 Hoes Lane Piscataway Township, NJ 08854; 6) Ernest Mario School of Pharmacy 160 Frelinghuysen Road Piscataway Township, NJ 08854; 7) Department of Physics, Jackson State University 1400 John R. Lynch St. Jackson, MS 39217; 8) Department of Neurology and Pediatrics, George Washington University 2121I St. NW Washington, DC 20052; 9) Neurodevelopmental Disorders and Neurogenetics, Children's National Medical Center 111 Michigan Ave NW Washington, DC 20010.

Introduction: 47, XXY is the most frequently occurring XY variation with an estimated prevalence of 1 in 660 men. Although this condition occurs with relative frequency, only 25% of cases are ever clinically ascertained. Features include eunuchoid body proportions, infertility, reduced bone mineralization, language-based learning disorders, intact spatial cognition and normal IQ. There were several prospective studies involving large prenatally diagnosed populations from the 1970s and 1980s. However, since then there have been great strides in longitudinal treatment to mitigate some of the deficits associated with the 47, XXY learning and endocrine profile. Greater use of NIPS will lead to an increased number of cases identified than before. Updated and comprehensive study of a large, prenatally diagnosed cohort of 47, XXY boys is necessary to provide these newly expecting and recently diagnosed families' information on the early childhood trajectory and potential effects of EHT. Statement of Purpose: To provide current study of neurodevelopment in a large, prenatally identified population of 47, XXY boys and to investigate the potential effects of EHT. Methods: 168 prenatally identified boys with 47, XXY were referred for comprehensive neurodevelopmental evaluations at least once between the ages of 0 & 5 years. The study population was segregated by EHT status with an EHT-treated group (n=46) and an untreated group (n=122). These groups were compared across scales of neurodevelopmental domains tailored to the patient's age at the time of their visit. This included the PLS, $4^{th}/5^{th}$ edition, auditory comprehension (AC) and expressive communication (EC) and the Bayley Scales of Infant Development, $2^{nd}/3^{rd}$ edition (Bayley) mental development index (MDI) and psychomotor development index (PDI). Results: The EHT-treated group showed a significant difference from the untreated group in EC scores on the PLS at 18 months of age and older (p=0.027). The EHT-treated cohort of 47, XXY boys had significant differences in cognition (MDI) on the Bayley between 18 and 36 months compared to the untreated cohort (p=0.031). Discussion: The EHT-treated group of 47, XXY boys had significant differences in EC and MDI. This study demonstrated positive effects of EHT on neurodevelopment supporting the presence of androgen deficiency in 47, XXY. Future study is warranted to elucidate the optimal window of opportunity and dosage for treatment during early childhood years.

95

The spectrum of loss of function intolerance in the human genome.

K.J. Karczewski[1,2], L.C. Francioli[1,2], K.E. Samocha[3], B.B. Cummings[1,2], D.P. Birnbaum[1,2], M.J. Daly[1,2], D.G. MacArthur[1,2], Genome Aggregation Database. 1) Massachusetts General Hospital, Boston, MA; 2) Broad Institute, Cambridge, MA; 3) Wellcome Trust Sanger Institute, Cambridge, UK.

Deciphering the function and essentiality of genes in the genome is a central problem in human genetics. While knockout generation is a workhorse of elucidating gene function in model organism genetics, obvious ethical and technical limitations prevent its use in humans. However, large-scale exome and genome sequencing panels allow us to identify naturally-occurring loss-of-function (LoF) variants that provide in vivo whole-organism models of gene inactivation in humans. The presence of LoF variants at high rates suggests a gene's redundancy, while a significant depletion compared to expectation suggests strong selective pressures against these variants, and thus, the gene's essentiality. Understanding exactly where each human gene lies along the spectrum between these extremes is important for prioritizing variants both as candidate disease genes and for the development of inhibitory therapeutics. Using a mutational model to generate expected numbers of variants for each gene in the genome, we previously found 3,230 genes were found to be significantly depleted (constrained) for LoF variation, even in the heterozygous state, in data from the Exome Aggregation Consortium (ExAC). Here, we expand this model using data from 15,496 genomes and 123,136 exomes from the Genome Aggregation Database (gnomAD). We have updated our mutational model using the non-coding regions of this expanded sample set and incorporated CpG methylation status for each base in the genome. We present a refined constraint model by incorporating variant prioritization algorithms such as LOFTEE, PolyPhen, and CADD, as well as transcript expression. We also incorporate allele frequency information to assess constraint against both recessive and dominant LoF variation. We apply this model to high-confidence LoF variants to these 123,136 exomes, in order to define a set of genes constrained against heterozygous and homozygous variation, substantially expanding the constrained list accessible using prior data sets. In addition, we use inferred haplotype phase to identify compound heterozygous LoF individuals, and define an updated list of over 2,000 genes that are homozygously inactivated in living individuals. This study substantially enhances our understanding of the impact of gene-inactivating variation across all human protein-coding genes, and will aid in the discovery of disease-associated genes and therapeutics.

96

When are predicted loss-of-function (LOF) mutations not LOF mutations?

Z. Coban Akdemir[1], J.J. White[1], Y. Bayram[1], S.N. Jhangiani[2], T. Gambin[3], E. Boerwinkle[2,4], R.A. Gibbs[1,2], C.M.B. Carvalho[1], J.R. Lupski[1,2,5,6]. 1) Molecular and Human Genetics, Baylor College of Medicine, HOUSTON, TX; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; 3) Institute of Computer Science, Warsaw University of Technology, Warsaw 00-665, Poland; 4) Human Genetics Center, University of Texas Health Science Center at Houston, TX 770530,USA; 5) Texas Children's Hospital, Houston, TX 77030, USA; 6) Department of Pediatrics, Baylor College of Medicine, Houston, Texas 77030, USA.

Nonsense-mediated decay (NMD) is a surveillance pathway that results in degradation of mRNA transcripts bearing a premature truncation codon (PTC). According to the most widely accepted model of NMD, mutant transcripts containing PTCs located in the last 50 bp of the penultimate exon and the entire last exon, can escape NMD resulting in translation of a mutant protein product. To select for transcripts with variants that may escape from NMD and possibly underlie human disease via a dominant-negative or gain-of-function mechanism, we examined the Baylor-CMG database of ~6000 exomes consisting of a wide variety of Mendelian phenotypes. First, we developed an algorithm to classify frameshift variants that are potentially subject to either degradation (NMD-competent) or escaping degradation (NMD-incompetent). Second, we designed an NMD-incompetency score metric to rank each gene based on the enrichment of NMD-incompetent vs. NMD-competent truncating variants using the ExAC (~60K exomes) and ARIC databases (~11K exomes). In total, this analysis revealed 363 genes significantly depleted (Bonferroni-corrected P value <= 0.05) for NMD-incompetent variants in control databases. Therefore, this set of genes could be potential candidates for causing disease only through escape from NMD. In fact, a subset of those genes presents at least two distinct NMD-incompetent truncating variants in patients with similar clinical phenotypes in the Baylor-CMG database. These include *DVL1* (autosomal dominant Robinow syndrome MIM #616331), in which truncated transcripts escape NMD and are hypothesized to mediate pathogenicity via a gain-of-function and/or dominant-negative mechanism (PMID: 25817016). Importantly, this analysis enabled the identification of a novel gene for Hereditary gingival fibromatosis (HGF). RE1-silencing transcription factor gene (*REST*) gene (OMIM *600571) was found to be depleted for NMD-incompetent truncating variants in the ARIC database; however, two distinct truncating mutations in unrelated patients presented with HGF (AJHG, under review). Collectively, we demonstrate that leveraging the knowledge of NMD and an NMD-incompetency score metric is an effective and efficient tool in discovery of novel disease genes due to mechanisms other than haploinsufficiency. Furthermore, our data suggest that gain-of-function and dominant-negative mutations are under-recognized in genomic analyses and likely contribute to a wide variety of human disease phenotypes.

97

Quantifying the impact of rare and ultra-rare coding variation across the phenotypic spectrum. *A. Ganna[1,2,3,4], F.K. Satterstrom[1,2,3], S.M. Zekavat[2,5], I. Das[6,7], C. Churchhouse[1,2,3], J. Alfoldi[1,2], A.R. Martin[1,2,3], A.S. Havulinna[8,13], A. Byrnes[1,2,3], W.K. Thompson[9,10], P.R. Nielsen[18,19], K.J. Karczewski[1,2], M.I. Kurki[1,2,8], M.A. Rivas[12], N. Gupta[2], J. Flannick[2,5], V. Salomaa[13], C. Hultman[8], S. Ripatti[8,14,15], O. Kuismin[11], P. Bo Mortensen[16,17], D. MacArthur[1,2], M.J. Daly[1,2,3], P.F. Sullivan[4,18], A.E. Locke[6,7], A. Palotie[1,2,3,8], J.C. Florez[2,5], A.D. Børglum[16,19], S. Kathiresan[2,5], B.M. Neale[1,2,3], GoT2D/T2D-GENES, SIGMA, Helmsley IBD Exome Sequencing Project, FinMetSeq Consortium, iPSYCH-Broad.* 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) Center for Genomic Medicine, Massachusetts General Hospital and Department of Medicine, Harvard Medical School, Boston, MA, USA; 6) McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA; 7) Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI, USA; 8) Institute for Molecular Medicine Finland, FIMM, University of Helsinki, Helsinki, Finland; 9) University of California San Diego, Department of Family Medicine and Public Health, La Jolla, CA, USA; 10) Institute of Biological Psychiatry, Copenhagen University, Copenhagen, Denmark; 11) Department of Clinical Genetics, Oulu University Hospital, Medical Research Center Oulu and PEDEGO Research Unit, University of Oulu, Oulu, Finland; 12) Department of Biomedical Data Science, Stanford University, Stanford, CA, USA; 13) Department of Health, THL-National Institute for Health and Welfare, Helsinki, Finland; 14) Department of Public Health, University of Helsinki, Helsinki, Finland; 15) Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; 16) iPSYCH - Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus University, Aarhus, Denmark; 17) National Centre for Register-based Research, School of Business and Social Sciences, Aarhus University, Aarhus, Denmark; 18) Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA; 19) Department of Biomedicine, Aarhus University, Aarhus, Denmark.

Protein truncating variants (PTVs) are likely to modify gene function and have been linked to hundreds of Mendelian disorders. However, the impact of PTVs on complex traits has been limited by the available sample size of whole-exome sequencing studies (WES) and remains an open question. Here we assemble WES data from 100,304 individuals to quantify the impact rare PTVs play on 13 quantitative traits and 10 diseases. We focus on those PTVs that occur in PTV-intolerant (PI) genes, as these are more likely to be pathogenic Carriers of at least one PI-PTV were found to have an increased risk of autism, schizophrenia, bipolar disorder, intellectual disability, ADHD (P-value (p) range: $5x10^{-3}$-$9x10^{-12}$). Interestingly, in controls with none of these disorders, we found that this burden associated with increased risk of mental, behavioral and neurodevelopmental disorders more broadly. Furthermore, carriers of PI-PTVs tended to be shorter ($p=2x10^{-5}$), have fewer years of education ($p=2x10^{-4}$) and tended to be younger ($p=2x10^{-7}$); the latter observation possibly reflecting reduced survival or study participation. While other gene-sets derived from in-vivo experiments did not show any associations with PTV-burden, gene sets implicated in GWAS of cardiovascular-related traits and inflammatory bowel disease showed a significant PTV-burden with corresponding traits, entirely driven by established genes involved in familial forms of these disorders. We leveraged population health registries from 14,093 individuals to study the phenome-wide impact of PI-PTVs and identified an increase in the number of hospital visits among PI-PTV carriers. In conclusion, we provide the most thorough investigation to date of the impact of rare deleterious coding variants on complex traits.

98

Loss of function *ABCC8* mutations in pulmonary arterial hypertension. *W.K. Chung[1], M.S. Bohnen[2], L. Ma[1], N. Zhu[1,4], H. Qi[3,4], C. McClenaghan[5], C. Gonzaga-Jauregui[6], F.E. Dewey[6], J.D. Overton[6], J.G. Reid[6], A.R. Shuldiner[6], A. Baras[6], K.J. Sampson[2], U. Krishnan[1], E.B. Rosenzweig[1], Y. Shen[3,4], C.G. Nichols[5], R.S. Kass[2].* 1) Department of Pediatrics, College of Physicians and Surgeons, Columbia University, New York, NY, USA; 2) Department of Pharmacology, College of Physicians and Surgeons, Columbia University, New York, NY, USA; 3) Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA; 4) Department of Systems Biology, Columbia University, New York, NY, USA; 5) Department of Cell Biology and Physiology, and the Centre for the Investigation of Membrane Excitability Diseases, Washington University School of Medicine, Washington University in St Louis, St Louis, MO, USA; 6) Regeneron Genetics Center, Regeneron Pharmaceuticals Inc. Tarrytown, NY, USA.

In pulmonary arterial hypertension, pathological changes in pulmonary arterioles progressively raise pulmonary artery pressure and increase pulmonary vascular resistance, leading to right heart failure and high mortality rates. Recently, the first potassium channelopathy in pulmonary arterial hypertension, due to mutations in *KCNK3*, was identified as a genetic cause and pharmacological target. Exome sequencing was performed to identify novel genes in a cohort of 99 pediatric and 134 adult onset group I pulmonary arterial hypertension patients without identifiable mutations in known pulmonary arterial hypertension genes. Novel rare variants in the gene identified were independently identified in a cohort of 680 adult onset patients. Variants were expressed in COS cells and function assessed by patch-clamp and rubidium flux analysis. We identified a *de novo* novel heterozygous predicted deleterious missense variant c.G2873A (p.R958H) in *ABCC8* (ATP-binding cassette, subfamily C, member 8), in a child with idiopathic pulmonary arterial hypertension. We then examined all individuals for rare or novel variants in *ABCC8* and identified seven additional heterozygous predicted damaging variants. *ABCC8* encodes sulfonylurea receptor 1, a regulatory subunit of the ATP-sensitive potassium channel. We observed loss of function for all *ABCC8* variants evaluated, and pharmacological rescue by the SUR1 activator, diazoxide. Novel and rare missense variants in *ABCC8* are associated with pulmonary arterial hypertension. Identified *ABCC8* mutations decreased potassium channel function, which was pharmacologically recovered.

99

Completing a human gene knockout catalog through accurate phasing of 15K rare, deleterious compound heterozygous mutations in 61K exomes. *J. Staples[1], N. Gosalia[1], E.K. Maxwell[1], C.G. Gonzaga-Jauregui[1], M.F. Murray[2], D. Carey[2], F.E. Dewey[1], O. Gottesman[1], G.R. DiscovEHR[1,2], L. Habeggar[1], J.G. Reid[1].* 1) Regeneron Genetics Center, Tarrytown, NY; 2) Geisinger Health System, Danville, PA.

A primary goal of human genetics is to better understand the function of every gene in the genome. Homozygous loss-of-function mutations (LoFs) are a powerful tool to gain insight into gene function by analyzing the phenotypic effects of these “human knockouts” (KOs). Rare (MAF <1%) homozygous LoFs have been highlighted in recent large-scale sequencing studies and have been critical in identifying many gene-phenotype interactions. While rare compound heterozygous mutations (CHMs) of two heterozygous LoFs are functionally equivalent to a rare homozygous KO, they are rarely interrogated in large sequencing studies. Accurate identification of rare CHMs of LoFs is valuable: (1) rare CHMs substantially increase the number of human gene KOs, improving statistical power; (2) rare CHMs KOs may involve extremely rare heterozygous mutations which may lack homozygous carriers; and (3) rare CHMs provide a more complete set of KOs for a “human KO catalog”. We performed a survey of rare CHMs among 61K whole exome sequenced individuals from the DiscovEHR cohort. First, we identified 39,459 high-quality putative CHMs (pCHMS) consisting of pairs of rare heterozygous variants that are either LoFs (i.e., nonsense, frameshift, or splice-site mutations) or missense variants with strong evidence of being deleterious. Second, we phased all pCHMs using a combination of allele-frequency-based phasing (EAGLE) and pedigree-based phasing. EAGLE phased all of the pCHMs with 91% accuracy based on trio validation. DiscovEHR cohort has >35K 1st and 2nd degree relatives involving 53% of the cohort that we used to phase nearly a third of the pCHMs with ~100% accuracy, reducing inaccurate phasing by 31%. In total, 38% of the pCHMs were phased in trans, yielding a high-confidence set of 15,032 rare, deleterious CHMs distributed among >11K individuals. Over 3,000 genes contain >=1 CHMs. When combined with 12,554 rare homozygous LoF and deleterious missense carriers, CHMs increased the number of genes with >=10 carriers from 181 to 629 (~250% increase). Only considering the 3,915 homozygous LoFs and the 1,307 LoF-LoF CHMs resulted in a 54% increase in the number of genes with >=10 carriers of gene KOs. In conclusion, a large number of rare deleterious CHMs can be accurately phased using population allele frequencies and cryptic relationships to significantly augment the number of human gene KOs. When coupled to phenotypic data, these KOs may inform on our understanding of gene function in humans.

100

Distribution and clinical impact of human gene knockouts from 61,000 whole exome sequences in the DiscovEHR study. *N. Gosalia[1], J. Staples[1], S. Balasubramanian[1], C. O'Dushlaine[1], V. Arunachalam[1], D.H. Ledbetter[2], M.D. Ritchie[2], D.J. Carey[2], J.D. Overton[1], J.G. Reid[1], T.M. Teslovich[1], N.S. Abul-Husn[1], L. Habegger[1], A.N. Economides[1], A. Baras[1], O. Gottesman[1], F.E. Dewey[1].* 1) Regeneron Genetics Center, Regeneron Pharmaceuticals Inc, Tarrytown, NY; 2) Geisinger Health System, Danville, PA.

Evaluating the phenotypic consequences of naturally occurring gene knockouts in humans directly informs our fundamental understanding of gene function. In the DiscovEHR collaboration, we performed whole exome sequencing of >61,000 individuals with longitudinal electronic health record (EHR) data. Identity-by-descent estimates show that 50.4% of the individuals have one or more first- or second-degree relatives, which we leveraged to phase predicted LOF (pLoF) variants with a minor allele frequency (MAF) of <1%. In combination with homozygous pLoFs with MAF <1%, we found 6,678 individuals harboring putative knockouts across 1,686 genes. To understand the phenotypic consequences of these knockouts, we performed association testing of 9,575 binary and 381 quantitative clinical phenotypes and also manually reviewed EHR data. Among the results, associations of *TMPRSS6* pLoFs with hemoglobin levels (β=-0.48, p=1.23e-06) and *TET2* pLoFs with thrombocytopenia (odds ratio 4.20, p=3.87e-16) were consistent with known biology. Additionally, we observed a novel association of pLoFs in *CRHR2*, encoding corticotropin releasing hormone receptor 2, with mean arterial pressure (β=2.88, p=4.74e-05) under a recessive model. Mouse models have demonstrated a role for CRHR2 and its ligand in cardiovascular homeostasis and response to stress. All three knockouts for *CRHR2* have diagnoses consistent with anxiety disorder and 2 of 3 have primary essential hypertension. To extend our results, we examined genotype array association data in the same sample set, and observed an association of rs4723008 (+40.57 kb from *CRHR2*) with anxiety states (p=2.23e-06). Review of EHR data for two compound heterozygotes for *GCKR* pLoFs revealed diagnoses of morbid obesity warranting bariatric surgery and chronic non-alcoholic liver disease, consistent with an extreme form of known associations between *GCKR* variation and obesity-related non-alcoholic fatty liver disease. Review of the EHR of an individual with a homozygous mutation in *CTNS* showed phenotypes consistent with cystinosis including: renal Fanconi syndrome, rickets, metabolic and nutrient deficiencies, and a prescription for cysteamine bitartrate. Finally, a compound heterozygous individual for *NPHP4* had diagnoses for medullary cystic kidney and secondary complications of renal failure. These findings, and ongoing work focused on systematic gene - phenotype association analyses, will further our understanding of gene function in humans.

101

Linked-read whole-exome sequencing identifies a mosaic deletion at the *NF1* locus resolving a previously intractable case of neurofibromatosis type 1. *M. Gonzalez[1], R. Pellegrino[1], F. Mafra[1], J. Garifallou[1], C. Kaminski[1], L. Fang[2], K. Wang[2], K. Wimmer[3], S. Wenzel[3], C. Kao[1], H. Hakonarson[1,4].* 1) Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA; 2) Columbia University Medical Center, Columbia University, New York, NY, 10032; 3) Division of Human Genetics, University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria; 4) Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104.

Neurofibromatosis type 1 (NF1) is an autosomal dominant condition that occurs in 1 in 3,000 individuals. Hallmarks of this disorder include café-au-lait spots and cutaneous neurofibromas, but complications can often extend to hearing loss, learning impairment, cardiovascular sequelae, and severe pain. NF1 is caused by mutations in the *NF1* gene on chromosome 17q11.2, which encodes neurofibrinin that has roles in tumor suppression and regulation of cell growth. Further, a variable range of expressivity and mosaicism has been shown to complicate and impact clinical presentation and disease severity. A patient was clinically diagnosed with a mild form of NF1, however, standard genetic testing techniques including cDNA sequencing, multiplex ligation-dependent probe amplification (MLPA), SNP array genotyping, and gDNA sequencing were all negative for known pathogenic mutations using a variant calling threshold of 10%. When thresholds were adjusted (>1%), a causative candidate was identified involving a genetic alteration within exon 54 of the *NF1* gene that was present in ~8% of reads, suggesting low level mosaicism in this patient. The discordant reads revealed that exon 54 of the *NF1* gene was joining with an Alu element, the repetitive nature of which made resolution of this alteration impossible with standard methods. We utilized the 10x Genomics Chromium platform to generate linked-reads coupled with whole-exome sequencing to resolve the suspected mosaic alteration. This method retains long-range information while maintaining the power, accuracy and scalability of standard short read sequencing by leveraging molecule specific barcode information. This allows high-resolution analysis of structural variants (SV) and other difficult to sequence regions of the genome. The Long Ranger pipeline (10x, v.2.1.3) was able to identify an SV with breakpoints in exon 54 of the *NF1* gene (GRCh37, Chr17:29684365) and IVS3 of *RAB11FIP4* (CRCh37, Chr17:29822454). Additionally, a novel algorithm was developed allowing more robust analysis and visualization of identified breakpoints in SVs. Sequencing has confirmed the presence of the *NF1-RABF11FIP4* breakpoint, consistent with a 138kb deletion, with level of mosaicism confirmed by ddPCR. This is a demonstration of how linked-read sequencing technologies can uniquely aid in the detection and resolution of complex structural variation that cannot be easily disentangled with traditional sequencing/genotyping technologies.

102

K-mer based reference-free detection of family-private variants reveal the genetic complexity of HHT. *A. Farrell[1], W. Wooderchak-Donahue[2,3], M. Velinder[1], A. Ward[1], P. Johnson[3], J. McDonald[2,4], P. Bayrak-Toydemir[1,2], G. Marth[1].* 1) Department of Human Genetics, USTAR Center for Genetic Discovery, University of Utah; 2) Department of Pathology, University of Utah; 3) ARUP Institute for Clinical and Experimental Pathology; 4) Department of Radiology, Hereditary Hemorrhagic Telangiectasia Center, University of Utah.

We have previously shown that our reference free, k-mer based, variant detection method RUFUS has extremely high specificity and sensitivity for *de novo* variations of all types including SNPs, INDELs, and structural variations. Here we present a substantial extension of this method to identify low population-frequency, familial inherited variations, which allows us to accurately track disease-causing mutations through pedigrees, and pinpoint family-private disease-causing variants that segregate with affected/unaffected status. We applied this novel method for analyzing patients with hereditary hemorrhagic telangiectasia (HHT), an inherited disease known to be caused primarily by mutations in the genes *ENG*, *ACVRL1*, and *SMAD4* (in addition to *BMP9*, which is associated with a phenotype similar to HHT). However, the genetic cause of the disease remains unexplained in approximately 15% of individuals identified as having HHT, despite extensive efforts to identify the causative variants with state-of-the-art existing tools. Here we present the results of our analysis of the 60X coverage Illumina whole genome sequencing data collected for 35 individuals from 13 distinct families, where previous causative variant identification methods have failed. To date, RUFUS was able to identify clear causative mutations in 7 of the 13 families: three families had a causative noncoding variant in the *ENG* or *ACVRL1* genes that was missed by previous analyses. Two families had a deleterious variant in *ACVRL1* intron 9 that ultimately disrupted splicing (confirmed by RNA sequencing), including one family with an *ACVRL1* intron 9:chromosome 3 translocation (confirmed by PCR). Further confirmations are currently underway to identify additional HHT causative genes and genetic modifiers in the remaining 6 families. This means that our method was able to “solve” over half of the non-diagnostic cases, with several additional, promising hits being currently pursued, including novel mobile element insertions and small INDELs, missed by other methods, that may be disrupting splicing and gene regulation. Our methodological advances also reveal that noncoding variation plays a larger role in HHT than previously appreciated, and this is the first report to show the role of chromosomal translocation as a mechanism for the development of HHT. .

103

GRAPHITE: A computational framework for structural variation adjudication through graph remapping and visualization. *A. Miller, D. Lee, G. Marth.* University of Utah, Salt Lake City, UT.

Multiple bioinformatic tools have been created that can perform structural variant (SV) detection (e.g. LUMPY, MANTRA, DELLY). However, these tools often produce divergent variant calls making it very difficult to reconcile the resulting variants into a single, accurate set. To assure confidence in these variants, further steps of validation are necessary which typically require expert manual curation and can be difficult and time-intensive to reduce false positives. Here we present a novel method to visualize sequencing read alignments to aid in variant curation. Our software algorithm, GRAPHITE (https://github.com/dillon/graphite), takes as input a set of variant calls from one or more detection tools and applies a novel "variant adjudication" procedure to discard false positives, while keeping true negatives. This is accomplished by constructing a graph from these variants (the Variant Graph) where the reference and alternate alleles are represented as different branches. GRAPHITE then applies a graph mapping algorithm (GSSW, a graph extension of the Smith-Waterman alignment algorithm) to re-map reads contributing to the candidate alleles. Reads are assigned to their respective branches on the graph based on their alignment scores produced by the mapping algorithm. Candidate variants not confirmed by re-mapping are then discarded from further analysis. This results in a call set that is highly specific due to the algorithm and highly sensitive by starting with a call set from previous SV caller tools. Visualization of the variant call set is a key method when confirming or rejecting novel candidate alleles. Current visualization techniques require the analyst to verify SV breakpoints by closely tracking deviations from the reference which is cumbersome and can lead to errors. GRAPHITE's output has been formatted to visualize sequencing read alignments against a Variant Graph, rather than a single reference string, which can be intuitively displayed by the popular IGV (Integrative Genome Viewer) alignment viewer program. With this format, each branch can be visually inspected with the appropriate reads realigned. This view greatly simplifies the effort needed to confirm the presence of SV's which can lead to a more accurate diagnosis.

104

Making the most of targeted sequencing: Detecting CNVs and homozygous regions using off-target reads with SavvyCNV. *M.N. Wakeling, E. De Franco, A.T. Hattersley, S. Ellard.* Medical School, University of Exeter, Exeter, England, United Kingdom.

Targeted short read sequencing is commonly used for the detection of SNVs, InDels and exon CNVs in known genes for specific diseases. Only a few million DNA fragments are sequenced, making the test cheap and sensitive. RNA baits are used to capture coding regions for sequencing. Although this provides an enrichment, 50-70% of the reads returned are "off-target". We investigated whether these off-target reads could be used to detect other clinically-relevant features, such as CNVs and homozygous regions. We sequenced 2591 samples referred for testing for monogenic diabetes or hyperinsulinism using a targeted short read sequencing panel of 75 genes. A median of 3.2 million reads were sequenced for each sample, with an off-target mean read depth of less than 0.1. To detect CNVs, we counted the reads in each 500kb region of the genome, normalised the counts between samples, and reduced noise using singular vector decomposition. We then used an adaptive hidden Markov model to detect regions of reduced or increased read depth. To detect homozygous regions, we searched for variant pairs in linkage disequilibrium where both variants have an informative read – each such pair was marked as concordant or discordant. We then used a hidden Markov model to detect regions lacking discordant variant pairs. Sensitivity and specificity were evaluated using a set of 117 samples sequenced by both targeted panel and whole genome sequencing. Sensitivity and specificity depend on the size of the feature, with larger features being easier to detect. All CNVs larger than 1Mb were correctly detected, and 40% of CNVs between 500kb and 1Mb were detected. Sensitivity and specificity to detect homozygous regions larger than 3Mb was 78%, increasing to 92% at 10Mb. Testing the cohort revealed 16 aneuploidies, 4 unbalanced translocations, 2 clinically-relevant deletions, and 9 1Mb duplications causative of neonatal diabetes, all confirmed using alternative methods. Excess homozygosity was detected in 20% of the samples. Homozygosity mapping assists with the discovery of recessive causes of disease. Analysis of off-target reads enables the detection of large CNVs, unbalanced structural variants, and homozygous regions. Previously, more expensive tests have been used to detect these features, such as whole exome/genome sequencing or ArrayCGH, but this software enables detection using only data from the existing diagnostic sequencing process. The software is freely available.

105

STIX: A scalable index for mining large whole-genome sequencing cohorts for reliable structural variant population allele frequency estimates. *R.M. Layer[1,2], B.S. Pedersen[1,2], A.M. Quinlan[1,2,3].* 1) Human Genetics, University of Utah, Salt Lake City, UT; 2) USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT; 3) Department of Biomedical Informatics, University of Utah, Salt Lake City, UT.

The population frequency of a SNV in a cohort like gnomAD is vital to our assessment of variant pathogenicity in an individual with a rare disease. Unfortunately, there is no equivalent resource for structural variants (SVs). This gap is due to the complexity inherent to identifying SVs. Unlike SNV detection, which considers every sequence for every individual at all 3 billion positions, the number of possible SV configurations (3 billion squared) makes it intractable to interrogate all possible SVs. Because of this, SV detection clusters evidence into a set of possible variants. These sets are then filtered to maximize true positives and minimize false positives. This filtering makes it difficult to draw conclusions about the prevalence of SVs observed in new samples since it is impossible to discern whether the variant is absent from the population or if it was indistinguishable from the noise (a false negative). We need a method that can provide a full accounting of SVs among WGS cohorts such as TOPMed and the Centers for Common Disease Genetics. Here we propose STIX, a structural variant index that searches the raw data from every discordant alignment across thousands of samples. For a given SV, STIX reports a per-sample count of all concurring evidence. From these counts we can, for example, conclude that an SV with high-level evidence in many samples is common and an SV with no evidence is rare and potentially pathogenic. By representing the raw signal, we avoid the previously described false negative issue. We indexed 2,504 genomes from the 1000 Genomes Project (1KG) and tested 14,146 cancer-related deletions from the COSMIC database. Each search took 0.1 seconds and found that 27% of SVs had some evidence in 1KG and 3% had evidence in >10% of the samples. We also use STIX to interpret SVs from families afflicted rare disorders. In one case, STIX identified an SV in a 1KG sample that was previously thought to be private to individuals affected by Treacher Collins. Since STIX retains all of the alternate evidence, it is useful for other analysis such as large-scale SV genotyping. With a compact representation of the reference evidence, STIX can quickly genotype new SVs across all indexed samples. This is vital to large sequencing projects that sequence cohorts in batches and must re-genotyped new SVs in all samples and existing SVs in the new batch. STIX can also empower population scale SV detection by jointly considering thousands of samples.

106

Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *M. Cretu Stancu[1], M.J. van Roosmalen[1], I. Renkens[1], M. Nieboer[1], S. Middelkamp[1], J. de Ligt[1], G. Pregno[2], D. Giachino[2], G. Mandrile[2], J.E. Valle-Inclan[1], J. Korzelius[1], E. de Bruijn[1], E. Cuppen[1], M.E. Talkowski[3,4,5], T. Marschall[6,7], J. de Ridder[1], W.P. Kloosterman[1].* 1) Genetics, Universitair Medisch Centrum Utrecht, Utrecht, Netherlands; 2) Medical Genetics Unit, Department of Clinical and Biological Sciences, University of Torino, Turin, Italy; 3) Molecular Neurogenetics Unit, Center for Human Genetic Research, Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA; 4) Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA; 5) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 6) Center for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany; 7) Max Planck Institute for Informatics, 66123 Saarbrücken, Germany.

Structural genomic variants (SVs) form a common type of genetic alteration underlying human genetic disease and phenotypic variation. Despite major improvements in sequencing technology and data analysis, the detection of SVs poses many challenges, particularly for high complexity SVs. Emerging long-read single-molecule sequencing technologies provide new opportunities for detection of SVs. Here, we demonstrate whole genome sequencing of two patients with congenital abnormalities using the ONT MinION. We developed a bioinformatic pipeline consisting of an SV calling and genotyping step - NanoSV - and a subsequent machine learning filtering step, in order to produce a high quality set of SVs. Using our approach, we readily detected all de novo chromothriptic rearrangements (40 and 29 breakpoints respectively), involving multiple chromosomes, in both our patients. Furthermore, we detected (and validated) 5 additional SVs fine-mapping the previously reported chromothripsis structure in one of the patients. We demonstrate the feasibility of readily applying our methods to other whole genome nanopore data such as the NA12878 dataset, that was previously generated. By comparing our SV calls to existing gold standard sets of SVs for the NA12878 sample, we estimate a sensitivity of 80-88% for our method, and extensive validation of SV calls from our patients show a precision of 92%. Genome-wide assessment of SVs in our two patients, revealed 3,253 (33%) novel variants that were missed in short-read data of the same sample, the majority of which are medium sized insertions ~300bp of SINE retro-transposon elements. Furthermore, the long-read specific SVs lie, on average, in higher GC-content regions ($p = 1.1e-10$), that were previously hard to evaluate. Long sequencing reads enabled efficient phasing of genetic variations, allowing the construction of genome-wide maps of phased SVs. We employed read-based phasing to show that all de novo chromothripsis breakpoints occurred on paternal chromosomes. Additionally, we use a haplotype aware assembly strategy for nanopore reads allowing us to resolve the long-range structure of the chromothripsis rearrangement. This work demonstrates that long-read sequencing technologies, will both improve as well as ease current genetic analyses. We demonstrate the value of long-read sequencing for scalable genetic analyses in life sciences research as well as for clinical diagnostics.

107

Clinical experience with a reflex-based testing algorithm for neurodevelopmental disorders. *L. Walters-Sen, J. Glass, E. Partack, E. Wakefield, L. Dyer, N. Leslie.* Human Genetics, Cincinnati Children's Hospital & Medical Center, Cincinnati, OH.

Neurodevelopmental disorders (NDDs) affect roughly one in six children in the US and are a common indication for genetic testing. Determining the etiology of a child's NDD is vital to ensuring early interventions and access to appropriate services. However, NDD testing can be expensive to families and put a strain on limited healthcare dollars. Historically, the majority of NDD patients at our institution had concurrent orders placed for Fragile X testing, chromosome analysis, and SNP microarray analysis. In addition, the low volume of *PTEN* sequencing indicated this test was underutilized in patients with NDD and macrocephaly. Therefore, our institution developed a reflex-based testing algorithm, the Neurodevelopmental Reflex (NDR), to improve physician test ordering and decrease the cost burden of NDD genetic testing. First, the NDR allowed clinicians to reflexively order chromosomes if microarray was denied by insurance. Second, the NDR ordered tests sequentially, starting with the test of highest expected yield. Patients were dichotomized based on head circumference. Macrocephalic patients received *PTEN* sequencing, followed by Fragile X testing, and finally SNP microarray analysis. Normocephalic patients began with SNP microarray analysis, followed by Fragile X testing. In the first eight months of offering the NDR, complete NDR results have been returned to 21 of 27 macrocephalic patients, with an average turn-around time of 48.1 days (range=9-77). One patient was found with a positive result of >200 repeats in *FMR1*. Complete NDR results have been returned to 96 of 114 normocephalic patients, with an average turn-around time of 44.7 days (range=10=147). Four patients were found with positive results on SNP microarray analysis. We also analyzed the effect of the NDR on testing costs. By giving clinicians the option to reflex to chromosome analysis only when microarray was denied, we realized a cost savings of $186,825 ($1325/patient). In addition, the ability to prevent unnecessary testing in the five patients with positive results has amounted to a cost savings of $7838 ($4506 for avoided SNP microarray and $3332 for four unneeded Fragile X testing cases). In total, we have saved $194,663 over eight months, with a projected savings of more than $300,000 per year. In summary, we have developed a reflex-based testing strategy that has the potential to reduce healthcare costs while still following clinical guidelines for NDD testing.

108

Defining the threshold for neurodevelopmental disorders in the context of the rare genetic background. *L. Pizzo[1], M. Jensen[1], A. Polyak[1], J. Yoon[1], D. Pazuchanics[1], E. Huber[1], V. Kumar[1], S. Zeesman[2], K. Mannik[3], A. Reymond[3], P. Stankiewicz[4], O. Pichon[5], P. Prontera[6], A. Renieri[7], D. Amor[8], E. Sistermans[9], C. Schwartz[10], C. Romano[11], S.W. Cheung[4], J. Rosenfeld[4], J. Andrieux[12], S. Girirajan[1].* 1) Department of Biochemistry and Molecular Biology. The Pennsylvania State University, University Park, PA; 2) Hamilton Health Sciences. Hamilton, Ontario; 3) Center for Integrative Genomics. University of Lausanne. CH-1015 Lausanne. Switzerland; 4) Department of Molecular and Human Genetics. Baylor College of Medicine. One Baylor Plaza, NAB 2015. Houston, TX 77030; 5) Service de Génétique Médicale (Cytogénétique). 9, quai Moncousu 44093 Nantes; 6) Azienda Ospedaliera di Perugia. SSD Neonatologia e Diagnosi Prenatale. CRR Genetica Medica. 06123 Perugia, Italy; 7) Medical Genetics, University of Siena, Siena, Italy; 8) Lorenzo and Pamela Galli Chair, Department of Paediatrics, University of Melbourne. Clinical Geneticist, Royal Childrens Hospital and VCGS; 9) VU University Medical Center. van der Boechorststraat 7, J376. 1081 BT Amsterdam; 10) JC Self Res. Inst.113 Gregor Mendel Circle. Greenwood, S.C. 29646; 11) Department of Laboratories. Pediatrics and Medical Genetics. Regional Center for Genetic Rare Diseases with Intellectual Disability or Brain Aging. IRCCS Associazione Oasi Maria Santissima. Via Conte Ruggero, 73. 94018 Troina. Italy; 12) Institut de Génétique Médicale. Hopital Jeanne de Flandre. CHRU de Lille.

We previously reported a two-hit model to explain inter and intra-familial phenotypic variability observed in individuals with rare pathogenic CNVs. To further explore the role of the genetic background in the manifestation of neurodevelopmental disorders (NDD), we evaluated 126 individuals with 16p12.1 deletion and available family members for phenotypic and genomic analyses. Probands manifested a range of phenotypes including speech delay (78%), DD (68%), ID (74%), and ASD (29%), while carrier parents showed subtle features, such as depression (41%), bipolar disorder (6%), and schizophrenia (24%). Whole-exome sequencing and SNP arrays showed that probands carried a higher burden of rare pathogenic SNVs and CNVs in functionally intolerant genes (CADD>25, RVIS<20%) compared to carrier parents (average number of mutations in proband=7.28 vs carrier parent=6.2, p=0.0392). We found that probands with a strong family history have a higher genetic burden (p=0.0016) and a more severe clinical outcome (p=0.035) compared to those with mild or no family history. These results underscore the functional tolerance of 16p12.1 deletion to second hits before surpassing the genetic threshold for severe NDD. We extended this model to include other pathogenic variants, and first analyzed 46 families from the Simons autism study (SVIP and SSC). We found that individuals with 16p11.2 deletion required fewer secondary mutations to manifest severe NDD compared to those with 16p12.1 deletion, indicated by a smaller change in burden between probands and parents (p=0.0002). Additional analysis in 47 probands with rare CNVs, including 16p11.2, 17q12, 3q29, 16p13.11, 1q21.1, and 7q11.23, showed that duplications were more tolerant to second hits (average mutations=9.1) compared to deletions (average=6.9) (p=0.055), consistent with reduced severity of duplications. Among the 404 individuals with IQ<70, we found no difference in the burden of secondary variants between carriers of inherited rare gene disruptive mutations (average=10.5) and those with *de novo* gene disruptive mutations in NDD genes (average =9.5, p=0.26); however, a higher burden was observed compared to individuals with rare pathogenic CNVs (average =7.9, p=0.007). Our results suggest that genetic variants associated with NDD confer differing sensitivities to disease, and therefore require differential burden of secondary variants to reach the threshold for manifestation of the phenotype.

109

Mitochondrial dysfunction in Smith-Magenis syndrome reveals aberrant respiration. *M.D. Fountain[1], S.V. Mullegama[1], J.P. Alaimo[1], C. Li[1], T. Donti[1,2], S.A. Behrendt-McLeroy[1], A. Besse[1], P.E. Bonnen[1], B.H. Graham[1], S.H. Elsea[1].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Greenwood Genetic Center, Greenwood, SC.

Smith-Magenis syndrome (SMS) is a neurodevelopmental disorder caused by reduced gene dosage of *retinoic acid-induced 1* (*RAI1*). Individuals with SMS typically manifest sleep disturbance, self-abusive and aggressive behaviors, craniofacial abnormalities, neurological abnormalities, and early-onset obesity. Mitochondrial dysfunction can result in a variety of clinical manifestations such as developmental delay, hypotonia, muscle weakness, cardiomyopathy, and others. Recent studies have begun examining the role of mitochondrial function in genetic neurodevelopmental disorders, such as Down syndrome (DS, Trisomy 21), Rett syndrome (RTT, MECP2), and fragile X syndrome (FXS, FMR1). Moreover, autism spectrum disorder has been similarly associated with mitochondrial dysfunction. While much is known about the clinical manifestation of SMS, multifactorial cellular defects of *RAI1* haploinsufficiency are not well understood. Utilizing fibroblasts from controls and individuals with SMS, we performed a battery of mitochondrial testing to identify significant dysregulation of mitochondrial-associated genes. We observed significantly reduced mitochondrial membrane potential equivalent to cells from patients with ABAT deficiency, highlighting the presence of mitochondrial dysfunction. Using Agilent's Seahorse XF technology for mitochondrial functional analysis, alterations in mitochondrial cellular respiration and oxygen consumption rates were also observed. Mitochondrial localization studies in SMS fibroblast lines using fluorescence microscopy show an aberrant perinuclear distribution. Taken together, these results demonstrate that mitochondrial function and integrity are compromised in SMS patients and supports a cellular model in which turnover and biogenesis of the organelle are aberrant. As well, mitochondrial dysfunction appears to be an under-appreciated co-morbidity of SMS. These results prompt further elucidation of the observed dysfunction in mitochondria. To further assess mitochondrial function in SMS and molecularly dissect the pathways contributing to this impairment, current studies are investigating the function of mitochondrial fission, fusion, transport, and turnover, as disorders such as Parkinson and Huntington diseases were recently shown to manifest alterations in these critical functions. Overall, these studies may identify a direct pathological cellular defect similar to other neurodevelopmental disorders, promoting comparable treatment approaches.

110

***FDXR* mutations cause sensorial neuropathies, a new mitochondrial Fe-S biogenesis disease.** *A. Paul[1], A. Drecourt[1], D. Dupin-Deguine[2], C. Vasnier[3], M. Oufadem[1], F. Petit[1], C. Masson[1], C. Bonnet[4], S. Masmoudi[5], I. Mosnier[6], L. Mahieu[7], D. Bouccara[6], J. Kaplan[1], G. Challes[8], C. Domange[9], F. Mochel[10], O. Sterkers[6], S. Gerber[1], P. Nitschke[1], C. Bole-Feysot[1], L. Jonard[11], S. Gherbi[12], I. Ben Aissa[12], S. Lyonnet[1], A. Rotig[1], A. Delahodde[3], S. Marlin[1,12,13].* 1) Institut Imagine, 24, boulevard du Montparnasse, 75015 Paris, Paris, France; 2) Service de Génétique Médicale, Hôpital Purpan, Toulouse, France; 3) Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France; 4) UMRS 1120, Institut de la Vision, Paris, France; 5) Laboratory of Molecular and Cellular Screening Processes, Center of Biotechnology of Sfax, Tunisia; 6) Service d'ORL, Hôpital Pitié-Salpêtrière, APHP, Paris, France; 7) Service d'ophtalmologie, Hôpital Rangueil, Toulouse, France; 8) Service d'ophtalmologie, Hôpital Pitié-Salpêtrière, APHP, Paris, France; 9) Service d'ORL, Hôpital Lariboisière, Paris, France; 10) Département de Génétique, Hôpital Pitié-Salpêtrière, APHP, Paris, France; 11) Service de Génétique, Laboratoire de Génétique Moléculaire, Hôpital Necker-Enfants, Malades, APHP, Paris, France; 12) Centre de Référence des Surdités Génétiques, Service de Génétique, Hôpital Necker-Enfants Malades, APHP, Paris, France; 13) Service de Génétique, Hôpital Necker-Enfants Malades, APHP, Paris, France.

Hearing loss and retinis pigmentosa have mostly genetic origins, some of them being related to sensorial neuronal defects. Here, we report eight subjects from four independent families presenting with auditory neuropathy and optic atrophy. Whole-exome sequencing revealed biallelic mutations in *FDXR* in affected subjects of each family. *FDRX* encodes the mitochondrial ferredoxin reductase, the sole human ferredoxin reductase which is implicated in the biosynthesis of iron-sulfur clusters (ISC) and in the heme formation. ISC proteins are involved in enzymatic catalysis and gene expression, DNA replication and repair. We observed deregulated iron homeostasis in *FDXR* mutant fibroblasts and indirect evidence of mitochondrial iron overload. Functional complementation in a yeast strain deleted for ARH1, the human *FDXR* counterpart, established the pathogenicity of these mutations. These data emphasize the wide clinical heterogeneity of mitochondrial disorders related to ISC synthesis. .

111

De novo mutations in protein kinase genes _CAMK2A_ and _CAMK2B_ cause intellectual disability. *S. Kury[1], G.M. van Woerden[2,3], T. Besnard[1], M.T. Cho[4], S. Sanders[5], H.A.F. Stessman[6], E.A. Sellars[7], J. Berg[8], J.L. Waugh[9], L. Robak[10], J.A. Bernstein[11], M. Deardorff[12], G.E. Hoganson[13], D.S. Johnson[14], T. Dabir[15], A. Sarkar[16], G. Lesca[17,18], P.A. Terhal[19], T.E. Prescott[20], D. Grange[21], A. Haeringen[22], C. Lam[23], G. Mirzaa[23,24], K. Helbig[25], J.A. Rosenfeld[10], P.B. Agrawal[26,27], S. Odent[28], S. Mercier[1], Y. Elgersma[2,3], S. Bezieau[1,29], CAMK2A/B Consortium.*
1) Service de Génétique Médicale, CHU Nantes, 9 quai Moncousu, 44093 Nantes Cedex 1, France; 2) Department of Neuroscience, Erasmus University Medical Center, 3015 CN, Rotterdam, The Netherlands; 3) ENCORE Expertise Center for Neurodevelopmental Disorders, Erasmus University Medical Center, 3015 CN, Rotterdam, The Netherlands; 4) GeneDx, Gaithersburg, Maryland, USA; 5) Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, California 94158, USA; 6) Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA; 7) Section of Genetics and Metabolism, Arkansas Children's Hospital, Little Rock, AR, USA; 8) Department of Clinical Genetics, Ninewells Hospital and Medical School, University of Dundee, UK; 9) Department of Neurology, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA; 10) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; 11) Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305, USA; 12) Department of Pediatrics, Division of Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; 13) Department of Pediatrics, University of Illinois at Chicago, College of Medicine, Chicago, Il, USA; 14) Sheffield Children's Hospital, Western Bank, Sheffield, S10 2TH, UK; 15) Northern Ireland Regional Genetics Centre, Belfast Health and Social Care Trust, Belfast City Hospital, Lisburn Road, Belfast, BT9 7AB, UK; 16) Nottingham Regional Genetics Service, City Hospital Campus, Nottingham University Hospitals NHS Trust, The Gables, Hucknall Road, Nottingham NG5 1PB, UK; 17) Service de génétique, Centre de Référence des Anomalies du Développement, Hospices Civils de Lyon, Lyon, France; 18) INSERM U1028, CNRS UMR5292, Centre de Recherche en Neurosciences de Lyon, Lyon, France; 19) Klinisch geneticus, Divisie Biomedische Genetica, Afdeling Genetica Universitair Medisch Centrum Utrecht, Kamernummer KC040872, Huispostnummer KC040842, The Netherlands; 20) Department of Medical Genetics, Telemark Hospital, 3710 Skien, Norway; 21) Division of Genetics and Genomic Medicine, Department of Pediatrics, Washington University School of Medicine, Saint Louis, Missouri, USA; 22) Department of Clinical Genetics, Leiden University Medical Center (LUMC), Leiden, The Netherlands; 23) Division of Genetic Medicine, Department of Pediatrics, University of Washington School of Medicine and Seattle Children's Hospital, WA 98105, USA; 24) Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, Washington, USA; 25) Ambry Genetics, 15 Argonaut, Aliso Viejo, CA, 92656, USA; 26) Division of Genetics and Genomics, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115, USA; 27) The Manton Center for Orphan Disease Research, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA; 28) CHU Rennes, Service de Génétique Clinique, CNRS UMR6290, Université Rennes1, Rennes, France; 29) CRCINA, Inserm, Université d'Angers, Université de Nantes, Nantes, France.

The alpha- and beta isoforms of calcium/calmodulin-dependent serine/threonine protein kinase II (CAMK2) play a pivotal role in neuronal function. Although CAMK2 was one of the first proteins shown to be essential for normal learning and synaptic plasticity in mice, its requirement for human brain development has not yet been established. Through a multi-center collaborative study based on a whole-exome sequencing approach, we identified 17 exceedingly rare *de novo CAMK2A* or *CAMK2B* variants in 21 unrelated individuals. The variants are either in the kinase domain or in the autoregulatory domain, suggesting they may change the kinetic function of the enzyme. Mutations were assessed for their effect on CAMK2 function and on neuronal migration. For both *CAMK2A* and *CAMK2B*, we identified mutations that decreased or increased CAMK2 auto-phosphorylation at Thr286/Thr287. We further found that all mutations affecting auto-phosphorylation also affected neuronal migration, highlighting the importance of tightly regulated CAMK2 auto-phosphorylation in neuronal function and neurodevelopment. Our data establish the importance of *CAMK2A* and *CAMK2B* and their auto-phosphorylation in human brain function, and expand the phenotypic spectrum of the disorders caused by variants in key players of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPAR) and N-methyl-D-aspartate receptor (NMDAR) signaling pathways.

112

***De novo*, deleterious sequence variants that alter the transcriptional activity of the homeoprotein PBX1 are associated with intellectual disability and pleiotropic developmental defects.** *A. Slavotinek[1,2+], M. Risolino[3+], M. Losa[3], M.T. Cho[4], K.G. Monaghan[4], D. Schneidman-Duhovny[5,6], S. Parisotto[7], J.C. Herkert[8], A.P.A Stegmann[9,10], K. Miller[11], N. Shur[11], J. Chui[12], E. Mueller[12], S.D. DeBrosse[13], J.O. Szot[14,15], G. Chapman[14,15], N.S. Pachter[16,17], D.S. Winlaw[18,19], B.A. Mendelsohn[1,2], H. Pedro[7], S.W. Dunwoodie[14,15], L. Selleri[3*], J. Shieh[1,2*].*
1) Dept Pediatrics, Division Genetics, Univ California, San Francisco, San Francisco, CA; 2) Institute of Human Genetics, University of California San Francisco, San Francisco, CA, 94143-2711, USA; 3) Program in Craniofacial Biology, Institute of Human Genetics, Departments of Orofacial Sciences and Anatomy, University of California San Francisco, San Francisco, CA, USA; 4) GeneDx, 207 Perry Parkway, Gaithersburg, MD, USA; 5) School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 9190401; 6) Department of Biochemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 9190401; 7) Division of Genetics, Department of Pediatrics, Hackensack University Medical Center, Hackensack, NJ, USA; 8) University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands; 9) Department of Clinical Genetics, Maastricht University Medical Center, Maastricht, the Netherlands; 10) Department of Genetics, Radboud University Medical Center (RUMC), Nijmegen, The Netherlands; 11) Department of Pediatrics, Albany Medical Center, Albany, NY, USA; 12) Clinical Genetics, Stanford Children's Health at CPMC, San Francisco, CA, USA; 13) Center for Human Genetics, University Hospitals Cleveland Medical Center, Cleveland, OH, USA; 14) Victor Chang Cardiac Research Institute, Developmental and Stem Cell Biology Division, Sydney, NSW, Australia; 15) University of New South Wales, Sydney, NSW, Australia; 16) Genetic Services of Western Australia, King Edward Memorial Hospital, Perth, WA, Australia; 17) School of Paediatrics and Child Health, University of Western Australia, Perth, WA, Australia; 18) University of Sydney, Medical School, Sydney, NSW, Australia; 19) Children's Hospital at Westmead, Heart Centre for Children, Sydney, NSW, Australia.

We present a novel intellectual disability syndrome in seven patients who were heterozygous for *de novo*, deleterious sequence variants in the *PBX1* gene. *PBX1* encodes a three amino acid loop extension (TALE) homeodomain transcription factor that forms multimeric complexes with other TALE and HOX proteins to regulate target gene transcription during development. As previously reported, *Pbx1* homozygous mutant mice (*Pbx1*$^{-/-}$) develop malformations and hypoplasia or aplasia of multiple organs, including the craniofacial skeleton, ear, branchial arches, heart, lungs, diaphragm, gut, kidneys, and gonads. Clinical findings similar to those reported in *Pbx* mutant mice were observed in all patients with varying expressivity and severity. These included external ear anomalies, abnormalities of branchial arch derivatives, heart malformations, diaphragmatic hernia with hypoplasia of the ipsilateral lung, renal hypoplasia, and cryptorchidism. All patients but one had developmental delays. The sequence variants included missense substitutions adjacent to the PBX1 homeodomain (p.Arg184Pro, p.Met224Lys, and p.Arg227Pro) or within the homeodomain (p.Arg234Pro, and p.Arg235Gln). Two sequence variants (p.Ser262Gln*fs**2, and p.Arg288*) yielded truncated PBX1 proteins. Functional studies on five of the *PBX1* sequence variants revealed abnormal cellular processes, including intrinsic perturbation of PBX-dependent transactivation ability, altered nuclear translocation, and abnormal interactions between mutant PBX1 proteins and wild-type TALE or HOX cofactors. It is thus likely that the mutations directly affect the transcription of PBX1 target genes and impact normal embryonic development. We conclude that deleterious sequence variants in *PBX1* cause an intellectual disability syndrome with pleiotropic developmental defects resembling those reported in *Pbx1* mutant mice, arguing for strong conservation of gene function between these two species. + = contributed equally to this work * = contributed equally to this work.

113

Clinical relevance of systematic phenotyping and exome sequencing in patients with short stature. *C.T. Thiel[1], N.N. Hauer[1], B. Popp[1], E. Schoeller[1], S. Schuhmann[1], K.E. Heath[2], A. Hisado-Oliva[2], P. Klinger[3], C. Kraus[1], U. Trautmann[1], M. Zenker[4], C. Zweier[1], A. Wiesener[1], R. Abou-Jamra[5], E. Kunstmann[6], D. Wieczorek[7], S. Uebe[1], F. Ferrazzi[1], C. Büttner[1], A.B. Ekici[1], A. Rauch[8], H. Stricht[9], H.-G. Dörr[10], A. Reis[1].* 1) Institute of Human Genetics, Friedrich-Alexander-University of Erlangen-Nürnberg, Erlangen, Bavaria, Germany; 2) Institute of Medical and Molecular Genetics and Skeletal dysplasia Multidisciplinary Unit Hospital Universitario La Paz Universidad Autónoma de Madrid IdiPAZ and CIBERER, Madrid, Spain; 3) Department of Orthopaedic Rheumatology Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany; 4) Institute of Human Genetics Otto-von-Guericke University Magdeburg, Magdeburg, Germany; 5) Institute of Human Genetics University of Leipzig, Leipzig, Germany; 6) Institute of Human Genetics University of Würzburg, Würzburg, Germany; 7) Institute of Human-Genetics University Düsseldorf, Düsseldorf, Germany; 8) Institute of Medical Genetics University of Zurich, Zurich, Switzerland; 9) Institute of Biochemistry Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany; 10) Department of Pediatrics and Adolescent Medicine Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany.

Short stature is a common condition of great concern to patients and their families. In most cases it is genetic in origin but the underlying cause remains elusive due to clinical and genetic heterogeneity. In an unbiased approach we carefully phenotyped 565 patients and randomly selected 200 for whole exome sequencing. Sequence variants were analyzed for pathogenicity and the affected genes characterized regarding their functional relevance for growth. All patients received extensive clinical and endocrinological examinations, careful clinical genetic phenotypic evaluation followed by targeted diagnostic assessment for suspected diagnoses. We identified a known disease-cause in only 14 % of patients, the most common causes being CNVs (7 %), followed by syndromic monogenic causes (5 %) and Turner syndrome (2 %). Whole exome sequencing identified additional mutations in known short stature associated genes (27) in 17 % of patients who manifested only part of the symptomatology precluding an early clinical diagnosis. Here, heterozygous carriers of recessive skeletal dysplasia alleles (*ACAN*, *NPR2*) were a surprisingly frequent cause of idiopathic short stature found in 3.5 % of cases. We next selected known short stature genes with mutations for pathway analyses of the affected proteins and found that 54 % are involved in the main functional categories cartilage formation, chromatin modification and Ras-MAPK signaling. In addition we identified 37 further strong candidate genes, of which seven had deleterious mutations in at least two families. Interestingly, 48 % of these candidate genes are involved in the 10 main functional categories already identified for the known short stature associated genes further supporting their pathogenicity. Finally, in 16 % of the 200 sequenced individuals our findings were of significant clinical relevance regarding preventive measures, symptomatic or even targeted treatment. These results demonstrated that systematic phenotyping combined with targeted genetic testing and whole exome sequencing is able to increase the diagnostic yield in short stature up to 31 % with concomitant improvement in treatment and prevention. Rigorous variant analysis considering phenotypic data further led us to the identification of further 37 probable novel candidate genes.

114

A low-frequency missense variant in *SLC39A8* associated with idiopathic scoliosis. *G. Haller[1], K. McCall[1], S. Jenkitkasemwong[2], C. Cruchaga[3], M. Harms[4], A. Goate[5], J. Morcuende[7], P. Giampietro[8], N. Miller[9], C. Wise[10,11,12], M. Knutson[2], M. Dobbs[1,13], C. Gurnett[1,6,14].* 1) Department of Orthopaedic Surgery, Washington University School of Medicine, Saint Louis, MO; 2) Food Science and Human Nutrition Department, University of Florida, Gainesville, FL; 3) Department of Psychiatry, Washington University, St. Louis, MO; 4) Department of Neurology, Columbia University, New York, NY; 5) Ronald M. Loeb Center for Alzheimer's disease, Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY; 6) Department of Pediatrics, Washington University, St. Louis, MO; 7) Department of Orthopaedics and Rehabilitation, University of Iowa, Iowa City, IA; 8) Department of Pediatrics, Drexel University, Philadelphia, PA; 9) Department of Orthopaedic Surgery, University of Colorado, Denver, CO; 10) Sarah M. and Charles E. Seay Center for Musculoskeletal Research, Texas Scottish Rite Hospital for Children, Dallas, TX; 11) Department of Orthopaedic Surgery, University of Texas Southwestern Medical Center at Dallas, Dallas, TX; 12) McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center at Dallas, Dallas, TX; 13) Shriners Hospital for Children, St. Louis, MO; 14) Department of Neurology, Washington University, St. Louis, MO.

The genetic factors predictive of severe adolescent idiopathic scoliosis (AIS) are largely unknown. To identify genetic variants associated with severe AIS, we performed a genome-wide association study of common exonic variants using whole exome sequence data of 457 severe AIS cases and 987 controls. A nonsynonymous missense SNP in the heavy metal ion transporter *SLC39A8* (p.Ala391Thr, rs13107325) was associated with severe AIS at exome-wide significance ($P = 1.60 \times 10^{-7}$; odds ratio (OR) = 2.01 (CI=1.54-2.62)). We replicated the association of this SNP with AIS in a second cohort (857 cases and 1095 controls) resulting in a combined $P = 3.64 \times 10^{-25}$; OR = 1.89. This pleiotropic *SLC39A8* missense variant was previously associated with a variety of human traits, including blood pressure, body-mass index and cholesterol and manganese level. Clinically, rs13107325 was associated with greater spinal curvature, decreased height, increased BMI and lower plasma manganese level in our AIS cohort. Functional studies demonstrate reduced manganese influx mediated by the SLC39A8 p.Ala391Thr variant compared to WT expressed *in vitro*. Further, *slc39a8* null zebrafish had abnormal fin folds, impaired growth, and abnormal movement compared to wild-type zebrafish. Our association of AIS pathogenesis with altered manganese homeostasis opens up the possibility of dietary intervention to prevent scoliosis progression.

115

Mutations in fibronectin cause a subtype of spondylometaphyseal dysplasia with "corner fractures". *P.M. Campeau[1], C.S. Lee[2], H. Fu[1], N. Baratang[1], J. Rousseau[1], H. Kumra[2], V.R. Sutton[3], M. Niceta[4], A. Ciolfi[4], G. Yamamoto[5], D. Bertola[5], C.L. Marcelis[6], D. Lugtenberg[6], A. Bartuli[4], C. Kim[7], J. Hoover-Fong[8], N. Sobreira[8], R. Pauli[9], C. Bacino[3], D. Krakow[10], A. Kariminejad[11], M.T. McDonald[12], M. Aracena Alvarez[13], E. Lausch[14], A. Superti-Furga[15], J.T. Lu[16], D.H. Cohn[17], M. Tartaglia[4], B.H. Lee[3], D. Reinhardt[2].* 1) CHU Sainte Justine Research Centre, University of Montreal, Canada; 2) Department of Anatomy and Cell Biology, McGill University, Canada; 3) Department of Molecular and Human Genetics, Baylor College of Medicine, TX; 4) Genetics and Rare Diseases Research Division, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy; 5) Instituto da Criança HC-FMUSP - Universidade de São Paulo, SP, Brazil; 6) Radboud University Medical Center, Nijmegen, The Netherlands; 7) Department of Pharmacology, Baylor College of Medicine, Houston, TX; 8) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 9) Midwest Regional Bone Dysplasia Clinic, University of Wisconsin, Madison, WI; 10) Department of Human Genetics, David Geffen School of Medicine at UCLA, CA; 11) Kariminejad-Najmabadi Pathology & Genetics Center, Tehran, Iran; 12) Department of Pediatrics, Division of Medical Genetics, Duke University Medical Center, Durham, NC; 13) Hospital Dr. Luis Calvo Mackenna, Santiago, Chile; 14) University of Freiburg, Freiburg, Germany; 15) Lausanne University Hospital (CHUV), Lausanne, Switzerland; 16) Helix, San Carlos, CA; 17) University of California Los Angeles, Los Angeles, CA.

Fibronectin is a master organizer of extracellular matrices (ECM), promoting assembly of collagens, fibrillin-1, and other proteins. It is also known to play roles in skeletal tissues through its expression in osteoblasts, chondrocytes and mesenchymal cells. Spondylometaphyseal dysplasias (SMD) comprise a diverse group of skeletal dysplasias often presenting with short stature, growth plate irregularities, and vertebral anomalies, such as scoliosis. By comparing the exomes of individuals with SMD with the radiographic appearance of "corner fractures" at metaphyses, we identified three individuals with novel variants in the fibronectin gene (*FN1*) affecting highly conserved residues. Furthermore, using matching tools and the SkelDys emailing list, we identified other individuals with *de novo FN1* variants and a similar phenotye. The severe scoliosis in most individuals and rare developmental coxa vara distinguishes individuals with *FN1* mutations from those with classical Sutcliffe type SMD. To study functional consequences of these *FN1* mutations, we introduced three disease-associated missense mutations (p.Cys87Phe, p.Tyr240Asp, p.Cys260Gly) in a recombinant secreted N-terminal 70 kDa fragment (rF70K) as well as in the full length fibronectin (rFN). The wild-type rF70K fragment and rFN were secreted into the culture medium, whereas all mutant proteins were either not secreted or secreted at significantly lower amounts. Immunofluorescence analysis demonstrated increased intracellular retention of the mutant proteins. In summary, mutations in *FN1* that cause defective fibronectin secretion are found in SMD, and we thus provide additional evidence for a critical function of fibronectin in cartilage and bone.

116

Multiple gene discoveries in Robinow syndrome identify perturbation in the balance between Wnt signaling pathways in humans. *C. Carvalho[1], J.J. White[1], J. Mazzeu[2,3], Z. Coban-Akdemir[1], Y. Bayram[1], V. Bahrambeigi[1,4], A. Hoischen[5,6], B. van Bon[5], A. Gezdirici[7], E. Gulec[7], F. Ramond[8], R. Touraine[8], M. Shinawi[9], E. Beaver[10], J. Heeley[10], J. Hoover-Fong[11], C. Durmaz[12], M. Duz[13], S. Price[14], B. Ferreira[2], A. Vianna-Morgante[15], S. Ellard[16,17], A. Parrish[16], K. Stals[16], J. Flores-Daboub[18], S. Jhangiani[19], R.A. Gibbs[19], H.A. Brunner[20,21], V.R. Sutton[1,22], J.R. Lupski[1,19,22].* 1) Dept Molecular Human Genetics, Baylor Col Medicine, Houston, TX; 2) University of Brasilia, Brasilia, Brazil; 3) Robinow Syndrome Foundation, Anoka, MN; 4) Graduate Program in Diagnostic Genetics, School of Health Professions, University of Texas MD Anderson Cancer Center, Houston, TX; 5) Department of Human Genetics, Radboud Institute of Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands; 6) Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, The Netherlands; 7) Department of Medical Genetics, Kanuni Sultan Suleyman Training and Research Hospital, Istanbul, Turkey; 8) Service de Génétique, CHU-Hôpital Nord, Saint-Etienne, France; 9) Division of Genetics and Genomic Medicine, Department of Pediatrics, Washington University School of Medicine, St. Louis, MO; 10) Mercy Clinic-Kids Genetics, Mercy Children's Hospital St. Louis, St. Louis MO; 11) Greenberg Center for Skeletal Dysplasias, McKusick-Nathans Institute for Genetic Medicine, The Johns Hopkins Children's Center, Baltimore MD; 12) Department of Medical Genetics, Ankara University, Ankara, Turkey; 13) Department of Medical Genetics, Cerrahpasa Medical School, Istanbul University, Istanbul, Turkey; 14) Oxford Centre for Genomic Medicine, Nuffield Orthopaedic Centre, Oxford, UK; 15) Departmetn of Genetics and Evolutionary Biology, Institute of Bioscience, Sao Paulo, SP, Brazil; 16) Department of Molecular Genetics, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK; 17) Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, UK; 18) Department of Pediatric Genetics, University of Utah School of Medicine, Salt Lake City UT; 19) Human Genome Sequencing Center, BCM, Houston TX; 20) Department of Human Genetics, Donders Institute for Brain, Cognition and Behavior, Radboud Univeristy Medical Center, Nijmegen, the Netherlands; 21) GROW School for Oncology and Developmental Biology, Maastricht University Medical Center, Maastricht, The Netherlands; 22) Texas Children's Hospital, Houston TX.

Genetic heterogeneity characterizes a variety of skeletal dysplasias often due to interacting or overlapping signaling pathways. Although genetic heterogeneity is typically viewed as an obstacle to gene discovery, genetically heterogenous syndromes provide an opportunity for insight into the biological underpinnings of disease. Robinow syndrome (RS) is a genetically heterogeneous disorder typically characterized by distinctive facial characteristics, mesomelic limb shortening, and genital hypoplasia. Pathogenic variants affecting *ROR2* in the autosomal recessive form, and *WNT5A* in a few autosomal dominant cases implicated the involvement of the noncanonical Wnt pathway in RS. Recently, in ~20-30% of the individuals, we identified a specific mutational mechanism: clustered protein-truncating variants affecting two out of the three human orthologues of the drosophila Dishevelled protein, DVL1 and DVL3, which are key mediators of the Wnt pathway. The RS associated variants in *DVL1* and *DVL3* are -1 frameshifting variants leading to the escaping from nonsense-mediated decay. We, therefore, hypothesized that other relevant variants in genes encoding proteins acting in the noncanonical Wnt pathway would underlie the remaining ~70% 'unsolved' RS cases. To address this hypothesis, we recruited 22 unrelated individuals diagnosed with Robinow or Robinow-like phenotypes to be studied in collaboration with the Baylor-Hopkins Center for Mendelian Genomics. In 20/22 (91%) kindreds we identified a likely pathogenic variant in novel, candidate or known genes, including indels affecting the C-terminus of DVL1/DVL3, confirmed to contribute to ~30% of the cases. Importantly, we detected five individuals with heterozygous variants in *FZD2*, and biallelic variants in *NXN* in the affected individual in two families, implicating two novel genes in dominant and *recessive* RS, respectively. Furthermore, all eight genes with causative variants - *ROR2*, *WNT5A*, *DVL1, DVL3, FZD2, NXN,* and the two candidate genes - *RAC3,* and *GPC4,* play a role in the Wnt pathway. In summary, our data suggest that Robinow syndrome results from perturbations in the ratio of canonical/noncanonical signaling during development. Collectively, our work reveals the contribution of distinct genes from the same pathway, elucidating the genetic basis for the locus heterogeneity in Robinow syndrome, and demonstrating the roles of various Wnt-related proteins in human development.

117

Alterations in NFkB signaling contribute to the bone fragility in osteogenesis imperfecta type V. *R. Marom, C. Lietman, A. Rajagopal, M. Jain, MM. Jiang, Y. Chen, E.M. Munivez, T.K. Bertin, R. Chen, B.H. Lee.* Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.

Osteogenesis Imperfecta (OI) type V is characterized by increased bone fragility, bone deformities, hyperplastic callus formation and calcification of the interosseous membranes. It is caused by a common mutation in the 5' UTR of *IFITM5* gene (c.-14C>T), which introduces a new start codon adding 5 amino acid residues to IFITM5 protein. The purpose of this study was to identify downstream signaling cascades that may be altered by this mutant protein, which is known to have normal localization at the cell membrane. We utilized RNA sequencing and proteomic studies via reverse phase protein-array (RPPA) method in bone of mutant transgenic mice and cell culture models. We have previously described a transgenic mouse model overexpressing the OI type V mutant *Ifitm5* in bone that showed severe bone deformities and perinatal lethality. Overexpression of the mutant *Ifitm5* caused abnormal mineralization, persistence of cartilage-like matrix throughout the limb, and altered cell morphology with chondrocyte-like cells in bone. RNA-sequencing in calvaria from transgenic mice suggested activation of NFkB signaling in the mutant model. The expression of *Ptgs2*, encoding the inflammatory mediator cyclooxygenase-2, was increased 6 fold in *Ifitm5* -mutant calvaria. In vitro mineralization assay in MC3T3 cells overexpressing mutant *Ifitm5* showed reduced alizarin red staining, and the application of an NFkB inhibitor partly rescued this phenotype. Interestingly, proteomic analysis using RPPA detected elevated levels of SOX9 in the mutant samples. NFkB signaling is reported to facilitate chondrogenic differentiation via regulation of SOX9 expression during endochondral bone ossification. Inflammation is known to negatively affect bone formation, and this effect is, at least partly mediated via NFkB pathway. We suggest that altered NFkB signaling plays a role in the pathogenesis, and contributes to the development of bone fragility in OI type V. .

118

Whole genome sequencing of Atacama skeleton shows novel mutations linked with dysplasia. *S. Bhattacharya[1], J. Li[2], A. Sockell[3], F. Bava[4], M. Kan[1], S. Chen[1], M. Avila-Arcos[5], X. Ji[6], N. Asadi[2], R. Lachman[7], H. Lam[2], C. Bustamante[3], A. Butte[1], G. Nolan[4].* 1) Institute for Computational Health Sciences, University of California San Francisco, San Francisco, CA; 2) Roche Sequencing Solutions, Belmont, CA; 3) Department of Genetics, Stanford University, Stanford, CA; 4) Baxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, Stanford University, Stanford, CA; 5) International Laboratory for Human Genome Research, National Autonomous University of Mexico (UNAM); 6) Human Immune Monitoring Center, Stanford University, Stanford, CA; 7) Department of Pediatric Radiology, Stanford University School of Medicine, Stanford, CA.

ABSTRACT Over a decade ago, the Atacama humanoid skeleton (Ata) was discovered in the Atacama region of Chile. The Ata specimen carried a strange phenotype – 6-inch stature, fewer than expected ribs, elongated cranium, and accelerated bone age, leading to speculation that this was a preserved non-human primate, human fetus harboring genetic mutations, or even an extraterrestrial. We previously reported that it was human by DNA analysis with an estimated bone age of about 6-8 years old at the time of demise. In order to determine the possible genetic drivers of the observed morphology, DNA from the specimen was subjected to whole genome sequencing using the Illumina HiSeq platform with an average 11.5x coverage of 101 base pair, paired-end reads. In total, 3,356,569 single-nucleotide variations (SNVs) were found as compared to the human reference genome, 518,365 insertions and deletions (InDels) were detected. Here, we present the detailed whole genome analysis showing that Ata is a female of human origin, likely of Chilean descent, and its genome is enriched for rare (*COL1A1, COL2A1*) and novel (*KMT2D, FLNB, ATR, TRIP11, PCNT*) mutations in genes previously linked with diseases of small stature, rib anomalies, cranial malformations, premature joint fusion, and osteochondrodysplasia (also known as skeletal dysplasia). The rare and novel variants were validated by targeted sequencing method. Together, these findings provide a molecular characterization of Ata's peculiar phenotype, which likely results from multiple known and novel putative gene mutations affecting bone development and ossification.

119

Detection of long repeat expansions from PCR-free whole-genome sequence data. *E. Dolzhenko[1], J.J.F.A. van Vugt[2], K. Ibáñez[3], G. Narzisi[4], M.A. Bekritsky[5], M. van Blitterswijk[6], A. Tucci[3,7], K.R. Smith[3], R. Rademakers[6], R. McLaughlin[8,9], W. Sproviero[10], A. Jones[10], A. Pittman[11], S. Morgan[11], O. Hardiman[8,9], A. Al-Chalabi[10], C. Shaw[10], K. Morrison[12], P.J. Shaw[13], C. Reeves[4], L. Winterkorn[4], N.S. Wexler[14], D.E. Housman[15], C. Ng[15], A. Li[15], R.J. Taft[1], L.H. van den Berg[2], D.R. Bentley[5], J.H. Veldink[2], 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) Genomics England, Queen Mary University London, Dawson Hall, London, EC1M 6BQ; 4) New York Genome Center, 101 Avenue of the Americas, New York, NY, USA; 5) Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, UK; 6) Department of Neuroscience, Mayo Clinic, Jacksonville, FL, USA; 7) Medical Cytogenetics and Molecular Genetics Lab, IRCCS Istituto Auxologico Italiano, Italy; Department of Medical Biotechnology and Translational Medicine, Università degli Studi di Milano, Italy; 8) Academic Unit of Neurology, Trinity College Dublin, Trinity Biomedical Sciences Institute, Dublin, Republic of Ireland; 9) Department of Neurology, Beaumont Hospital, Dublin, Republic of Ireland; 10) Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, King's College London, London, UK; 11) Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK; 12) University of Southampton, Southampton, UK; 13) Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield, UK; 14) Columbia University, 1051 Riverside Drive, New York, NY USA; Hereditary Disease Foundation, 3960 Broadway, 6th floor, New York, NY, USA; 15) Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA, USA.

Identifying large repeat expansions such as those that cause amyotrophic lateral sclerosis (ALS) and fragile X syndrome (FXS) is challenging for short-read (100-150 bp) whole genome sequencing (WGS) data. A solution to this problem is an important step towards integrating WGS into precision medicine. To this end, we developed a software tool called ExpansionHunter that, using PCR-free WGS short-read data, can genotype repeats at a locus of interest even if the expanded repeat is larger than the read length. To test our method we applied ExpansionHunter to a set of 144 samples harboring repeat expansions associated with Huntington's disease, fragile X syndrome, Friedreich's ataxia and five other genetic disorders. All repeat expansions in these samples were assessed correctly even though many of the repeats were significantly longer than the read lengths. We also tested the ability of our method to genotype short repeats by comparing our results to fragment length analysis of the *C9orf72* repeat on 860 samples and determined that our genotypes were >95% accurate. Motivated by this success, we used ExpansionHunter to analyze repeat sizes in two large cohorts of samples. First, we applied our algorithm to WGS data from 3,001 ALS patients who have been tested for the presence of the *C9orf72* repeat expansion with repeat-primed PCR (RP-PCR). ExpansionHunter correctly classified all (212/212) of the expanded samples as either expansions (208) or potential expansions (4). Additionally, 99.9% (2,786/2,789) of the wild type samples were correctly classified as wild type by this method with the remaining three identified as possible expansions. We next applied ExpansionHunter to WGS data from 4,298 individuals with unidentified genetic diseases that were sequenced as part of the 100,000 Genomes Project Rare Disease Programme. Our findings included pathogenic expansions in *C9orf72*, *HTT*, and *FMR1* genes that were subsequently validated. In one family, the mother had the *FMR1* premutation (spanning 102 repeat units) and the child had the full expansion (223 repeat units). We will present validation of ExpansionHunter and the results from ongoing discovery efforts in the 100,000 Genomes Project Rare Disease Programme sequencing project. In addition, we will demonstrate how ExpansionHunter can be used to discover novel pathogenic repeat expansions.

120

A spinocerebellar ataxia mapping to the *SCA37* locus is caused by a pentanucleotide ATTTC repeat insertion in the noncoding region of the *DAB1* gene. *J.R. Loureiro[1,2,3], A.I. Seixas[1,2], C. Costa[4], A. Ordóñez-Ugalde[5], H. Marcelino[1,2], C.L. Oliveira[1,2], J.L. Loureiro[6], A. Dhingra[7], E. Brandão[6], V.T. Cruz[6], A. Timóteo[4], B. Quintáns[5], G.A. Rouleau[8], P. Rizzu[7], A. Carracedo[5], J. Bessa[1,2], P. Heutink[7], J. Sequeiros[1,2,3], M.J. Sobrido[5], P. Coutinho[1,2], I. Silveira[1,2].* 1) IBMC-Institute for Molecular and Cell Biology, Universidade do Porto, Portugal; 2) I3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal; 3) ICBAS, Universidade do Porto, Portugal; 4) Department of Neurology, Hospital Fernando Fonseca, Amadora, Portugal; 5) Instituto de Investigación Sanitaria and Fundación Pública Galega de Medicina Xenómica, Centro para Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Santiago de Compostela, Spain; 6) Department of Neurology, Hospital São Sebastião, Feira, Portugal; 7) DZNE- German Center for Neurodegenerative Diseases, Tübingen, Germany; 8) Montreal Neurological Institute, and department of Neurology and Neurosurgery, McGill University, Montréal, Canada.

The spinocerebellar ataxias (SCAs) are a heterogeneous group of neurodegenerative diseases, usually characterized by late-onset of gait, limb and speech ataxia, resulting from progressive cerebellar degeneration. For SCAs 28 genes have been identified, but only 5 originate from noncoding mutations. The first noncoding pathogenic repeat insertion in a polymorphic repeat was described for SCA31, in 2009, hinting an emergent new type of mutations. Over half of the SCA families, however, remain without molecular diagnosis. To identify the molecular basis of an unknown SCA genetic type in 3 large families, we performed genome-wide linkage analysis. We obtained LOD scores of 5.1, 4.4 and 2.2 on chromosome 1p32, a region previously linked to *SCA37*. Haplotype analysis established a candidate region of 8 Mb; fine-mapping allowed us to narrow this region down to 2.8 Mb. We next performed NGS on 6 affected individuals and 4 unaffected relatives, but this strategy failed to detect the causative mutation. However, the analysis of 278 repeats led us to identify an $(ATTTC)_n$ insertion in the middle of an ATTTT repeat tract in the noncoding *DAB1, reelin adaptor protein* (*DAB1*) gene. This insertion segregates with the disease and is absent in 260 normal individuals. Sequence analysis showed that normal alleles ranged from 7-400 ATTTT repeats without any insertion, whereas affected individuals carried an allele with a complex $[(ATTTT)_{60-79}(ATTTC)_{31-75}(ATTTT)_{58-90}]$ configuration. The insertion is present in a distinct haplotype and showed increased size bias upon transmission correlating with earlier onset. The $(ATTTC)_n$ is located in the *DAB1* 5'-UTR introns of cerebellar-specific transcripts arising from alternative promoter usage, during human fetal brain development and maintained in adult cerebellum. Transfection of $(ATTTT)_{57}(ATTTC)_{58}(ATTTT)_{73}$, but not the $(ATTTT)_7$ or $(ATTTT)_{139}$, led to nuclear RNA aggregation. Zebrafish embryos injected with $(AUUUU)_{57}(AUUUC)_{58}(AUUUU)_{73}$ had a significant increased lethality when compared with zebrafish embryos injected with $(AUUUU)_7$ or $(AUUUU)_{139}$. Upon analysis of the repeat flanking regions by transgenesis assays, we uncovered a promoter driving antisense expression upstream of the repeat, suggesting that it is transcribed in both orientations. Altogether, we establish an (ATTTC)n insertion as the cause of this SCA through RNA-mediated toxicity. Based on the genetic evidence, we propose this insertion is the molecular basis for SCA37.

121

Identification of genetic modifiers of FXTAS by combining whole genome sequencing with fly genetics. *H.E. Kong, J. Lim, E.G. Allen, D.J. Cutler, M.E. Zwick, S.L. Sherman, S.T. Warren, T.S. Wingo, P. Jin.* Emory University, Atlanta, GA.

Fragile X-associated tremor/ataxia syndrome (FXTAS) is an adult-onset neurodegenerative disorder caused by the premutation CGG repeat expansion (55-200 repeats) within the 5'UTR of *FMR1*. A significant proportion of male premutation carriers develop the FXTAS phenotype later in adulthood, which includes intention tremor, cerebellar ataxia, progressive neurodegeneration, parkinsonism and cognitive decline, while other male carriers do not exhibit disease at all. Importantly, a mechanistic understanding of FXTAS disease pathogenesis is yet to be completely elucidated, and as a result, we do not have an explanation for the significant variability in the onset and phenotypic presentation of FXTAS. In an effort to tackle this conundrum, we set out to identify genetic modifiers that modulate CGG toxicity and that may account for the variable phenotype and onset of disease. We performed whole genome sequencing on both FXTAS patients and premutation carrier controls. We identified 2660 and 2878 variants associated with early-onset and late-onset FXTAS cases, respectively. To further test the potential roles of these genes in CGG repeat-mediated toxicity, we employed the Drosophila model of expanded CGG repeat that we developed previously, which demonstrated that the expression of the expanded CGG repeat RNA was sufficient to cause neurodegeneration. Among the 96 genes that we tested, 76 RNAi lines displayed either enhanced or suppressed neuronal toxicity associated with rCGG repeats. Among them are *RNF157, PABPC1L, PSMB5* and *PURG*. Taken together, our analyses suggest the presence of multiple genetic modifiers that could modulate the age-of-onset in FXTAS patients. We show that combining whole genome sequencing with a *Drosophila* genetic screen could facilitate the identification of novel genetic modifiers of human diseases.

122

Dissecting the causal mechanism of X-linked dystonia-Parkinsonism by integrating genome and transcriptome assembly. *R. Yadav[1,2,3], T. Aneichyk[1,2,3], W.T. Hendriks[2,4,5], D. Shin[2,4,5], D. Gao[1,2,3], C.A. Vaine[2,4,5], R.L. Collins[1,3,6], A. Stortchevoi[1,2], B. Curral[1,2], H. Brand[1,2,3], C. Hanscom[1,2], C. Antolik[1,2], M. Dy[2,4], A. Ragavendran[1,2], P. Acuña[2,4], C. Go[7], Y. Sapir[2], B. Wainger[2], D. Henderson[2,4], J. Dhakal[2,4], N. Ito[2,4,5], N. Weisenfeld[8], D. Jaffe[8], N. Sharma[2,4], X.O. Breakefield[2,4,5,9], L.J. Ozelius[1,2,4], D.C. Bragg[2,4,5], M.E. Talkowski[1,2,3,4,10,11].* 1) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 2) Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA; 3) Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA; 4) The Collaborative Center for X-linked Dystonia Parkinsonism, Massachusetts General Hospital, Charlestown, MA; 5) Harvard Brain Science Initiative, Harvard Medical School, Cambridge, MA; 6) Program in Bioinformatics and Integrative Genomics, Division of Medical Sciences, Harvard Medical School, Boston, MA; 7) Jose Reyes Memorial Medical Center, Manila, 1003, Philippines; 8) Genome Sequencing and Analysis Program, Broad Institute, Cambridge, MA; 9) Department of Radiology, Massachusetts General Hospital, Boston, MA; 10) Departments of Psychiatry and Pathology, Massachusetts General Hospital, Boston, MA; 11) Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA.

Introduction: X-linked Dystonia Parkinsonism (XDP) is an adult-onset neurodegenerative disorder that is indigenous to the Philippines and exhibits features of both dystonia and parkinsonism in a characteristic temporal progression. Over the last two decades, conventional genetic approaches have linked XDP to a 410 Kb founder haplotype that included the same seven markers shared by all probands, five single nucleotide variations known as Disease Specific Changes (DSCs), 2627 bp sine-VNTR-Alu (SVA) retrotransposon and 48bp deletion; yet, the causal variant and pathogenic mechanism have remained unknown. **Materials and Methods:** We integrated *de novo* genome and transcriptome assembly methods using short-read, multiple long-read, linked-read and targeted sequencing technologies among XDP probands, carriers, and unaffected family members (n=120). We characterized transcriptomes from fibroblasts among 46 subjects (probands, carriers, and unaffected family members) as well as a subset of 24 clones from iPSC-derived neural stem cells (NSCs) and induced neurons. **Results:** These analyses identified a set of 47 alleles defining the initial founder haplotype as well as three independent recombination events that narrowed the putative causal genomic segment to 219 Kb including only the *TAF1* gene. Intriguingly, the *de novo* transcriptome assembly in NSCs revealed a striking expression signature involving aberrant splicing and intron retention (IR) of the *TAF1* gene, and partial retention of intronic sequence proximal to the SVA insertion within intron 32 of *TAF1* that has never been observed in controls. Canonical *TAF1* transcripts were significantly reduced in XDP probands as compared to unaffected relatives in iPSC-derived NSCs, and this reduction was driven by decreased exon usage 3' to exon 32. Both the aberrant splicing and reduced *TAF1* expression signatures were rescued following CRISPR/Cas9 excision of the SVA in patient-derived NSCs. Transcriptome-wide analyses also revealed expression alterations of neurodegenerative disorders genes as well as co-expression of a highly interconnected network of genes associated with synaptic transmission and neurodevelopment. **Conclusions:** These data implicate aberrant splicing and intron retention as a consequence of a noncoding SVA insertion into *TAF1* as a possible pathogenic mechanism in XDP, and propose a potential roadmap for integrated, reference-free genome and transcriptome assemblies in genomic studies of population isolates.

123

Functional prioritization of Huntington's disease onset modifier genes in the Hdh^{Q111}/+ mouse. *J. Loupe[1], T. Gillis[1], M. Kovalenko[1], J. Mysore[1], A. Nowell[1], R. Mauro Pinto[1], V. Wheeler[1], J. Gusella[1,2], J. Lee[1], M. MacDonald[1,2], GeM-HD Consortium.* 1) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 2) Broad Institute of Harvard and MIT, Cambridge, MA.

Huntington's Disease (HD) is a dominantly inherited neurodegenerative disease, featuring motor signs, psychiatric disturbances and intellectual decline. The HD mutation is an unstable expanded CAG trinucleotide repeat in *HTT* whose size is negatively correlated with the age at onset. A genome wide association study (GWAS) to identify genetic modifiers of age at onset of motor symptoms, consisting of four thousand HD individuals, identified genome-wide significant loci on CHR15 and CHR8 that hasten onset by 6 years and 1.6 years, respectively. Each locus harbors two viable candidates: *FAN1* and *MTMR10* at CHR15, along with *RRM2b* and *UBR5* at CHR8. To guide the prioritization of the human modifier candidates at each locus, we have tested the hypothesis that a loss of function mutation in the true modifier gene may influence CAG length-dependent somatic instability of the CAG repeat, a prominent phenotype in the striatum and liver of Hdh^{Q111}/+.B6 CAG repeat knock-in mice. To test this, we created seven lines of mice with CRISPR/Cas9-targeted inactivating mutations for each modifier gene candidate. These mice were mated with Hdh^{Q111}/+.B6 mice to generate cohorts of progeny with the Hdh^{Q111} CAG expansion mutation that are wild-type or alternatively mutated in an individual candidate modifier gene. Somatic instability of the CAG repeat in genomic DNA extracted from striatum and liver at 2.5 or 5 months of age was assessed by PCR amplification and conversion of the ABI 3730XL fragment sizes to a standardized CAG repeat instability index. The results of these comparisons demonstrated that while *Mtmr10* and *Ubr5* inactivating mutations did not affect CAG repeat size, the inactivating mutations in the alternative genes - *Fan1* and *Rrm2b* - significantly altered the size of the CAG repeat in both tissues. An inactivating mutation in *Fan1* led to greater CAG repeat instability, potentially by increasing the burden of unresolved interstrand cross-links that are subject to error-prone DNA repair. *Rrm2b* inactivating mutation yielded decreased instability, perhaps by decreasing nucleotide pools necessary for DNA repair. These findings implicate CHR15 *FAN1* and CHR8 *RRM2B* as the true HD onset modifier genes and, furthermore, strongly suggest that their effects may reflect loss of function and gain of function variants, respectively, in increasing the size of the primary HD mutation in the brain and other tissues as a driver of the rate of onset of clinical motor symptoms.

124

Insights into the molecular pathogenesis of Huntington's disease via multidimensional data analysis of the OVT73 sheep model. *E.R. Mears[1], R.R. Handley[1], S.J. Reid[1], J.F. Gusella[5], M.E. MacDonald[5], S.R. Rudiger[4], S.C. Bawden[4], S. Patassini[2,6], P. Maclean[3], R. Brauning[3], H.J. Waldvogel[2], R.L.M. Faull[2], R.G. Snell[1].* 1) School of Biological Sciences, The University of Auckland, Auckland 1010, New Zealand; 2) Centre for Brain Research, Faculty of Medical and Health Science, The University of Auckland, Auckland 1023, New Zealand; 3) Invermay Agricultural Centre, AgResearch Ltd., Mosgiel 9053, New Zealand; 4) Molecular Biology and Reproductive Technology Laboratories, South Australian Research and Development, Adelaide, SA 5350, Australia; 5) Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston MA 02114, United States of America; 6) Centre for Advanced Discovery and Experimental Therapeutics (CADET), Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Sciences Centre, Manchester M13 9WL, UK; Institute of Human Development, Faculty of Medical and Human Sciences, The University of Manchester,.

Huntington's disease (HD) is a neurodegenerative disorder caused by an expanded CAG repeat in the *huntingtin* gene. Despite considerable effort, the underlying biochemical mechanism leading to cell dysfunction and death is yet to be discovered. We have taken a multidimensional data analysis approach to identify the early disease associated changes using information gained from our prodromal HD sheep model (OVT73). The OVT73 transgenic sheep line carries full length huntingtin cDNA with an expanded polyglutamine coding repeat of 73 units. This large mammalian model of HD shows no overt symptoms even at 10 years old, but does develop huntingtin positive inclusions characteristic of HD. Our laboratory has generated an extensive catalogue of molecular data from a single cohort of 5 year old OVT73 and matched control sheep. The data ranges from basic biometric measurements, to high-throughput transcriptomic, metabolomic and proteomic data, from blood, brain and other tissues. Very significant changes have been observed when comparing transgenic and control animals, within these single datasets. However this approach does not give insight into the relationships across the phenotypic features. Using the programming language, R, we have now collated and integrated all the existing datasets from the 5 year old sheep cohort into a single platform. Multivariate analysis of this data has revealed very significant changes in the urea pathway, including increased urea levels and upregulation of the urea transporter, *SLC14A1* in the brain, providing good evidence of a perturbation prior to cell death. Further analyses have revealed disease-specific changes in osmotic regulation and membrane transport, (*RHCG*, *SLC5A7* *SLC12A2*, *AVPR1A*). We also demonstrate that factors involved in metabolic pathways show striking differential correlations in the liver. Taken together, our findings imply a possible mechanism of salt imbalance and/or a metabolic dysfunction. This is not unexpected given the considerable weight loss observed in pre-manifest HD patients. Other HD datasets have been included in the database, including human brain expression and mouse allelic series. Results from these cross-species comparative analyses confirm our findings and validate the reliability of OVT73 as a HD model. Our final aim is to make this database available to the HD research community in a queryable format whereby a range of multivariate methods can be applied for further exploration.

125

Evidence-based assessments of clinical actionability in the context of secondary findings: Updates from ClinGen's Actionability Working Group. *J.E. Hunter[1], E.M. Webber[1], K. Lee[2], K.R. Muessig[1], L.G. Biesecker[3], A.H. Buchanan[4], N. Lindor[5], C.L. Martin[6], J.M. O'Daniel[2], E.M. Ramos[7], A. Slovotinek[8], N. Sobreira[9], M.A. Weaver[10], M.S. Williams[4], J.P. Evans[2], K.A.B. Goddard[1].* 1) Center for Health Research, Kaiser Permanente Northwest, Portland, OR; 2) Department of Genetics, University of North Carolina, Chapel Hill, NC; 3) Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute (NHGRI), National Institutes of Health, Bethesda, MD; 4) Genomic Medicine Institute, Geisinger Health System, Danville, PA; 5) Department of Medical Genetics, Mayo Clinic, Scottsdale, AZ; 6) Autism & Developmental Medicine Institute, Geisinger Health System, Danville, PA; 7) Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 8) Department of Pediatrics, University of California, San Francisco, San Francisco, CA; 9) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 10) American College of Medical Genetics and Genomics, Bethesda, MD.

The ClinGen Actionability Working Group (AWG) has developed a structured, evidence-based framework for assessing the clinical actionability of genes and associated disorders in the context of pathogenic secondary findings in adults. For these assessments, the AWG has defined actionability as clinical interventions that could improve future health outcomes, including patient management, surveillance, family management, and circumstances to avoid. The AWG framework addresses four domains that may impact the actionability of interventions for each gene-disorder pair: 1) the nature of the threat to health; 2) the likelihood of the outcome (penetrance); 3) the effectiveness of the intervention to prevent harm; and 4) the risk/burden of the intervention to the individual. As part of the framework, each domain is scored using a semi-quantitative metric that takes into account the knowledge base for the likelihood and effectiveness domains. The AWG has assessed 124 gene-disorder pairs and scored 288 outcome-specific interventions. The gene-disorder pairs include the list of 59 genes recommended for return as secondary findings by the American College of Medical Genetics and Genomics (ACMG SF v2.0). As expected, the scores associated with the ACMG genes are higher for the 4 domains of clinical actionability compared to scores associated with genes not included in the ACMG recommendations. Nevertheless, the AWG framework has identified non-ACMG genes that receive high scores across multiple domains of clinical actionability. We will present several interesting patterns that have emerged across the four domains of actionability. In general, common features of gene-disorder pairs with higher actionability scores include cardiovascular conditions that can lead to sudden death and disorders for which non-invasive imaging or monitoring leads to reduction in morbidity or mortality. Factors that frequently negatively impact actionability scores include limited documentation of the likelihood of the outcome or effectiveness of the intervention. Additionally, a subset of disorders has lower actionability scores due to moderate likelihood and non-invasive interventions with unknown effectiveness. The framework developed by the AWG provides support to the research and clinical community for making clear, streamlined, and consistent determinations of clinical actionability based upon transparent criteria to guide analysis and reporting of variation in genome-scale sequencing.

126

Secondary (incidental) findings in whole exome and genome sequencing. *A. Maksimovic[1], Z. Yüksel[1], A. M. Bertoli-Avella[1], O. Brandau[1], N. Nahavandi[1], A. Mohamed Saeed Al Shamsi[2], M. Alfadhel[3], M. A. Albalwi[4], N. Abbas Al-Sannaa[5], S. Kishore[1], P. Bauer[1,6], A. Rolfs[1,7].* 1) Centogene AG, Rostock, Mecklenburg-Vorpommern, Germany; 2) Department of Pediatrics, Tawam Hospital, Al-Ain, United Arab Emirates; 3) King Abdullah International Medical Research Centre, King Saud bin Abdulaziz University for Health Science, Genetics Division, Department of Pediatrics, King Abdulaziz Medical City, Riyadh, Saudi Arabia; 4) Pathology and Laboratory Medicine, King Abdulaziz Medical City, Riyadh, Saudi Arabia; 5) Johns Hopkins Aramco Health Care, Pediatric Services, Dhahran, Saudi Arabia; 6) Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany; 7) Albrecht-Kossel-Institute, University of Rostock, Rostock, Germany.

Introduction: Reporting secondary findings (SF) aims to prevent or significantly reduce the morbidity and mortality of the disorders that are not directly related to the actual genetic test request. In order to promote a standardized reporting of actionable information from clinical exomic/genomic sequencing the American College of Medical Genetics and Genomics (ACMG) published a society consensus list of genes to be reported as incidental or SF. In CENTOGENE, we report SF according to the up-to-date ACMG guidelines (Kalia *et al.* 2017, *Genet Med.* 19(4):484). **Patients and Methods:** We present SF results for whole exome sequencing (WES) and whole genome sequencing (WGS) requests between 01/29/2016 and 17/03/2017. When SFs are requested, the ACMG listed genes are actively checked for known pathogenic and likely pathogenic variants for reporting. For Trio analyses, SFs are only analyzed for the index patient. Parental status for SF identified in the index can be requested with a new consent of the parents which are then issued on individual reports. SFs are not reported for fetal samples or samples from deceased individuals. **Results:** Among 4010 requests of WES (93.6%) (n=3754; Trio=2357, Solo=1397) and WGS (6.4%) (n=256; Trio=225, Solo=31), SFs were requested in 1670 (41.6%) cases of which 1585 were WES (95%; Trio=907, Solo=678) and 85 were WGS requests (5%; Trio=72, Solo=13). Among 1670 SF requests, 38 (2.2%) had a reportable variant. SFs were reported in 18 WES (Solo=10, Trio=6) and 2 WGS (Solo=2) cases in the genes *BRCA2*, *PMS2*, *TNNT2*, *DSC2* (2 times), *LDLR* (2 times), *MUTYH* (2 times), *BRCA1* (3 times), *MYBPC3* (3 times) and *SCN5A* (4 times). SFs were reported in 20 out of 38 (52%) cases as a primary finding as the phenotype of interest was the main indication for the testing. Variants in *APC*, *ATP7B*, *LDLR*, *MSH6*, *OTC*, *SMAD4*, *TNNI3*, *TNNT2*, *TSC1*, *TSC2*, *FBN1* (3 times), *PTEN* (3 times), *RYR1* (4 times) were among those reported as primary findings. **Conclusion:** The up-to-date SF ACMG list includes 59 medically-actionable genes. Curating the gene list for SF is an ongoing and dynamic process. Our results are supportive of the latest recommendations about adding four other genes (*ATP7B*, *BMPR1A*, *SMAD4*, and *OTC*) as variants in three of them (*ATP7B*, *OTC* and *SMAD4*) were reported as a primary finding. Therefore, we would like to encourage the medical community to contribute to updating the SF gene list and if possible for establishing an international consensus.

127

Our experience with incidental findings in the CAUSES Research Clinic: A pediatric sequencing study in British Columbia. *S. Adam[1], C. du Souich[1], A.M. Elliott[1], J.C. Mwenifumbo[2], C.D. van Karnebeek[3], T.N. Nelson[4], A.M. Lehman[1], J.M. Friedman[1].* 1) Department of Medical Genetics, Children's & Women's Hospital, University of British Columbia, Vancouver, BC, Canada; 2) BC Children's Hospital Research Institute, Vancouver, BC; 3) Division of Biochemical Diseases, Department of Pediatrics, University of British Columbia, Vancouver, BC; 4) Department of Pathology and Laboratory Medicine, BC Children's Hospital, Vancouver, BC.

Background: The CAUSES Clinic, a 3 year translational research study at BC Children's Hospital, is performing diagnostic genome-wide trio sequencing in 500 children with suspected genetic conditions. Incidental findings (IFs) are managed according to Canadian College of Medical Geneticists' recommendations: "…we do not endorse the intentional clinical analysis of disease-associated genes other than those linked to the primary indication…". Our policy is to consider for return only highly penetrant, pathogenic and medically actionable variants that are discovered incidentally in the analysis of the genomic data. Parents are *always* informed of IFs in their child for childhood onset disorders. Parents can *choose* to learn about medically actionable adult-onset IFs for themselves. IFs related to adult-onset conditions in the child are not returned. IFs are classified on a case-by-case basis with input from the multidisciplinary study team composed of clinical geneticists, genetic counsellors, pediatric subspecialists, bioinformaticians and clinical molecular geneticists. *Results:* IFs have been found in 8 (4%) of the first 200 families analyzed. In two families, the IF was returned for the proband – one, a variant in *G6PD* and the other, a variant in *LDLR*. Seven IFs were seen in parents. There were two *BRCA2* variants seen in fathers and one in a mother; two *DPYD* variants in fathers and one in a mother; and one *LDLR* variant in a mother. Families receive pre-test counselling about IFs. Most families are very accepting of the IFs disclosure policy, but some would prefer information on all variants including adult onset conditions for their child. Eighteen parents have chosen not to receive IF information for themselves. Disclosure of IFs for probands is accompanied by genetic counselling, appropriate medical guidance, and/or referral to a specialist. Disclosure of IFs to the parents includes explanation of the condition, and possible intervention/ prevention strategies, referral recommendations, and discussion about informing other family members. *Summary:* The CAUSES IF protocol has resulted in an incidental finding rate of 4%, and has allowed us to disclose clinically actionable, high confidence, Sanger-validated IFs without spending excessive study resources on variants unrelated to the primary indication. Preliminary data indicates families have found the information returned useful, and have appreciated the opportunity to learn more about their personal health.

128

Identification of secondary genetic variation in the HudsonAlpha CSER project. *M.L. Thompson[1], C.R. Finnila[1], K.M. Bowling[1], S.M. Hiatt[1], M.D. Amaral[1], K.B. Brothers[2], K.M. East[1], D.E. Gray[1], J. Lawlor[1], W.V. Kelley[1], M. Neu[1], N.E. Lamb[1], E.J. Lose[3], C.A. Rich[2], S. Simmons[3], R.M. Myers[1], G.S. Barsh[1], E.M. Bebin[3], G.M. Cooper[1].* 1) HudsonAlpha Institute for Biotechnology, Huntsville, AL; 2) University of Louisville, Louisville, KY, USA; 3) University of Alabama at Birmingham, Birmingham, AL, USA.

We performed exome and genome sequencing for 434 families (1227 total individuals) enrolled as part of the HudsonAlpha Clinical Sequencing Exploratory Research (CSER) project. This study was designed to genetically diagnose children with developmental delay and/or intellectual disability (DD/ID), primarily through sequencing affected children and their unaffected parents. In addition to genetic variation related to the child's DD/ID, we also identified pathogenic/likely pathogenic secondary findings in 58 of the 766 parent participants (7.7%). We did not return genetic variants of uncertain significance as secondary findings in our study. Within this 7.7%, a portion (1.3%) had actionable secondary findings defined by the ACMG. We also performed a limited carrier screen assessment and 4.6% of total parent participants had variation in HBB, HEXA, or CFTR. Fourteen parents (1.8%) received a genetic diagnosis as they harbored likely pathogenic/pathogenic variation in a known disease gene (OMIM database) - some of these individuals already had a clinical diagnosis. These diagnoses include, but are not limited to, spherocytosis, carnitine deficiency, and polycystic kidney disease. When available, we assessed parent participants as mate pairs and identified genes (using an OMIM gene list) in which both parents were heterozygous carriers of pathogenic/likely pathogenic alleles for recessive disease. One parental pair (among 349 total) was identified as carriers for Wilson disease. In this report, we discuss cases that highlight the clinical significance of returning secondary findings. With a rise in the number of healthy individuals seeking genetic screening to identify disease risk, the identification of secondary findings in seemingly healthy individuals is becoming more prevalent. We suggest that there is benefit to reporting secondary findings, and demonstrate that doing so will aid in the management or prevention of current and future health-related issues, respectively.

129

Frequency of pathogenic variants in Fanconi/BRCA pathway genes in ten thousand clinical exomes referred for non-cancer indications. *S.E. Plon[1,2], A.K. Petersen[1,2], C.M. Eng[2], Y. Yang[2].* 1) Pediatrics, Baylor College of Medicine, Houston, TX; 2) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.

Background: With diagnostic rates of 25-30%, the utilization of clinical whole exome sequencing (WES) has undergone rapid expansion, generating concerns regarding the frequency of non-diagnostic secondary findings. These include monoallelic pathogenic variants in cancer genes (often with both dominant and recessive phenotypes), which can present a unique challenge to interpretation when found in patients referred for WES without a cancer indication. Clinicians must consider when to recommend additional testing (functional studies, arrayCGH, and/or targeted gene sequencing) to look for a second allele. Variants in DNA repair genes from WES have also been reported in multiple pediatric cancer cohorts with little comparative data in non-cancer cohorts. **Methods:** Here, we investigated the frequency in non-cancer patients of monoallelic pathogenic variants in genes within the BRCA/Fanconi pathway, through analysis of clinical exomes from one laboratory. We evaluated pathogenic variants in the Fanconi anemia (FA)/BRCA pathway, as important regulators of genomic stability, potential therapeutic targets and genes reported as secondary findings. We evaluated the frequency in each of 14 FA genes (*FANCA, B, C, D1/BRCA2, D2, E, F, G, I, J/BRIP1, L, N/PALB2, O/RAD51C, P/SLX4*) and, given its functional role in the FA pathway, *BRCA1*. **Results:** This postnatal cohort (n=9986) has a median age of 6 years (88% in pediatric age range). Monoallelic pathogenic variant frequencies for individual genes ranged from 0% for *FANCB* to 0.7% for *FANCL*. *BRCA1, BRCA2* and *PALB2* pathogenic variant frequencies were 0.28%, 0.37% and 0.13%, respectively. The combined frequency for the 15 FA/BRCA pathway genes was 3.4%. Ten percent of *BRCA1* and 5% of *BRCA2* pathogenic variants were missense alleles, whereas all other variants in FA genes were predicted to be truncating. Conversely, while 53% of the pathogenic variants in the FA genes were not previously reported, 90% of *BRCA1* and 92% of *BRCA2* mutations were previously reported in the literature. **Conclusion:** These data underscore that approximately 3% of a primarily pediatric cohort undergoing WES for non-cancer indications carry monoallelic pathogenic variants in Fanconi/BRCA pathway genes including 0.75% with *BRCA1/2* alleles, the latter recommended for reporting as secondary findings. These data should be considered when performing pre- and post-test WES counseling and for comparison with ongoing sequence analysis of pediatric cancer cohorts.

130

Secondary findings after virtual panels: A new frontier in incidental findings. *E.D. Esplin, S. Yang, E. Haverfield, S. Aradhya, R.L. Nussbaum.* Invitae, San Francisco, CA.

Background ACMG recommendations for secondary findings in whole exome or genome sequencing (WES/WGS) indicate pathogenic variants can be medically actionable regardless of the test indication. Previous studies have estimated the number of secondary findings in individuals using WES/WGS (Natarajan et al. PMID:27831900). It is now possible to perform diagnostic testing using assay platforms covering hundreds of genes, from which virtual panels can be derived by clinical indication. We report the estimated prevalence of secondary findings in over 3000 individuals undergoing hereditary cardiovascular genetic testing, in a dataset generated on a virtual, multi-gene panel. Methods Per an IRB approved protocol, we analyzed de-identified sequence data for 47 cancer-risk genes in 3679 patients referred for hereditary cardiovascular genetic testing. These genes were deemed medically actionable by an expert panel and include nearly all of the cancer genes recommended by the ACMG. After removing known non-pathogenic loss-of-function (LOF) variants, we classified variants as pathogenic if they had already been classified by us as pathogenic or if they were predicted to be LOF because of frameshift, nonsense, or splice-site disruption mutations. Results In 3679 individuals with personal or family history of hereditary cardiovascular conditions, we observed 141 PPVs in cancer-risk genes, for a prevalence rate of 6.03%. This includes PPVs with established management guidelines in MUTYH (64), CHEK2 (38), ATM (20), BRCA2 (17), MITF (17), FH (12), PALB2 (9), PMS2 (6), BRCA1 (5). When restricted to ACMG recommended genes, overall prevalence of PPVs was 2.72%. Conclusions Using a virtual panel strategy, this study estimates the prevalence of secondary findings for cancer-risk in individuals with hereditary cardiovascular risk at up to 6%, which is higher than recent studies. This may be due to analyzing a larger number of cancer-risk genes and the inclusion of variants which confer modest increased cancer risk, such as heterozygous MUTYH variants. However, it is also likely to be an underestimate as we did not assess novel missense or copy number variants. The secondary findings observed in this >3,000 individual study provides an estimate of overall prevalence rates and suggests that secondary findings of potential clinical utility could be gleaned from virtual multi-gene panels, a situation not currently addressed by the ACMG recommendations on secondary findings in WES/WGS.

131

Identification transcriptomic and epigenetic mediators in Alzheimer's disease: Bayesian inference and causal mediation analysis of regulatory programs in GWAS statistics. *Y. Park[1,2], A. Sarkar[1,2], L. He[1,2], M. Kellis[1,2].* 1) CSAIL, MIT, Cambridge, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA.

Summary-based transcriptome-wide studies (sTWAS) have proven to be a powerful tool to identify genes associated with human complex diseases by aggregating cis-regulatory genetic effects on gene expression. However, we noticed several limitations. Firstly, sTWAS rely on building predictive models of gene expression, which are sensitive to the sample size on which they are trained. Moreover, the multivariate normality assumption of the true GWAS effect sizes does not necessarily hold across all genome and potentially inflates sTWAS statistics. Lastly, sTWAS-significant genes could be interpreted in many different ways–mediation, pleiotropy and reverse causation. It is important to distinguish mediating genes from the others to understand regulatory mechanisms of disease progression. Here, we improve sTWAS approach by addressing the limitations in novel computational framework. (1) We reduced generalization error of predictive models controlling sparsity of model parameters by novel Bayesian inference algorithm on spike-slab prior. Our black-box inference algorithm works for versatile link functions of generalized linear models. (2) We examined GWAS summary statistics on genes and DNA methylation levels and performed non-parametric significance tests through 5 million QTL permutations. Unlike the original sTWAS statistics (Gusev et al. 2016), our permutation test is immune to localized bias of GWAS statistics that violates standard normal assumptions. (3) We extracted true mediation effects from horizontal pleiotropy (colocalization of QTLs with GWAS SNPs), by fitting a complete joint model that can distinguish mediation from direct marginal effects by variable selection. We trained polygenic expression and methylation QTL models on ROS/MAP postmortem Brain samples to carry out sTWAS and summary-based methylome-wide studies (sMWAS) on GWAS statistics of neurodegenerative disorders. Our extensive simulation confirmed eQTL and mQTL models easily generalize without over-fitting to reference training data. We restricted downstream analyses on 4,708 heritable genes and 11,638 CpG sites with permutation FDR < 5%. In sTWAS we found 18 significant genes and 12 CpGs at FDR < 5%. Remarkably 8 genes were found to be genuine transcriptomic mediators: MRPL24, THNSL2, SULT1C2, HLA-DOB, SPG7, CHRNE, SPATA20, TOMM40. Moreover, 2 CpGs, cg14991358 and cg25010880, were found to be epigenetic mediators.

132

Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *L.S. Chen, F. Yang, J. Wang, B. Pierce, GTEx consortium.* University of Chicago, Chicago, IL.

The impact of inherited genetic variation on gene expression in humans is well-established. The majority of known expression quantitative trait loci (eQTLs) impact expression of local genes (cis-eQTLs). More research is needed to identify effects of genetic variation on distant genes (trans-eQTLs) and understand their biological mechanisms. One common trans-eQTLs mechanism is “mediation” by a local (cis) transcript. Thus, mediation analysis can be applied to genome-wide SNP and expression data in order to identify transcripts that are “cis-mediators” of trans-eQTLs, including those “cis-hubs” involved in regulation of many trans-genes. Identifying such mediators helps us understand regulatory networks and suggests biological mechanisms underlying trans-eQTLs, both of which are relevant for understanding susceptibility to complex diseases. The multi-tissue expression data from the Genotype-Tissue Expression (GTEx) program provides a unique opportunity to study cis-mediation across human tissue types. However, the presence of complex hidden confounding effects in biological systems can make mediation analyses challenging and prone to confounding bias, particularly when conducted among diverse samples. To address this problem, we propose a new method: Genomic Mediation analysis with Adaptive Confounding adjustment (GMAC). It enables the search of a very large pool of variables, and adaptively selects potential confounding variables for each mediation test. Analyses of simulated data and GTEx data demonstrate that the adaptive selection of confounders by GMAC improves the power and precision of mediation analysis. Application of GMAC to GTEx data provides new insights into the observed patterns of cis-hubs and trans-eQTL regulation across tissue types.

133

Casual inference in imaging-genetic data analysis. *N. Lin[1], Z. Hu[2,1], R. Jiao[1], L. Luo[3], V. Calhoun[3], M. Xiong[1,2].* 1) University of Texas School of Public Health, Houston, TX; 2) Fudan University School of Life Science, Shanghai,China; 3) University of New Mexico in Albuquerque, Albuquerque,NM.

Next generation of genomic, sensing and image technologies will produce a deluge of DNA sequencing, transcriptomes and imaging data. Standard imaging- genetic data analysis is to test the association of single variants with imaging signals. Despite significant progress in dissecting the genetic architecture of imaging signal variation by association analysis, understanding the etiology and mechanism of complex diseases remains elusive. The purpose of this talk is to develop novel statistical methods for paradigm changes in big genomic, and imaging data analysis from low dimensional data analysis to high dimensional data analysis and from association analysis to causal analysis. We develop novel structural causal models coupled with integer programming as a new framework for inferring large-scale causal networks of genomic-brain images. The proposed method for large-scale genomic-imaging causal network analysis was applied to the MIND clinical imaging consortium's schizophrenia image-genetic study with 142 series of diffusion tensor MRI images (DTI) and 50,725 genes (731,961SNPs) typed in 64 schizophrenia patients and 78 healthy controls. Images were segmented into 23 regions. The imaging signals in each region were summarized by three dimensional FPC scores. A region was taken as a node. The sparse SEMs were used to compute score of image node and network regularized logistic regression will be used to compute score function of the SCZ node. IP was used to search the optimal causal graph. Additive noise models (ANMs) were also used to examine the causal relationships between imaging region and SCZ. Most image regions (22 out of 23) were associated with the SCZ. Linear SEMs algorithm with IP identified 5 image regions causing SCZ. The ANM narrowed down 5 regions to 3 causal regions: Frontal_R, Occipital_R , and Occipital & Pareital_Sup. We identified 176 (6.83e-8) SNPs that were associated with imaging signal variation, 82 SNPs significantly associated with SCZ and 27 SNPs that cause imaging signal variation. Finally, we infer the genotype-imaging-disease causal networks and estimated direct and indirect effects of the SNP on the imaging regions and SCZ as well. We shift the paradigm of big genomic and imaging data analysis from association studies to causal inference and provide powerful tools for unravelling causal chain of mechanisms of psychiatric disorders, delivering new therapeutic targets and biomarkers for precision medicine.

134

Gene x environment interactions and causal relationships between obesity and depression. *T. Frayling, A.R. Wood, H. Yaghootkar, R. Beaumont, S.E. Jones, M.A. Tuke, K.S. Ruth, R.M. Freathy, A. Murray, M. Weedon, J. Tyrrell.* University of Exeter Medical School, University of Exeter, Exeter, Devon, United Kingdom.

Statement of Purpose Obesity is associated with a higher risk of depression and vice versa but the causal nature is unknown. We also do not know if depression interacts with BMI to accentuate the genetic risk of obesity and until recently most gene x "environment" interactions of clinical phenotypes have been underpowered. We aimed to use genetic data from the UK Biobank to test a) whether there are bidirectional causal relationships between depression and obesity and b) perform interaction analyses to test whether or not depression accentuates an individual's genetic risk of obesity. *Methods* Firstly, we created genetic risk scores (GRS) for depression and BMI in 120,000 individuals from the UK Biobank, using genetic variants identified in the most recent genome-wide association scans of these traits. We performed Mendelian Randomization (MR) analyses to test for causal relationships between BMI and depression (N=7,781 cases based on major depressive disorder) and vice versa. We also investigated the association of the BMI GRS with BMI in depressed and non-depressed individuals and tested for an interaction. Sex-specific analyses and sensitivity analyses were also performed. *Results* Our Mendelian randomisation analyses suggested a causal effect of depression on higher BMI: a 1.10 odds ratio of depression was causally associated with a 0.09 kg/m^2 (95%CI: 0.01, 0.16; p=0.032) higher BMI. These results were similar in both sexes. In contrast, there was only nominal evidence of a causal relationship from BMI to depression in all individuals. Sex-stratified MR provided no evidence of a relationship in males, but borderline evidence in females with a one SD higher BMI associating with a 1.34 (95%CI:1.00,1.86; p=0.048) higher odds of depression. Depression accentuated the genetic risk of obesity; within depressed individuals, a high genetic risk of obesity (defined as carrying 10 additional BMI-raising alleles, weighted by effect size) was associated with 5.0 kg extra weight, but only 3.3 kg of extra weight in non-depressed individuals of average height ($P_{interaction}=2x10^{-4}$). *Conclusions* Initial analyses in 120,000 UK Biobank participants provides some evidence for a bidirectional causal relationship between BMI and depression, especially in women and the role of depression in accentuating genetic risk to obesity. Further genetic analyses in 500,000 individuals will provide more insight into this complex relationship between obesity and mental health.

135

Large scale application of Mendelian randomization to electronic health records yields novel causal inferences. *J.S. Weinstock[1], E.M. Schmidt[1], L.G. Fritsche[1,2], S. Kheterpal[3], C.M. Brummett[3], G.R. Abecasis[1].* 1) Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA; 2) K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, N-7491 Trondheim, Norway; 3) Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, MI 48109, USA.

Mendelian randomization is an approach for making causal inferences between modifiable exposures and outcomes, using genetic variants as instrumental variables. The approach is less prone to confounding and reverse causation than observational epidemiological studies. Here we present the results of Mendelian randomization applied to the Michigan Genomics Initiative, a biorepository effort of the University of Michigan Health System. We included a total of 18,267 unrelated individuals of recent European ancestry that were genotyped on a customized Illumina HumanCoreExome array (~ 550,000 variants) and imputed additional variants using the Haplotype Reference Consortium panel (8 million variants with minor alleles frequency >1%). We first performed 232 genome-wide association studies (GWAS) on health related quantitative measurements, which we treat as modifiable exposures. The GWASs used a linear wald test, and included age, sex, and four principal components as covariates. We excluded quantitative measurements that displayed high degrees of zero-inflation (> 5%). We identified multiple independent genome-wide significant SNPs in 56 of these GWAS that we treat as instruments for modifiable exposures. We next applied the PheWAS R package to available ICD9 codes in the electronic health records and defined 1,448 cases control studies, each with at least 20 cases, which we used as our outcomes. We applied Mendelian randomization to the GWAS summary statistics across each pair of 56 exposures and 1,448 outcomes. To estimate the causal effects of the exposures on the outcomes we applied three methods that operate on the summary statistics: egger regression, fixed effects meta-analysis, and inverse-variance weighted average. We included egger regression to test for directional pleiotropy to ensure that the pleiotropy assumptions of Mendelian randomization were satisfied. We then calculated q-values for each causal effect estimate. Our results complement the existing body of literature on the relationship between lipids and cardiovascular outcomes, replicating known associations between low-density lipoprotein and ischemic heart disease. We also observed an effect of high-density lipoprotein on ill-defined descriptions and complications of heart disease. Results also indicated a potentially novel causal signal between lymphocyte counts and anxiety disorders. Ongoing work includes continued interrogation of the Mendelian randomization causal effect estimates.

136

Distinguishing genetic correlation from causation across 37 diseases and complex traits. *L.J. O'Connor[1], A.L. Price[1,2].* 1) Harvard T.H. Chan School of Public Health, Boston, MA; 2) Broad Institute, Cambridge, MA.

Mendelian randomization (MR) is widely used to identify causal relationships among heritable traits, but can be confounded by genetic correlations reflecting shared etiology. We propose a Latent Causal Variable (LCV) model, under which the genetic correlation between traits A and B is mediated by a latent causal variable. We define trait A as *partially genetically causal* for trait B if the latent variable has a larger effect on A than on B; by comparing the size of these effects we define the *genetic causality proportion* (gcp) of A on B, which is 0 when there is no partial causality and 1 when trait A is equal to the causal variable. We fit this model using mixed fourth moments $E[\beta_A^2 \cdot \beta_A \beta_B]$ and $E[\beta_B^2 \cdot \beta_A \beta_B]$ of marginal effect sizes, exploiting the fact that if trait A is causal for trait B then SNPs with large effects on A will have correlated effects on B, but not vice versa. We performed simulations under a wide range of genetic architectures to compare LCV to MR, MR-Egger (Bowden et al. 2015 IJE) and bidirectional MR (Pickrell et al. 2016 Nat Genet). In simulations with genetic correlation but no partial causality, LCV produced well-calibrated p-values and gcp estimates, while MR and MR-Egger produced false positives in many settings (e.g. FDR>0.8 at rg=0.25 and q=1%/1%/1% of SNPs causal for A only/ B only/ both traits) and bidirectional MR produced false positives in the presence of differential polygenicity (FDR=1 at rg=0.25, q=1%/10%/1%) or differential power (FDR=0.99 at rg=0.25, N_A/N_B=5). Applying LCV to GWAS summary statistics for 7 cardiovascular-related traits, we identified relationships consistent with the published literature but at higher resolution and power: a causal effect of LDL on cardiovascular disease (CVD) ($p<10^{-15}$ for partial causality, gcp 95% CI [0.87,1]) but not of HDL on CVD (p=0.5), and causal effects of HDL and triglycerides on hypertension (e.g. $p=3\times10^{-10}$, gcp 95% CI [0.74,1] for HDL). Expanding our analysis to 55 traits (avg N=109k), we identified strong evidence of partial causality for 34 trait pairs, all of which are biologically plausible and many of which are novel in the genetics literature. In particular, we identified a causal effect for LDL on bone mineral density ($p=3\times10^{-13}$, gcp 95% CI [0.60,1]), consistent with clinical trials of statins in osteoporosis. Our results provide etiological insights and demonstrate that it is possible to distinguish between correlation and causation using genetic data.

137

Co-occurrence network modeling reveals disease-specific configurations of microbiome community structure across 2,500 twins. *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.

The human gut microbiome is associated with a large number of complex traits and diseases. Although certain bacterial taxa have been identified that correlate with disease, it is commonly thought that alterations in microbial community dynamics are indicative of a disease state. Therefore, investigating the structure of the microbiota using systems biology approaches has the potential to yield insight into the rules that govern community dynamics in disease, or dysbiosis. Although methodology for generating microbial co-occurrence networks exists, it has largely been applied to data sets containing healthy individuals or in studies examining one particular disease of interest. Whether there are consistent alterations in co-occurrence across diseases remains an open question. Here, we address this gap by examining how microbial community structure shifts between healthy and disease states for ~180 immune-related diseases and quantitative traits in a cohort of 2,500 twins from the United Kingdom for which 16S rRNA gene sequencing data is available. First, we identified trait-associated taxa using linear models that took into account relatedness among twin pairs. Subsequently, we built co-occurrence networks using all individuals in the dataset, regardless of disease status. Although significant for certain traits (permutation $P < 0.05$), node-level properties, such as degree, betweenness, and constraint, were not consistently different between disease-associated and non-disease associated taxa. Next, we built networks for healthy and diseased individuals separately for ten of the immune-related phenotypes. General network properties, such as the number of edges, transitivity, and modularity, again did not differ consistently between healthy and diseased networks. These results highlight that microbiome community structure is altered in a disease-specific manner. In addition to disease, we identified other factors contributing to microbial co-occurrence. First, bacterial families from the same phylum tend to co-occur more often than expected by chance (permutation $P < 0.0001$), pointing to the role of phylogeny in determining co-occurrence patterns. Additionally, host genetics influences community structure, as the co-occurrences between certain taxa are a heritable property of the human host (Falconer $h^2 > 0.2$). Using these data, we have conducted one of the first large scale comparisons of microbiome community dynamics across health and disease.

138

Biome-explainability: Quantifying microbiome-phenotype associations while accounting for host genetics. *O. Weissbrod[1,2], D. Rothschild[1,2], E. Barkan[1,2], T. Korem[1,2], D. Zeevi[1,2], P. Costea[1,2], A. Godneva[1,2], I. Kalka[1,2], N. Bar[1,2], N. Zmora[3,4,5], D. Israeli[6], N. Kosower[1,2], G. Malka[1,2], B.C. Wolf[1,2], T. Avnit-Sagi[1,2], M. Lotan-Pompan[1,2], A. Weinberger[1,2], Z. Halpern[5,7], S. Carmi[8], E. Elinav[3], E. Segal[1,2].* 1) Computer Science Department, Weizmann Institute of Science, Rehovot, Israel; 2) Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel; 3) Immunology Department, Weizmann Institute of Science, Rehovot, Israel; 4) Internal Medicine Department, Tel Aviv Sourasky Medical Center, Tel Aviv ,Israel; 5) Research Center for Digestive Tract and Liver Diseases, Tel Aviv Sourasky Medical Center, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel; 6) Day Care Unit and the Laboratory of Imaging and Brain Stimulation, Kfar Shaul Hospital, Jerusalem Center for Mental Health, Jerusalem, Israel; 7) Digestive Center, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel; 8) Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel.

Heritability estimation is a common technique for quantifying the fraction of phenotypic variance explained by genetic factors. In recent years, the gut microbiome gained considerable attention as having fundamental roles in multiple aspects of human physiology and health, and is often referred to as our second genome. However, the fraction of phenotypic variance explained by the microbiome as compared to host genetics has not been evaluated to date. Here we define the term biome-explainability, which, analogously to genetic heritability, quantifies the fraction of phenotypic variance explained by the microbiome, while accounting for host genetics. Using a cohort of 696 individuals for whom we obtained information on genotypes, gut microbiome composition, numerous anthropometric and blood phenotypes, and dietary habits, we (1) demonstrate that biome-explainability can be estimated far more accurately than host genetics heritability; (2) demonstrate that accurate bacterial kinship can be estimated via bacterial gene abundance; and (3) find significant biome-explainability levels of 16-37% for body mass index, fasting glucose, glycemic status, high density lipoprotein cholesterol, waist circumference, waist-hip-ratio, and lactose consumption. As another comparison between host genetics and microbiome, we demonstrate that using both microbiome and host genetics substantially improves human phenotype prediction compared to models utilizing only one of these data sources. These results suggest that the microbiome should be routinely considered in studies aimed to explain the variance of human phenotypes.

139

Microbiome and host genetics in inflammatory bowel disease and irritable bowel syndrome. *A. Zhernakova[1], A. Vich Vila[1,2], F. Imhann[1,2], V. Collij[1,2], S. Jankipersadsing[1], Z. Mujagic[3], T. Gurry[4], A. Kurilshikov[1], MJ. Bonder[1,5], X. Jiang[4], L. Franke[1], G. Dijkstra[2], E.A.M. Festen[2], R. Xavier[6], E.J. Alm[4], D. Jonkers[3], J. Fu[1], R.K. Weersma[2], C. Wijmenga[1].* 1) Departmenet of Genetics, University Medical Center Groningen, Groningen, Netherlands; 2) Department of Gastroenterology and Hepatology, University Medical Center Groningen, Groningen, Netherlands; 3) Division Gastroenterology-Hepatology, NUTRIM School for Nutrition, and Translational Research in Metabolism, Maastricht University Medical Center, Maastricht, The Netherlands; 4) Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 5) European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; 6) Massachusetts General Hospital, Boston, USA.

Irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD; including Crohn's disease (CD) and Ulcerative colitis (UC)) are two of the most common gastrointestinal (GI) disorders, affecting respectively 7-21% and 0.3%-0.5% of the global population. Next to host genetics, microbiome plays an important role in disease pathology. Here we present the largest gut microbiome case-control analysis in both IBD and IBS to date, using metagenomic shotgun sequences of stool samples from 1025 Healthy Controls 355 IBD and 412 IBS patients. For all patients and controls (n=1792) information on diet, medication and gut diseases and complaints were extensively collected and included in the analysis. Taxonomy was determined for bacteria, viruses and micro-eukaryotes. Bacterial pathways were determined using HUMAnN2. In addition, bacterial strain diversity and growth rates and abundance of antibiotic resistance genes and virulence factors were inferred from the sequencing data. In the case-control analyses, after correcting for 25 previously identified microbiome-modifying factors including diet and medication, we observed 157 differentially abundant species associated with CD, 87 species associated with UC and 125 species for IBS. IBS and IBD patients showed strong significant differences in abundance of virulence factors, antibiotic resistance genes and bacterial growth rates. Prediction models for differentiating between IBS and IBD based on microbiome data show a best predictive value up to 94.6%, which is significantly higher than prediction based on a combination of genetics, clinical and inflammatory markers (including fecal calprotectin levels, which was measured in all participants, and is commonly used as IBD marker). In addition, we identified alterations of the gut microbiota of healthy individuals with a high host genetic risk for IBD. In this study, we present a high-resolution characterization of changes in the microbiome, and its functional implications, in patients with IBD or IBS, which can be used in the disease prediction.

140

Systematic analysis of association of blood circulating proteins with genome and microbiome. *A. Kurilshikov[1], D. Zhernakova[1,2], T. Le[1], B. Atanasovska[1,3], M.J. Bonder[1], S. Sanna[1], R. Boer[4], F. Kuipers[3], L. Franke[1], C. Wijmenga[1], A. Zhernakova[1], J. Fu[1,3].* 1) Department of Genetics, UMCG, Groningen, Netherlands; 2) Theodosius Dobzhansky Center for Genome Bioinformatics, SpBU, St. Petersburg, Russia; 3) Department of Pediatrics, UMCG, Groningen, Netherlands; 4) Department of Cardiology, UMCG, Groningen, Netherlands.

Blood circulating proteins play an important role in human metabolism and are often measured as biomarkers for metabolic disorders and diseases, including cancer, immune diseases and cardiovascular diseases (CVD). A better understanding of the factors that influence the variability of such proteins can pinpoint to causal components of associated diseases. We therefore investigated the variation of 92 CVD-related proteins in 1294 individuals from the LifeLines-DEEP population cohort and linked this variation to their genome and gut microbiome profiles. To estimate the effect of the genome, we performed, on each protein, a genome-wide association study on 8 million imputed SNPs. To explore the impact of microbiome, we correlated protein levels with 340 taxonomic and 702 functional categories (microbial pathways) identified in the same individuals through metagenomic sequencing. On FDR of 5%, we observed 73 out of 92 proteins to be under control of host genetics, by both in cis- and in trans- protein QTLs (pQTLs). Interestingly, many of the observed cis- and trans- pQTLs do not have an effect at the transcript level (defined as showing an eQTL effect in whole-blood); we hypothesized that these proteins are expressed in other tissues and then secreted to the blood. Using the same significance threshold, we observed that 41 proteins were associated with one or more microbiome features, with 32 being under control of both microbiome and host genetics. We identified a number of novel associations with strong biological implications, including the information on known disease markers. For the proteins associated with both genetics and microbial features, we performed interaction analysis to explore if genetic background can shape the interaction between protein and microbiome. We observed 2 cases of such interaction. Although the direct biological meaning of this interaction is not clear yet, these examples show that genetics may shape not only the baseline level of circulating proteins but also the efficiency of the host response to external factors. Our study shows that genetics and microbiome shape a significant proportion of variation of disease-relevant proteins circulating in the blood. In addition to independent linear relationships, there is an evidence of complex interactions between SNPs, proteins and gut bacteria, and this complexity has to be taken into account when interpreting circulating proteins as disease markers.

141

Population structure of the human gut microbiome across ethnically diverse sub-Saharan Africans. *M.A. Rubel[2], M.E.B. Hansen[1], A.G. Bailey[3], J.R. Dave[5], A. Ranciaro[1], S.R. Thompson[1], M. Campbell[1], W.R. Beggs[1], S.W. Mpoloka[6], G. Mokone[7], M.M.M. Bolaane[8], T. Nyambo[9], F.D. Bushman[3], S.A. Tishkoff[1-4].* 1) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 2) Department of Anthropology, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA; 3) Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA,; 4) Department of Biology, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA; 5) The Commonwealth Medical College, Scranton, PA 18509; 6) Department of Biological Sciences, University of Botswana, Gaborone, Botswana; 7) Department of Biomedical Sciences, University of Botswana School of Medicine, Gaborone, Botswana; 8) History Department, University of Botswana, Gaborone, Botswana; 9) Department of Biochemistry, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania.

Global populations with different degrees of industrialization have shown marked differences in their gut microbial (GM) composition. Notably, "westernized" microbiomes from individuals on processed diets largely derived from industrial agropastoralism show decreased diversity as compared to rural, non-industrialized groups practicing traditional forms of subsistence. Prior research has tended to focus on comparing individual populations from geographically disparate regions, which ignores the complex relationships with neighboring groups. Here, we report the GM bacterial compositions of seven ethnically diverse rural populations in Tanzania and Botswana (N=114) and one western cohort from Philadelphia, U.S.A. (N=12), using 16S ribosomal sequence (16s rRNA V1-V2) classification. The seven African populations include two current or recent hunter-gatherer populations, three small-scale agropastoralist communities, and two current or recent pastoralist groups. A subset of Africans were genotypes on a 5M Omni SNP array, and this was used to test the relative impact of host genetic relatedness on compositional fitness. We find that GM diversity distances are strongly associated with host genetic similarity after controlling for geographic distance. Partitioning the bacterial abundance variation using linear mixed models indicates that several bacteria, most notably *Ruminococcus,* associate significantly with host relatedness. Some Botswanan Bantu agropastoralist GMs cluster close to U.S. samples, which could be the result of factors such as increased urbanization and industrialization. The mean bacterial diversity per host decreases from Tanzania to Botswana to the U.S., while the mean bacterial phylogenetic distances between individuals in the same population increases from Tanzania to Botswana to U.S. Although the GMs of all populations have a similar rank order of most abundant bacteria, with Prevotellaceae and Ruminococcaceae as the two most abundant known taxonomic families, there is diversity in their relative abundances in the presence/absence of lowly abundant bacteria. There is no clear signal of GM compositional "types" associated with subsistence lifestyle, and the evidence for unique single taxa that can distinguish pastoralism or hunter-gatherer lifestyle is marginal. Instead, we find evidence that the range of GM compositions is influenced by geographic region and host genetics.

142

Inter-species variation in the gut microbiota controls host gene regulation in primates. *R. Blekhman[1], A.L. Richards[2], A. Muehlbauer[1], A. Alazizi[2], M. Burns[1], A. Gomez[3], J. Clayton[4], K. Petrzelkova[5], C. Cascardo[2], R. Pique-Regi[2], F. Luca[2].* 1) Genetics, Cell Biology, and Development, University of Minnesota; 2) Center for Molecular Medicine and Genetics, Wayne State University; 3) J. Craig Venter Institute; 4) BioTechnology Institute, University of Minnesota; 5) Institute of Vertebrate Biology, Czech Academy of Sciences.

The gut microbiota can regulate and train host immune response, perform important metabolic functions, produce nutrients, and protect against pathogen infection. Exciting new research has catalogued extreme variation in the composition of gut microbial communities across primate species. However, we know very little about how this inter-species variation affects host health. An important mechanism by which the microbiome can affect host physiology is by altering gene expression in proximal colonic epithelial cells. Thus, it has been hypothesized that variation in the microbiome can control species-specific gene expression and potentially affect human disease. Here, we used a novel experimental system based on colonic epithelial cells co-cultured with live microbiomes extracted from four primate hosts (human, chimpanzee, gorilla, and orangutan), with 4-8 individuals from each species. This allowed us to dynamically profile host gene expression changes (via RNA-seq) that are directly modulated by the microbiome. We found a conserved signature whereby 958 genes consistently respond to microbiomes from all four species. These genes are involved in cell adhesion, metabolic functions (such as cholesterol biosynthesis), and immune pathways, including interleukin signaling. In addition, we identified species-specific host gene response, with ~2500 genes that respond to microbiomes from one primate species and not the others. Host genes that respond specifically to human microbiomes are significantly enriched with genes that have been previously identified in GWAS studies as associated with microbiome-related health conditions, such as HDL cholesterol, obesity-related traits, celiac disease, and inflammatory bowel disease. Our study provides the first evidence that symbiosis with gut microbial communities affects species-specific gene regulation in primates, and demonstrates how human-specific microbiomes may influence host health. .

143

Understanding the genetic architecture of migraine using a collection of 1,214 families from Finland. *P. Gormley[1,2,3], M.I. Kurki[1,2,3], M.E. Hiekkala[4], K. Veerapen[1,2,3], P. Häppölä[5], A. Mitchell[6], D. Lal[1,2,3,7], P. Palta[5], I. Surakka[5], M.A. Kaunisto[5], E. Hämäläinen[5], P. Jousilahti[8], V. Salomaa[8], V. Artto[9], M. Färkkilä[9], H. Runz[6], M.J. Daly[1,2,3], B.M. Neale[1,2,3], S. Ripatti[5], M. Kallela[9], M. Wessman[4,5], A. Palotie[1,2,3,5].* 1) Massachusetts General Hospital, Boston, MA, USA; 2) Harvard Medical School, Boston, MA, USA; 3) Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4) Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland; 5) Institute for Molecular Medicine Finland FIMM, HiLIFE, University of Helsinki, Helsinki, Finland; 6) Merck Research Laboratories, Merck and Co., Inc., Boston, MA, USA; 7) Cologne Center for Genomics, University of Cologne, Cologne, Germany; 8) National Institute for Health and Welfare, Helsinki, Finland; 9) Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland.

It has long been observed that complex traits such as migraine often aggregate in families (PMID:9029065), but the underlying genetics are not fully understood. Two competing hypotheses emphasize either rare or common genetic variation. For example, familial aggregation could be explained by highly penetrant rare variants that segregate according to Mendelian inheritance or rather by the polygenic burden of many common variants of small effect. We investigate this in a collection of 1,214 migraine families from Finland. Individual diagnoses in this collection include migraine without aura (MO, n=2,357), migraine with aura (MA, n=2,420), hemiplegic migraine (HM, n=540), and their unaffected relatives (n=3,002). For comparison we used 14,764 population controls from FINRISK (PMID:19959603). All individuals were genotyped on the Illumina CoreExome or PsychArray and imputed to a Finnish reference panel of 6,962 haplotypes. Polygenic risk scores (PRS), representing the common variant burden in each individual, were calculated using weights from the latest genome-wide association study of migraine (PMID:27322543). To account for family structure in our analyses we used a mixed-model approach, adjusting for the genetic relationship matrix as a random effect. We found that polygenic burden significantly contributed to all migraine subtypes, particularly HM which had significantly higher burden than MO (P=4.3×10^{-5}, β=0.15, SE=0.04) but showed no difference compared to MA (P=0.11, β=0.06, SE=0.04). MA was also significantly higher compared to MO (P=4.6×10^{-5}, β=0.09, SE=0.02). Further, using FINRISK as a reference, HM cases were 3.8 times more likely to be in the upper PRS quartile than the lower quartile (P=2.5×10^{-17}, OR=3.84, 95CIs=3.52-4.15), compared with 3.0 times more likely for MA (P=9.0×10^{-35}, OR=3.02, 95CIs=2.85-3.20), and 2.2 times more likely for MO (P=1.4×10^{-19}, OR=2.2, 95CIs=2.03-2.37). Interestingly, we found that higher polygenic burden corresponded to earlier age of headache onset (P=8.3×10^{-4}, OR=0.84, 95CIs=0.76-0.93). Finally, although rare variants have been suggested as the primary cause for HM, we only found 18 out of 540 individuals (3.33%) with a causal mutation in a known gene. In summary, our results demonstrate a substantial contribution of common polygenic variation to familial aggregation in migraine. The findings also suggest that individuals with aura symptoms (either typical sensory or rare motor aura) tend to have higher polygenic burden.

144

A meta-analysis of genome-wide association studies identifies novel Parkinson's disease risk loci. *D. Chang[1], M. Nalls[2,3], I. Hallgrímsóttir[4], J. Hunkapiller[1], M. van der Brug[1], F. Cai[1], G. Kerchner[1], G. Ayalon[1], B. Bingol[1], M. Sheng[1], D. Hinds[1], T. Behrens[1], A. Singleton[2], T. Bhangale[1], R. Graham[1], International Parkinson□s Disease Genomics Consortium, UKDPDSBB, 23andMe Research Team.* 1) Genentech, South San Francisco, CA; 2) Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD; 3) Data Tecnica International, Glen Echo, MD; 4) 23andMe Inc., Mountain View, CA.

Genome-wide association studies (GWAS) have identified >24 Parkinson's disease (PD) risk loci. However, the narrow-sense heritability explained by these loci is low (0.227), suggesting more genetic loci remain to be discovered. In order to uncover additional missing heritability defined common variants we first carried out a new PD GWAS of 6,476 PD cases and 302,042 controls (PDWBS). Applying LD score regression to PDWBS, we found that regions contributing to PD heritability were significantly enriched for H3K27 acetylation as well as histone marks specific to central nervous system, adrenal and pancreas cell-types. We next carried out a meta-analysis between PDWBS and publicly available data from PDGene, a recent study of over 13,000 cases and 95,000 controls at 9,830 overlapping variants. Thirty-five loci reached a significance level of P<1x10^{-6}, and were carried forward into the replication phase with an additional 5,851 cases and 5,866 controls. In a final meta-analysis between all three studies, 17 novel risk loci reach a genome-wide significant threshold (P<5x10^{-8}). Using a neuro-centric strategy, we assigned candidate genes to each of the 17 novel as well as 24 previously reported loci. We found either cis-eQTLs or protein altering variants in linkage disequilibrium with the index variant for 30 of the 41 PD risk loci. Of the candidate genes identified for novel risk loci, five genes map to the either the lysosomal, autophagy or mitochondrial pathways, which are known to play a role in PD pathology.

145

Parkinson's disease gene identification using differential gene expression analysis of iPSC generated neural stem cells. *S. Kumar[1], J.E. Curran[1], D.M. Lehman[2], R. Duggirala[1], D. Glahn[3,4], J. Blangero[1].* 1) South Texas Diabetes and Obesity Inst., School of Medicine, University of Texas Rio Grande Valley, Edinburg - Brownsville, TX, USA; 2) Department of Medicine, University of Texas Health Science Center, San Antonio, TX, USA; 3) Olin Neuropsychiatry Research Center, The Institute of Living, Hartford, CT, USA; 4) Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA.

Parkinson's disease (PD) is a movement disorder associated with the degeneration of nigral dopaminergic (DA) neurons. The past two decades of genetic research have identified several monogenetic forms of the disorder and a number of genetic risk loci that may increase the risk of the disease. The genetics of monogenic forms are well established, however they collectively account for only 30% of the familial and 3% - 5% of the sporadic PD cases. Unfortunately, in most cases (particularly in late-onset PD), the etiology of PD is multifactorial, plausibly resulting from a complex interplay of gene-gene and gene-environment interactions. To identify gene(s) influencing late onset PD, we performed whole genome differential gene expression (DGE) analysis of iPSC generated neural stem cells (NSCs) of 2 late onset PD cases and their 10 normal blood relatives. All 12 subjects have participated in our Genetics of Brain Structure Study (originally recruited into our San Antonio Mexican American Family Studies). Using their previously established lymphoblastoid cell lines (LCLs), integration free iPSCs were reprogrammed and then differentiated into NSCs. Both generated iPSC and NSC lines were confirmed by immunocytochemistry and DGE analysis. Total RNA from the generated 12 NSC lines underwent deep mRNA enriched sequencing yielding on average 30 million reads per samples. Known mRNAs having a normalized read count *(NRC)* ≥ *20* in one or more samples and after filtering for sex specific differences, were analyzed for DGE. Following the criteria moderated *t* statistics *p* value ≤ *0.05* and fold change-absolute ≥ *4.0*, a total of 89 genes were significantly differentially expressed (DE). The functional enrichment analysis using IPA shows that 14 of the 23 genes, which were significantly upregulated in PD cases, were enriched in nervous system disorders, and 6 of these genes (*HES5, LGR5, NRG1, PRMT8, SCUBE2, SH3GL2*) are known to be associated with neurological movement disorders. The elevated levels of *SH3GL2* and *NRG1* have been reported in PD cases. Interestingly 4 genes (*GRIK3, NRG1, PCDH8, PHOX2B*) known to be associated with Schizophrenia were also significantly upregulated in our PD cases. Of the 66 genes which were significantly down-regulated in PD cases, 28 were enriched in skeletal and muscular disorder and 7 were enriched in nervous system development and function. These results highlight the utility of iPSC derived NSCs for the identification of PD risk genes.

146

Examining the genetic architecture of the human cortex: Results from the ENIGMA consortium GWAS meta-analyses of cortical thickness and surface area. *S. Medland on behalf of the The Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) Consortium.* Psychiatric Genetics, QIMR Berghofer, Brisbane, QLD, Australia.

At the individual level, there is substantial variation in cortical thickness and surface area which has been implicated in a wide range of psychiatric and neurological traits. While cortical thickness and surface area are both strongly heritable, the two processes are genetically independent and there is little known about the loci influencing these morphological characteristics. We will present results from GWAS meta-analyses of the thickness and surface area of cortical regions of interest derived from magnetic resonance imaging (MRI) scans from a meta-analysis of ~28,000 individuals (comprising a first round of ~18,000 individuals and a second round of ~10,000). Across cohorts, structural T1-weighted MRI brain scans were analysed locally using harmonized analysis and quality-control protocols (http://enigma.ini.usc.edu/protocols). Cortical parcellations were performed with freely available and validated segmentation software. Cortical measurements were averaged across the hemispheres resulting in the average thickness and surface area of 34 gray matter regions. We also analyzed two summary measures, average cortical thickness across regions and total surface area. Corrections for the summary measures were included in the regional measures to account for the omnibus effects of brain size. Results from the meta-analyses demonstrate that common variation substantially influences the architecture of the human cortex and supports the findings from twin studies that genetic influences on thickness and surface area are largely orthogonal. Unlike our previous work on the subcortical structures, we find the effects of genetic variants often impact a number of neighboring regions, reflecting regions with related functions within the cortex. For example, variants in 15q14 region significantly influenced surface area in the precentral region ($p = 5.5 \times 10^{-7^4}$) with signal also influencing surrounding motor and somatosensory areas. Results of multivariate analyses support these patterns of genetic variation across regions. Results are interpreted with reference to known biological and developmental processes and genetic covariation with psychiatric, neurological and cognitive phenotypes.

147

Low frequency coding variation in *CYP2R1* has large effects on Vitamin D level and risk of multiple sclerosis. *D. Manousaki[1,2], T. Dudding[3], S. Haworth[3], Y. Hsu[4,5,6], C. Liu[7], C. Medina-Gomez[8,9,10], T. Voortman[9,10], N. van der Velde[8,11], H. Melhus[12], C. Robinson-Cohen[13], D.L. Cousminer[14,15], M. Nethander[16,17], L. Vandenput[16], R. Noordam[18], V. Forgetta[1,2], L. Lind[13], E.S. Orwoll[19,20], D.O. Mook-Kanamori[21,22], K. Michaelsson[23], B. Kestenbaum[13], C. Ohlsson[16], D. Mellstrom[16,24], L.C.P.G. de Groot[25], S.F.A. Grant[14,26], D.P. Kiel[4,5,6,27], M.C. Zillikens[8], F. Rivadeneira[8,9,10], S. Sawcer[28], N.J. Timpson[3], J.B Richards[1,2,29,30].* 1) Department of Human Genetics, McGill University, Montreal, Canada; 2) Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montreal, Canada; 3) Medical Research Council Integrative Epidemiology Unit (IEU) at the University of Bristol, Bristol, UK; 4) Institute for Aging Research, Hebrew SeniorLife, Boston, USA; 5) Harvard Medical School, Boston, USA; 6) Broad Institute of MIT and Harvard, Boston, USA; 7) Department of Biostatistics, Boston University School of Public Health, Boston, USA; 8) Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The; 9) The Generation R Study Group, Erasmus Medical Center, Rotterdam, The Netherlands; 10) Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands; 11) Department of Internal Medicine, Section of Geriatrics, Academic Medical Center, Amsterdam, The Netherlands; 12) Department of medical sciences, Uppsala university, Uppsala, Sweden; 13) Kidney Research Institute, Division of Nephrology, University of Washington, Seattle, WA, USA; 14) Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; 15) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; 16) Centre for Bone and Arthritis Research, Department of Internal Medicine and Clinical Nutrition, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; 17) Bioinformatics Core Facility, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; 18) Department of Internal Medicine, section of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, the Netherlands; 19) Bone and Mineral Unit, Oregon Health & Science University, Portland, USA; 20) Department of Medicine, Oregon Health & Science University, Portland, USA; 21) Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, the Netherlands; 22) Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, the Netherlands; 23) Department of surgical sciences, Uppsala university, Uppsala, Sweden; 24) Geriatric Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; 25) Department of Human Nutrition, Wageningen University, Wageningen, The Netherlands; 26) Division of Endocrinology, Children's Hospital of Philadelphia, Philadelphia, PA, USA; 27) Beth Israel Deaconess Medical Center, Boston, USA; 28) University of Cambridge, Department of Clinical Neurosciences, Box 165, Cambridge Biomedical Campus, Hills Road, Cambridge, UK; 29) Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom; 30) Department of Medicine, McGill University, Montreal, QC, Canada.

Introduction: Vitamin D insufficiency is common, correctable and influenced by genetic factors, and it has been associated to risk of several diseases. We sought to identify low-frequency genetic variants that strongly increased the risk of vitamin D insufficiency and tested their effect on risk of multiple sclerosis, a disease influenced by low vitamin D concentrations. **Methods/Results**: We used whole-genome sequencing data from 2,619 individuals through the UK10K program and deep imputation data from 39,655 genome-wide genotyped individuals. Meta-analysis of the summary statistics from 19 cohorts identified a low-frequency synonymous coding variant (rs117913124[A], minor allele frequency=2.5%) in the *CYP2R1* gene which conferred a large effect on 25-hydroxyvitamin D (25OHD) levels (-0.43 standard deviations of standardized natural log-transformed 25OHD, per A allele, P-value = 1.5×10^{-88}). The effect on 25OHD was four-times larger and independent of the effect of a previously described common variant near *CYP2R1*. By analyzing 8,711 individuals we showed that heterozygote carriers of this low-frequency variant have an increased risk of vitamin D insufficiency (OR=2.2, 95% CI 1.78-2.78, $P=1.26 \times 10^{-12}$). Individuals carrying one copy of this variant had also an increased odds of multiple sclerosis (OR=1.4, 95%CI 1.19-1.64, $P=2.63 \times 10^{-5}$) in a sample of 5,927 cases and 5,599 controls. **Conclusions:** We describe a novel low-frequency coding variant in the *CYP2R1* gene, which exerts the largest effect upon 25OHD levels identified to date in the general European population. Since *CYP2R1* is known to encode a critical enzyme in the production of the active form of vitamin D, these findings implicate vitamin D in the etiology of multiple sclerosis.

148

Genome-wide association study: Pandemrix-induced narcolepsy in Sweden. *M. Wadelius[1], N. Eriksson[1,2], H. Smedje[3], Q-Y. Yue[4], P.K.E. Magusson[5], P. Hallberg[1] on behalf of Swedegene.* 1) Department of Medical Sciences & Science for Life Laboratory, Uppsala University, Uppsala, Sweden; 2) Uppsala Clinical Research Center, Uppsala, Sweden; 3) Division of Child and Adolescent Psychiatry, Karolinska Institutet, Stockholm, Sweden; 4) Medical Products Agency, Uppsala, Sweden; 5) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

Background: Narcolepsy is an auto-immune disease characterized by an inability to control sleep and wakefulness that may lead to learning difficulties, depression and metabolic disruption. The number of children and youngsters diagnosed with narcolepsy rose sharply in Sweden following vaccination with Pandemrix against H1N1 influenza in 2009–2010. The most frequent form of narcolepsy, type I, is caused by loss of hypocretin (orexin) neurons. Narcolepsy type I is associated with *HLA-DQB1*0602*, but only 0.02 % of carriers develop narcolepsy, and it is probable that a combination of other genetic and external risk factors are required. **Method**: 48 cases of Pandemrix-induced narcolepsy collected by the Swedish adverse drug reaction biobank Swedegene (www.swedegene.se) were genotyped on Illumina HumanOmni2.5. They were compared with 4891 controls from TwinGene genotyped on Illumina HumanOmniExpress 700K. The merged genotyped set contained 600K single nucleotide polymorphisms (SNPs). After phasing and imputation, the dataset contained 8.6 million SNPs. We corrected for principal components 1-4. The genome-wide significance p-value threshold was set to $p<5x10^{-8}$. **Results:** Pandemrix-induced narcolepsy was significantly associated with >100 SNPs in the HLA region on chromosome 6. The top hit in the HLA region had an odds ratio (OR) of 7.0 [95% confidence interval (CI) 4.6, 10.7], $p=5.4x10^{-19}$). After correction for the top hit, three genes on chromosomes 5, 8 and 16 were significantly associated on a genome-wide level. The associated genes were *GDNF-AS1* (OR=7.4 [95% CI 3.6, 15.0], $p=3.7x10^{-8}$), *MYOM2* (OR=9.3 [95% CI 4.2, 20.6], $p=4.3x10^{-8}$), and *ABCC1* (OR=9.3 [95% CI 4.2, 20.6], $p=4.9x10^{-8}$). **Conclusion:** After correction for HLA, narcolepsy induced by Pandemrix was associated with the non-coding RNA gene *GDNF-AS1,* which is located head to head with *GDNF* on the opposite chromosomal strand (1). *GDNF* encodes an essential neurotrophic factor that supports neuronal survival. *GDNF* is potentially regulated by *GDNF-AS1*, and dysregulation of *GDNF* has been shown in Alzheimer's disease (1,2). We also found an association with the blood-brain-barrier transporter *ABCC1,* which has been implicated in Alzheimer's disease (3). These findings may increase the understanding of disease mechanisms underlying narcolepsy. **References** 1. Modarresi F, et al. Nat Biotechnol 2012;30(5):453-9. 2. Airavaara et al. J Biol Chem 2011;286(52):45093-102. 3. Pahnke J, et al. Neurobiol Dis 2014;72PtA:54-60.

149

High-resolution dissection of regulatory function for millions of predicted enhancers in human. *M. Kellis[1,2], X. Wang[1,2,3], L. He[1,2], S. Goggin[2], A. Saadat[2], M. Claussnitzer[2,4].* 1) MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) MIT Department of BIology, Cambridge, MA; 4) Harvard Medical School, Boston MA.

Genome-wide profiling of epigenomic marks allows rapid prediction of regulatory regions, including promoters and enhancers. However, experimental validation that predicted regions indeed drive gene expression, and high-resolution dissection of driver sequence elements within them, have remained unfeasible in high throughput. [alt] Here, we describe ATAC-STARR, a high-throughput episomal assay for enhancer activity across millions of DNA fragments extracted from open-chromatin regions. We extracted transposase-accessible regions, size-selected fragments 50-1000 nucleotides long, cloned them into a plasmid library downstream of a common reporter gene, and used short-read sequencing to quantify reporter expression for each fragment, using the fragment itself as the barcode. We tested libraries with up to 67 million fragments, a 300-fold increase compared to array-based massively-parallel assays. We also tested fragment lengths up to 1.5kb, a 6-fold increase compared to array-based assays, and found that longer fragments tended to show higher activity. Applying our method to the GM12878 lymphoblastoid cell line revealed hundreds of significantly up-regulated fragments, which are enriched for H3K27ac, for other active histone marks, for known regulatory motifs, and for binding of immune transcription factors. We also developed a new algorithm, Sharpr2, for deconvolving fragment-level activity measurements into segment-level activity measurements, enabling nucleotide-level resolution of activity patterns, and pinpointing individual driver regulatory motifs underlying enhancer activity patterns, and individual nucleotide variants that underlie genetic association hits with complex traits. Our results suggest that ATAC-STARR can be a general strategy for experimentally assaying and dissecting the DNA regulatory landscape in the context of human biology and disease.

150

Measuring enhancer activity at the human genome scale: Comprehensive and quantitative assessment of steady-state and induced regulatory activity following glucocorticoid stimulation. *G.D. Johnson[1,2], C.M. Vockley[1,2], L.C. Bartelt[1,2], N. Clark[2,3], S.M. Leichter[1,2], G.E. Crawford[2,3], T.E. Reddy[1,2].* 1) Department of Biostatistics and Bioinformatics, Duke University; 2) Center for Genomic and Computational Biology, Duke University; 3) Division of Medical Genetics, Department of Pediatrics, Duke University.

Glucocorticoids (GCs) are one of the most widely used pharmaceuticals due to their potent suppression of inflammation and immunity. The GC response is also a paradigmatic model for studying basic mechanisms of gene regulation that contribute to many of the genetic associations with human traits, diseases, and drug responses. To comprehensively quantify the regulatory effects of GCs, we completed high-throughput reporter assays with >10^8 unique fragments that together cover the human genome at >50x. With those assays, we quantified steady-state and induced regulatory activity in A549 cells over five durations of GC exposure (0, 1, 4, 8, & 12 hours). To demonstrate that our measurements of regulatory activity were predictive of changes in gene expression and correlated with genomic features including modified histones and transcription factor (TF) binding across the time course, we integrated our results with over 1,000 previously completed assays of the same model system. Glucocorticoid receptor (GR) binding was often coincident but not strictly requisite at GC-inducible enhancers, demonstrating the potential for non-GR-bound GC-responsive enhancers. The GC-responsive sites that lack appreciable GR binding had delayed activation and were bound by other TFs throughout the time course. The presence of additional putative enhancer-bound TFs was supported by motif analysis and their combinatorial effect on cis-regulatory function assessed. The functional relevance of enhancers identified in our high-throughput episomal reporter assay was supported by the physical interaction of their corresponding genomic regions and target genes as demonstrated by Hi-C. Together, these results represent the first comprehensive and agnostic quantitative map of enhancer activity at the human genome scale.

151

An ultra-high resolution Capture-C promoter 'interactome' implicates causal genes at SLE GWAS loci. *M.E. Johnson[1], E. Manduchi[1,2], C. Le Coz[3], M.E. Leonard[1], S. Lu[1], K.M. Hodge[1], N.D. Romberg[3,4], A. Chesi[1], A.D. Wells[5], S.F.A. Grant[1,4,6,7,8].* 1) Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA; 2) Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA; 3) Division of Allergy Immunology, Children's Hospital of Philadelphia, Philadelphia, PA; 4) Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 5) Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 6) Division of Diabetes and Endocrinology, The Children's Hospital of Philadelphia, Philadelphia, PA; 7) Institute of Diabetes, Obesity and Metabolism, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 8) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

Genome Wide Association Studies (GWAS) have been successful in yielding >60 loci for Systemic Lupus Erythematosus (SLE). However, it is known that GWAS just reports genomic signals and not necessarily the precise localization of culprit genes, with eQTL efforts only able to infer causality to a minority of such loci. Thus, we sought to carry out physical and direct 'variant to gene mapping'. Chromatin conformation capture-based techniques that detect contacts between distant regions of the genome offer a powerful opportunity to understand GWAS signals that principally reside in non-coding regions, and thus likely acting as regulatory elements for neighboring genes. To move beyond analyzing one locus at a time and to improve on the low resolution of available Hi-C data, we developed a massively parallel, ultra-high resolution Capture-C based method to simultaneously characterize the genome-wide interactions of all human promoters in any cell type. We applied this method to study the promoter 'interactome' of primary human T Follicular Helper (TFH) cells from tonsils of healthy volunteers, a model relevant to SLE as TFH operate upstream of the activation of pathogenic autoantibody-producing B cells during the disease. We designed a custom Agilent SureSelect library targeting both ends of *DpnII* restriction fragments that overlap promoters of protein-coding, noncoding, antisense, snRNA, miRNA, snoRNA and lincRNA transcripts, totaling 36,691 RNA baited fragments. Each library was sequenced on multiple lanes of an Illumina HiSeq 4000 and then significant interactions were determined. We also generated ATAC-seq open chromatin maps from the same TFH samples to determine informative proxy SNPs for each of the 63 SLE GWAS loci (*Nat Genet* 48, 940-946, 2016). By intersecting our sub-1kb promoter 'interactome' data with SNPs from 33 candidate loci provided by the ATAC-seq experiments, we observed consistent contacts for at least 20 loci. Some 'nearest' genes to the sentinel SNP were supported e.g. *ARID5B* and *IKZF3*, while at other loci more distant genes were implicated e.g. *LCLAT1* at the '*LBH*' locus and the master TFH transcription factor *BCL6* at the '*LPP-TPRG1*' locus. In conclusion, we observed consistent contacts to at least 30% of SLE GWAS loci using the highest resolution promoter 'interactome' to date in a single, disease-relevant cell type. Only by establishing which genes such loci regulate in the correct cellular context can one truly translate GWAS findings.

152

High throughput functional prioritization of candidate genes from large-scale sequencing and GWAS studies. *P. Heutink, J. Taeger, E. Lara Flores, M. Bedi, N. Alves Fernandes, J. Simon-Sanchez, P. Rizzu, The International Parkinsons Disease Genomics Consortium (IPDGC).* Genome Biology of Neurodegenerative Diseases, DZNE-Tuebingen, Tuebingen, Germany.

Large scale Whole-Exome/Whole Genome Sequencing (WES/WGS) and Genome Wide Association Studies (GWAS) generate large numbers of candidate genes for Neurodegenerative Diseases for which it is difficult to obtain convincing prove of causality. In population based WES/WGS studies the identified variants are often extremely rare and additional affected family members are often not available. We recently performed a systematic functional prioritization of candidate genes from a large WES study on Parkinson's Disease (Jansen et al. genome Biol. 2017). We now report a similar approach for GWAS data. GWAS studies identify “tagging”-SNPs that are often located in non coding regions of the genome that point to an approximate genomic location of the risk variant but the association peaks often contain multiple genes. The large majority of GWAS signals are located in non coding regions of the genome and we hypothesize that a large proportion of causal variants are located in regulatory elements of flanking genes or noncoding RNAs with regulatory functions, increasing or decreasing the expression levels of flanking genes. In a recent meta-analysis (Nalls et al. Nat. Genet 2014) we identified 24 risk loci for Parkinson’s Disease. In order to understand the biological consequences of the unknown causal variants we used LD measures to identify all genes underlying the identified risk loci and selected multiple shRNA clones per gene to knock-down each gene in cellular models mimicking a negative regulatory effect of the unknown causal variants. We then performed fully automated high-throughput/high-content cellular RNAi screens on all susceptibility gene candidates from our GWAS studies. By integrating a series of screening readouts varying from mRNAseq, protein expression to high-content cellular imaging we systematically investigated all genes underneath the association curves of the identified genetic risk loci to identify those genes that influence pathways and cellular phenotypes that are relevant in the pathogenesis of Parkinson's disease. In addition, genes affecting the same functional pathways are considered to have functional and genetic interactions which can be examined in our GWAS data for epistasis and validated by performing double knockdown experiments. Our approach highlights a powerful experimental strategy with broad applicability for future studies of disorders with complex genetic etiologies. .

153

Functional interrogation of common genetic variation uncovers regulators of hematopoiesis and therapeutic targets. *S.K. Nandakumar[1,2], S.K. McFarland[1,2], L.M. Mateyka[1,2], J.C. Ulirsch[1,2], G.S. Cowley[2], X. Yang[2], J.G. Doench[2], D.E. Root[2], V.G. Sankaran[1,2].* 1) Division of Hematology/Oncology, Boston Children's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA; 2) The Broad Institute of MIT and Harvard, Cambridge, MA.

Genome-wide association studies (GWAS) have identified >36,000 single nucleotide polymorphisms (SNPs) associated with human diseases and traits. However, the majority of GWAS-identified SNPs are in non-coding regions of the genome and numerous functional approaches, which can be difficult to scale, are necessary to identify target genes and gain insight into the underlying biology. Here, rather than focusing on causal variants, we have undertaken a high-throughput loss-of-function screen to interrogate candidate genes contained in loci implicated through a GWAS of blood cell traits. We utilized a relevant functional screen with a pool of lentiviral short hairpin RNAs (shRNAs) and assessed depletion or enrichment during the synchronous process of differentiation of primary human hematopoietic stem/ progenitor cells (HSPCs) into mature red blood cells over the course of 16 days. At a total of 75 loci associated with red blood cell traits, we were able to define a set of 390 distinct genes contained in the linkage-disequilibrium blocks harboring all putative causal variants. Each of the candidate genes was targeted with 5-7 distinct shRNAs, along with a large set of positive and negative controls. Enrichment or depletion of shRNAs was quantified across 6 distinct time points in HSPCs from 3 different donors. Using an innovative and robust linear mixed modeling approach, we were able to resolve genes showing significant hairpin enrichment or depletion in the screen. We identified a total of 53 genes across 29 GWAS loci (p-value <0.05 and $\beta < -0.15$ or > 0.12) with a single or two candidate genes identified at 25 loci. In depth follow up of our screen has identified several previously unreported regulators of blood cell production, which may be promising targets for therapeutic manipulation. For example, our screen identified the *TFR2* gene as a negative regulator of red blood cell production and our follow up results demonstrate how this gene normally constrains erythropoietin (EPO) signaling. Targeted manipulation of TFR2 could be valuable in anemias characterized by ineffective erythropoiesis, which are ordinarily refractory to EPO treatment. We will discuss several examples of the key biological and therapeutic insights gained from this screen. More generally, this functional screen provides a paradigm for gene-centric follow up of numerous GWAS for a variety of human diseases and traits.

154

High-throughput functional analysis of *PTEN* variants reveals genotype-phenotype relationships. *T.L. Mighell[1,2], S. Evans[2], B.J. O'Roak[1,2].*
1) Neuroscience Graduate Program, Oregon Health & Science University, Portland, OR; 2) Department of Molecular & Medical Genetics, Oregon Health & Science University.

The tumor suppressor *PTEN* is a critical antagonist of the pro-survival, pro-growth PI3K/mTOR pathway. Loss-of-function *PTEN* mutations are extremely common in somatic cancers, and germline *PTEN* mutations are also observed in individuals with autism and macrocephaly, as well as various overgrowth syndromes collectively known as PTEN hamartoma tumor syndrome (PHTS). Currently our ability to accurately predict a phenotype based on *PTEN* genotype is limited, complicating the interpretation of novel mutation events. We predict that characterizing the effects of all single-residue *PTEN* mutations would inform our understanding of the underlying biology of PTEN-associated phenotypes. To enable this pursuit, we have adapted a humanized yeast model (Rodríguez-Escudero *et al*., 2005) to permit massively parallel analysis of thousands of synthetically generated *PTEN* mutations (Melnikov *et al*., 2014) with DNA sequencing as proxy readout for phosphatase function. In this system, expression of human hyperactive *PIK3CA* results in toxicity through depletion of essential phosphatidylinositol (4,5)-bisphosphate (PIP_2). Co-expression of PTEN rescues yeast growth by lipid phosphatase-mediated restoration of the essential PIP_2 pool; therefore, the catalytic competence of a PTEN variant is coupled to cell growth. We generated saturation mutagenesis libraries encompassing 93% of all missense, nonsense, and single residue deletion mutations in *PTEN* (7,891 total mutations). Using our model, we first estimated the phosphatase activity of 399 variants overlapping the highly conserved phosphatase motif. Our data conform to basic tenets of biochemistry: nonsense mutations are damaging, variants within well-ordered secondary structures are damaging, and mutations to proline are generally damaging. We further demonstrate a series of solvent-exposed amino acids (residues 113-118) are remarkably tolerant to mutation, which is in agreement with their low evolutionary conservation. Comparing our scores of variants associated with human phenotypes, we find preliminary evidence that autism-associated variants retain higher phosphatase activity than those associated with the PHTS Cowden Syndrome. We are now actively analyzing the effects of the remaining PTEN variants in order to discover additional insights into genotype-phenotype relationships, which in turn can be leveraged to inform the care and treatment of individuals with *PTEN* mutations.

155

Population-wide whole-genome sequencing in a Greek isolate reveals rare variant burdens associated with multiple quantitative traits. *A. Gilly[1], D. Suveges[1], K. Kuchenbaecker[1], L. Southam[1,2], K. Hatzikotoulas[1], T. Bjørnland[3], E.V.R. Appel[4], E. Casalone[5], G. Melloni[6], K.B. Kilian[1], N.W. Rayner[1,2], A.-E. Farmaki[7], E. Tsafantakis[8], M. Karaleftheri[9], G. Dedoussis[7], E. Zeggini[1].*
1) Human Genetics, Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 2) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; 3) Department of Mathematical Sciences, Norwegian Institute of Science and Technology, Trondheim, Norway; 4) Section for Metabolic Genetics, Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark; 5) Human Genetics Foundation, University of Torino, Torino, Italy; 6) Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA; 7) Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Greece; 8) Anogia Medical Centre, Anogia, Greece; 9) Echinos Medical Centre, Xanthi, Greece.

Population-scale deep whole-genome sequencing can capture genetic variation across the full allele frequency spectrum in complex trait association studies. Isolated populations offer power gains in detecting associations with rare and low-frequency variants. Here, we have whole genome sequenced 1,457 individuals at an average depth of 22.6x from an isolated population cohort from Crete, Greece. We test 13,449,852 SNPs with minor allele count (MAC) >10 for association with 48 quantitative traits, and report 29 independent signals across 24 traits at the $5x10^{-8}$ significance level, including 6 previously reported associations with haematological and lipid traits. For gene-based approaches, we test exonic and regulatory variants associated with 19,025 protein-coding genes reported by GENCODE V25 (GRCh38). We benchmark 12 different pipelines using different regions of interest (exonic, exonic and regulatory and regulatory only), variant weights and filters. Changing the region of interest gives rise to different signals, highlighting the importance of running genome-wide burden tests under multiple conditions. We report 29 genome-wide significant ($P<1.3x10^{-7}$) rare variant burden signals not driven by a single variant, including for known loci, such as *ADIPOQ* for adiponectin ($P=9.1x10^{-8}$), *APOA1* and *APOC3* for HDL ($P=2.12x10^{-20}$ and $3.96x10^{-20}$, respectively), *UGT1A10* for bilirubin ($P=1.2x10^{-8}$), as well as *HBB* and *HBE1* for multiple haematological traits ($P<10^{-50}$). Some signals at known loci are fully recapitulated by known common variant associations despite independence based on LD (such as *UGT1A10* for bilirubin and the intronic rs887829). Others, such as *GGT1* for gamma-glutamyltransferase and *ADIPOQ* for adiponectin, remain significant after adjusting for single-point associations in the region. Finally, we identify associations at novel loci, such as between thyroxine levels and variants in the *PTK2B* gene ($P=1.0x10^{-7}$). We demonstrate that whole genome sequencing-based burden tests complement single-point associations when exploring the allelic architecture of human traits including at established loci with common variant associations, and provide a first comprehensive view of the rare variant burden landscape for complex traits in a population cohort.

156

Inferring compound heterozygotes from large-scale exome sequencing data. *L.F. Francioli[1,2], M.H. Guo[2,3,4], K.J. Karczewski[1,2], B.B. Cummings[1,2], M. Lek[1,2], V. Thaker[5], M.J. Daly[1,2], J.N. Hirschhorn[2,3,4], D.G. MacArthur[1,2], Genome Aggregation Database.* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Division of Endocrinology, Boston Children's Hospital, Boston, MA; 4) Department of Genetics, Harvard Medical School, Boston, MA; 5) Division of Molecular Genetics, Columbia University Medical Center, New York, NY.

Short-read sequencing technologies have enabled sequencing of the exomes and genomes of hundreds of thousands of people. However, most standard sequencing technologies do not readily provide phase, that is, assigning variants to individual haplotypes. The absence of phase information particularly complicates the identification of compound heterozygous variants in the diagnosis of recessive disorders, where mutations of both copies of a given gene are necessary to develop disease. While the phase of common variants can be accurately inferred using imputation, phasing rare variants is challenging, especially from exome sequencing due to the sparse coverage of each haplotype. Here, we leverage large-scale whole-exome sequencing data from the genome aggregation database (gnomAD) which contains 123,136 individuals, to infer haplotype frequencies between pairs of rare variants (allele frequency < 1%) located in the same gene. We demonstrate that these frequencies can be used to infer compound heterozygote status for variants within a gene with high accuracy (92.3%) from whole-exome sequencing data using 1,494 trios for which true phase can be determined by allele transmission. Notably our approach performs very well even with a single observation of each allele in gnomAD (allele frequency ~10^{-5}). Interestingly, 88.2% of the pairs of private variants (both variants absent from gnomAD) are located in close proximity (median distance of 8.5bp) on the same haplotype and exhibit a mutational spectrum with reduced transitions at CpG sites, suggesting a different mutational mechanism. Applying our approach to the 123,136 individuals in gnomAD, we compare the burden of rare protein-altering compound heterozygous variants and homozygous variants in each gene. We show the clinical utility of our method when applied to several rare disease cases where we could either exclude or prioritize putative causative compound heterozygous variants (validated by sequencing the proband's parents). Finally, we will report the results of integrating phase information in rare variant burden testing for recessive gene discovery using a cohort of 1,300 probands with severe limb girdle muscular dystrophy.

157

Novel dual-indexing strategy enables high scale sample multiplexing and eliminates the impact of index-swapping on Illumina sequencers. *M. Costello[1], M. Fleharty[2], S. Ferriera[1], T. Howd[1], J. Abreu[1], Y. Farjoun[2], T. Desmet[1], S. Dodge[1], N. Lennon[1], S. Gabriel[1].* 1) Genomics Platform, Broad Institute, Cambridge, MA; 2) Data Sciences and Data Engineering, Broad Institute, Cambridge, MA.

"As sequencing costs decline…" is a common phrase heard as plans are put in place to increase the size and scope of human genetics projects. The HiSeqX provided ~1000 Gb of data per flowcell and cut costs by two thirds, enabling large scale projects like TOPMed to sequence 75,000 samples, and future projects like the All of Us initiative will look to collect DNA from millions of individuals. Sequencing genomes at this scale requires a highly streamlined workflow that makes the most efficient use of sequencer yield and eliminates the effects of lane to lane variability. To achieve this sample multiplexing on sequencer has become a necessity. To address this challenge, we have developed a dual indexing scheme utilizing a set of non-redundant indexes (comprising 96 unique i7 and 96 unique i5) that have been validated across multiple workflows and Illumina sequencer models, including NovaSeq. Using this approach, we can identify and eliminate reads that have been "index-swapped" due to a known fail mode on patterned flowcells (HiSeqX, HiSeq 4000, and NovaSeq)[1], reducing the rate of contamination due to sample read misassignment from 2-5% down to <0.1%. We have also characterized the swapping phenomenon in more detail, observing trends in GC content, insert size, and variability between various sample preparation methods. With the NovaSeq, flowcell yields will eventually increase to 3 Tb and sample multiplexing will be required to truly realize reduced sequencing costs. In particular targeted gene panels, RNA-seq, and single cell applications will require the ability to pool to at much higher plexity to take advantage of the NovaSeq's output. Unique dual indexing is necessary to eliminate the impact of index swapping, but currently vendors, including Illumina, only provide solutions that allow pooling up to 8 samples with unique dual indexes. Our strategy allows for unique plexing of up to 96 samples per pool, and we are in the process of increasing the number of available indexes to allow up to 384 samples per pool. 1. Sinha et al. bioRxiv 125724; doi: https://doi.org/10.1101/125724.

158

Recessive coding variants make only a minor contribution to undiagnosed developmental disorders in the United Kingdom. *H.C. Martin[1], W. Jones[1,2], J. McRae[1], D.R. Fitzpatrick[3], H.V. Firth[1,4], M. Hurles[1], J. Barrett[1], Deciphering Developmental Disorders Study.* 1) Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom; 2) Clinical Genetics Department, Great Ormond Street Hospital for Children NHS Foundation Trust, Great Ormond Street, London, United Kingdom; 3) MRC Human Genetics Unit, MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; 4) East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom.

We have analyzed 7,833 trios in the Deciphering Developmental Disorders (DDD) study to estimate the overall burden of recessive coding causes in the cohort. To identify individual recessive genes, we tested whether the number of biallelic (homozygous or compound heterozygous), putatively damaging variants observed in each gene was significantly greater than expected by chance given the background rates of such variants in healthy controls. There was a significant enrichment for known developmental disorder (DD) genes amongst those with $p<1\times10^{-4}$ ($p=5\times10^{-8}$; odds ratio = 128), supporting the validity of our approach. We identified two interesting new genes. Nine probands were homozygous for the same missense variant (frequency=0.1% in Europeans) in the translation initiation factor *EIF3F*, a highly significant enrichment over chance expectation ($p<8\times10^{-7}$). We also identified four probands with biallelic loss-of-function (LoF) variants in the histone demethylase *KDM5B* ($p=1.8\times10^{-5}$). Curiously, this gene is also enriched for *de novo* LoF and missense mutations in the DDD ($p=5.1\times10^{-7}$), but it had previously been excluded as a DD gene because *de novo* LoF mutations had been observed in healthy individuals. We observed 30 *KDM5B* LoF variants amongst the DDD parents, of which 22 were transmitted (p=0.01 on a transmission disequilibrium test), and also found a significant enrichment of LoFs in the DDD parents compared to controls from ExAC ($p=1.5\times10^{-12}$; odds ratio=6.8). This suggests that *KDM5B* has a more complex mode of inheritance, in which heterozygous LoF mutations show incomplete penetrance and homozygous LoFs are fully penetrant. In our recessive gene-discovery and burden analyses, we have found two factors to be paramount: having an appropriate ancestry-matched control dataset produced and processed using the same methods as the cases, and having a sufficient number of controls to accurately estimate the background frequencies of very rare variants. Previous work has shown that ~40% of the DDD can be explained by *de novo* coding mutations. Here, we estimate that the proportion of unexplained cases attributable to recessive coding variants is only ~2% for cases of European ancestry (rising to ~5% for those with affected siblings), but higher in those of South Asian ancestry (~21%) due to elevated levels of autozygosity. Thus, a large proportion of cases remain to be explained by other factors, such as noncoding variants and polygenic risk.

159

Fine-scale structure of rare variants in 18K genomes from the TOPMed Consortium and its implications for study design. *T. O'Connor[1,2], D. Harris[1], M. Kessler[1], A. Shetty[1], B. Mitchell[2], D. Taliun[3], D. Nickerson[4], R. Hernandez[5], G. Abecasis[3], TOPMed Consortium.* 1) Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD; 2) Department of Medicine, University of Maryland School of Medicine, Baltimore, MD; 3) Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI; 4) Department of Genome Sciences, University of Washington, Seattle, WA; 5) Department of Bioengineering and Therapeutic Sciences University of California, San Francisco, CA.

Rare variants comprise the vast majority of human genomic variation, and understanding their distribution among closely related populations is critical for how we design large-scale association analyses. This is especially true when collecting samples from multiple sources to conduct studies with sufficient power. The Trans-Omics for Precision MEDicine (TOPMed) consortium is a NHLBI sponsored collaboration that created a high coverage whole genome catalogue from individuals with multiple ancestries; deep phenotype data for heart, lung, blood, and sleep disorders; and a broad range of omics characterizations. In this study we analyzed rare variants in the first 18K samples to explore patterns of fine-scale population structure. We find varying degrees of diversity across the genomes, with the extremes being represented by the Old Order Amish and African ancestry groups. Even after correcting for differences in heterozygosity, African ancestry individuals share a greater proportion of rare variants than European ancestry individuals. This is likely due to European American cohorts having multiple European sources, whereas African ancestry in African American cohorts primarily comes from a few regions of West Africa. We also developed a new metric to quantify and contrast the population or cohort specificity of different variant categories (e.g. coding vs non-coding) while controlling for allele frequency. Using this metric we can separate relatively older vs younger variants, which have the same allele frequency, and can measure these differences across categories that represent varying degrees of purifying selection (i.e. PhyloP, GERP, CADD, PolyPhen, and Ensemble gene model annotations). To do this, we contrast variants in regions more likely to be neutral (e.g. unconserved or intergenic) to those predicted to be functional or deleterious (e.g. conserved or nonsense). We find that variants which are annotated as functional are more likely to be younger and cohort specific than variants in non-functional regions, even among only European ancestry cohorts. These results have implications for study designs where cases, or extreme phenotypes, are generated from one sampling location and controls are used from a collective sampling strategy, as genome-wide estimates of relatedness will underestimate the structure of the variants that we are most likely to prioritize from a disease association perspective giving them greater probability for spurious associations.

160

Large-scale identity-by-descent mapping discovers rare haplotypes of large effect. *S. Shringarpure, D. Hinds, V. Vacic, S. Pitts, R. Gentleman, A. Auton, 23andMe Research Team.* 23andMe Inc., Mountain View, CA.

The power of genome-wide association studies (GWAS) to discover rare associations (< 0.1% frequency) is limited by imputation accuracy. Current panels such as the Haplotype Reference Consortium or 1000 Genomes can impute variants with MAF > 0.1% well (r^2 > 0.5) but do not perform well for rarer variants. Here, we demonstrate the use of identity-by-descent (IBD) mapping to discover rare associations of large effect. We have developed a method for IBD mapping on cohorts of hundreds of thousands of individuals. We use an approach similar to DASH (Gusev et al. 2011) to cluster cases and controls using IBD segments at a locus. We then test IBD haplotype clusters for enrichment of cases (for causal haplotypes) or controls (for protective haplotypes). Our method improves on DASH in terms of scalability to large samples while also correcting for covariates in the association test. We validated our method on a set of 23andMe consented research participants of European ancestry in scenarios where rare functional SNPs were assayed on the 23andMe genotyping arrays. In an association analysis for breast/ovarian cancer (N=114,789), we find an IBD cluster of 60 individuals (cluster freq=0.05%, p=1.3e-12 , OR=10.73) overlapping the BRCA2 gene. We find that the cluster tags rs397507813, a rare PTV with a MAF of 0.04%, with 92% of cluster members heterozygous for the rare PTV. We validated the approach similarly for Parkinson's disease in individuals of Ashkenazi Jewish ancestry (N=43,173). We find an IBD cluster of 674 individuals overlapping the LRRK2 gene (freq=1.5%, p=3.5e-68, OR=9.42). This cluster tags the G2019S variant (rs34637584), which has known associations with Parkinson's disease. The G2019S snp has a MAF of 1.2% in the GWAS sample but 98% of the IBD cluster members carry the GS2019S variant. We analyzed 48 disease phenotypes using our method and found a number of known and novel associations. Among others, we replicated associations related to genes HFE (hemochromatosis, freq=0.2%, p=7.8e-10, OR=4.83), ATM (non-skin cancers, freq=0.4%, p=5.5e-7, OR=1.8), SLC45A2 (basal cell carcinoma, freq=0.02%, p=8.3e-8, OR=0), and JAG1 (prostate cancer, freq=0.05%, p=1.5e-7, OR=8.22). Our results suggest that IBD mapping on large cohorts is a viable way of finding rare associations that are inaccessible from imputed genotype data. Reference: DASH: A Method for Identical-by-Descent Haplotype Mapping Uncovers Association with Recent Variation. Alexander Gusev, et al. AJHG 2011.

161

Using induced pluripotent stem cells to identify downstream effects of genetic risk factors linked to cancer. *M.J. Bonder[1], D. Seaton[1], B. Miraut[1], H. Kilpinen[1], J. Korbel[1,2], O. Stegle[1], The HipSci consortium.* 1) European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, UK; 2) European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany.

Genome-wide association studies in large population cohorts have yielded a compendium of genetic variants that are associated with human diseases, including cancer. However, understanding the molecular consequences of these associations remains challenging, mainly because the majority of human disease variants are in intergenic regions. One strategy for linking these risk variants to genes are expression quantitative trait loci (eQTL) studies. However, disease-causing variants tend to have tissue- and context-specific effects. To date, most eQTL studies have been performed in lymphoblast cell lines, blood cell types, or in post-mortem collected tissues, with limited relevance for cancers. To address this, we here map eQTLs in the largest panel of human induced pluripotent stem cells (iPSCs) considered to date. iPSCs have recently emerged as a promising model for dedifferentiated and stem-like cells, since iPS reprogramming in part resembles the dedifferentiation processes in tumors. Using the latest release of data from the HipSci project (http://www.hipsci.org), we investigated genetic effects on gene expression traits in a set of over 550 IPSC lines, derived from close to 400 donors. Gene expression was quantified using RNA-sequencing, providing an opportunity to carry out genetic analyses at the level of individual genes, but also to map QTLs for exons, transcript levels and alternative splicing, as well as assessing allele-specific expression. We identified significant *cis*-eQTLs for more than 45% of expressed genes (FDR<5%), 13% of which were not identified in the largest study of blood eQTLs to date. Notably, our set of *cis* eQTLs includes regulatory variants for 176 cancer genes catalogued in cosmic (out of 404 genes tested), and our data link 53 cancer risk variants identified through GWAS to downstream gene expression changes. This set of genes is enriched for cancer-implicated pathways, including KEGG pathways for pancreatic cancer, leukemia and lung cancer. We are currently extending our study by including additional lines generated by other large-scale iPS generation efforts. We anticipate that this even larger dataset will be sufficiently powered to detect genes that are regulated through *trans* networks, thereby providing a comprehensive picture of *cis*- and *trans*-eQTLs in pluripotent human cells.

162

A spectrum of rare Mendelian to common GWAS variants underlying familial prostate cancer at 8q24. *J. Smith, J. Breyer, W. Dupont.* Vanderbilt University Medical Center, & VA Tennessee Valley Healthcare System, Nashville, TN.

Prostate cancer has the greatest estimated heritable risk (58%) of all common cancers. One in 39 men will die from it. Our study investigates the etiology of familial prostate cancer, men with a relatively greater potential genetic load for the disease. We employed an extreme-contrast association study design based upon family history in order to detect a spectrum of risk variants from common small-effect to rare large-effect. The study compares cases with a strong family history of prostate cancer to controls. As a training set, we explored array data of the International Consortium for Prostate Cancer Genetics (2,568 cases, 1,422 controls) with imputation against 32K genomes of the Haplotype Reference Consortium. We identified loci of greatest significance and effect size for subsequent replication tests by direct assay within the independent Nashville Familial Prostate Cancer Study (930 cases, 928 matched controls without a personal or family history of the disease). Here we present comprehensive fine-mapping of a locus at 8q24. We identified numerous independent risk variants across a 2 MB interval with individual ORs up to 10 and at genome-wide significance among subjects of European descent. Haplotype analyses further detected combinations of these that confer even greater risk for prostate cancer when jointly inherited. This approach successfully detected rare Mendelian risk variants as well as common GWAS variants. Variance analysis estimates that the 8q24 locus accounts for 6% of familial prostate cancer heritability.

163

Cis-regulatory drivers in chronic lymphocytic leukaemia and skin cancer. *H. Ongen, O. Delaneau, M. Stevens, C. Howald, E.T. Dermitzakis.* Department of Genetic Medicine and Development, University of Geneva, 1211 Geneva, Switzerland.

While the role of coding somatic mutations in tumorigenesis is well understood, the contribution of the non-coding genome is poorly studied. We previously indirectly detected putative somatic regulatory drivers in colorectal cancer using perturbations in allele specific expression (Ongen et. al. Nature 2014). However, direct assessment of excess somatic mutations in the non-coding regulatory regions (NRRs) from whole genome sequencing (WGS) analyses is challenging due to the difficulty of defining the NRRs of genes and the short length of NRRs. We solve this problem by identifying modules (sets) of coordinated NRR of genes using the correlation between 3 histone modifications, assayed by CHiP-seq, in immortalized B-cells (LCLs, 320 samples) and skin fibroblasts (80 samples). Using hierarchical clustering we build a tree of correlated marks, from which we define modules of NRRs, which we correlate with gene expression to identify the genes regulated by the module. Since we can define modules of NRRs in B-cells and skin fibroblasts we focused on chronic lymphocytic leukaemia (CLL) and skin cancer, which are affecting these cell types. Using publicly available WGS data and accounting for differential mutation rate across the genome and between open and closed chromatin, we identify modules of NRRs that show a significant excess of somatic mutations. In CLL we find 149 significant modules at 5% FDR, 53 of which are regulating 114 genes, and in skin cancer there are 465 significant modules 55 of which regulate 73 genes. The genes that accumulate significantly more somatic mutations in their NRRs in both cancers include known drivers of cancer and in CLL are enriched for B cell receptor signalling pathway, the main pathway involved in CLL development. We find that the mutations in significant modules are hitting bases that are more likely to be functional based on the LINSIGHT method than ones in the non-significant modules. Further, we observe a significant change in the mutational signatures observed in NRRs vs. the signatures in their spacers. Both of these results indicate selection during tumorigenesis. Finally, we examined the transcription factor binding sites (TFBS) that are disturbed by the somatic mutations in NRRs. We find significant enrichments for TFBS that are involved in tumorigenesis like p53 and NFKB signalling. In conclusion, we are describing a new powerful approach to discover non-coding regulatory somatic mutations driving tumorigenesis.

164

Recurrently altered enhancers in colorectal cancer identify known and novel predisposition loci. *S.A. Bien[1], A. Saiakhova[2], T.A. Harrison[1], C. Qu[1], J.R. Huyghe[1], H.M. Kang[3], G.R. Abeçasis[3], G. Casey[5], D.A. Nickerson[6], L. Hsu[1], S.B. Gruber[4], P. Scacheri[2], U. Peters[1] on behalf of GECCO, CCFR and CORECT.* 1) Public Health Genetics, Fred Hutchinson Cancer Research Center, Seattle, WA; 2) Department of Genetics School of Medicine, Case Western Reserve University, Cleveland, OH; 3) Department of Biostatistics, University of Michigan, Ann Arbor, MI; 4) Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA; 5) Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA; 6) Department Genome Sciences, University of Washington, Seattle, WA.

Previous genome-wide association studies (GWAS) of Colorectal Cancer (CRC) have identified 61 common genetic risk variants across 45 regions. However, much of the estimated risk attributable to heritable factors remains unexplained. Recently, the generation of epigenomic reference maps for 7 normal colon samples and 35 primary CRC cell lines has enabled the identification of enhancers that are recurrently activated or lost across a disproportionately high number of primary CRC cell lines. We have found that half of the known CRC risk variants are in these recurrent Variant Enhancer Loci (VELs). To further evaluate whether any novel associations between genetic variants residing in recurrently gained or lost VEL and CRC risk, we employed a “functionally informed” GWAS focusing on only these variants in 22,994 CRC cases and 14,407 controls. Genetic variants that were not present on the GWAS platforms were imputed using whole genome sequences from the haplotype reference consortium panel release 1, and analyses were restricted to variants with a minor allele frequency greater than 1% and imputation $R^2 \geq 0.3$. Association with CRC risk was evaluated for both recurrently gained and lost VEL (N variants = 145,768 and 28,160, respectively) through multivariate logistic regression models, including sex, age, study, principal components of ancestry, and previously identified CRC risk variants. We identified two novel, statistically significant (0.05/# variants), associations with *SMAD3*–rs11071933 (gained VEL p-value = 2.4×10^{-7}) and *CDKN2B-AS1*–rs10733376 (lost VEL p-value=1.3×10^{-6}). Bioinformatic follow-up showed that both loci were positioned in intronic regions and that rs11071933 is associated with expressed levels of *SMAD3*. Interestingly, rs10733376 was predicted to alter the binding efficiency of the DNA motif for *SMAD3*. These findings are of particular interest given that *SMAD3* is a tumor driver gene and recent evidence that chemoresistance to 5-Fluorouracil treatment is mediated through *SMAD3* activation. The cancer risk locus, *CDKN2B-AS1*, encodes a long non-coding RNA thought to repress other genes clustered in this region (*p15/CDKN2B-p16/CDKN2A-p14/ARF*) through its role in the polycomb repressive complex 1. Results from this study highlight the utility of integrating functional genomics data to guide novel discovery of biologically and potentially clinically relevant loci.

165

Assessing the gene regulatory landscape in 1,188 human tumours. *K. Lehmann[5,9], A. Brazma[1,9], C. Calabrese[1,9], S. Dentro[2,3,9], S. Erkek[4,9], N. Fonseca[1,9], A. Kahles[5,9], H. Kilpinen[1,2,6,9], J. Korbel[4,9], F. Liu[7,9], J. Markowski[8,9], G. Raetsch[5,9], RF. Schwarz[1,8,9], O. Stegle[1,9], L. Urban[1,9], P. Van Loo[3,9], S. Waszak[4,9], D. Wedge[5,9], Z. Zhang[7,9], PanCancer Analysis Working Group 3, PanCancer Analysis Working Group 8.* 1) European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK; 2) Wellcome Trust Sanger Institute, Hinxton, UK; 3) The Francis Crick Institute, London, UK; 4) European Molecular Biology Laboratory, Heidelberg, Germany; 5) Department of Computer Science, ETH Zurich; 6) UCL Great Ormond Street Institute of Child Health, University College London, UK; 7) BIOPIC and College of Life Sciences, Peking University, China;; 8) Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany; 9) All authors are ordered alphabetically.

To systematically assess regulatory effects of germline and somatic variants on a genome-wide scale, we analyze matched whole-genome sequencing and RNA-seq data from 1,188 cancers from the TCGA/ICGC PanCancer Analysis of Whole Genomes (PCAWG) . We map expression quantitative trait loci (eQTL) of germline variants, and contrast their effects to the corresponding normal tissues. This analysis allows us to identify 3,396 genes with a germline effect (FDR <10%), including prominent cancer genes (e.g. MAP3K6, MDM1, TERT). A comparison of this eQTL map to regulatory variants in matched normal tissues from GTEx, resulted in around 10% of cancer-specific eQTLs that cannot be recapitulated in GTEx. Using this germline map, we assess the regulatory effect of somatic mutations using allele-specific expression (ASE) analyses and QTL mapping at multiple scales, considering individual aberrations, localised mutational burdens and global signatures of mutational processes. In particular, we map fine-grained spatial categories of somatic mutational burden on ASE, and identify 1,176 eGenes (FDR<10%) that are regulated by these variants using somatic eQTL analysis. To assess global effects of mutational processes on gene regulation, we extend 28 conventional mutational signatures with de-novo predictions of somatic effects on predicted epigenetic marks and transcription factor binding sites using deep neural networks. 1,373 genes are associated with at least one of 70 signatures (FDR<10%), including 47 cancer genes. Regulated genes are enriched for categories that are consistent with the individual signatures and underscore interdependencies between somatic mutation profiles and expression states. Associations with global somatic signatures further allow for in-depth functional annotations of genome-wide patterns of somatic variation in a cancer-type specific manner. This study is the first integrated analysis of germline, somatic SNVs and structural aberrations across different cancer types on a whole-genome scale.

166

Sex-specific differences in expression dysregulation across The Cancer Genome Atlas studies. *A.D. Skol[1,2,3], R. Dohn[1,2], E. Lipschultz[1,2], Z. Zhang[3], R.L. Grossman[3], B.E Stranger[1,2].* 1) Section of Genetic Medicine, Department of Medicine, University of Chicago; 2) The Institute for Genomics and Systems Biology, University of Chicago; 3) Center for Data Intensive Science, University of Chicago.

Sex is an important clinical variable known to be associated with incidence, age of onset, and disease severity in many cancers. For example: thyroid cancer is 2.5x more common in women than men, whereas bladder cancer is 4x more common in men; mean age of onset in melanoma is 8 years earlier in women than men; and female nasopharyngeal carcinoma patients have a significant survival advantage over their male counterparts. Previous studies have shown sex-specific differences in patterns of tumor mRNA and miRNA expression, methylation, and somatic mutations in a subset of The Cancer Genome Atlas (TCGA) cancers. And while these studies have revealed many interesting relationships between tumor genomics and biology, a great deal more can be garnered from these data. In an attempt to do so, we have focused on four genomic data types in 26 TCGA cancer datasets: mRNA expression, methylation, miRNA expression, and somatic mutations. One of the challenges in conducting these analyses is correctly accounting for sex-specific differences in expression and methylation observed in normal samples of the appropriate tissue. In our analyses, we implemented a model-building framework that identifies sex-specific differences in gene expression and its regulation, using normal sex- and tissue-specific differences to avoid false positives and improve power. As an example, when we focus only on mRNA expression in the 26 TCGA cancers, we are able to reduce the number of sex-differentially expressed genes an average of 5.7% by taking into account sex differences in expression observed in non-cancer GTEx tissues (range of reduction 0.9% - 25.9%). The mean proportion of differentially expressed genes observed only in the TCGA cancers is 5.25% and range from 0.2% to 20.4%. Approximately 56% of these genes are unique to the cancer in which it was identified, while 9% are shared by more than half the cancers. We will present additional details that highlight how integrating multiple genomic signals can help improve our understanding of sex-differences in cancer biology. Our hope is that by focusing on cancer-specific genomic differences between men and women, we will gain an improved understanding of how sex-specific genetic differences or response to the environment can lead to sex-disparities in clinical measures and outcomes.

167

Inherited nodopathies and defective axoglial interactions are responsible for arthrogryposis multiplex congenita. *J. Melki[1], J. Maluenda[1], S. Xue[2], G. Ravenscroft[3], C. Manso[4], L. Quevarec[1], A. Vivanti[1], F. Marguet[5], M. Gut[6], I. Gut[6], M. Tawk[7], N.G. Laing[3], J. Devaux[4], B. Reversade[2], A. Laquérriere[5].* 1) Institut National de la Santé et de la Recherche Médicale (Inserm) UMR-1169, University Paris Sud, Le Kremlin Bicêtre, 94276, France; 2) Institute of Medical Biology and Institute of Molecular and Cell Biology, A*STAR, 138648, Singapore; 3) Harry Perkins Institute of Medical Research, Centre for Medical Research, University of Western Australia, Nedlands, Western Australia, Australia, 6009; 4) Aix-Marseille Université, Centre National de la Recherche Scientifique, CRN2M-UMR-7286, 13444 Marseille, France; 5) Pathology Laboratory, Rouen University Hospital and Normandie University, UNIROUEN, NéoVasc, Rouen, 76000, France; 6) CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, 08028, Spain; 7) UMR-1195, Inserm, University Paris Sud, University Paris-Sud, Le Kremlin-Bicêtre, France.

Arthrogryposis multiplex congenita (AMC) is a developmental condition characterized by multiple joint contractures resulting from reduced or absent fetal movements. The overall incidence is 1 in 3000 live births. Non-syndromic AMCs are genetically heterogeneous and include a large spectrum of diseases which arise as a result of mutations in genes encoding components required for the formation or the function of neuromuscular junctions, skeletal muscle, the survival of motor neurons or myelination of peripheral nerve. Through whole or targeted exome sequencing, we identified 5 new genes involved in the node of Ranvier formation (*CNTNAP1* and *GLDN*) or essential for axoglial interaction in the peripheral nervous system (*ADCY6*, *ADGRG6* and *LGI4*, Laquerrière et al. 2014, Ravenscroft et al. 2015, Maluenda et al. 2016, Xue et al. 2017). The node of Ranvier, the flanking paranodal junctions and the juxtaparanodes underlie saltatory conduction of action potentials along myelinated axons, an essential process for neuronal function. We showed that biallelic loss of function mutations of *GLDN* or *CNTNAP1* encoding essential components of the nodes of Ranvier and paranodes, respectively, lead to inherited nodopathies, a disease characterized by marked lengthening of the nodes which represent a new disease entity among peripheral neuropathies. We showed that mutation of *ADCY6*, which encodes an adenylate cyclase that synthesizes cAMP, causes AMC, suggesting that ADCY6 acted in the GPR126-cAMP pathway, which drives the differentiation of promyelinating Schwann cells by elevating cyclic AMP levels. Therefore, GPR126 was regarded as a strong candidate for human AMC. Indeed, we showed that mutations of *GPR126* are responsible for AMC associated with defective myelination of the peripheral axons during fetal development. More recently, biallelic loss-of-function mutations in *LGI4* were shown to be responsible for AMC. Morphological analysis of the sciatic nerve revealed a lack of myelin and functional tests revealed that the germline mutations impaired the secretion of truncated LGI4 protein. LGI4 is secreted by Schwann cells and binds to axonal Adam22 to drive the differentiation and myelination of Schwann cells. Defective node of Ranvier formation or axoglial interaction represent novel disease mechanisms in severe AMC.

168

World's largest gene-panel sequencing effort reveals genetic landscape of limb-girdle muscular dystrophies and potential multi-genic inheritance. *B. Nallamilli[1], S. Chakravorty[1], L. Rufibach[2], M.P. Wicklund[3], M. Harms[4], T. Mozaffar[5], M. Hegde[1].* 1) Human Genetics, Emory University, Atlanta, GA; 2) Jain Foundation; 3) Neurology, University of Colorado Denver; 4) Neurology, Columbia University; 5) University of California Irvine.

Limb-girdle muscular dystrophies (LGMDs) are an important genetically and clinically heterogeneous group of neuromuscular disorders (NMDs) with >25 different sub-types predominantly involved in proximal muscle weakness with an autosomal-recessive or -dominant inheritance. The clinical-genetic heterogeneities in LGMD make disease diagnosis complicated and expensive. To overcome the diagnostic hurdle encountered by LGMD patients, recently MDA/Jain-Foundation jointly launched the world's largest gene-panel sequencing program in Emory University with the 36 gene LGMD NGS-panel, where >5000 patients were sequenced. 30% of the patients had a definitive molecular diagnosis which is higher than exome sequencing in NMDs, with major pathogenic contributors in 5 genes: *CAPN3*, *DYSF*, *FKRP*, *ANO5, DMD*. Our finding of pathogenic variants' prevalence in *DMD* gene suggests a convergence of LGMD with Becker's muscular dystrophy. Interestingly, in our LGMD-suspected cohort, we successfully identified >27 genetically diagnosed (*GAA* gene) late-onset Pompe cases, a rare, but treatable lysosomal-storage disorder, often showing LGMD-like phenotype. We also identified high number of unique pathogenic variants in *TTN* gene even being one of the least contributors showing a broader variant spectrum of LGMD. We identified >28 cases of *DNAJB6* gene-associated dominant subtypes of LGMD which is relatively lesser studied gene suggesting natural history studies are required for new LGMD genes. Most interestingly, 50% of all the patients harbored variants of uncertain significance (VUS), and >15% of all the patients, pathogenic variants were detected in more than one LGMD gene, both aspects increase diagnostic hurdle and needs clinical functional assays to overcome. For example, one patient clinically diagnosed with LGMD2L but showing abnormal progression had homozygous pathogenic variants in both *ANO5* and *SGCA* genes. Another patient harboring heterozygous pathogenic variants in both *GAA* and *ANO5* genes show abnormal disease progression with both Pompe and limb-girdle phenotypes. Similar multigenic combinations of pathogenic variants were detected in many other individuals suggesting a possible role of "synergistic heterozygosity" and/or "digenic/multigenic contribution" to disease presentation and progression. Overall, this large-scale LGMD sequencing project has tremendously improved understanding of the different LGMD subtypes and potentially new disease inheritance modalities.

169

ATP1A1 represents a significant novel dominant Charcot-Marie-Tooth disease gene. *A.P. Rebelo[1], L. Lassuthova[2], G. Ravenscroft[3], P. Lamont[3], M. Baxter[3], R. Ong[3], M. Davis[3], F. Manganelli[7], F. Tao[1], C. Saghira[1], L. Abreu[1], Y. Bai[6], D. Isom[4], N. Laing[3], B.O. Choi[5], P. Seeman[2], M. Shy[6], L. Santoro[7], S. Zuchner[1].* 1)) Dr. John T. Macdonald Foundation Department of Human Genetics, John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, USA; 2)) DNA Laboratory, Department of Paediatric Neurology, 2nd Faculty of Medicine, Charles University in Prague and University Hospital Motol, Prague, Czech Republic; 3) Centre for Medical Research, University of Western Australia and Harry Perkins Institute of Medical Research, Nedlands, Australia; 4) Department of Pharmacology, Sylvester Comprehensive Cancer Center, and Center for Computational Sciences, University of Miami, Miami, USA; 5) Department of Neurology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea; 6) Department of Neurology, Carver College of Medicine, University of Iowa, Iowa City, USA; 7) Department of Neuroscience, Reproductive Sciences and Odontostomathology, Naples, Italy.

Despite more than 90 Charcot-Marie-Tooth disease (CMT) genes identified today, at least 50% of CMT2 patients do not carry a mutation in any of these genes. Progress in gene identification in recent years suggests that there are still many more CMT2 disease genes to be discovered; however, it has become rare to gather support from multiple large families that allow for conclusive linkage analysis. By employing global gene matchmaking (Korea, Italy, Czech Republic, Australia) we have identified multiple extended dominant CMT2 families with linkage support for a gene at chromosome 1p13.1. Originally a Czech family yielded a two point LOD score of 2.4 at this locus and a family from Southern Italy showed a LOD score of 3.2. Whole exome sequencing of multiple family members identified missense mutations in the gene ATPase Na+/K+ Transporting Subunit Alpha 1 (ATP1A1). ATP1A1 has not been associated with human diseases thus far. Expression studies on teased nerve fibers revealed predominant ATP1A1 expression at the Schmidt-Lanterman incisures, which are cytoplasmic pockets in the Schwann cells that facilitate ion exchange between axons and surrounding tissues. Through the GENESIS data sharing platform and collaborative efforts in the Inherited Neuropathy Consortium we identified five additional multigenerational families via exome or Sanger sequencing resulting in a total of seven unique segregating missense changes: Leu48Arg, Ile592Thr, Ala597Thr, Asp601Phe, Pro600Ala, Pro600Thr and Asp811Ala. Five of these mutations fall into a remarkably narrow motif associated with the sodium binding structure of ATP1A1, flanking the flexible hinge motif. Patch clamp measurements on Xenopus oocytes demonstrated significant reduction in Na current activity in the ouabain-insensitive ATP1A1 mutants confirming a loss of function defect of the Na, K-pump. Taken together, we show strong support for a major new dominant CMT2 gene, ATP1A1. This finding represents a new pathway and an attractive new target for therapy development in axonal CMT. .

170

Rare variant burden analysis deciphers genetic architecture of inherited peripheral neuropathies. *D.M. Bis[1,2], F. Tao[1,2], L. Abreu[1,2], P. Sleiman[3], H. Hakonarson[3], S. Zuchner[1,2], Inherited Neuropathy Consortium.* 1) J.T. MacDonald Department of Human Genetics, University of Miami, Miami, FL; 2) Hussman Institute for Human Genomics, University of Miami, Miami, FL; 3) Center for Applied Genomics, the Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.

Inherited peripheral neuropathies, also known as Charcot-Marie-Tooth (CMT) disease, are rare, clinically and genetically heterogeneous diseases that lead to distal muscular atrophy and sensory loss. Mendelian high-penetrance alleles in over one hundred different genes have been shown to cause CMT; yet, more than 50% of patients with the axonal type of CMT do not receive a genetic diagnosis. A more comprehensive spectrum of genes and alleles is warranted, including causative and risk alleles, as well as oligogenic inheritance. Exome studies in the international Inherited Neuropathy Consortium are beginning to be sufficiently powered to perform rare variant burden analysis. Our approach compared the frequency of damaging alleles at the gene unit in exomes of 343 CMT cases and 935 controls. Initially, we explored rare (ExAC MAF<0.01) and damaging (non-synonymous and loss-of-function consequences) variant burden in known CMT genes. In 76 axonal CMT genes, we saw that cases carried on average 2.08 rare, damaging variants, while unrelated non-neuropathy controls harbored 1.69 variants (p-value=3.13×10^{-5} MW U-test). This result was achieved despite prior exclusion of cases carrying a mutation in a known CMT gene from exome sequencing. Thus, enrichment of damaging variants in CMT disease genes in such a cohort suggests the presence of additional 'risk' alleles in these genes, potentially supporting oligogenic inheritance models after further exploration. To expand upon this result, we performed an unbiased exome-wide rare variant burden analysis. We tested 17,637 protein coding loci for association using the C-alpha test. After filtering results by the PLINK/SEQ i-statistic and applying Bonferroni multiple-testing correction, three genes, *KDM5A* (p-value= 9.9×10^{-7}, OR=3.6), *EXOC4* (p-value= 6.9×10^{-6}, OR=2.1), and *CEP78* (p-value= 2.3×10^{-5}, OR=4.4), reached genome-wide significance (p-value=2.3×10^{-5}, alpha=0.05). Interestingly, several known CMT genes achieved nominal p-values <0.05, serving as a 'positive control' for the ability to identify both risk and causative genes. We are currently performing molecular genetics and cell biology follow up studies and also working towards enlarging our sample size. In summary, statistical methods, traditionally reserved for more common phenotypes, are becoming increasingly available for rare disease genetics such as CMT and will help to comprehensively define the genetic architecture of complex rare neurodegenerative disorders.

171

Pilot study of population-based newborn screening for spinal muscular atrophy in New York state. *D.M. Kay[1], J.N. Kraszewski[2,3], C.F. Stevens[1], C. Koval[2], B. Haser[2], V. Ortiz[2], L. Cohen[4], R. Jain[1], S.P. Andrew[5], N.M. LaMarca[5], S. Dunaway Young[5], D.C. De Vivo[2,5], M. Caggana[1], W.K. Chung[2,6].* 1) Division of Genetics, Wadsworth Center, New York State Department of Health, Albany, NY; 2) Department of Pediatrics, Columbia University, New York, NY; 3) Department of Epidemiology, Columbia University, New York, NY; 4) Department of Pediatrics, Weill Cornell Medical College, New York, NY; 5) Department of Neurology, Columbia University, New York, NY; 6) Department of Medicine, Columbia University, New York, NY.

Spinal muscular atrophy (SMA) is a degenerative neuromuscular condition and is the most common genetic cause of death in children under age two years. SMA is an autosomal recessive disorder caused by a deletion in the Survival Motor Neuron 1 (*SMN1*) gene in the majority of affected individuals across race/ethnic groups. The *SMN1* exon 7 deletion can be readily detected using real-time qPCR. In late 2016, the FDA approved the first and only effective treatment for SMA, Spinraza, an antisense oligonucleotide designed to increase production of full length SMN protein. Because of the progressive loss of motor neurons in SMA, early detection and proactive treatment, ideally before symptom onset, will be necessary for optimal outcome. To assess the feasibility and utility of newborn screening for SMA, we validated an assay using dried blood spots and piloted screening for SMA in New York State using an informed consent model. During the first 12 months, we screened 3,826 newborns and returned heterozygous and homozygous *SMN1* deletion results to infants' families and their providers. The overall SMA carrier frequency was 1.5%. We identified one newborn with a homozygous *SMN1* deletion and two copies of *SMN2*, which strongly suggests the severe type 1 SMA phenotype. The infant was enrolled in the NURTURE clinical trial and was first treated with Spinraza at age 15 days. She is now age 12 months, meeting all developmental milestones, and free of any respiratory issues. Our pilot study demonstrates the feasibility of population-based screening, the acceptance by families (93% opt-in rate), and the benefit of newborn screening for SMA. We suggest that SMA should be considered for addition to the national recommended uniform screening panel (RUSP).

172

Recessive mutations in the fusiogenic protein myomaker cause Carey-Fineman-Ziter syndrome. *S.A. Di Gioia[1,5,6], S. Connors[10], N. Matsunami[12], J. Cannavino[14], M.F. Rose[1,2,5,9,15,26], N.M. Gilette[1,5], P. Artoni[1,5,6], N.L. de Macena Sobreira[16], W.M. Chan[1,5,6,22,26], B.D. Webb[21], C.D. Robson[3,7], C. Van Ryzin[17], A. Ramirez-Martinez[14], P. Mohassel[18,19], T. Hartman[23], I.M. Hayes[24], D.M. Markie[11], A. Swift[17], P.S. Chines[17], C.E. Speck-Martins[25], F.S. Collins[17,20], E.W. Jabs[16,21], C.G. Bönnemann[18,19], E.N. Olson[14], J.C. Carey[13], S.P. Robertson[10], I. Manoli[17], E.C. Engle[1,4,5,6,8,9,22,26], Moebius Syndrome Research Consortium.* 1) Department of Neurology, Boston Children's Hospital, Boston, MA; 2) Department of Pathology, Boston Children's Hospital, Boston, MA; 3) Department of Radiology, Boston Children's Hospital, Boston, MA; 4) Department of Ophtalmology, Boston Children's Hospital, Boston, MA; 5) F.M. Kirby Neurobiology Center, Boston Children's Hospital, Boston, MA; 6) Departments of Neurology, Harvard Medical School, Boston, MA; 7) Departments of Radiology, Harvard Medical School, Boston, MA; 8) Departments of Ophthalmology, Harvard Medical School, Boston, MA; 9) Medical Genetics Training Program, Harvard Medical School, Boston, MA; 10) Departments of Women's and Children's Health, University of Otago, Dunedin, New Zealand; 11) Department of Pathology, University of Otago, Dunedin, New Zealand; 12) Departments of Genetics,University of Utah School of Medicine, Salt Lake City, UT; 13) Departments of Pediatrics,University of Utah School of Medicine, Salt Lake City, UT; 14) Department of Molecular Biology and Neuroscience, and Hamon Center for Regenerative Science and Medicine, The University of Texas Southwestern Medical Center, Dallas, TX; 15) Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, U.S.A; 16) McKusick-Nathans Institute of Genetic Medicine, Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD; 17) Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 18) Neuromuscular and Neurogenetic Disorders of Childhood Section, National Institutes of Health, Bethesda, MD; 19) Neurogenetics Branch, National Institute of Neurological Disorders and Stroke,National Institutes of Health, Bethesda, MD; 20) Office of the Director, National Institutes of Health, Bethesda, MD; 21) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York City, NY; 22) Howard Hughes Medical Institute, Chevy Chase, MD 20815, U.S.A; 23) Department of Pediatrics, Dartmouth-Hitchcock Medical Center, Geisel School of Medicine, Hanover, NH; 24) Genetic Health Services New Zealand, Auckland City Hospital, Auckland, New Zealand; 25) SARAH Network of Rehabilitation Hospitals, Brasilia, DF; 26) Broad Institute of M.I.T. and Harvard, Cambridge, MA.

Carey-Fineman-Ziter syndrome (CFZS) was first described in 1982 in two siblings with marked bilateral facial weakness, Robin sequence (mandibular hypoplasia, hypoglossia, cleft palate), generalized muscle hypoplasia with hypotonia and mild proximal weakness, failure to thrive, delayed motor milestones, and scoliosis. To identify the genetic cause of CFZS, exome sequence data were generated and analyzed from the original CFZS family and independently from two other unrelated families with closely overlapping phenotypes. The affected members of all three families harbored compound heterozygous missense variants in *MYMK,* formerly known as *TMEM8C*. Targeted screening of more than 300 additional probands with various forms of congenital facial weakness revealed two additional families carrying *MYMK* compound heterozygous or homozygous variants. Five out of 6 families carried a common rare variant (p.Pro91Thr, rs776566597) annotated in EXAc with a cumulative MAF 0.0013. Haplotype analysis identified a common origin for this allele. *MYMK* encodes for Myomaker, a conserved plasma membrane protein required for myoblast fusion to form multinucleated myotubes in mouse, chick, and zebrafish. We hypothesized that the recessive *MYMK* missense variants cause CFZS through two hypomorphic or a combination of one hypomorphic and one null allele that reduce MYMK function and myoblast fusion below a threshold but not to zero. To address this hypothesis we ectopically expressed human *MYMK* alleles in HeLa and fibroblast cell lines. Alleles predicted to be hypomorphic had reduced membrane expression in HeLa cells but when overexpressed in fibroblasts retained the ability to fuse to myoblasts similar to the WT allele. Conversely, we did not observe membrane expression or retained fusiogenic activity with predicted loss-of-function alleles. To model the variants *in vivo,* we used CRISPR to generate tmem8c loss-of-function zebrafish and documented lack of muscle fusion in homozygotes. We then ectopically expressed the human wildtype and mutant alleles in the *tmem8c* null background. Wildtype human *MYMK* rescued the lack of fusion phenotype, while by contrast the predicted hypomorphic alleles only partially rescued and the predicted null alleles failed to rescue this phenotype. Collectively, these data establish that MYMK activity is necessary for normal muscle development and maintenance in humans, and expand the spectrum of congenital myopathies to include cell-cell fusion deficits.

173

Direct reconstruction of human genomes capturing highly divergent regions including MHC. *N.I. Weisenfeld, P.N. Shah, V. Kumar, S. Williams, C. Catalanotti, N. Keivanfar, D.M. Church, D.B. Jaffe.* 10x Genomics, Inc., Pleasanton, CA.

The ability to routinely reconstruct the genome of a patient would be a powerful tool for understanding their individual biology and disease state. Two major methods have been devised for this purpose. One compares sequence reads to a reference genome, reporting out the observed differences, while the other compares reads to one another, assembling the genome de novo. Reference-based methods excel when both alleles are similar to each other and to the reference, but struggle when one or both alleles differ significantly from it. Conversely, de novo methods excel at elucidating novel sequence, but nearly all such methods yield a collapsed representation of the genome in which parental alleles are merged into a single consensus sequence that matches neither parental haplotype. Thus the holy grail has been to reconstruct the patient's genome as a 'diploid assembly', in which the two alleles appear as completely separate sequences. In a few cases, this has been carried out, at great expense. We previously described an approach, Supernova™, for diploid, de novo assembly of individual genomes using 10x Linked-Read™ technology (PMID 28381613). Here we describe new algorithms that build upon this work, correctly reconstructing separate alleles, even for highly divergent regions such as MHC. Linked-Reads are barcoded short reads that each localize to a group of several long (~100 kb) molecules. These data are progressively assembled to greater degrees of resolution, using kmers, read pairs, and finally barcodes to create ~10^5 local assemblies, which are merged and then phased into separate alleles. Our new push-button assembly process runs locally or in the cloud. We tested our method using data from six cell lines, blood from a donor to the Human Genome Project (HGP), and a pseudo-diploid constructed by mixing DNA from two hydatidiform mole cell lines. These assemblies are highly-contiguous and phased at multi-megabase distances. For example, in each sample the entire 5 Mb MHC region is phased, with many complex duplications resolved. Using long read assemblies of the moles, we validate that a single phased scaffold covers ~97% of the MHC region across both moles. For the HGP sample, our assembly closely matches the sequence of finished clones that cover more than 10% of the genome. The N50 length of perfect matches is 25 kb. The accuracy, speed, simplicity and cost of our method make it possible to routinely study the biology of an individual genome.

174

Full resolution HLA typing of 273 individuals from deep whole-genome sequencing data enables genetic studies of human 6p21.3. *J. Reyna[1], N. Nariai[3], D. Jakubosky[2], E. Smith[1], K. Frazer[1,3].* 1) Department of Pediatrics and Rady Children's Hospital, University of California, San Diego, La Jolla, CA, USA; 2) Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, CA, USA; 3) Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA, USA.

The human leucocyte antigen (HLA) gene complex located in a 4 Mb region on chromosome 6p21.3 is the most variable region in the human genome and genetic variation at this locus is strongly associated with many autoimmune and infectious diseases. However, due to the genetic complexity of the locus, determining HLA types at full resolution (eight-digit) is challenging from whole genome sequence data. Therefore, a method that effectively identifies complex haplotypes from sequence data is needed to accurately determine HLA types at full resolution in order to examine the genetic basis of many immune-based human diseases. Further, as full resolution HLA types contain noncoding variants, their identification may facilitate the identification of variants that influence gene expression. We conducted HLA typing for HLA class I genes and class II genes of 273 individuals who participated in the iPSCORE (iPSC Collection for Omic Research) resource using high-coverage whole-genome sequencing data (52x on average). For HLA typing, we used HLA-VBSeq, which optimizes read alignment to reference HLA sequences in the IPD-IMGT/HLA database. We observed a high degree of diversity for both HLA class I and class II genes, identifying between 29 to 68 unique alleles for each of the six genes at full resolution. As the iPSCORE collection contains monozygotic twins and related individuals, we evaluated the HLA typing calls by calculating twin concordance and Mendelian inheritance error in trios. We observed 96.7% and 94.0% twin concordance rates, and 8.8% and 10.8% trio Mendelian error rates for HLA class I and II, respectively. These data show that HLA alleles can be consistently predicted at full resolution using HLA-VBSeq and high-coverage WGS data. We next tested whether the HLA types showed differential gene expression by estimating allele-specific gene expression levels for HLA class I genes in 128 iPSC RNA-seq samples from unrelated iPSCORE individuals. Reads were aligned to individual-specific reference cDNA sequences based on the determined HLA types for each sample. We show that allele-specific gene expression of HLA class I genes in iPSC were associated with HLA types of the corresponding individuals. Our results suggest that whole genome sequence data can be used to consistently identify full resolution HLA types and that these haplotypes can be associated with changes in gene expression levels.

175

The MHC Diversity in Africa Resource: A roadmap to understanding HLA diversity in Africa. *M.O. Pollard[1,2], A.J. Mentzer[3], T. Porter[1,2], A.T. Dilthey[6], C. Pomilla[1,2], S. Peacock[4], N. Careb[5], S. Lule[7], A. Diarra[8], N. van Niekerk[9,10] on behalf of the MDAP Investigators.* 1) Global Health and Populations, Human Genetics, Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 2) Department of Medicine, University of Cambridge, Cambridge, United Kingdom; 3) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK; 4) Histocompatibility and Immunogenetics Laboratory, Cambridge University Teaching Hospitals NHS Foundation Trust, Addenbrooke's Hospital; 5) Histogenetics Inc., Ossining, New York, USA; 6) National Institutes of Health, Bethesda, USA; 7) MRC/UVRI Uganda Research Unit on AIDS, P.O. Box 49, Entebbe, Uganda; 8) Centre National de recherche et de Formation sur le Paludisme, Burkina Faso; 9) Department of Science and Technology/National Research Foundation, Vaccine Preventable Diseases, University of the Witwatersrand, South Africa; 10) Medical Research Council, Respiratory and Meningeal Pathogens Research Unit, University of the Witwatersrand, South Africa.

The MHC Diversity in Africa Project (MDAP) aims to generate the largest HLA haplotype resource for medical genetics in Africa through characterising HLA diversity in African populations specifically including diverse indigenous populations. To achieve our goals, we have developed and validated methods for accurate HLA typing specifically aimed at African populations. This work will improve our understanding of immune related adaptive evolution, population history, response to vaccines and infectious diseases in this region, as well as to facilitate fine mapping of causative variants from GWAS. Here, we describe the MDAP resource, and key findings from the examination of more than 1500 individuals from 13 populations across Africa. HLA typing was carried out using multiple approaches, to assess consistency and accuracy of high resolution HLA types in African populations. We carried out PacBio full gene sequencing using GenDX primer based approaches for HLA class I and II among 125 samples from 5 populations (25 individuals per population), and Histogenetics MiSeq exon based typing for 1422 individuals. PacBio types were evaluated against SBT for validation. We also developed a laboratory and bioinformatics workflow for accurate typing both suitable for use in the clinic setting and for research using PacBio sequences; with our results showing that this method is highly concordant with our clinical types with fewer errors post validation. In the largest HLA reference panel from the African continent to date, we showed high levels of genetic diversity among the 13 populations in the HLA class I and II regions, both among African populations, and between Europeans and Africans. Of these, approximately three hundred are full length PacBio sequenced types, whilst the remainder are exon phased MiSeq derived HLA types. We also identified a large number of HLA haplotypes not previously discovered, both in class I and Class II MHC genes. We noted high levels of differentiation in population allele frequencies which in some cases is likely to be related to adaptive selection such as enrichment for the HLA B*53:01:01 allele in West Africa, which has previously been associated with resistance to malaria. Finally, we show that there are marked gains in HLA imputation accuracy that can be achieved by using diverse HLA panels from Africa, providing an important resource for medical genetics in Africa and globally.

176

GWAS of canker sores implicates Th-1 differentiation and signaling pathway and shared genetic architecture with inflammatory bowel disease. *F. Sathirapongsasuti, S. Pitts, D. Hinds, V. Vacic, R. Gentleman.* 23andMe Inc., Mountain View, CA.

Immune-mediated diseases (IMDs), which manifest along a spectrum of severity and prevalence from rare and severe to common and benign, often share common genetic components. Canker sore, or recurrent aphthous stomatitis (RAS), is a common condition characterized by mouth ulcers and largely unknown etiology. We surveyed 23andMe consented research participants and found that among 247,219 participants of European ancestry, 74% reported having had canker sores at least once in their lifetime and 34% within the past year, making it one of the most common IMDs. We performed the first genome-wide association study of canker sores and found 48 associated loci. The strongest associations include a SNP in a regulatory region upstream of the T cell-stimulating factor *IL12A* (rs17753641, OR=1.36, p=3.9e-165), an intronic variant in the immunoregulatory cytokine *IL10* (rs1518110, OR=1.21, p=3.2e-116), and an intronic variant in the infection-related *IFNGR1* (rs2234711, OR=0.92, p=2.4e-34). A number of the associated genes belong to the T-cell and Th1-cell differentiation and signaling pathways, including cytokines (*IL10, IL12A, IL12B, IL18R1, TNFSF15*), transcription regulator (*STAT4*), and TNF-β (*LTA*). Comparing the results with GWASes of other IMDs, we found that genetic architecture of canker sore is similar to that of other autoimmune diseases such as inflammatory bowel disease. Interestingly, some risk variants for other IMDs appear to be protective for canker sores, and vice versa. One of the strongest protective associations is a loss of function variant (LoF) in *NOD2* gene known to increase risk of Crohn's disease. We note an intriguing parallel between smoking and the NOD2 LoF in that while they increase the risk of Crohn's disease (OR_smoking=1.33, p=1.6e-35; OR_*NOD2*=2.07, p=5.3e-222), they are protective against canker sores (OR_smoking=0.66, p=9.2e-140; OR_*NOD2*=0.84, p=6.1e-48). Cigarette smoke extract has been found to delay NOD2 expression and negatively interact with the *NOD2* LoF in Crohn's disease. We replicated the interaction in both Crohn's (p=0.00016) and canker sores (p=0.0053) in the 23andMe cohort. This first GWAS of canker sores and its connection with other IMDs help elucidate the molecular pathways involved in this very common yet poorly understood condition.

177

Using exome sequencing to expand the genetic architecture of IBD. *M. Daly[1,2], M. Rivas[1,2,3], C. Stevens[2], B. Avila[1,2], J. Koskela[1,2,4], T. Ahmad[5], G. Atzmon[6,7], S.R. Brant[8], J. Cho[9], M. Farkkila[10], A. Franke[11], B. Glaser[12], K. Kontula[13], S. Kugathasan[14], D. McGovern[15], A. Palotie[4,16], J. Rioux[17], T. Segal[18], H. Sokol[19], D. Turner[20], R.K. Weersma[21], H. Winter[22], R. Xavier[23].* 1) Analytic and Translational Genetics, Massachusetts Gen Hosp, Boston, MA; 2) The Broad Institute of Harvard and MIT; 3) Biomedical Data Science, Stanford University; 4) Institute for Molecular Medicine Finland (FIMM); 5) University of Exeter; 6) Department of Human Biology, Faculty of Natural Science, University of Haifa, Haifa, Israel; 7) Department of Medicine and Genetics Albert Einstein College of Medicine, Bronx, NY 10461, USA; 8) Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, School of Medicine and Department of Epidemiology, School of Public Health, Johns Hopkins University, Baltimore, MD, USA; 9) Charles Bronfman Institute for Personalized Medicine Icahn School of Medicine at Mount Sinai; 10) Martti A Färkkilä, Helsinki University, Helsinki University Hospital, Clinic of Gastroenterology; 11) Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany; 12) Endocrinology and Metabolism Service, Hadassah-Hebrew University Hospital, Jerusalem Israel; 13) Helsinki Univ. Central Hospital, Finland; 14) Emory University; 15) F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90025; 16) Psychiatric & Neurodevelopmental Genetics Unit, Massachusetts Gen Hosp, Boston, MA; 17) Montreal Heart Institute and Université de Montréal, Montreal, Quebec, Canada, H1T 1C8; 18) Univ. College London; 19) Gastroenterology department, Saint-Antoine Hospital, APHP, Paris, France; 20) Shaare Zedek Medical Center Jerusalem, Israel; 21) Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands; 22) Pediatric Medical Services (MGHfC) / Pediatric Gastroenterology & Nutrition, Massachusetts Gen Hosp, Boston, MA; 23) Gastrointestinal Unit, Massachusetts Gen Hosp, Boston, MA.

Crohn's disease (CD) and ulcerative colitis (UC) are debilitating, inflammatory diseases of the gastrointestinal tract, collectively known as the inflammatory bowel diseases (IBD). Among complex diseases, genetics has been particularly successful in IBD, with genome-wide association studies (GWAS) over the past decade defining confirmed association to 250 gene loci. A handful of these associations have led to specific validated functional variants highlighting intracellular response to microbes and regulation of adaptive immunity in the pathogenesis of IBD. For the vast majority, however, the specific implicated gene and causal functional variants have not been identified, limiting near-term insights into pathogenesis and longer-term ability to convert associations into actionable therapeutic hypotheses. This limitation is commonplace – the emerging challenge for human genetics is no longer discovering genetic associations, it is deducing how identified genes and corresponding alleles exert influence on biology in health and disease. With support from the Helmsley Charitable Trust and partnership with IBD researchers around the world, we launched an exome sequencing initiative in 2014 with a goal to define the full allelic spectrum of protein-altering variation in genes associated to CD and/or UC, assess their role in clinical course and response to therapy, and to determine whether truncating variants confer risk or protection in each IBD gene in order to highlight opportune therapeutic targets. We have completed exomes of 13,000 IBD cases providing a high-resolution view of coding variation at each GWAS hit and demonstrated a convincing excess of rare exome signal. The cases are drawn from individual substudies focusing on isolated populations (Ashkenazi, Finnish, French-Canadian), admixed populations and clinical extremes, each providing unique opportunities for discovery. Early findings from the effort include novel protective truncating variants, the complete allelic series (including unique founder population alleles), non-additive inheritance models at known genes such as NOD2, overlooked low-frequency coding variants that explain GWAS hits and novel alleles for thiopurine-induced myelosuppression. The integration of low frequency and rare functional variants with GWAS is moving us closer to a complete genetic architecture of IBD. Association results for all variants are available as analyses are completed at ibd.broadinstitute.org. .

178

Genotype and phenotype analyses revealed novel susceptibility genes and new clinical classification for psoriasis. *B.-J. Feng[1,2], S. McCarthy[5], H. Li[2], K. Praveen[5], J. Walsh[3], J. Hawkes[1], M. Milliken[1], D.E. Goldgar[1,2], J.G. Reid[5], J.D. Overton[5], F. Dewey[5], C. Gonzaga-Jauregui[5], S.L. Guthery[4], K. Callis-Duffin[1], G.G. Krueger[1].* 1) Department of Dermatology, University of Utah, Salt Lake City, Utah, USA; 2) Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA; 3) Department of Rheumatology, University of Utah, Salt Lake City, Utah, USA; 4) Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA; 5) Regeneron Genetics Center, Regeneron Pharmaceuticals Inc., Tarrytown, NY, USA.

Psoriasis (Ps) is a complex multigenic disorder, with heritability estimated to be 60-90%. Although several large-scale genome-wide association studies (GWAS) and linkage studies have been performed, there is still a large proportion of the familial relative risk not explained by the known loci. To search for the missing heritability of Ps, we performed whole exome sequencing (WES) and array genotyping of 1133 Ps cases and 1176 healthy controls. Among the Ps cases there were 73 pedigrees ascertained through the Utah Population Database, a genealogy database of 8 million individuals with up to 12-generation pedigrees. We jointly analyzed the pedigrees and case-control samples by the framework PERCH, which quantitatively integrates the deleteriousness of variants, quality of variant calls, co-segregation of variants with disease within pedigrees, association of variants with disease among case-controls, and the connection of each gene with known Ps genes within a gene-gene interaction network. The results demonstrated several candidate genes for Ps but none of them reached genome-wide significance. To create a homogeneous subset of samples, we then performed a latent class analysis (LCA) to identify groups of patients with similar clinical phenotypes. This analysis yielded three classes. Class 3 was associated with a younger age of Ps initiation, more severe disease, and trauma or injury associated disease onset or worsening. Familial risk analysis further showed that this class of patients had the highest level of familial aggregation. Our second round of genomic analysis restricted to this group of patients resulted in the identification of *ADAMTS9* with genome-wide significance (p=0.0000008), a locus that was previously reported to have a higher deletion rate in psoriatic arthritis than healthy controls. Class 2 Ps showed a moderate familial clustering and is not sensitive to sunlight treatment. Pedigree-informed genomic analyses identified novel candidate genes segregating in families. Network analysis further suggested a role for some candidate genes in epithelial immune response. These results highlight the importance of phenotypic classification of cases in a complex disease research and the possibility that new susceptibility loci for Ps are discoverable using a combination of family-based and cohort approaches such as PERCH. They also suggest a new classification scheme for Ps that is relevant to disease etiology and clinical management.

179

Functional characterization of *BRIP1* missense alleles: Distinguishing cancer susceptibility alleles from benign polymorphisms. *C. Moyer[1], J. Gillespie[1], R. Doberstein[1], J. Ivanovich[2], G.M. Collett[3], M. Harrell[3], E. Swisher[3], P.J. Goodfellow[1].* 1) Ohio State University, James Comprehensive Cancer Center, Department of Obstetrics and Gynecology, Columbus, OH; 2) Indiana University Medical Center, Indianapolis, IN; 3) University of Washington, Department of Obstetrics and Gynecology, Seattle, WA.

Purpose: BRCA1 interacting protein C-terminal helicase 1 (*BRIP1*) is a member of the Fanconi anemia pathway. Individuals homozygous for loss-of-function mutations present with Fanconi anemia. Cells lacking BRIP1 have impaired DNA interstrand crosslink repair. *BRIP1* is the third most frequently mutated ovarian cancer (OvCA) susceptibility gene with stop, splice and frameshift mutations identified in nearly 1.4% of cases. *BRIP1* may also contribute to breast cancer (BrCA) risk, particularly among those patients who develop disease at an early age. The role that *BRIP1* missense mutations play in cancer risk is poorly understood. We identified a cluster of rare missense variants in *BRIP1* in the C-terminal helicase domain. The objective of this study is to functionally characterize rare *BRIP1* missense variants identified in OvCA and early-onset BrCA patients, focusing on the altered protein's ability to repair interstrand crosslink damage. Methods: Next generation sequencing for 1543 OvCA patients and 2143 early-onset BrCA patients was used to identify candidate missense alleles. To determine the function of missense alleles, *BRIP1*-null and mutant HeLa clones were created using CRISPR-Cas9 genome editing of the C-terminal helicase domain. Sensitivity to interstrand crosslinking agent, Mitomycin C, was determined by cell growth assay and karyotype. This robust, isogenic system allows us to introduce vectors containing wild type *BRIP1* or candidate mutants into null and wildtype clones via transient transfection to determine loss of function and dominant negative variants. Key Results: Sequencing revealed 12 OvCA and 13 early-onset BrCA patients carry a rare (<0.05% minor allele frequency) or novel, constitutional, missense variant in the C-terminal helicase motifs IV-VI of *BRIP1*. Among the identified alleles, Q540L, I691L, Q740H and A745T overlap between the OvCA and BrCA cohorts. Novel variants, also clustered in this helicase domain, are predicted to alter protein function. HeLa clones with deletion of a single amino acid at position A778 are as sensitive to interstrand crosslinking agents (Mitomycin C treatment) as *BRIP1*-null cells, emphasizing the importance of this helicase domain. These findings suggest that novel, missense variants within the helicase domain of *BRIP1* may confer risk for both breast and ovarian cancer. Functional testing for 21 candidate missense variants in the C-terminal helicase domain is ongoing.

180

Functional characterization of variants of uncertain significance in BRCA2: Fifty shades of BRCA2 deficiency. *R.L.S. Mesman[1], F.M.G.R. Calleja[1], G. Hendriks[1], P. Devilee[1], C.J. van Asperen[2], H. Vrieling[1], M.P.G. Vreeswijk[1].* 1) Human Genetics, LUMC, Leiden, Zuid-Holland, Netherlands; 2) Clinical Genetics, LUMC, Leiden, Zuid-Holland, Netherlands.

During the last 20 years, genetic testing to identify pathogenic variants in *BRCA1* and *BRCA2* has become routine clinical practice. However, quantitative assessment of the cancer risk associated with the occurrence of a variant turned out to be less straight forward as anticipated. For many intronic and missense variants lack of clinical and family data prevents reliable estimation of their cancer risk (variants of uncertain significance (VUS)). Furthermore, recent findings challenge the existing dogma that truncating variants and variants in the canonical splice sites are always associated with high cancer risk. We have optimized and validated a previously developed (Kuznetsov et al., 2008; Hendriks et al., 2014) mouse embryonic stem cell (mES) based model system that allows functional analysis of all types of BRCA2 variants, including variants that may affect RNA splicing. The procedure involves the generation of a desired mutation in *BRCA2* present in a bacterial artificial chromosome (BAC) and its subsequent introduction into conditionally knock out *mBRCA2* mES cells. The performance of the assay was validated using a series of clinically proven pathogenic (Class 4/5; n=15) and neutral (Class 1/2; n=20) missense variants. Of 60 clinically relevant missense VUS, eight variants did not complement the lethality conferred by removal of the endogenous *mBrca2* gene. The ability to perform homologous recombination, one of the key functions of BRCA2, varied among *BRCA2* variants that were capable of complementing *mBrca2* deficiency between 30-120% of wild type. Recently, a number of the hypomorphic *BRCA2* variants were shown to be associated with moderate risks of breast cancer in a large scale case control study (Shimelis et al., 2017). To assess whether naturally occurring *BRCA2* mRNA isoforms might be able to rescue the deleterious effects predicted to occur for exon deletions or nonsense variants, we are currently assessing the functional implications of single exon deletions and nonsense variants. Preliminary data indicate for some exons that nonsense variants are not only not lethal but still display considerable levels of homologous recombination. Ongoing efforts are now focused on establishing the relationship between functional results and the associated cancer risk. .

181

Incorporation of tumor RNA expression data in pathogenicity classification of germline variants. *C. Kesserwan[1], D. Hedges[2], S. Newman[2], K. Hamilton[1], R. Mcgee[1], E. Quinn[1], R. Nuccio[1], J. Valdez[1], M. Rusch[2], S. Foy[2], J. Nakitandwe[3], L. Harrison[1], A. Ouma[1], S. Hines-Dowell[1], S. Shurtleff[3], E. Azzato[3], D. Ellison[3], J. Downing[3], J. Zhang[2], K. Nichols[1].* 1) Oncology, St. Jude Children's Research Hospital, Memphis, TN; 2) Computational Biology, St. Jude Children's Research Hospital, Memphis, TN; 3) Pathology, St. Jude Children's Research Hospital, Memphis, TN.

INTRODUCTION: The adoption of high throughput sequencing has sharply increased the number of germline variants requiring classification, the majority of which lack experimental evidence to ascertain their functional significance. Genomes for Kids (G4K) is a research study addressing the feasibility of integrated clinical whole genome (WGS), whole exome (WES) and transcriptome sequencing of tumor samples and WGS and WES of germline samples from oncology patients at St. Jude Children's Research Hospital. Paired analysis of tumor data holds potential to enhance germline variant interpretation and resolve variants of uncertain significance (VUS). Here we assess the extent to which tumor transcriptome data facilitated classification of germline variants proximal to splice junctions. **METHODS**: 248 germline samples were analyzed to date for mutations in 63 cancer predisposition genes. Variants meeting quality and population frequency criteria were adjudicated for pathogenicity following ACMG (2015) guidelines. To examine impact of tumor transcriptome data on germline variant classification, we selected variants falling within +/- 10bp of canonical splice junctions. Out of 42 variants meeting these criteria, 36 were within samples with accompanying tumor RNA-sequencing data. We re-examined these variants to assess the impact of RNA-sequencing data on classification outcomes. **RESULTS:** Of the 36 variants evaluated, 13 provided splice profile evidence concordant with final classification. For 21 variants, RNA-sequencing data were insufficient to draw conclusions. Classifications were significantly impacted in two instances: *BAP1* (c.256-3C>A) and *APC* (K150R). For the *BAP1* variant, evidence of intron retention among variant alleles resulted in a Likely Pathogenic (LP) classification. For the *APC* variant, interpretation of LP from an external laboratory was called into question due to lack of evidence for aberrant splicing in tumor RNA-seq, supporting our VUS categorization and prompting the external laboratory to re-classify the variant as VUS. **CONCLUSION:** The availability of tumor RNA-seq can impact interpretation of germline variants. Although the total number of cases clearly impacted (2/36; 5.5%) was modest, the distinction between reporting a germline VUS and an LP variant in patients with splice junction proximal variants cannot be overstated. These results support ongoing efforts to integrate tumor RNA-seq data into germline variant interpretation guidelines.

182

Splicing mutation risk analysis in hereditary breast and ovarian cancer exomes. *E.J. Mucaki[1], B.C. Shirley[2], S.N. Dorman[1], P.K. Rogan[1,2].* 1) Biochemistry, University of Western Ontario, London, Ontario, Canada; 2) CytoGnomix Inc, London, Ontario, Canada.

Genetic testing of patients with inherited cancer frequently reveals variants of unknown significance (VUS). We have presented an Information Theory (IT) framework to predict and prioritize coding and non-coding VUS in hereditary breast and ovarian cancer (BRCA) patients, including effects on mRNA splicing[1,2]. We investigated the exome wide distribution of predicted mRNA splicing mutations in a large BRCA cohort. Predicted splicing mutations in IT-based splicing analysis of all variant data from AmbryShare BRCA exome (n=11,416; with 1.2 million VUS) and the control genome Aggregation Databases (gnomAD; n=138,632) were identified using the Shannon splicing mutation software pipeline[3]. IT-flagged variant frequencies (decreasing R_i values [in bits] of either leaky or inactivated natural splice sites [ΔR_i >4 bits and $R_i \leq 1.6$] or strengthened cryptic splices sites with an R_i exceeding that of adjacent natural sites) were compared for each gene using odds ratios (OR). OR^A is defined as the ratio of frequencies of the same flagged variants in a gene in AmbryShare relative to gnomAD. OR^P is based on the ratio of frequencies of all flagged variants in a gene in AmbryShare relative to all flagged variants in that gene in gnomAD. A greater number of IT-flagged variants were present in AmbryShare than in gnomAD among 2012 genes with severe splicing mutations. Increasing the ΔR_i threshold disproportionally decreases the number of flagged variants in gnomAD due to fewer severe splicing mutations. Variants that abolish natural splice sites flagged known inherited breast cancer genes with respectively increased OR^A and OR^P in *ATM* (493, 407), *BARD1* (407, 407), *BRCA1* (19, 14), *BRCA2* (54, 54), *CDH1* (549, 549), *MLH1* (303, 303), *MUTYH* (95, 11), and *PALB2* (233, 116). Other flagged breast cancer-related genes with high OR include *AAMP*, *C1QTNF6*, *CDK3*, *FOLR1*, *PRLR*, *RAD50*, *RING1*, *S100A2*, *SRGN*, *TMSB10*, *TYRO3*, and *VIM.* Notable highly mutated genes from other cancers include *GKN1* (gastric), *C1orf61* (hepatocellular), *CREM* (prostate), *PNKP* (multiple)*, PPP1CA* (gastric) and *ZFAND2B (*myeloid). Flagged genes not known to be linked to cancer include *ATP1A4*, MFF, *PACSIN1*, *PTS*, and *USH1C.* Severe splicing mutations occur more frequently in inherited and somatic breast cancer genes as well as in other genes in BRCA populations.

[1]Mucaki et al. BMC Med. Genom. 9:19, 2016; [2]Caminsky et al. Hum. Mut. 37:640, 2016; [3]Shirley et al. Genom. Prot. Bioinf. 11:75, 2013.

183

Early diagnosis of lung cancer: Identification of miRNAs in the blood as non-invasive biomarkers. *L. Shi[1,2], B. Song[1,2], Z. Yi[1], W. Zhang[1].* 1) Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Horace Mann School, Bronx, NY.

In 2015, lung cancer was responsible for 1.69 million deaths, making it a leading cause of death worldwide; this is partially due to late diagnosis. As such, early diagnosis is crucial and can allow for much higher chances of curing patients. microRNA (miRNA), which are non-coding RNA molecules that play a role in the silencing and post-transcriptional regulation of gene expression, have been used as stable biomarkers for diagnosis or prognosis of various diseases. Blood miRNA profiling of lung cancer has been researched in previous studies, but a consensus signature has not been reached. This study sought to determine a common miRNA signature that can function as non-invasive biomarkers for an accurate early cancer diagnostic from a simple blood test. We applied a bioinformatic and statistical approach to analyze blood miRNA expression profiles of a total 552 individuals (373 cases vs 179 controls) in 5 datasets. Out of the total 1353 miRNAs, robust meta-analysis yielded 397 meta-signatures at an adjusted p-value less than 0.05. By searching for these miRNAs in 2 independent blood datasets and 7 tumor tissue datasets in order to identify pathogenetic miRNAs associated with lung cancer, we finally identified 22 miRNAs that were significantly dysregulated ($p<0.05$) in most of these datasets. We found that hsa-mir-210, an important regulator of the cellular response to hypoxia, showed particular prevalence across the datasets, with significant decreased expression in blood and increased expression in tumor tissues across all validation datasets. Using the logistic regression model, we observed that these 22 miRNAs can be used to accurately detect lung cancer with an average area under the curve (AUC) of 0.853. More importantly, they are also significantly associated with overall survival in lung cancer patients (log rank p value of 0.000335, 0.0258, and 0.00332 in three tissue miRNA datasets), suggesting that these miRNAs have both diagnostic and prognostic values for lung cancer. Therefore, the 22 blood miRNAs identified in this study can be employed as non-invasive markers in the early diagnosis of lung cancer patients, allowing for cancer patients to be treated before metastasis in order to improve survival rates.

184

Assessing the feasibility of using whole genome mate pair sequencing in detecting diagnostic/prognostic chromosomal abnormalities seen in patients with acute myeloid leukemia. *U. Aypar[1], G. Vasmatzis[2,3], S. Johnson[2], J. Smadbeck[2], S. Smoley[1].* 1) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN; 2) Center for Individualized Medicine - Biomarker Discovery, Mayo Clinic, Rochester, MN; 3) Department of Molecular Medicine, Mayo Clinic, Rochester, MN.

Acute myeloid leukemia (AML) is the most common form of acute leukemia affecting both children and adults. Once recognized as a single disease, the World Health Organization now groups AML into eight subtypes based on genetic abnormality. This is based on the ability of recurrent genetic abnormalities to predict response to therapy, relapse risk and overall survival. Currently most testing of AML patients occurs by karyotype analysis, FISH or PCR. We propose using a next generation sequencing approach to better classify and accurately detect abnormalities associated with AML. To assess the feasibility of using whole genome mate pair sequencing (MPseq) to test patients with AML/MDS, 41 samples previously tested by conventional karyotyping and/or FISH were run with our MPseq assay. Samples were chosen based on reason for referral and were not chosen based on reported abnormalities. DNA was processed using the Illumina Nextera Mate Pair library preparation kit, multiplexed at 2 samples/lane and sequenced on the Illumina HiSeq with an average bridged coverage of 55x. Data was aligned to the reference genome using BIMAv3 and abnormalities were identified using SVAtools, both in-house developed tools. A targeted analysis approach was utilized using a well-defined list of recurrent chromosomal abnormalities associated with known diagnostic and prognostic significance as well as targeted therapies in AML. The karyotype, FISH and MPseq results were compared to determine concordance. FISH and/or karyotype detected a total of 49 abnormalities targeted by our AML panel. MPseq detected 43/49 of those abnormalities. Of the 6 targets that were not identified by MPseq, all were seen in less than 13% of the sample based on FISH or 2/20 cells (10%) by karyotype analysis. In addition, MPseq identified additional abnormalities not seen by karyotype or FISH and further characterized several complex abnormalities. In conclusion, whole genome mate pair sequencing is a feasible methodology to assess diagnostic patients with AML. Due to the limitations in detecting very low level abnormalities, MPseq would not be recommended for follow-up post therapy or minimal residual disease testing unless the coverage is increased with deeper sequencing.

185

Identifying complex traits under polygenic selection in the UK Biobank. *X. Liu, P.R. Loh, L. O'Connor, A. Schoech, S. Gazal, A.L. Price.* Epidemiology, Harvard Chan TH School of Public Health, Boston, MA.

The genetic architecture of human complex traits is highly polygenic, motivating efforts to detect polygenic selection involving a large number of loci (Turchin et al 2012 Nat Genet, Robinson et al. 2015 Nat Genet). A further goal is to quantify the genetic component of population differences in phenotype. We developed a method that identifies complex traits under polygenic selection and quantifies the genetic component of population differences in phenotype by analyzing genome-wide summary association statistics together with principal component (PC) SNP loadings and LD information from target or reference samples. Our method estimates the genetic component of observed phenotypic correlations along top PCs by leveraging the non-zero correlation between PC SNP loadings and causal effect sizes, which are estimated by multiplying marginal effect size estimates by the inverse of a regularized, banded LD matrix (Yang et al. 2012 Nat Genet). Simulations using UK Biobank genotypes, PCs and LD information showed that our method provides unbiased estimates of the genetic component of phenotypic correlations along top PCs. We analyzed summary association statistics for 12 complex traits, computed from ~113k British-ancestry samples from the UK Biobank. We determined that skin pigmentation, hair pigmentation, height and BMI showed clear evidence of polygenic selection. In particular, genetic effects explained the bulk of the phenotypic correlation between PC1 and skin pigmentation (95% CI: 80-167%, of 5.9% phenotypic correlation), hair pigmentation (95% CI: 55-161%, of 5.5% phenotypic correlation), height (95% CI: 46-94%, of 9.8% phenotypic correlation) and BMI (95% CI: 98-245%, of 3.2% phenotypic correlation). The very small amount of drift reflected in PC1 (corresponding to F_{ST} less than 0.0002) implies that these findings are best explained by polygenic selection. In conclusion, we have identified evidence of polygenic selection for several UK Biobank traits and have specifically shown that genetic effects account for the bulk of observed phenotypic correlations with PC1 for these traits.

186

Widespread signatures of negative selection in the genetic architecture of human complex traits. *J. Zeng[1], R. de Vlaming[2,3], Y. Wu[1], M. Robinson[1,4], L. Lloyd-Jones[1], L. Yengo[1], C. Yap[1], A. Xue[1], J. Sidorenko[1], A. McRae[1], J. Powell[1], G. Montgomery[1], A. Metspalu[5], T. Esko[5], G. Gibson[6], N. Wray[1,7], P. Visscher[1,7], J. Yang[1,7].* 1) Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia; 2) Department of Complex Trait Genetics, VU University, Amsterdam, HV, The Netherlands; 3) Erasmus University Rotterdam Institute for Behavior and Biology, Rotterdam, PA, The Netherlands; 4) Department of Computational Biology, University of Lausanne, Lausanne, Switzerland; 5) Estonian Genome Center, University of Tartu, Tartu, Estonia; 6) School of Biology and Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, GA; 7) Queensland Brain Institute, The University of Queensland, Brisbane, QLD, Australia.

Estimation of the joint distribution of effect size and minor allele frequency (MAF) for genetic variants is important for understanding the genetic basis of complex trait variation and can be used to detect signature of natural selection. We develop a Bayesian mixed linear model that simultaneously estimates SNP-based heritability, polygenicity (i.e. the proportion of SNPs with nonzero effects) and the relationship between effect size and MAF for complex traits in conventionally unrelated individuals using genome-wide SNP data. We apply the method to 28 complex traits in the UK Biobank data (N = 126,752), and show that on average across 28 traits, 6% of SNPs have nonzero effects, which in total explain 22% of phenotypic variance. We detect significant ($p < 0.05/28 = 1.8\times10^{-3}$) signatures of natural selection for 23 out of 28 traits including reproductive, cardiovascular, and anthropometric traits, as well as educational attainment. We further apply the method to 27,869 gene expression traits (N = 1,748), and identify 30 genes that show significant ($p < 2.3\times10^{-6}$) evidence of natural selection. All the significant estimates of the relationship between effect size and MAF in either complex traits or gene expression traits are consistent with a model of negative selection, as confirmed by forward simulation. We conclude that natural selection acts pervasively on human complex traits shaping genetic variation in the form of negative selection.

187

High-throughput inference of pairwise coalescent times identifies signals of selection and enriched disease heritability. *P. Palamara[1,2], J. Terhorst[3], Y. Song[3], A. Price[1,2].* 1) Harvard School of Public Health, Boston, MA; 2) Broad Institute of Harvard and MIT, Cambridge, MA; 3) University of California, Berkeley, Berkeley, CA.

Interest in reconstructing demographic histories has motivated the development of tools to estimate locus-specific pairwise coalescent times from whole-genome sequence (WGS) data, including PSMC (Li & Durbin 2011 Nature) and SMC++ (Terhorst et al. 2017 Nat Genet). Here, we developed a new tool, ASMC, that requires only SNP array data and is orders of magnitude faster than previous methods. We determined via coalescent simulations that ASMC is nearly as accurate (r^2=0.87) as methods that require WGS data (r^2=0.95), and 2-4 orders of magnitude faster than previous methods when WGS data is available. We were thus able to apply ASMC to 113,851 phased British samples from UK Biobank, aiming to detect recent positive selection by identifying loci with unusually high density of very recent coalescent times. We detected 12 genome-wide significant signals, including 6 loci with previous evidence of positive selection (including LCT, HLA and TLR) and 6 novel loci (including the STAT4 autoimmune disease locus), consistent with coalescent simulations showing that our approach is well-powered to detect recent selection on standing variation. We also applied ASMC to Genome of the Netherlands WGS data (N=769) to detect background selection at deeper time scales and finer genomic resolution, relying on the fact that loci under background selection have smaller effective population size and lower average coalescent time. As expected, we observed highly significant correlations between average coalescent time and other measures of background selection (e.g. McVicker B-statistic, r=-0.28; nucleotide diversity, r=0.50). We investigated whether this signal translated into an enrichment in disease and complex trait heritability by analyzing summary association statistics from 20 independent traits (average N=84K) using stratified LD score regression (Finucane et al. 2015 Nat Genet). Our average coalescent time annotation was strongly enriched for heritability ($p=6\times10^{-106}$) in a joint analysis with annotations from the baselineLD model (Gazal et al. biorxiv), meta-analyzed across traits. We detected a 0.25 conditional increase in per-SNP heritability per 1 s.d. increase in our annotation, with 4x more heritability in SNPs in the top 20% of the annotation compared to the bottom 20%, the largest effect among all annotations related to background selection. These results underscore the widespread effects of background selection on disease and complex trait heritability.

188

Polygenic selection underlies evolution of brain structure volumes and behavioral traits. *B.E. Stranger[1,2,3], E.R. Beiter[4], E.A. Khramtsova[1,2], C. Van Der Merwe[5], E.R. Chimusa[5], C. Simonti[6], D.J. Stein[7], J.A. Capra[6,8], J.A. Knowles[9], P. Straub[6,10], L.K. Davis[6,10,11].* 1) Department of Medicine, University of Chicago, Chicago, IL; 2) Institute of Genomics and Systems Biology, University of Chicago, Chicago, IL; 3) Center for Data Intensive Science, University of Chicago, Chicago, IL; 4) Department of Biology, Washington University, St. Louis, MO; 5) Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa; 6) Vanderbilt Genetics Institute; Vanderbilt University Medical Center, Nashville, TN; 7) Department of Psychiatry and MRC Unit on Risk & Resilience in Mental Disorders, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa; 8) Department of Biological Sciences, Vanderbilt University, Nashville, TN; 9) Department of Cell Biology, State University of New York Downstate Medical Center, Brooklyn, New York; 10) Division of Medical Genetics, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; 11) Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN.

Inter-individual variation in neuropsychiatric and behavioral traits is present across diverse human populations and has persisted through history. Many characteristics of psychiatric disorders have led to the question of how risk alleles have persisted throughout history. One hypothesis is that neuropsychiatric traits have experienced weak polygenic adaptation-directly or through selection on correlated traits. Given recent advances in detecting polygenic adaptation, we evaluated evidence for selection across twenty-five complex traits: ten neuropsychiatric disorders, three personality traits, total intracranial volume, seven subcortical brain structure volume traits, and four complex traits with no known neuropsychiatric associations. We tested trait-associated variants for evidence of rapid evolution since divergence from Neanderthal occurring ~600 kya (Neanderthal selective sweep score, NSS), partial sweeps which could occur as early as the first divergence between modern human populations ~150 kya (extreme population differentiation, F_{st}), classical hard sweeps dating to ~25-30 kya (integrated haplotype scores, iHS), ancient polygenic selection within the past ~30,000 years (Qx scores), and very recent polygenic selection within the past 2,000 years (trait singleton density scores, tSDS). SNPs associated with schizophrenia, neuroticism, and intracranial brain volume are enriched in regions of the genome under selection since divergence from Neanderthal (p_{SCZ} = 0.002, p_{NEU} < 0.002, p_{ICV} = 0.002). Variants associated with schizophrenia, extraversion, subjective well-being, hippocampus volume, and putamen volume are enriched for signatures of ancient polygenic adaptation (p_{SCZ} < 0.001, p_{EXT} = 0.001, p_{SWB} < 0.001, p_{HIP} < 0.001, p_{PUT} < 0.001). Finally, signatures of very recent polygenic adaptation are enriched among schizophrenia-protective alleles and total intracranial volume-increasing alleles, but putamen, amygdala, and pallidum volume-decreasing alleles alleles (p_{SCZ}=0.002, p_{ICV} = 0.002, p_{PUT} < 0.001, p_{AMY} < 0.001, p_{PAL} < 0.001). No trait displayed evidence of classical hard sweeps. Our results suggest that alleles associated with neuropsychiatric, behavioral, and brain volume phenotypes have experienced both ancient and recent polygenic adaptation. Our results provide the first genome-wide evidence from humans in support of the mosaic theory of brain evolution wherein individual subcortical brain structure volumes are able to evolve independently. .

189

An approximate full-likelihood coalescent method for detecting genomic sites under selection. *A.J. Stern[1], R. Nielsen[1,2].* 1) Dept. of Integrative Biology, UC Berkeley, Berkeley, CA., USA; 2) Dept. of Statistics, UC Berkeley, Berkeley, CA., USA.

A major goal of genetics research is to use polymorphism data to identify selected sites throughout the genome with high statistical power. Determining which variants have been subject to selection can reveal important functions such as implications in human disease. Furthermore, identifying selected sites helps us better understand the effects of selection on neighboring genetic diversity via selective sweeps. While there exist numerous methods to detect signatures of selective sweeps from polymorphism data, these methods are limited in several ways: lack of a generative model, conflation of selection with factors such as demography, and restricted power outside of detecting the particular type of sweep to which an individual method is tuned. To this end, we present an approximate full-likelihood test for selection using importance sampling of the latent ancestral recombination graph (ARG). Using simulated data, we show that asymptotically under reasonable conditions, our approach gives nearly optimal power to detect selective sweeps. We apply our method to inferring selection strength and the genomic location of the selected allele to the aforementioned dataset. We also apply our method to infer the time that a selective sweep fixed in both simulated data and the FOXP2 gene in humans.

190

Fine-mapping the favored mutation in a positive selective sweep. *A. Akbari[1], A. Iranmehr[1], M. Bakhtiari[2], S. Mirarab[1], V. Bafna[2].* 1) Electrical and Computer Engineering University of California San Diego; 2) Computer Science and Engineering University of California San Diego.

Motivation. Methods that scan population genomics data to identify signatures of selective sweep have been actively developed, but mostly do not identify the specific favored mutation. The selective sweep signal can extend to large, linked regions, as far as 1Mbp on either side of the favored allele. These 'soft-shoulders' of sweeps are helpful in identifying the region under selection, but make it harder to pinpoint the favored mutation. **Method**. We present a method, iSAFE (integrated Selection of Allele Favored by Evolution), that uses coalescent-based signals and a boosting approach to pinpoint the favored mutation within a 5Mbp around the region under selection. The iSAFE technique is motivated by boosting with weak classifiers and exploits the soft shoulders. **Result**. iSAFE was tested on simulated data and known sweeps in human populations using the 1000GP data. In extensive simulations, the median iSAFE rank of the true favored mutation is 5 out of ~20,000 candidate variants in 5Mbp. Table 1 shows the excellent performance of iSAFE on 8 well-characterized sweeps with known favored mutation. We also examined other sweeps, with some evidence for the favored mutation, and identified previously unreported mutations as being the favored. For example, GRM5-TYR region is associated with light skin pigmentation and known to be under selection in CEU population. iSAFE ranks mutation rs672144 at the top and very well separates it from rest of the mutations in 5Mbp around it. Interestingly, this variant was the top-ranked mutation not only in CEU (p-value = 1.3e-8), but also in EUR, EAS, AMR, and SAS (p-value << 1.3e-8). It may not have been previously reported because it is near fixation (Frequency > 0.94) in all populations of 1000GP except for AFR (Frequency = 0.27). **Software and preprint.** https://github.com/alek0991/iSAFE.

Table 1: iSAFE on 8 well characterized selective sweeps in 5Mbp (~20,000 SNPs).

Gene	Target Population	Candidate SNP ID	Frequency	Selective Advantage	iHS Rank	iSAFE Rank	iSAFE P-value
SLC24A5	CEU	rs1426654	1	Light skin pigmentation	24	1	<1.3e-8
EDAR	CHB+JPT	rs3827760	0.87	Hair and teeth	512	1	<1.3e-8
LCT/MCM6	FIN	rs4988235	0.59	Lactase persistence	205	1	<1.3e-8
TLR1	CEU	rs5743618	0.77	Sepsis, leprosy, tuberculosis	2963	1	1.00E-5
ACKR1/ DARC	YRI	rs2814778	1	Malaria resistance	1591	1	2.80E-5
ABCC11	CHB+JPT	rs17822931	0.93	Cold climate, earwax, body odour	106	2	<1.3e-8
HBB	YRI	rs334	0.14	Malaria resistance	19153	4	1.60E-4
G6PD	YRI	rs1050828	0.21	Malaria resistance	268	13	7.30E-6

191

In their own words: Adolescent attitudes about deferring genetic testing for adult-onset conditions. *A. Rahm[1], L. Bailey[1], O. D'Accordo[2], Y. Munishor[2], C. Miller[2], E. Davidson[2], L. Hercher[2], A.J. Young[5], K. Pulliam[3], H. Zhang[3], M. Dougherty[3,4], M. Williams[1].* 1) Genomic Medicine Institute, Geisinger Health System, Danville, PA; 2) Joan H. Marks Graduate Program in Human Genetics, Sarah Lawrence College, Bronxville, NY; 3) Education Dept, ASHG, Bethesda, MD; 4) Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO; 5) Biomedical and Translational Informatics Institute, Geisinger Health System, Danville, PA.

Introduction: When to test children for adult-onset conditions is an ongoing issue in genetics, and information is lacking on the attitudes and opinions of adolescents themselves. **Methods:** Essays submitted by 9th-12th grade students to the American Society of Human Genetics' 2016 national DNA Day Essay Contest were analyzed for adult condition discussed and why testing should be deferred to adulthood or not. Demographic information including gender, grade, school type, and location were also analyzed. Summary statistics and univariate analyses were conducted for variables associated with choice to defer/not-defer testing. Additional thematic coding for reasons to defer/not-defer was conducted. **Results:** 1241 student essays were submitted from 44 U.S. states (87%) and other countries (13%). Over 100 adult-onset conditions were discussed by students; most commonly discussed conditions were Huntington's disease (38%), BRCA-related breast or ovarian cancer (16%), and Alzheimer's disease (9%). Overall, students were evenly split whether they believed testing should be delayed until adulthood or not; however, more agreed to defer testing for Alzheimer's (64%) or Huntington's (62%) than BRCA (46%). Disease chosen was significantly associated with agreement to defer testing, as was whether the condition was actionable or non-actionable. Agreement to defer testing was not associated with school type (public/private), grade level, personal experience with condition, or geographic variables related to socioeconomic status. Initial analysis of reasons to defer testing suggest a close match to reasons cited in clinical guidelines. Additional thematic analyses regarding reasons for choosing to defer/not defer are ongoing. **Conclusions**: Adolescents have opinions about learning genomic risk information for adult-onset conditions and their choices differ by condition. They also display nuanced reasoning regarding choice to defer predictive genetic testing. This research contributes to our understanding of healthy adolescents' opinions about learning genetic information for adult onset-conditions and may inform the development of future clinical guidelines.

192

International attitudes of genetics professionals toward human gene editing. *A.J. Armsby[1], Y. Bombard[2], N.A. Garrison[3,4], B.L. Halpern-Felsher[5], K.E. Ormond[1,6].* 1) Stanford University School of Medicine, Department of Genetics, Stanford, CA 94305, USA; 2) Li Ka Shing Knowledge Institute of St. Michael's Hospital, University of Toronto, Institute of Health Policy, Management and Evaluation, Toronto, ON M5B 1W8, Canada; 3) Trueman Katz Center for Pediatric Bioethics, Seattle Children's Hospital and Research Institute, Seattle, WA 98101, USA; 4) Division of Bioethics, Department of Pediatrics, University of Washington, Seattle, WA 98101, USA; 5) Division of Adolescent Medicine, School of Medicine, Stanford University, Palo Alto, CA 94304, USA; 6) Stanford University School of Medicine, Department of Genetics and Stanford Center for Biomedical Ethics, Stanford, CA 94305, USA.

New gene editing technologies have made targeting and changing genes easier, more efficient, and more precise than ever before. This has sparked recent global discussion on acceptable uses of human gene editing and the need for harmonized regulations. There is limited data on how individuals trained in genetics (e.g. clinicians, clinical laboratory scientists, research scientists, ELSI researchers, and educators) view gene editing. These genetics professionals develop gene editing technologies and work with patients and families who may wish to utilize them in the future, and their perspectives are important to inform policy and practice in this field. We designed an online survey to describe the attitudes of genetics professionals across the globe toward somatic and germline gene editing. Participants were recruited through email distributions to genetics organizations and snowball sampling. We achieved participation from across the globe (631 responses collected, N=500 eligible completed responses included in analysis), although the majority of respondents originated from (63.5%) and/or resided (74.9%) in North America. Most respondents were clinicians (physicians, genetic nurses or genetic counselors, other clinicians; 57.5%). The majority had been practicing for <10 years (60.5%), with a mean age of 35-45 years. Virtually all our respondents are supportive of somatic gene editing in basic science research (99.2%) and clinical research (87.4%) on non-reproductive human cells. Compared to somatic gene editing research, respondents are currently less supportive of germline gene editing basic science research (57.2%) and clinical research (31.9%) using viable embryos ($p<0.001$). While most are in favor of using somatic (96.6%) and germline (77.8%) gene editing for clinical therapeutic purposes in the future, there is little support for enhancement applications (somatic: 13.0%; germline: 8.6%). There is decreased support for gene editing for diseases with lower penetrance, later age of onset, less significant impact on lifespan, and lesser degree of disability due to condition ($p<0.001$). Our data suggest that, while more strongly supportive of somatic gene editing, the majority of respondents approve of both somatic and germline gene editing. This study is the first of its kind to describe attitudes toward human gene editing from genetic professionals around the world, and contributes to ongoing discourse and policy guidance in this domain.

193

Consent to genome sequencing in research: A randomized controlled trial comparing two consent interventions. *E. Turbitt[1], P.P. Chrysostomou[2], A.R. Heidlebaugh[1], H.L. Peay[3], L.M. Nelson[2], B.B. Biesecker[1].* 1) Social and Behavioral Research Branch, National Human Genome Research Institute, Bethesda, MD; 2) Reproductive and Adult Endocrinology, National Institute of Child Health and Development, Bethesda, MD; 3) RTI International, Chapel Hill, NC.

Despite the widespread use of genome sequencing in research, evidence-based procedures for obtaining participants' consent to enter such studies are lacking. Prior research has found wide variation in length and reading level of genome sequencing consent material. Increasingly, these studies include return of secondary findings, which are results unrelated to the primary reason the sequencing was obtained. While the likelihood for detecting secondary findings among any one individual is low, participants should be made aware of this possibility during the consent process. This study evaluated the efficacy of a novel, evidence-based consent among women affected with primary ovarian insufficiency eligible to participate in an NIH sequencing study. Participants were randomized to receive either the novel or the standard consent document. A mixed methods approach involved data collection with questionnaires at baseline, immediate, and six-week follow up. Differences in quantitative outcomes were assessed using independent samples t-tests; thematic content analysis was used to analyze qualitative data. Of the 387 women contacted, 212 were recruited and randomized (response rate=55%), with complete data available for 188 participants. At six weeks, there were no differences between the two consent type groups in genome sequencing benefits knowledge (d=0.12, 95%CI: -0.03,0.27), genome sequencing limitations knowledge (d=0.04, 95%CI: -0.13,0.21), expected personal benefits (d=-0.01, 95%CI: -0.26,0.23), or decisional conflict regarding the choice to enroll (d=0.04, 95%CI: -0.14,0.21). Overall, participants had high expectations to learn information of personal benefit because of being in the study, and had positive attitudes and intentions toward receipt of secondary findings. These analyses demonstrate a lack of difference in outcomes between the longer, standard consent and the simplified, novel consent suggesting that a more concise, evidence-based consent is as effective to use when enrolling patients in research using genome sequencing. Participants have high expectations for receiving sequencing results of personal benefit, and may be at risk of misunderstanding the intent of the research (to create generalizable knowledge). Future research should include evaluation of interventions to ensure appropriateness of participants' expectations. Such progress is vital to ethical implementation of such technologies given the rapidity of the field of genomics.

194

Beyond uncertainty: Experiences of patients who participate in variant of uncertain significance reclassification research. *S. Makhnoon[1], L. Garrett[2], W. Burke[3], D. Bowen[3], B. Shirts[2].* 1) Institute of Public Health Genetics, University of Washington, Seattle, WA; 2) Department of Laboratory Medicine, University of Washington, Seattle WA; 3) Department of Bioethics and Humanities, University of Washington, Seattle WA.

Background: Patients' understanding of a genetic variant of unknown clinical significance (VUS) are likely to influence beliefs about risk implications, consequent medical decisions, and other actions such as involvement in research. Learning about VUS understanding of patients who chose to participate in a variant classification research study may shed light on the cognitive and affective beliefs and associated heuristics underlying their motivation to participate in such research. **Methods:** We interviewed 26 self-selected participants with a clinically identified VUS before they enrolled into a VUS reclassification study. Semi-structured interviews addressed topics including motivation to get genetic test, experience with the VUS result, affective responses to receiving VUS, and perceived effect of VUS and reclassification on medical care. An inductive thematic analysis approach was used to interpret the interview data. **Results:** We found that family and personal history of disease were the most prevalent motivators for getting a genetic test. Participants demonstrated mixed understanding of VUS. Most expressed negative affect on learning of their VUS result and uncertainty about its impact on clinical management, while others were relieved to not have a deleterious variant. Most expected reclassification efforts to benefit their family members but not themselves. Some expressed distrust of their providers following a VUS result. **Discussion:** Participation in the VUS reclassification study appeared to be motivated by three factors– negative affect about VUS, uncertainty about its impact on clinical management, and concern for family members' well-being. Perhaps the direct acknowledgement and appraisal of uncertainty as a means of coping was missing in some pre-test counseling experienced by our participants and thus they were not psychologically prepared for atypical VUS results. Frustration and distrust towards health care providers was expressed when VUS results could not confirm patients' intuition that there was a genetic cause underlying their disease, as a result, patients seemed to doubt the information that contradicted their beliefs. The finding of VUS induced provider distrust suggests a need for careful consideration of appropriate pre- and post-test counseling about VUS.

195

Importance of returning research results to exome participant families. *A. Warman, A. Incorvaia, M. Angrist, S. Katsanis.* Duke Initiative for Science & Society, Duke University, Durham, NC, USA.

The Duke Task Force for Neonatal Genomics (TFNG) uses exome sequencing to investigate causes of unexplained medical conditions in neonates and young pediatric patients. The TFNG undertakes genomic and variant functional analyses of various organ defects including central nervous system, renal, cardiac, craniofacial, skeletal, vascular, and skeletal muscle abnormalities. A major premise of the project is to return genome-based research results related to participants' unexplained medical conditions to families, whether causal or inconclusive. The TFNG team has endeavored to improve communication among researchers, clinicians, and families through annual family forums, guest speakers, a documentary film (*Rarefied,* in which nine families participated), and social science research. To examine the impact and utility of returning post-exome results to families and the impact of the interaction between researchers and families, we conducted semi-structured interviews with 13 family members and nine clinicians. Transcript coding followed by theme analysis revealed that participant recall of genetic variants associated with probands' conditions was limited to what was relayed at the return of results meeting. All of the families indicated a positive experience with the research experience and satisfaction with the results returned, including those receiving inconclusive results. Family participants noted that the return of positive results led to (A) improved treatments of their children's conditions; (B) networking with other families with similar genetic conditions; and (C) assistance with future family planning. Family participants with inconclusive results expressed hope that the condition may not be genetic and did not regret the exome sequencing process despite the inconclusive outcome. Clinician participants cited benefit from this research including their ability to guide families toward a diagnosis that might have remained elusive through other methods. The inclusion of the research team in the return of results session enabled communication of information to families that was outside of the clinicians' expertise. A common negative theme among the families and clinicians was the length of time to receive results. Families and clinicians had almost the same expectations and concerns for return of results and agreed on the importance of exome research results to the families' wellbeing and children's medical care.

196

Participant characteristics, motivations, healthcare utilization, and perceived utility in ostensibly healthy adults undergoing genome sequencing: Early findings from the PeopleSeq Consortium. *E.S. Zoltick[1,2], M.D. Linderman[3,4], L.S. Pais[1,5], M.A. McGinniss[6], E. Ramos[6], M.P. Ball[7,8], G.M. Church[8,9,10], D.G.B. Leonard[11], S. Pereira[12], A.L. McGuire[12], T.C. Caskey[13], S.C. Sanderson[4,14], E.E. Schadt[4], S.D. Crawford[15], R.C. Green[1,5,16,17], The PeopleSeq Consortium.* 1) Section of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA; 2) Section of Preventive Medicine and Epidemiology, Department of Medicine, Boston University School of Medicine, Boston, MA; 3) Department of Computer Science, Middlebury College, Middlebury, VT; 4) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 5) The Broad Institute of MIT and Harvard, Cambridge, MA; 6) Illumina Inc., San Diego, CA; 7) Open Humans Foundation, Boston, MA; 8) Harvard Personal Genome Project, Harvard Medical School, Boston, MA; 9) Department of Genetics, Harvard Medical School, Boston, MA; 10) Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA; 11) Department of Pathology and Laboratory Medicine, Robert Larner, M.D. College of Medicine of the University of Vermont, Burlington, VT; 12) Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX; 13) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 14) Department of Behavioural Science and Health, University College London, London, UK; 15) SoundRocket, Ann Arbor, MI; 16) Harvard Medical School, Boston, MA; 17) Partners Healthcare Personalized Medicine, Cambridge, MA.

Purpose: To describe participant characteristics, healthcare utilization, and perceived utility following genomic sequencing in the PeopleSeq Consortium, a collaboration of academic and commercial predispositional personal genome sequencing projects, which aims to examine outcomes of sequencing ostensibly healthy individuals. **Methods:** Web-based surveys were administered to ostensibly healthy adults from four cohorts before and/or after receiving personal sequencing results. Surveys inquired about sociodemographics, motivations, concerns, behavioral/medical responses to sequencing results, and perceived utility. Descriptive statistics were used. Logistic regression was used to examine possible predictors of perceived utility, including sociodemographics and healthcare utilization. **Results:** To date, 560 individuals have enrolled in the Consortium and 456 participants completed a survey after disclosure of their genomic results. Participants had a mean age of 53 years, 62% were men, and most were White (89%) and college graduates (96%). "Curiosity about my genetic make-up" and "interest in finding out about my personal disease risk" were endorsed as the most important motivators for pursuing sequencing. Less than 13% of participants reported being very concerned about privacy or insurance discrimination prior to sequencing. Over 80% of participants reported discussing their results with someone over a median of 11.8 months from genomic results disclosure; 80% reported discussions with a family member and 51% with a healthcare provider. Less than 20% of participants reported making/planning to make an appointment with a healthcare provider and 13% reported having any medical tests, exams, or procedures because of their results. A quarter of participants reported that what they learned from their results would help reduce their chances of getting sick. Men and those who discussed their results with a healthcare provider were more likely to endorse this statement ($p<0.05$). Decisional regret regarding undergoing sequencing was low (<3%). **Conclusion:** These preliminary results provide insight into the early adopters utilizing genomic sequencing as a screening tool. Participants report discussing results with their healthcare provider and receiving results that they consider useful. Longer follow-up may uncover additional downstream consequences of sequencing. The PeopleSeq Consortium continues to enroll additional participants and administer annual follow-up surveys.

197

Recurrent *de novo* heterozygous mutations disturbing the GTP/GDP binding pocket of RAB11B cause intellectual disability and a distinctive brain phenotype. *M.R.F. Reijnders[1], I.J.C. Lamers[2], H. Venselaar[3], A. Kraus[4], S. Jansen[1], L.B.A. de Vries[1], G. Houge[5], G. Aasland Gradek[5], J. Seo[6], M. Choi[6], J. Chae[7], S. Letteboer[2], S. van Beersum[2], S. Dusseljee[2], H.G. Brunner[1,8], D. Doherty[9], T. Kleefstra[1], R. Roepman[2].* 1) Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands; Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands; 2) Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands; Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands; 3) Centre for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands; 4) Yorkshire Regional Genetics Service. Chapel Allerton Hospital, Leeds, UK; 5) Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, N-5021, Norway; 6) Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, Republic of Korea; 7) Department of Pediatrics, Seoul National University College of Medicine, Seoul, Republic of Korea; 8) Department of Clinical Genetics and School for Oncology & Developmental Biology (GROW), Maastricht University Medical Center, Maastricht, 6229 ER, The Netherlands; 9) Department of Pediatrics, Seattle Children's Research Institute and University of Washington, Seattle, WA 98195, USA.

The Rab GTPase family comprises of approximately 70 GTP-binding proteins that function in vesicle formation, transport and fusion. They are activated by a conformational change induced by GTP-binding to interact with downstream effector proteins. The tightly regulated spatiotemporal activity of Rabs is controlled by guanine nucleotide exchange factors (GEFs) that catalyze the GDP/GTP-exchange, and GTPase activating proteins which catalyze the hydrolysis of GTP into GDP. Here, we report six patients with two recurrent *de novo* missense mutations in Rab GTPase family member *RAB11B;* p.(Val22Met) in four patients and p.(Ala68Thr) in two patients. An overlapping neurodevelopmental phenotype, including severe intellectual disability with absent speech, epilepsy and spasticity was observed in all affected individuals. Additionally, visual problems, musculoskeletal abnormalities, and microcephaly were present in the majority of cases. Re-evaluation of brain MRI images of four patients showed a shared distinct brain phenotype, consisting of severely decreased white matter volume, thinned corpus callosum, hypoplasia of the cerebellar vermis, optic nerve hypoplasia and mild ventriculomegaly. To study the functional effect of the identified *RAB11B* variants, and to compare this with known inactive GDP- and active GTP-bound RAB11B mutants (p.(Ser25Asn) and p.(Gln70Leu), respectively), we modeled the variants on the three-dimensional protein structure and performed subcellular localization studies. We predicted that both patient variants alter the GTP/GDP binding pocket, and found that they both resulted in disturbed Golgi localization and abolished membrane association. In line with these findings, evaluation of the affinity of RAB11B mutants to a series of binary interactors, showed enhanced affinity to the GEF SH3BP5 for both patient variants. Interestingly, these observations were comparable to the known GDP-bound inactive mutant, suggesting that the *RAB11B* mutations in patients resulted in a predominantly inactive state of the protein. In conclusion, we report two recurrent dominant mutations in *RAB11B* leading to a neurodevelopmental syndrome with distinct brain abnormalities, potentially caused by disturbed protein localization, enhanced GEF affinity and altered GDP/GTP binding. .

198

Deficient activity of genes associated with amino acid metabolism underlies an autosomal recessive syndrome of microcephaly and hypomyelination. *T. Nakayama[1,2], A. Al-Maawali[1,2,3], Q. Ouyang[4], J. Wu[5], D.J. Vaughan[1,2], M. El-Quessny[1,2], A. Rajab[6], S. Khalil[7], S. Niaz[8], M. Gul Butt[8], S. Imran Murtaza[8], A. Javed[8], H. Rashid Chaudhry[8], A.A. AlZahrani[9], P. Galvin-Parton[10], J. Weiss[10], M.R. Andriola[10], S.M. Amudhavalli[11], L. Cross[11], O. Baytas[4], K. Schmitz-Abe[1,2], K. Markianos[1,2], R.S. Hill[1,2], J.N. Partlow[1,2], B.J. Barry[1,2], M. Al-Saffar[1,2], A.J. Barkovich[12], E.M. Morrow[4], J. Ling[5], G.H. Mochida[1,2,13].* 1) Genetics and Genomics, Boston Children's Hospital, Boston, MA; 2) Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA; 3) Department of Genetics, College of Medicine and Health Science, Sultan Qaboos University, Muscat, Oman; 4) Department of Molecular Biology, Cell Biology and Biochemistry; and Institute for Brain Science, Brown University, Laboratories for Molecular Medicine, Providence, RI; 5) Department of Microbiology and Molecular Genetics, Medical School, University of Texas Health Science Center, Houston, TX; 6) National Genetics Center, Directorate General of Health Affairs, Ministry of Health, Muscat, Oman; 7) Department of Pediatrics, Al-Makassed Islamic Charitable Society Hospital, Jerusalem; 8) Pakistan Psychiatric Research Centre, Fountain House, Lahore, Pakistan; 9) King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia; 10) Department of Pediatrics, Stony Brook University Medical Center, Stony Brook, New York, NY; 11) Department of Genetics, Children's Mercy Hospital, Kansas City, MO; 12) Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA; 13) Department of Neurology, Massachusetts General Hospital, Boston, MA.

Microcephaly is an important cause of neurological morbidity in children and commonly has a genetic etiology. Underlying mechanisms of genetic microcephaly syndromes are highly diverse, with causative genes being implicated in many biological processes. Through genetic evaluation of 88 families with undiagnosed microcephaly, we identified likely pathogenic mutations in 46 pedigrees (52%). 25 pedigrees (28%) had mutations in 19 previously known microcephaly genes and 21 pedigrees (24%) had mutations in 17 novel disease gene candidates. Though the identified genes in both categories implicate many biological pathways, metabolic genes, particularly those involved in amino acid metabolism, were highly represented. These genes include *PYCR2* [MIM 616406] (4 families), which is involved in proline biosynthesis; *ASNS* [MIM 108370] (3 families), which is essential for asparagine synthesis; *GPT2* [MIM 138210] (2 families), which is important for transamination between alanine and 2-oxoglutarate to form pyruvate and glutamate; and *AARS* [MIM 601065] (1 family), which is needed for the ligation of alanine to the specific transfer RNA (tRNA). The onset and severity of microcephaly observed in all affected individuals varied but the presence of cerebral white matter abnormalities was often noted. Notably, their metabolic screening tests, including amino acid profiles in blood and/or cerebrospinal fluid, were negative. We applied clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 genome editing to create cellular models of PYCR2, GPT2 and ASNS deficiency. Cell lines deficient in PYCR2 and GPT2 both showed increased apoptosis under oxidative stress. ASNS-deficient cell lines showed reduced viability due to cell cycle defects and increased apoptosis, which were successfully rescued by asparagine supplementation of the culture medium. The mutations in AARS decreased aminoacylation efficiency, and one of the mutations also abolished editing activity required for hydrolyzing misacylated tRNAs. Our findings suggest that mutations in genes involved in amino acid metabolism form a unique subgroup of undiagnosed neurodevelopmental disorders with microcephaly and negative clinical biochemical testing.

199

Apoptosis drives the pathogenicity in a *Drosophila* model of 3q29 deletion. *M.D. Singh, E. Huber, L. Pizzo, B. Lifschutz, I. Desai, A. Kubina, S. Sunder, M. Jensen, S. Girirajan.* Department of Biochemistry and Molecular Biology, The Pennsylvania State University, STATE COLLEGE, PA.

Rare CNVs such as the 3q29 microdeletion contribute significantly to neurodevelopmental disorders, including schizophrenia, autism, and intellectual disability. We used the powerful genetic system of *Drosophila melanogaster* and a series of quantitative methods to assay the phenotype, function, and interactions of fly homologs of 3q29 genes. Using the *UAS*-Gal4 system and RNA interference, we evaluated phenotypes for 38 fly lines (representing 13 homologs) in a tissue-specific manner. For example, neuronal knockdown of *MFI2*, *UBXN7*, *SENP5*, and *WDR53* caused larval lethality, and neuromuscular defects were observed in *NCBP2*, *BDH1*, *PAK2,* and *DLG1* knockdown flies. Using the *Drosophila* eye for testing genetic interactions within the developmental and neuronal system, we identified rough eye phenotypes for several lines, including *NCBP2*, *BDH1, PAK2, PIGZ,* and *DLG1*. Further, markers for cellular phenotypes showed significant increases in apoptosis as well as cellular proliferation with knockdown of *NCBP2* ($p<0.01$)*, PIGX* ($p<0.01$)*,* and *DLG1* ($p<0.01$), for example. The increase in the number of mitotic cells may be due to compensatory proliferation to maintain a stable number of cells. We then tested 95 interactions by knocking down pairs of 3q29 homologs using recombinant lines, and found that the phenotypes of all tested homologs were enhanced when paired with *NCBP2*. For example, knockdown of both *NCBP2* and *DLG1* led to an increase in eye phenotype severity ($p=0.0043$), with necrotic patches on the eye surface. This phenotype was consistent with increased apoptosis ($p=3.11\times10^{-4}$) and no change in cellular proliferation in the two-hit knockdown compared to knockdown of *NCBP2* or *DLG1* alone. We also tested 152 two-hit models for *NCBP2*, *DLG1*, *PIGZ* and *PAK2* with known neurodevelopmental genes, and identified significant modifiers such as *SCN1A*, *SHANK3, UBE3A* and *CHD8*. Interestingly, the severe eye phenotypes of *NCBP2* and *DLG1* knockdown flies were rescued with overexpression of *DIAP1* (*Drosophila* Inhibitor of Apoptosis), and the level of apoptosis and proliferation were significantly reduced in both *NCBP2* ($p<0.01$) and *DLG1* ($p<0.01$) knockdown flies. These results suggest an additive model for 3q29 deletion, where multiple genes including *DLG1, PAK2, PIGX*, and *NCBP2* are sensitive to dosage imbalance and act in apoptosis pathways. The collective effects of these genes are in turn enhanced by reduced expression of *NCBP2* and modulated by other neurodevelopmental genes.

200

Dominant *RORA* variants cause an intellectual disability syndrome associated with epilepsy, autistic features or cerebellar ataxia. *X. Latypova[1,2], C. Guissart[3], T.N. Khan[1], P. Rollier[4], K. Õunap[5,6], L. Schema[7], M. Cho[8], K. Retterer[8], G. Lesca[9,10,11], S. Pajusalu[5,6], M.H. Wojcik[12,13], H. Stamberger[14,15,16], T. Deconinck[14,15], S. Weckhuysen[14,15,16], P. De Jonghe[14,15,16], L. Al-Gazali[17], S. Sanders[18], S. Sasorith[3], N. Leboucq[19], F. Rivier[20], C.M. Freitag[21], A.G. Chiocchetti[21], S. Kjaergaard[22], N. Katsanis[1], S. Bézieau[2], M. Koenig[3], S. Küry[2], E.E. Davis[1], L. Pasquier[4,23].* 1) Center for Human Disease Modeling, Duke University Medical Center, Durham, NC 27701, USA; 2) Service de Génétique Médicale, CHU Nantes, 9 quai Moncousu, Nantes Cedex 1, France; 3) EA7402 Institut Universitaire de Recherche Clinique, and Laboratoire de Génétique Moléculaire, University Hospital, Montpellier, France; 4) Service de Génétique Clinique, Rennes University Hospital, 35203 Rennes, France; 5) Department of Clinical Genetics, United Laboratories, Tartu University Hospital, Tartu, Estonia; 6) Department of Clinical Genetics, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia; 7) University of Minnesota Medical Center; 8) GeneDx, Gaithersburg, MD, USA; 9) Department of Genetics, Lyon University Hospitals, Lyon, France; 10) Claude Bernard Lyon I University, Lyon, France; 11) Lyon Neuroscience Research Centre, CNRS UMR5292, INSERM U1028, Lyon, France; 12) Division of Genetics and Genomics, Department of Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA; 13) Broad Institute of Harvard and MIT, Cambridge, MA, USA; 14) Neurogenetics Group, Department of Molecular Genetics, VIB, Antwerp, Belgium; 15) Laboratory of Neurogenetics, Institute Born-Bunge, University of Antwerp, Antwerp, Belgium; 16) Division of Neurology, University Hospital Antwerp (UZA), Antwerp, Belgium; 17) Department of Pediatrics, College of Medicine and Heath Sciences United Arab Emirates UniversityAl-Ain, United Arab Emirates; 18) Department of Psychiatry, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94158, USA; 19) Department of Neuroradiology, Montpellier University Hospital, Montpellier, France; 20) Department of Neuropediatrics and CR Maladies Neuromusculaires, Montpellier University Hospital, France; 21) Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Autism Research and Intervention Center of Excellence, JW Goethe University Frankfurt, Deutschordenstraße 50, Frankfurt am Main, D-60528, Germany; 22) Chromosome Laboratory, Department of Clinical Genetics, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark; 23) CNRS UMR 6290, Université de Rennes 1, 2 Avenue du Professeur Léon Bernard, 35043 Rennes, France.

Genetic factors contribute strongly to the etiology of intellectual disability (ID), and comprise a substantial proportion of the morbid human genome. Through a multi-center international collaborative effort, we identified 12 ID cases with *de novo* or dominant variants in the gene encoding ROR-alpha, the retinoic acid receptor (RAR)-related orphan nuclear receptor alpha (*RORA*), by whole exome sequencing or chromosomal microarray. Affected individuals harbor a mutational spectrum that includes four copy number variations (two *de novo* deletions, one dominant deletion and one *de novo* duplication) and five *de novo* single nucleotide variants. ID is the predominant feature in individuals with *RORA* variants (11/12), and is accompanied by seizures (8/12), autistic features (3/12) and cerebellar hypoplasia or atrophy (3/12). These data are consistent with previously reported phenotypes associated with a microdeletion on 15q22.2, for which the minimal region of overlap included *RORA* and the NMDA receptor-regulated 2 gene (*NARG2*). Furthermore, individuals with truncating variants have ID associated with autistic features, while two patients with *de novo* missense mutations altering the DNA binding domain of ROR-alpha present with ataxia and cerebellar atrophy. RORA plays a critical role in cerebellar development, established through phenotyping of a spontaneous murine mutant discovered two decades ago to result from homozygous intragenic *Rora* deletions. This model, the *staggerer* mutant, presents with an ataxic gait caused by massive neurodegeneration of Purkinje cells in the cerebellum. To investigate the relevance of *RORA* disruption to neurodevelopmental phenotypes in humans, we abrogated the *D. rerio* ortholog, *roraa,* through either transient suppression or CRISPR/Cas9 based genome editing. Using acetylated alpha tubulin immunostaining of zebrafish larvae at 3 days post fertilization we show that *roraa* disruption causes a significant reduction of cerebellar volume. Together, our results suggest that *RORA* variants lead to a neurodevelopmental disorder characterized by intellectual disability, seizures, autistic features and cerebellar defects.

201

***PRESO1* mutations cause X-linked intellectual disability by disrupting dendritic spine morphogenesis.** *J. Piard[1], J.H. Hu[2,3], P. M Campeau[4], S. Rzońca[5], H. Van Esch[6], E. Vincent[2], M. Han[2], E. Rossignol[7], J. Castaneda[5], J. Chelly[8], C. Skinner[9], V. Kalscheuer[10], R. Wang[2], E. Lemyre[3], J. Kosińska[5], P. Stawinski[5], J. Bal[5], D. Hoffman[3], C. Schwartz[9], L. Van Maldergem[1,11], T. Wang[2], P. Worley[2].* 1) Centre de génétique Humaine, CHU Besançon, Besançon, France; 2) Department of Neuroscience, Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 3) Molecular Neurophysiology and Biophysics Section, Program in Developmental Neuroscience, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, USA; 4) Department of Pediatrics, University of Montreal, Montreal, QC, Canada; 5) Institute of Mother and Child, Warsaw, Poland; 6) Department of Human Genetics, University Hospitals Leuven, Belgium; 7) Department of Neurosciences, University of Montreal, Montreal, QC, Canada; 8) CNRS UMR7104, Institut de génétique, biologie moléculaire et cellulaire, Illkirch, France; 9) Greenwood Genetic Center, Greenwood, South Carolina, USA; 10) Research Group Development and Disease, Max Planck Institute for Molecular Genetics. Berlin, Germany; 11) Centre of Clinical Investigation 1431, National Institute for Health and Medical Research (INSERM), Université de Franche-Comté, Besançon, France.

PRESO1, alias FRMPD4 (FERM and PDZ Domain Containing 4) is a neural scaffolding protein that interacts with PSD-95 to positively regulate dendritic spine morphogenesis, and with mGluR1/5 and Homer to regulate mGluR1/5 signaling. We report the genetic and functional characterization of 4 *PRESO1* deleterious mutations that cause a new X-linked intellectual disability (ID) syndrome. These mutations were found to be associated with ID in ten affected male patients from four unrelated families, following an apparent X-linked mode of inheritance. Mutations include deletion of an entire coding exon, a nonsense mutation, a frameshift mutation resulting in premature termination of translation, and a missense mutation involving a highly conserved amino acid residue neighboring the PRESO1-FERM domain. Clinical features of these patients consisted of moderate to severe ID, language delay and seizures along with behavioral and/or psychiatric disturbances. An in-depth functional study of the frameshift mutation p.Cys618ValfsX8 indicated it disrupted PRESO1 binding with PSD-95 and HOMER1. It also appeared to reduce spine density in transfected hippocampal neurons. Behavioral studies of Preso1-KO mice identified hippocampus-dependent spatial learning and memory deficits using the Morris Water maze test. These findings point to an important role of *PRESO1* in normal cognitive development and function in humans and mice, and support the hypothesis that *PRESO1* mutations cause ID by disrupting dendritic spine morphogenesis in glutamatergic neurons.

202

***USP9X* mutations cause a spectrum of neurodevelopmental disorders underpinned by a disruption of multiple signalling pathways that control brain development.** *L.A. Jolly[1], B.V. Johnson[1], R. Kumar[1], N. Dikow[2], A. Goldstein[3], S. Asher[4], P. VanHasselt[5], M. Perry[6], S. Mahmutoglu[7], S. Grøborg[8], P. Zwijnenburg[9], M. Weiss[9], C. Reis[10], M. Koenig[11], L. Pasquier[12], M. Lines[13], C. Keegan[14], C. Lopez-Otin[15], A. Fernández Jaén[16], H. Lefroy[17], B. Keren[18], M. Raynaud[19], S. Küry[20], M. Reijnders[21], T. Kleefstra[21], T. Pierson[22], T. Burne[23], M. Piper[24], S.A. Wood[25], J. Gecz[1].* 1) Robinson Research Institute, University of Adelaide, Adelaide, South Australia, Australia; 2) Institute of Human Genetics, Heidelberg University, Heidelberg, Germany; 3) Children's Hospital of Pittsburgh, Pittsburgh, PA, USA; 4) Translational Medicine & Human Genetics, Hospital of the University of Pennsylvania, Philadelphia PA USA; 5) Department of Metabolic Diseases, University Medical Center Utrecht, 3584 Utrecht, The Netherlands; 6) Jane and John Justin Neuroscience Center, Cook Children's Medical Center, Fort Worth, TX, USA; 7) Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, Ontario, Canada; 8) Center for Rare Diseases, Clinical Genetics Department, University Hospital Copenhagen, Copenhagen, Denmark; 9) Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands; 10) Medical Genetics Unit, Hospital Pediátrico, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal; 11) Department of Pediatrics, University of Texas Medical School at Houston, Houston, TX, USA; 12) Service de Génétique Clinique, Centre de Référence Maladies Rares CLAD-Ouest, CHU Hôpital Sud, Rennes, France; 13) Children's Hospital of Eastern Ontario, Ottawa, Canada; 14) Division of Genetics, Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA; 15) Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain; 16) Unidad de Neurología Infantil, Hospital Universitario Quirón Madrid, Spain; 17) Oxford Centre for Genomic Medicine, Oxford University Hospitals NHS Foundation Trust, UK; 18) Hôpital de la Pitié-Salpêtrière, Département de Génétique, Paris, France; 19) Centre Hospitalier Régional Universitaire, Service de Génétique, Tours, France; 20) Service de Génétique Médicale, CHU Nantes, Nantes, France; 21) Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands; 22) Department of Pediatrics, Cedars-Sinai Medical Center, Los Angeles, CA, USA; 23) Queensland Brain Institute, University of Queensland, Brisbane, Australia; 24) School of Biomedical Sciences, University of Queensland, Brisbane, Australia; 25) Griffith Institute for Drug Discovery, Griffith University, Brisbane, Australia.

Mutations in the X-linked gene *USP9X* have been associated with intellectual disability (ID) in both males and females. Nineteen mutations causing haploinsufficiency of *USP9X* in females with ID, congenital malformations and recognisable brain abnormalities have been reported. In males, only three missense mutations associated with ID had been reported, and another two associated with seizures, and as such the involvement of *USP9X* in male ID remained less certain. We report 26 additional *USP9X* missense mutations associated with male ID, with 21 mutations considered strong candidates for pathogenicity based on segregation and in-silico metrics. We describe an evolving phenotypic spectrum associated with *USP9X* missense mutations in males. In addition to ID and developmental delay, we found speech delay, hypotonia, seizures, autistic behaviour, aggressiveness and visual impairment were frequently identified (64-100% of cases). Brain structural imaging showed evidence of disrupted white matter, thin corpus callosum and cortical malformations. Our *USP9X* knockout mouse model displayed overlapping brain structural features, and we now resolve severe learning and memory deficits highlighting its utility to understanding mechanisms of pathology. USP9X is a deubiquitylating enzyme capable of protecting substrates from proteasomal degradation. In embryonic brains of *USP9X* knockout mice, we show altered levels of multiple key substrates belonging to signalling pathways, and as such defective mTOR, WNT, NOTCH and TGFβ signalling is observed. Furthermore, we found these key substrates are disrupted in patient-derived fibroblast cells lines, suggesting defective signalling underlies pathology. Collectively, our data demonstrate the involvement of *USP9X* in male ID and other neurodevelopmental disorders, and identify plausible mechanisms of pathogenesis.

203

Polygenic risk scores identify novel relationships between complex traits. *R.L. Kember[1], S. Damrauer[2,3], A. Small[4], M. Bucan[1], D. Rader[1,5,6].* 1) Department of Genetics, University of Pennsylvania, Philadelphia, PA; 2) Department of Surgery, University of Pennsylvania, Philadelphia, PA; 3) Department of Surgery, CPL. Michael J. Crescenz VA Medical Center, Philadelphia, PA; 4) Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 5) Department of Medicine, University of Pennsylvania, Philadelphia, PA; 6) Department of Pediatrics, University of Pennsylvania, Philadelphia, PA.

Pleiotropy occurs when a gene or variant has an effect on multiple traits. At the level of the individual, this can lead to disease co-morbidity, complicating clinical presentation of disease. For instance, individuals with mood disorder have higher levels of early mortality than the general population, and a substantial proportion of this may be linked to increased risk for co-morbid diseases. The recent adoption of electronic health records (EHR) in a research setting has paved the way for phenome-wide association studies (PheWAS), which can assess the effect of genetic variants across a broad range of available phenotypes. The collaboration between the Penn Medicine BioBank and the Regeneron Genetics Center aims to combine whole exome and SNP genotype data on 12,000 subjects with EHR from the University of Pennsylvania Health System to conduct EHR-based phenome analysis. In order to quantify the extent of genome-wide pleiotropy between mood disorders and medical co-morbidities, we combined polygenic risk scores (PRS) with medical phenotypes available in EHR to identify cross-phenotype associations. We generated PRS for multiple psychiatric traits, including bipolar disorder, anxiety, and depression, in addition to common, complex traits such as cardiovascular disease, diabetes, and blood lipid levels, and conducted a PheWAS for each risk score. As expected, risk scores were associated with relevant traits (e.g. Bipolar disorder PRS with mood disorders, $p=3.7x10^{-4}$, Cardiovascular disorder PRS with Coronary atherosclerosis, $p=2.3x10^{-23}$). The PheWAS also recapitulates other known phenotypic associations (e.g. CAD PRS with Type 2 diabetes, $p=9.3x10^{-6}$). Finally, we performed hierarchical clustering analysis to identify phenotypic relationships between psychiatric disorders and other medical co-morbidities. Interestingly, both bipolar disorder and depression clustered with HDL. Traits shared between these disorders included dermatitis, psoriasis, lipoprotein disorders, psychosis, and thyroiditis. Many of these traits have been independently associated with both mood disorders and HDL levels, but here we provide evidence within a single population for a cross-phenotypic association. Collectively, our results demonstrate the reliability of EHR to recapitulate an individual's genetic risk for disease, provide evidence for genetic pleiotropy across disease categories, and provide a genetic basis for known cross-phenotype associations.

204

Overtransmission of polygenic risk alleles for migraines in 2,048 Finnish trios. *K. Veerapen[1,2,3], M.E. Hiekkala[4], P. Gormley[1,2,3], M.I. Kurki[1,2,3], A. Mitchell[5], H. Runz[6], P. Häppölä[7], P. Palta[7], E. Hämäläinen[7], M.A. Kaunisto[7], V. Arrto[8], M. Färkkilä[8], B. Neale[1,2,3], M. Daly[1,2,3], M. Wessman[4,5,6,7], M. Kallela[8], A. Palotie[1,2,3,7,9,10].* 1) Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA; 2) Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA; 3) The Stanley Centre for Psychiatric Research, The Broad Institute of Harvard and MIT, Cambridge, MA, USA; 4) Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland; 5) Discovery Data Science Unit, Eisai Inc., Andover, MA; 6) Department of Genetics & Pharmacogenomics, Merck Research Laboratories, Merck and Co., Inc., Boston, MA; 7) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; 8) Helsinki University Central Hospital, Helsinki, Finland; 9) Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA; 10) Department of Neurology, Massachusetts General Hospital, Boston, MA, USA.

A recently published migraine genome wide association study (GWAS) (*N*=375,000) identified 38 genomic loci (PMID:27322543). However, these loci explain only a fraction of the heritability (~1%) and familial risk. As a common disease affecting up to 15-20% of the population, migraine is a complex disease that aggregates within families and twins (h^2~0.4-0.6) (PMID:14624726) but little is known of the genetic risk transmission. Therefore, we hypothesize the transmission of polygenic risk alleles contributes to familial migraine. Additionally, we hypothesize that the more common forms of migraine (migraine without aura (MO) and migraine with aura (MA)) have a larger polygenic transmission compared to the rarer hemiplegic migraine (HM), which is considered to have a greater Mendelian mutation component. Using polygenic risk scores (PRS), we investigate polygenic transmission in a Finnish migraine family cohort (*N* = 1,214 families) -- the largest migraine family collection. A total of 8,319 individuals (2,357 MO, 2,420 MA, 540 HM, and 3,002 unaffected family members) were genotyped on either Illumina CoreExome or PsychArray and imputed against a Finnish population reference panel (*N*=1,941 WGS) using IMPUTE2. The PRS were computed using effect sizes estimated for SNPs from the GWAS of migraine (*N*=366,681; PMID:27322543); and SNPs were binned for inclusion in PRS based on different *p*-value thresholds. Using the bin that had the largest variance explained ($p<0.1$), we computed transmission in 2,048 possible trios (609 MO, 665 MA, 211 HM) using PRS-Transmission Disequilibrium Test (pTDT). Overall, we found that affected offspring, received a significant overtransmission of polygenic risk from their unaffected parents ($N_{families}$=167, deviation=0.31 sd, 95%-CI=0.16-0.46, p=0.00009); no significant overtransmission was observed in unaffected offspring in similar trio structures (p =0.78). When analysing migraine subtypes, MO had the lowest overtransmission ($N_{families}$=87, deviation=0.26 sd, 95%-CI=0.087-0.43, p=0.004) compared to MA ($N_{families}$=53, deviation=0.28 sd, 95%-CI=-0.03-0.59, p=0.09) and HM ($N_{families}$=27, deviation=0.57, 95%-CI=0.1-1.03, p=0.02) therefore implying a higher transmission of polygenic risk in the comparatively rare forms of migraine. Further filtering based on offspring sex, an overtransmission in affected daughters in all migraine subtypes ($p < 0.05$) was observed; supporting the literature where migraines occur 2-3 times more frequently in women.

205

Association of polygenic risk scores for multiple cancers in a phenome-wide study: Results from The Michigan Genomics Initiative. *L.G. Fritsche[1,2,3], S.B. Gruber[4], Z. Wu[1,5], E.M. Schmidt[1], S.E. Moser[6], V.M. Blanc[7], C.M. Brummett[6,8], S. Kheterpal[6,8], G.R. Abecasis[1,2], B. Mukherjee[1,2,5,9,10].* 1) Center for Statistical Genetics, University of Michigan School of Public Health; 2) Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA; 3) K.G. Jebsen Center for Genetic Epidemiology Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology; 4) USC Norris Comprehensive Cancer Center, Los Angeles, CA, USA; 5) Michigan Institute of Data Science, University of Michigan, Ann Arbor, MI, USA; 6) Division of Pain Medicine, Department of Anesthesiology, University of Michigan Medical School, Ann Arbor; 7) Central Biorepository, University of Michigan Medical School, Ann Arbor, MI, USA; 8) Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, MI, USA; 9) Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA; 10) University of Michigan Comprehensive Cancer Center, University of Michigan, Ann Arbor, MI, USA.

Introduction: In recent years, there has been an increasing interest in phenome-wide association studies (PheWAS) – simultaneous explorations of the association between genetic variants and broad spectrums of physiological/clinical phenotypes – to uncover novel association and/or relevant cross-phenotype associations. Current studies that employ PheWAS utilize curated groups of "computable phenotypes" defined and validated by experts using a combination of International Classification of Diseases (ICD) codes. Standard PheWAS have primarily focused on correlating genetic variants, one at a time, but when each variant is associated with a small effect size, these studies can only provide limited insights. Here, we introduce the new concept of PheWAS based on a polygenic risk score (PRS) of published SNP associations instead of a single genetic variant. **Methods:** We construct effect-size weighted PRS for multiple cancers using SNP summary statistics reported in the NHGRI EBI GWAS catalog. We then evaluate their predictive performance for their underlying trait using the EHR-based phenome of the Michigan Genome Initiative (MGI, https://www.michigangenomics.org). Application of the PheWAS R package (https://github.com/PheWAS/PheWAS) on 2.9 million cumulative code days of 18,267 genotyped samples of European ancestry yielded a total of 1,815 case-control studies of which 1,448 with ≥20 cases were used for further analyses. We applied Firth's bias corrected logistic to PRS models and also included sensitivity analysis on matched case-control studies or on unweighted risk allele counts. **Results:** Several of cancer-specific PRS revealed high predictive power for their underlying trait, e.g., skin cancer (OR=1.8, $P=1.9\times10^{-19}$), female breast cancer (OR=2.5, $P=1.4\times10^{-21}$) and prostate cancer (OR=2.9, $P=2.5\times10^{-19}$). Phenome-wide significant associations were observed between PRS and many diagnoses including elevated prostate specific antigen levels, erectile dysfunction, and urinary incontinence in patients with high prostate PRS, supporting the constellation of primary and secondary diagnoses that arise in patients managed in a large health system with genetically identifiable risk. **Conclusion:** Electronic health records and genomic data available from large health systems like Michigan Medicine provide opportunities to accurately stratify patients' risk of cancer and to discover and validate networks of expected and unexpected relationships of risk factors and diagnoses.

206

A powerful approach to estimating annotation-stratified genetic covariance using GWAS summary statistics. *Q. Lu[1], B. Li[2], D. Ou[2], M. Erlendsdottir[2], R. Powles[2], T. Jiang[2], Y. Hu[2], D. Chang[2], C. Jin[2], W. Dai[2], Q. He[3], Z. Liu[3], S. Mukherjee[4], P. Crane[4], H. Zhao[1].* 1) University of Wisconsin-Madison, Madison, WI; 2) Yale University, New Haven, CT; 3) Shanghai Jiao Tong University, Shanghai, China; 4) University of Washington, Seattle, WA.

Despite success of genome-wide association studies (GWAS), our understanding of complex traits' genetic architecture is incomplete. Jointly modeling multiple traits' genetic profiles has provided insights into the shared genetic basis of many complex traits. However, large-scale inference sets a high bar for both statistical power and biological interpretability. Here we introduce a principled framework to estimate annotation-stratified genetic covariance between complex traits using GWAS summary statistics. Through theoretical and numerical analyses we demonstrate that our method provides substantially more accurate covariance estimates and more powerful statistical inference than LD score regression. Among 50 complex traits with publicly accessible GWAS summary statistics (n=4.5 million), we identified 175 pairs with statistically significant correlations, including associations between traits regarding which the literature is either equivocal or absent, such as correlations of serum uric acid with type-II diabetes (p=1.2E-17), triglycerides (4.6E-9), and many other markers of metabolic syndrome. In contrast, LD score regression can only identify 127 significant correlations. In particular, we performed an in-depth, annotation-driven analysis on Alzheimer's disease (AD; n=54,162) and amyotrophic lateral sclerosis (ALS; n=36,052), two major neurodegenerative diseases. We identified a novel genetic correlation between AD and ALS (p=2.0E-4). In addition, we demonstrate that 83% of the total genetic covariance between AD and ALS is concentrated in the 32% of the genome predicted to be functional (p=8.2E-5); 54.6% of the covariance can be explained by the most common variants as indicated by the highest quartile of minor allele frequencies (p=0.005) while genetic covariance in the lowest quartile is nearly negligible. Furthermore, the genetic covariance between AD and ALS is concentrated in immune-related functional genome (p=0.014) while the covariance in brain-related DNA elements is non-significant. AD and ALS showed distinct patterns of correlation with other traits. We identified significant negative correlations between AD (but not for ALS) and cognitive traits including cognitive function (p=1.0E-11). ALS (but not for AD) is positively correlated with immune-related diseases including multiple sclerosis (p=4.6E-4), hinting at an autoimmune component for ALS. Our findings shed light onto the shared and distinct genetic architecture of complex traits.

207

Phenotype connectivity map across human diseases derived from phenome-wide association study on 38,682 samples. *A. Verma[1,2], L. Bang[1], J.E. Miller[1], Y. Zhang[3], M.T.M. Lee[3], D.J. Carey[4], M.D. Ritchie[1,2], S.A. Pendergrass[1], D. Kim[1,2] on behalf of the DiscovEHR collaboration.* 1) Biomedical and Translational Informatics Institute, Geisinger Health System, Danville, PA; 2) The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA; 3) Genomic Medicine Institute, Geisinger health System, Danville, PA; 4) Weis Center for Research, Geisinger Clinic, Danville, PA.

Pleiotropy is when a given locus (i.e. SNP or gene) influences two or more different phenotypes or traits. Phenome-Wide Association studies (PheWAS) have been commonly used to test associations between genetic variations and multiple complex traits or diagnoses. Linking these associations between a variant or a genomic region and various different phenotypes into a network offers a powerful resource for investigating cross-phenotype association and discovering potential pleiotropic effects. We utilized associations from one of the largest PheWAS on Electronic Health Record (EHR) derived phenotypes, using 38,682 unrelated samples data from the MyCode Community Health Initiative of Geisinger Health System (GHS) through the DiscovEHR project. Associations between 632,574 variants and 541 disease diagnosis based ICD-9 codes, as well as 25 clinical laboratory measures were calculated. Using these associations, two types of networks were constructed: (1) a *Disease-Disease* network where pairs of diseases are connected based on shared associations with an SNP, representing a potential pleiotropic SNP; (2) a *SNP-SNP* network where pairs of SNPs are connected through shared association with a disease or trait, representing a polygenic feature of disease/trait. The network of diseases provides a landscape of intra-connections within the same disease class as well as interconnections across other disease classes. We identified clusters of diseases with known biological connections, such as endocrine and metabolic disorders (Type 1 Diabetes, Hyperlipidemia, and Gout) and nervous system disorder (Multiple Sclerosis). We also identified various potential novel clusters of diseases with similar underlying genetic patterns. For example, a cluster between Macular degeneration, Essential Hypertension, and Hypercoagulable state disorder, where hypertension is a risk factor for both macular degeneration and hypercoagulable state disorder. We also investigated the pleiotropic effect of independent variations on a given disease using networks of SNPs. We will aggregate the SNPs by gene and pathway based on biological annotations to expand the network. Using this approach, novel interactions between different diseases can be investigated due to the shared potential pleiotropic SNPs.

208

Quantifying the shared genetic components of complex traits and Mendelian phenotypes. *M. Kumar, V. Arboleda, H. Shi, N. Mancuso, B. Pasaniuc.* Pathology and Lab Medicine, David Geffen School of Medicine, UCLA, Los Angeles, CA.

Although recent studies provide evidence for common genetic origins of multi-factorial complex traits and Mendelian diseases, a thorough assessment of their overlap in a trait-specific manner is currently missing. Here, we investigate the shared genetic basis of 37 complex traits with their Mendelian forms (e.g., height vs. skeletal dysplasias). We use publically available genome-wide association study (GWAS) summary statistics across 37 complex traits to identify putative risk genes, and quantify the overlap with genes that are known to cause similar Mendelian phenotypes. Of 592 pairs of complex and Mendelian traits, we find 20 pairs with significant enrichment of Mendelian genes at known risk GWAS loci for a trait-matched complex phenotype. In addition, we observe 13 pairs of phenotype-unmatched complex and Mendelian with significant gene overlap suggestive of shared biological mechanism yet to be examined. We further find a significant enrichment of heritability within 50kb of trait-matched Mendelian genes in 17 complex traits. We also identify examples of associated SNPs found at the transcription start site of these phenotypically-relevant Mendelian disease genes as candidates for functional follow-up for the corresponding complex trait. In addition to providing insight into genetic mechanisms of complex traits, these results can be leveraged to prioritize candidate genes at significant GWAS loci and identify modifiers of rare disease phenotypes in both large-scale association studies and clinical exome sequencing.

209

Single cell transcriptome atlas of the mouse kidney reveals important cell diversity. *J. Park, R. Shrestha, C. Qiu, A. Kondo, S. Li, K. Suszták.* Renal Electrolyte and Hypertension Division, University of Pennsylvania, Philadelphia, PA.

A revolution in cellular measurement technology is under way: For the first time, we have the ability to monitor global gene regulation in thousands of individual cells in a single experiment. They overcome fundamental limitations inherent in measurements of bulk cell population that have frustrated efforts to resolve precise cellular states. These methods also provide a stunningly high-resolution view of transitions between states. Single-cell transcriptomics will allow us to identify cell type specific expression changes, discover novel disease associated cell types and trace cell composition changes in complex diseases. Using droplet-based single-cell barcoding and sequencing methods, here, we cataloged mouse kidney cell types in an unbiased manner. We have developed a novel cell isolation method and individually profiled more than 30,000 cells from mouse kidneys. Computational analysis included normalization, quality control, dimension reduction and clustering followed by identification of cell types using known markers and bulk RNAsequencing data of kidney segments. In addition using time-series analysis we have generated cell trajectory data for individual cell clusters. Main clustering analysis identified 15 major cell populations in normal mouse kidneys; three distinct ureteric bud derived and seven metanephric mesenchyme derived epithelial cell clusters, in addition to endothelial cells, fibroblasts, different immune cell types and two novel epithelial cell populations that have not been described before. Cell trajectory analysis highlighted novel cell type conversion trajectories in the collecting duct. Furthermore we have developed methods for single cell marker based in silico deconvolution of previous bulk RNAseqeuncing datasets. We found that most transcript level changes previously reported in kidney disease samples are resulted from cell type proportion changes in kidney samples, such as accumulation of immune cells and fibroblast and decrease of tubule epithelial cells in disease development. Finally, we identified key cell type specific transcription factors, mapped GWAS candidate genes, known drug target genes, and nephrotic syndrome genes in our single cell cluster showing their cell type-specific expression patterns. In conclusion, our first single cell transcriptome of the entire mouse kidney could have a transformative impact to understand transcriptional networks maintaining cell identity and development of kidney disease.

210

Characterizing the landscape of somatic mutations in normal aging human tissues using a novel single-cell whole-genome sequencing approach. *X. Dong, L. Zhang, M. Lee, A. Maslov, J. Vijg.* Department of Genetics, Albert Einstein College of Medicine, Bronx, NY.

Single-cell sequencing for analyzing DNA mutations across the genome in somatic tissues is critically important for studying development, cancer and aging. However, current procedures are prone to artifacts and to date a reliable protocol for single-cell somatic mutation analysis remains to be developed. We address the two largest sources of artifacts, i.e., DNA denaturation-related cytosine deamination and allelic bias-driven amplification errors. We first reconfigured multiple displacement amplification (MDA) into an efficient protocol for whole genome amplification of single cells without cytosine deamination artifacts, i.e., Single Cell MDA (SCMDA). We then developed a new single-cell SNV caller (SCcaller) that distinguishes real somatic mutations and amplification errors by utilizing a SNP-based localized estimate of allelic amplification bias. The procedure was validated by comparing SCMDA-amplified single cells with unamplified clones derived from single cells from the same population. Using this highly accurate single-cell whole-genome sequencing method we analyzed human B lymphocytes from donors varying in age from birth to over 100 year, studying both genome distribution and functional impact of base substitution mutations. Mutations per cell were found to increase with age from less than 500 in cord blood to over 3,000 in cells from individuals over 100. While overall mutations were randomly distributed across the genome with all chromosomes equally affected, 24 hotspot regions were identified, five of which were part of Immunoglobulin variable regions subject to somatic hypermutation. Age-related mutation accumulation was found to be significantly slower in genomic sequences directly involved in cellular function, such as exons and gene regulatory sequences. Still, on average, B lymphocytes from aged individuals contained 3-15 damaging mutations in the transcribed part of the exome identified by RNA-seq, plus 15-50 mutations in transcription factor binding sites identified by ATAC-seq. We also identified copy number variation, the frequency of which also increased with age. These results strongly suggest that spontaneous somatic mutations accumulating with age reach high enough levels to contribute to age-related functional decline, such as the well-documented cell-intrinsic changes in B lymphocytes. Taken together, our single-cell sequencing method provides a firm foundation for analyzing cellular genetic heterogeneity in normal human tissue.

211

Dissecting the microenvironment of multiple tumor types using 5' and 3' single cell RNA-seq. *S.C. Boutet, V. Giangarra, G.X.Y. Zheng, A.M. Barrio, L. Montecarlos, J. Lee, S. Marrs, K.J. Wu, P. Ryvkin, T. Mikkelsen, D.M. Church.* 10x Genomics, Pleasanton, CA., United States.

A critical component of personalized cancer treatments is understanding the tumor microenvironment, in particular, the complex interactions between tumors and the immune cells. We describe an approach that couples single-cell transcriptional profiling of the cellular component of the tumor with high-resolution characterization of tumor infiltrating lymphocytes (TILs). To explore tumor heterogeneity, we used a fully integrated, droplet-based system for single cell RNA sequencing (scRNA-seq) on three types of tumor, metastatic melanoma (MM), primary colorectal cancer (CRC) and primary clear cell renal carcinoma (CCRC), from multiple patients. Each tumor varies in type and proportion of its cellular components, noticeably in the proportion of TILs. The MM cells, obtained from an axillary lymph node, consist of T cells (30%) and an expanded B cell population (70%). In contrast, tumor cells in primary CRC and CCRC are mostly epithelial in nature. Whereas TILs from CRC are predominantly CD4+ cells (80% CD4+, 9% CD8+), TILs from CCRC display a balanced proportion of CD4+ and CD8+ cells (47% CD4+, 41% CD8+). The CD4/CD8 population ratio can be informative on relative risk of metastasis and overall survival in patients. We then applied targeted 5' scRNA-seq to characterize full length paired T cell receptor alpha (TCRA) and beta (TCRB) chains. To assess the sensitivity and accuracy of the 5' scRNA-seq method, T cells activated with Epstein-Barr virus peptide and peripheral blood mononuclear cells (PBMCs) from a healthy donor were mixed at ratios ranging from 1% to 50% and profiled. Correct, full length, and paired TCRA/TCRB clonotypes were detected at expected ratios and demonstrated a detection sensitivity of clonal expansion as low as 1% for 1000 cells observed.This assay revealed a ~10% clonal expansion in TILs of CCRC relative to the T cells from PBMCs from the same patient. The same clonal expansion was observed in the unenriched tumor samples, implying the possibility to directly perform 5' scRNA-seq from unenriched tumor samples, significantly reducing the complexity of experimental set-up. In contrast, no clonal expansion was observed in CRC, suggesting different molecular programs in these tumors. While we have applied this technology to study 3 different tumor types, we envision its broad application to characterize the complex ecosystem of many different tumor types in many different patients and to be key to better understand tumor immunity.

212

Scalable single-cell DNA methylation sequencing by combinatorial indexing. *R. Mulqueen[1], D. Pokholok[2], S. Norberg[2], A. Fields[1], J. Shendure[3,4], C. Trapnell[3], B.J. O'Roak[1], Z. Xia[5], F. Steemers[2], A. Adey[1,6].* 1) Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR; 2) Advanced Research Group, Illumina, Inc., San Diego, California, USA; 3) Department of Genome Sciences, University of Washington, Seattle, Washington, USA; 4) Howard Hughes Medical Institute, Seattle, Washington, USA; 5) Department of Computational Biology, Oregon Health & Science University, Portland, Oregon, USA; 6) Knight Cardiovascular Institute, Portland, Oregon, USA.

DNA methylation shows cell type-specificity, is actively modified in developing tissues, and has been linked to neurodevelopmental disorders and cancer. Despite this, the progression of methylation throughout development remains elusive. This is largely due to the challenges of separating cell state transitions from bulk data and the inability to assess rare cell types. Recent advances have enabled high cell count, low coverage strategies to interrogate various properties of single cells. However, DNA methylation has lagged behind due to the chemical challenges of bisulfite sequencing. To overcome these challenges and allow comprehensive cataloguing of methylation cell types at scale, new approaches are needed. We developed a novel single-cell combinatorial indexing method for methylation assessment (sci-MET), which for the first time allows truly scalable production of whole genome methylation data from single cells. Combinatorial indexing strategies leverage multiple rounds of random sampling and barcoding on pooled nuclei to identify library molecules at single-cell resolution. Our sci-MET method adapts this strategy, using bisulfite-resistant tagmentation barcodes, random primer incorporation of a reverse adaptor, and barcoded PCR to achieve a throughput that far exceeds previous single-cell, single-well methods. Final cell count is readily scalable by additional barcodes at any stage. We demonstrate the ability of sci-MET to produce high quality single-cell bisulfite sequencing libraries in bulk samples with minimal barcode collisions. We produced a total of 1,830 sci-MET libraries passing quality control, exceeding the count of all previously published single-cell methylomes by 4.5-fold. Furthermore, our libraries show improved quality metrics compared to those from previous protocols. Single-cell alignment rates exceed existing methods (69% +/- 7% v. 25% +/- 20% for previous single-cell methods), making them comparable to bulk cell preparations. Libraries exhibited high complexity and are projected to saturate at >2x10^{6} unique CG dinucleotides sampled per cell, comparable to other low throughput studies. We extend our assay to a proof-of-principle experiment where we successfully deconvolve cells from a mix of three human cell lines based on their methylation profiles using a naïve clustering approach, demonstrating its potential utility to assess sub-populations of cells in complex tissues.

213

X chromosome inactivation in human single cells. *C. Borel[1], M. Garieri[1], G. Stamoulis[1], E. Falconnet[1], P. Ribaux[1], F. Santoni[1,2], S.E. Antonarakis[1,2,3].* 1) University of Geneva, Geneva, Switzerland; 2) University Hospitals of Geneva, Geneva, Switzerland; 3) iGE3, Geneva, Switzerland.

X chromosome inactivation (XCI) is a phenomenon where in each female cell one of the two X chromosomes is randomly silenced. However, some genes on the silenced X chromosome escape from XCI and are expressed from both X chromosome alleles. To date, the majority of X inactivation studies were performed in populations of cells (bulk) from organs and tissues. The ability to study the single cell transcriptome (scRNA-seq) now provides an unprecedented opportunity to revisit the X-inactivation in single cells. We have studied 902 single cell fibroblasts from five female individuals and performed scRNA-seq in order to investigate XCI at single cell resolution. To this aim we developed a computational and statistical framework integrating single cell transcriptome and whole genome sequencing to robustly eliminate confounding artifacts and identify genes which escape X inactivation. We identified 55 genes as escapees (34 known and 21 novel genes significantly escaping X inactivation). Among cells coming from the same individual, we observed that each gene exhibited a variable propensity to escape XCI in each single cell. Unexpectedly, through the calculation of an Inactivation Score as the mean of the allelic expression profiles of the escapees per cell, we discovered some cells being “inactive”, i.e. exclusively expressing the escaping genes from the active allele and others being “escapers”, i.e. expressing the escapee from both alleles. A possible mechanism to explain this cellular heterogeneity is the single cell variability of *XIST* transcription which we revealed being positively correlated (rho=0.45, p<10^{-5}) with the Inactivation Score. These results provide evidence of an unexpected cellular heterogeneity of the mechanism of X-inactivation. CB, FS, SEA contributed equally to this work.

214

SAVER: Gene expression recovery for single cell RNA sequencing. *M. Huang[1], J. Wang[1], E.A. Torre[2], H. Dueck[3], S.M. Shaffer[2], R. Bonasio[4], J.I. Murray[3], A. Raj[2], M. Li[5], N.R. Zhang[1].* 1) Department of Statistics, University of Pennsylvania, Philadelphia, PA; 2) Department of Bioengineering, University of Pennsylvania, Philadelphia, PA; 3) Department of Genetics, University of Pennyslvania, Philadelphia, PA; 4) Department of Cell and Developmental Biology, University of Pennsylvania, Philadelphia, PA; 5) Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA.

Rapid advances in massively parallel single cell RNA sequencing (scRNA-seq) is paving the way for high-resolution single cell profiling of biological samples. In most scRNA-seq studies, only a small fraction of the transcripts present in each cell are sequenced. The efficiency, that is, the proportion of transcripts in the cell that are sequenced, can be especially low in highly parallelized experiments where the number of reads allocated for each cell is small. This leads to unreliable quantification of lowly and moderately expressed genes, resulting in extremely sparse data and hindering downstream analysis. To address this problem, we propose SAVER (Single-cell Analysis Via Expression Recovery), an expression recovery method for scRNA-seq that borrows information across genes and cells to impute the zeros as well as improve the expression estimates for all genes. SAVER first employs a regularized Poisson regression across genes to obtain prior predictions. It then performs adaptive shrinkage estimation based on the predictions to arrive at a posterior gamma distribution for each gene in each cell. The SAVER estimate is the posterior mean, and estimation uncertainty is quantified by the posterior distribution. We show that the SAVER estimate is a weighted average of the prediction and the observed expression, where the weights are adaptively learned from the data. Through data down-sampling simulations, we show that true expression patterns, such as gene-to-gene and cell-to-cell correlations as well as expression differences between conditions, are weakened by technical noise as efficiency decreases. SAVER is able to recover the correlations and improve power in differential expression analysis. Next, we apply SAVER to a melanoma cell population sequenced by Drop-seq and profiled by RNA fluorescence in situ hybridization (FISH). RNA FISH is widely considered as a gold standard in gene expression quantification. SAVER performs better than using the observed Drop-seq data in recovering gene-to-gene correlations and gene distribution characteristics such as the Gini coefficient (RMSE = 0.19 vs 0.45). Finally, we apply SAVER to a differentiating mouse embryonic stem cell dataset to demonstrate its ability to detect known relationships between pluripotency factors and differentiation markers such as *Sox2* and *Krt8*. These results show that SAVER recovers true expression patterns and captures biological variation while removing technical noise. .

215

Clinical utility and cost effectiveness of rapid whole genome sequencing in the neonatal and pediatric intensive care unit. *S. Chowdhury, S. Nahas, M. Bainbridge, S. Batalov, J. Cakici, S. Caylor, Y. Ding, L. Farnaes, J. Friedman, K. Gil, A. Hildreth, R. Hovey, L. Puckett, L. Salz, L. van der Kraan, N. Veeraraghavan, S. White, M. Wright, C. Yamada, N. Sweeney, D. Dimmock, S. Kingsmore.* Genomics, Rady Children's Institute for Genomic Medicine, San Diego, CA.

Introduction: Initial studies have shown that rapid whole genome sequencing (rWGS) in the NICU has resulted in reduced time to diagnosis and improved diagnostic rates compared to the current standard of care. Children with severe chronic illness, often due to genetic disease, represent 70% of healthcare costs. Thus, this testing holds potential to impact healthcare economics worldwide. We sought to investigate the clinical utility of rWGS in a broad range of infant inpatients at Rady Children's Hospital. **Methods:** Following nomination by a treating physician and parental consent, blood samples were collected from ill inpatient infants and parents when available ideally within the first 48-72 hours of admission. PCR-free rWGS was performed to 40-45X coverage. Phenotypic features of the proband were translated into Human Phenotype Ontology terms and mapped to potentially causative genetic diseases. DNA sequences were aligned and variants identified using commercial tools. Variants were curated and prioritized changes reviewed by board certified geneticists. Clinical confirmation was performed for all clinically significant results. Precision medicine recommendations were verbally communicated to clinicians. **Results:** After 10 months of testing, 94% of families approached enrolled. WGS was interpreted in 98 families, yielding diagnostic information in 34 families (35%). On average, diagnosis occurred within one week (fastest 37 hours). Changes in management as a result of diagnosis were identified in 28 families (80% of diagnosed patients, 28% of all sequenced patients). The changes in management ranged from specific medications targeted to the underlying disease, changes in surgical interventions, to palliative care guidance. Among the first 42 infants, rWGS provided over $1.3M in net cost savings over projected standard care. **Conclusion:** Consistent with other studies, rWGS has a high diagnostic yield and reduces time to diagnosis. This study demonstrates that early diagnosis changes acute management. rWGS improves clinical care, preventing disability and unnecessary procedures, while simultaneously reducing acute care costs among a broad cohort of quaternary childrens hospital inpatient infants. .

216

Utility of exome sequencing for infants in intensive care units: Ascertainment of severe single-gene disorders and impact on medical management. *L. Meng[1,2], M. Pammi[3], A. Saronwala[1], P. Magoulas[1], A. Ghazi[1], F. Vetrini[2], W. He[2], A. Dharmadhikari[2], C. Qu[2], X. Ge[1], M. Tokita[1], T. Santiago-Sim[1], H. Dai[1], H. Smith[1], M. Azamian[1], M. Wangler[1,4], D. Scott[1,5], J. Belmont[1], X. Wang[1,2], M. Leduc[1,2], R. Xiao[1,2], P. Liu[1,2], C. Shaw[1,2], M. Walkiewicz[1,2], W. Bi[1,2], F. Xia[1,2], B. Lee[1,2], C. Eng[1,2], Y. Yang[1,2], S. Lalani[1,2].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Baylor Genetics Laboratory, Houston, Texas; 3) Department of Pediatrics, Section of Neonatology, Baylor College of Medicine, Houston, Texas; 4) Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, Texas; 5) Department of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston Texa.

Importance: While congenital malformations and genetic diseases are a leading cause of early infant death, the contribution of single-gene disorders in this group is undetermined. **Objective:** To determine the diagnostic yield and utility of clinical exome sequencing in critically ill infants. **Design, setting, participants:** Clinical exome sequencing was performed on 278 unrelated infants within the first 100 days of life (median age: 28 days), admitted to Texas Children's Hospital in Houston, over a period of five years, between December 2011 and January 2017. Exome sequencing performed on the infants included proband exome, trio exome, as well as critical trio exome, a rapid genomic assay for seriously-ill infants. **Main outcomes and measures:** Indications for testing, diagnostic yield of clinical exome sequencing, turnaround time, molecular findings, and impact on medical management in a group of critically ill infants suspected to have genetic disorders. **Results:** Clinical indications for exome sequencing included a wide range of medical concerns such as multiple congenital anomalies, liver failure, neonatal seizures, syndromic congenital heart defects, cardiomyopathy, brain malformations, skeletal dysplasia, immunodeficiency, neonatal hypotonia, and lactic acidosis. Overall, molecular diagnosis was achieved in 102/278 infants by clinical exome sequencing with a diagnostic yield of 36.7%. Noticeably, critical trio exome revealed a molecular diagnostic rate of 50.8% (32/63) with turnaround time (TAT) of 13.0 ± 0.4 days. The most frequent diagnoses in this group were Kabuki and Noonan syndromes, with neonates presenting with left ventricular outflow tract obstructive defects and cardiomyopathy, respectively. Four out of the 102 diagnosed infants were found to have a dual molecular diagnosis (3.9%). Overall, genetic diagnosis affected medical management in 53/102 (52.0%) of infants, with substantial impact on informed redirection of care, initiation of new subspecialist care, medication/dietary modifications, and furthering life-saving procedures in select patients. Of the deceased infants (n=81), genetic disorders were molecular diagnosed in 39 (48.1%) by exome sequencing with implications for recurrence risk counseling. **Conclusions and relevance:** Exome sequencing is a powerful tool for the diagnostic evaluation of critically ill infants with suspected monogenic disorders in the neonatal and pediatric ICUs, leading to notable impact on clinical decision-making.

217

Clinical benefit of whole genome sequencing: A report on 300 families. *A. Narravula[1], A.M. Bertoli-Avella[1], Z. Yüksel[1], O. Brandau[1], J. Balázs[1], J.M. Garcia-Aznar[1], C. Baldi[1], O. Paknia[1], M. Weber[1], M.E.R. Weiss[1], H. Yavuz[1], S. Franzenburg[1], K.K. Kandaswamy[1], D. Trujillano[1], N. Nahavandi[1], A. Al Shamsi[2], M. Alfadhel[3], M.A. Albalwi[4], N. Al-Sannaa[5], S. Kishore[1], P. Bauer[1,6], A. Rolfs[1,7].* 1) CENTOGENE AG, Rostock, Germany; 2) Department of Pediatrics, Tawam Hospital, Al-Ain, United Arab Emirates; 3) King Abdullah International Medical Research Centre, King Saud bin Abdulaziz University for Health Science, Genetics Division, Department of Pediatrics, King Abdulaziz Medical City, Riyadh, Saudi Arabia; 4) King Abdullah International Medical Research Centre, King Saud bin Abdulaziz University for Health Science, Department of Pathology and Laboratory Medicine, King Abdulaziz Medical City, Riyadh, Saudi Arabia; 5) Johns Hopkins Aramco Health Care, Pediatric Services, Dhahran, Saudi Arabia; 6) Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany; 7) Albrecht-Kossel-Institute, University of Rostock, Rostock, Germany.

Introduction: Clinical practice in medical genetics has dramatically changed in the last decade with the introduction of next generation sequencing. Today, whole genome sequencing (WGS) holds the potential to bring human genetics to a higher level by ending the long and arduous diagnostic odyssey for previously undiagnosed patients including those who have had previous negative whole exome sequencing (WES) analysis. **Patients and Methods:** Samples from 300 families underwent clinical WGS analysis at Centogene, using Illumina's technology. Diagnostic yield was calculated for the total group and subgroups with (n=212) or without (n=88) prior whole exome sequencing (WES). Most of the families had undergone WES previously (70.7%). **Results:** Overall, pathogenic or at least likely pathogenic variants were identified in 54 index cases from the 300 families (18%). Additionally, in 16% of the cases (n=48) a variant of uncertain significance (VUS) was reported, for a total diagnostic yield of 34%. Importantly, in 31 of the 212 cases with prior negative WES analysis, a genetic diagnosis was achieved by WGS (14.6%). In 28 additional cases (13.2%) a VUS was identified (total 27.8%). Positive WGS diagnosis following a negative WES was due to deep intronic variants known to be causative, exon deletions or previously uncovered regions. For the group without prior WES (n=88) the diagnostic yield was lower than expected with 26.1% (n=23). Additionally, in 22.7% (n=20) a VUS was reported. **Conclusion:** These results highlight the diagnostic strength of WGS in previously undiagnosed, complex cases and WGS should be considered as the first line test for these cases. The lower than expected yield for cases without a prior WES can likely be attributed to selection bias from cases with limited clinical information or due causative variants in genes yet to be associated with disease amongst others. As demonstrated here, WGS is able to diagnose many patients in whom all previous genetic testing, including WES, failed to identify the suspected genetic cause.

218

The breadth of genomic variation detected by clinical whole genome sequencing: An iHope Program Cohort summary. *E. Thorpe[1], A. Scocchia[1], J. McEachern[1], M.C. Jones[2], D. Masser-Frye[2], D. Henry[3], R. Ortiz[4], S.S. Ajay[1], M. Bennett[1], K. Bluske[1], C.M. Brown[1], N.J. Burns[1], A. Chawla[1], A.J. Coffey[1], M.L. Cremona[1], M. Eberle[1], V.G. Gainullin[1], A. Gross[1], R.T. Hagelstrom[1], W.L. Li[1], A. Malhotra[1], D.L. Perry[1], M. Rajan[1], V. Rajan[1], J.W. Belmont[1], D.R. Bentley[1], R.J. Taft[1], ICSL software development team; ICSL variant curation team.* 1) Illumina Inc., San Diego, CA; 2) University of California San Diego, San Diego, CA; 3) Department of Pediatrics, University of California San Francisco, San Francisco, CA; 4) Rare Genomics Institute, Downey, CA.

Clinical whole genome sequencing (cWGS) combines the simultaneous detection of single nucleotide variants (SNVs), indels, copy number variants (CNVs), aneuploidy, and other chromosomal changes. Through Illumina's iHope program, a philanthropic initiative aimed at identifying the genetic cause of rare disease in children, the TruGenome Undiagnosed Disease™ cWGS test was donated to partner sites including the Rare Genomics Institute, UCSF Benioff Children's Hospital, and the Hospital Infantil de las Californias. We present a case series from this partnership to date. Forty-six patients and their families (eight duos, 36 trios, and two quads) pursued cWGS. Genomic findings with clinical congruence to the reported phenotype were identified in 52% (24/46) of patients. Results spanned a range of variation type, including chromosomal aneuploidy (n=2), uniparental disomy (n=1), CNVs > 10 kb (n=6) and SNVs or small indels < 12bp (n=12). Compound heterozygous variant pairs involving SNVs and larger indels (18bp and 46bp deletions) were also reported (n=2), and a dual diagnosis including aneuploidy and a SNV was obtained (n=1). Variants of unknown significance for clinical consideration were reported in four additional patients. Some findings were particularly notable. In one patient, a heterozygous, likely mosaic, *de novo* deletion was detected in a non-coding, differentially methylated region upstream of an imprinted gene. A similar deletion is described in the literature in a single patient with Kagami-Ogata syndrome. In two patients, the detection of multiple large CNVs suggested the presence of unbalanced translocations, where analysis of short-read sequencing data at the breakpoints for one family supported a balanced form in the unaffected father and sister. In one patient with a broad differential, a compound heterozygous variant pair including a SNV and 18bp deletion was detected in the NUP93 gene, conferring a molecular diagnosis of steroid-resistant nephrotic syndrome and resulting in the patient being referred for transplant evaluation. This iHope case series demonstrates the breadth of genomic variation detectable through a single cWGS test, including case examples that may not be identified until a multi-platform testing approach is deployed. The iHope initiative has recently expanded to involve a network of collaborating WGS testing institutions, which will likely provide further insight into the benefits of a comprehensive cWGS approach.

219

Increased rates of diagnosis and precision medicine with genomic sequencing compared to chromosomal microarray: A meta-analysis of 19,714 infants and children with likely genetic diseases. *D. Dimmock[1], M.M. Clark[1], Z. Stark[2], L. Farnaes[1,3], T.Y. Tan[2,4], S.M. White[2,4], S.F. Kingsmore[1].* 1) Rady Children's Institute for Genomic Medicine, San Diego, CA; 2) Murdoch Children's Research Institute, Melbourne, Australia; 3) Department of Pediatrics, University of California San Diego, San Diego, CA; 4) University of Melbourne, Melbourne, Australia.

Genetic diseases are a leading cause of childhood mortality. Timely establishment of an etiologic diagnosis in children with likely genetic diseases is critically important for effective treatment (precision medicine) and optimal outcomes. Whole genome sequencing (WGS) and whole exome sequencing (WES) have started to gain broad use for the diagnosis of children with suspected genetic diseases. By allowing concomitant examination of all or most genes in a differential diagnosis, these methods have the potential to permit timely ascertainment of genetic diseases. During the past six years, WGS and WES methods have improved substantially but guidelines for their use by clinicians are lacking. To improve understanding of the current evidence, we conduct a systematic review of the literature published since 2011 for studies of diagnostic or clinical utility of WGS, WES, or chromosomal microarray (CMA) in children with suspected genetic diseases and perform a random-effects meta-analysis. Among 25 published studies comprising 19,714 affected children, the rate of diagnosis by WGS (0.40, 95% CI 0.28-0.53) and WES (0.33, 95% CI 0.30-0.36) were significantly greater than CMA, the current first tier-test for several types of genetic diseases (0.11, 95% CI 0.09-0.13, p < .0001). There was no significant difference between WGS and WES diagnostic rates. The clinical utility of WGS (0.37, 95% CI 0.22-0.54) and WES (0.20, 95% CI 0.11-0.32) were higher than CMA (0.06, 95% CI 0.05-0.07, *P* < .002). Among studies that sequenced trios and singletons, the diagnostic rate of trio sequencing (0.33, 95% CI 0.29-0.37) was higher than singleton sequencing (0.23, 95% CI 0.20-0.26, p < .0001). Moreover, the diagnostic rate of genomic sequencing with a deep proband phenotype (0.41, 95% CI 0.34-0.48, hospital setting) was higher than that of reference laboratories (0.29, 95% CI 0.27-0.31, p=0.0007). Affected infants had a higher rate of diagnosis than other groups. In two studies of infants, early diagnosis (day of life [DOL] 49) led to greater implementation of precision medicine (0.65) than later diagnosis (DOL 374, 0.35, P < .02), particularly with regard to palliative care guidance (0.30 vs 0.00, P < .01). This meta-analysis, the first of its kind, indicates that genome sequencing should be considered as a first tier test for infants and children with likely genetic diseases. Additional studies are needed to examine determinants of clinical utility.

220

An automated reanalysis pipeline for clinical exome data reveals novel diagnoses. *S.W. Baker[1], J. Murrell[1], B. Krock[1], A. Santani[1,2].* 1) The Children's Hospital of Philadelphia, Philadelphia, PA; 2) Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA.

Clinical Exome Sequencing (CES) is a diagnostic test for patients with suspected rare genetic conditions and has a reported diagnostic yield of 20% to 30%. As new discoveries of disease causing genes and disease causing variants are made, the diagnostic utility of CES is expected to further increase. Each year, hundreds of gene-disease and thousands of variant-disease associations are reported, highlighting the necessity for reevaluating nondiagnostic CES cases using the most up-to-date literature. With rapidly growing CES cohorts and literature, performing reanalysis in a time and cost-effective manner represents a significant challenge for clinical laboratories. Here, we present an automated CES data reanalysis method to address this challenge. Our method utilizes annotations that are automatically derived from PubMed abstracts, the Online Mendelian Inheritance in Man catalog, the ClinVar archive, and the Human Gene Mutation Database and significantly reduces the time required for CES test reanalysis. Using a cohort composed of the first 300 CES cases analyzed by the Division of Genomic Diagnostics at the Children's Hospital of Philadelphia, we demonstrate that our reanalysis method identifies the diagnosis in 100% of previously diagnostic cases and reveals novel diagnostic findings in 8.8% of previously non-diagnostic cases. Importantly, these results reflect diagnostic gain made through reinterpretation of existing CES data using newly available information. Additionally, to estimate the workload required for regular CES case reanalysis, we retrospectively examined the number of variants in non-diagnostic cases with novel annotations when reanalyses were performed monthly between January 2017 and April 2017. On average per month, previously non-diagnostic cases had 0 to 13.67 (mean=0.97) novel annotations. Importantly, there were no novel annotations for >55% of the non-diagnostic cases when reanalysis was performed at monthly intervals and 38.75% of non-diagnostic cases were not assigned novel annotations across all three consecutive re-analyses. Overall, our results illustrate both the utility of automating reanalysis of CES data, and that implementation of automated CES reanalysis strategies can benefit both patients and clinical laboratories.

221

Returning carrier status from genomic sequencing in newborns: Early observations from the BabySeq Project. *C.A. Genetti[1], G.E. VanNoy[1], S. Fayer[2], W. Betting[2], O. Ceyhan-Birsoy[3,4], K. Machini[3], J. Murry[3], M. Lebo[3], T. Yu[1,5,6], P.B. Agrawal[1,5,7], R.B. Parad[5,8], I.A. Holm[1,5], S. Pereira[9], A.L. McGuire[9], H.L. Rehm[3,5,10,11], R.C. Green[2,5,11], A.H. Beggs[1,5], The BabySeq Project.* 1) Division of Genetics and Genomics, The Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA; 2) Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA; 3) Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA; 4) Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY; 5) Harvard Medical School, Boston, MA; 6) Department of Neurology, Boston Children's Hospital, Boston, MA; 7) Division of Newborn Medicine, Boston Children's Hospital, Boston, MA; 8) Department of Pediatric Newborn Medicine, Brigham and Women's Hospital, Boston, MA; 9) Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX; 10) Department of Pathology, Brigham & Women's Hospital, Boston, MA; 11) The Broad Institute of MIT and Harvard, Cambridge, MA.

The BabySeq Project is the first randomized clinical trial assessing the impact of providing genomic sequencing (GS) information on newborns to their parents. Information in 3 categories is reported back to families: monogenic disease risk, carrier status for recessive conditions, and limited PGx risk. Here we report preliminary results of returning carrier status. Two cohorts, healthy infants and sick infants in the ICU, are being enrolled. All families receive state-mandated newborn screening and a family history assessment. Half of the infants are randomized to receive GS via exome sequencing. The GS analysis pipeline targets genes with strong evidence of association with pediatric onset disorders. Pathogenic or likely pathogenic variants are reported if penetrance is estimated to be high or moderate with actionability. Result disclosure is performed in-person by a genetic counselor and study physician. Of the first 107 newborns sequenced, 88% were enrolled from a well-baby nursery and 12% from an ICU. Carrier status was identified in 94 cases (88%). 207 variants were reported in 148 genes associated with autosomal recessive conditions, with an average of 2 variants per subject (range 0-6). The genes most frequently identified with carrier status variants were: *BTD* (biotinidase deficiency, n=11), *GJB2* (non-syndromic hearing loss, n=9), *RBM8A* (thrombocytopenia absent radius syndrome, n=9), *CFTR* (cystic fibrosis, n=6), *ABCA4* (Stargardt macular degeneration n=4), and *MUTYH* (*MUTYH*-related attenuated familial adenomatous polyposis, n=4). Approximately 70% of the carrier status variants identified on GS would not have been identified on commonly offered commercial expanded carrier screening panels. Post-disclosure surveys reveal that parents utilized their child's results to inform their own genetic screening. One family pursued preimplantation genetic diagnosis after follow-up sequencing found that both parents were carriers of variants in the same gene. In addition to early detection of a wide range of health risks, since GS detects a greater number of carrier status variants than panel-based carrier screening, parents are provided with a greater depth of risk information. Return of carrier status from newborn GS provides potentially beneficial information to two generations, allowing for immediate and future use in reproductive planning.

222

Genetic screening for healthy individuals: Preliminary results from a medically actionable genetic screening panel. *E. Haverfield[1], E.D. Esplin[1], S. Aguilar[1], K.E. Ormond[2], A. Hanson-Kahn[2], P. Atwal[3], S. Macklin[3], C. Sak[4], S. Bleyl[5], C. Fine[5], A. Lynch[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, Monterey, CA.

Introduction Motivated by the American College of Medical Genetics and Genomics (ACMG) policy statement on secondary findings, health-related genetic information is increasingly available to healthy individuals through their healthcare professionals. This information can identify hereditary disease risk and may lead to earlier detection and prevention. However, it must be accompanied by adequate educational support for clinicians and genetic counseling for patients. We report data on our early experience and initial findings related to the frequency of important variants found in a medically actionable genetic screening panel in healthy individuals. We will also describe several cases that highlight the value of different results and underscore the importance of genetic counseling when incorporating such results into routine healthcare. **Methods** Under an IRB-approved protocol, we analyzed de-identified data from 165 healthy individuals who had genetic screening with an expanded panel of up to 139 medically actionable genes. Clinician-documented health information, if provided, was also reviewed. **Results** Pathogenic/likely pathogenic (P/LP) results were observed in 17.6% (29 of 165) of cases, with findings in cancer-related genes (51.7%), cardiovascular-related genes including those associated with hereditary thrombophilia (36.4%), and bi-allelic P/LP variants in genes causing other medically actionable disorders (34.5%), such as hereditary hemochromatosis or alpha-1-antitrypsin deficiency. When we restricted our evaluation to the 56 genes originally recommended by the ACMG, the positive rate was 7.9% (13 of 165) or 3.0% (5 of 165) depending on whether *MUTYH* heterozygosity and other well-known moderate risk alleles are reported, or not, respectively. In cases where carrier status was reportable, 45.5% of individuals were identified as carriers for an autosomal recessive condition, predominantly hereditary hemochromatosis or alpha-1-antitrypsin deficiency. **Conclusions** Proactive health-related genetic information represents a unique and expanding area in which healthcare providers can educate those pursuing screening for genetic risks. This type of testing offers healthy individuals who would not otherwise have met diagnostic testing criteria, based on personal or family history, the opportunity to learn more about clinically significant genetic risks for certain types of hereditary disorders. .

223

Physicians' perspectives on returning unsolicited genomic results to patients and health care providers. *D.B. Pet[1], I.A. Holm[2,3], J.L. Willaims[4], M.F. Myers[5], L.L. Novak[7], K.B. Brothers[6], G.L. Wiesner[7], E.W Clayton[7,8].* 1) Vanderbilt University School of Medicine, Nashville, TN; 2) Boston Children's Hospital, Boston, MA; 3) Harvard Medical School, Boston, MA; 4) Geisinger Health System, Danville, PA; 5) Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 6) University of Louisville Medical School, Louisville, KY; 7) Vanderbilt University Medical Center, Nashville, TN; 8) Vanderbilt University School of Law, Nashville, TN.

Introduction: Genomic screening for disease risk in healthy individuals, for example in the context of research studies or commercial testing, can uncover unexpected pathogenic variants. When patients seek answers and treatment, physicians must address results they themselves did not order or "unsolicited genomic results" (UGR). Prior studies report physician discomfort with UGR and disagreement about who is medically responsible. A better understanding of physicians' views on UGR is crucial for integrating genomics into patient care. This study aimed to identify physicians' perspectives on actionability of and responsibility for UGR, impacts on patients and workflow, and support needs. **Methods:** Semi-structured interviews were conducted with adult and/or pediatric primary care and subspecialty physicians at four hospitals involved in a large-scale return-of-results project led by the Electronic Medical Records and Genomics (eMERGE) Consortium. Interviews occurred before eMERGE results were returned, eliciting views on forthcoming eMERGE results and UGR in general. Seventeen physicians were interviewed across five major domains: 1) actionability, 2) impacts on patients, 3) health care workflow, 4) return of results process/support, and 5) responsibility for results. The *Dedoose* qualitative analysis platform was used for analysis. **Results:** Physician participants questioned current utility of genomic screening, expressing concern that they will not know what to do with UGR. Physicians endorse the requirement of actionability and the need for clear, evidence-based "paths" or guidelines for action coupled with clinical decision support to ensure patient benefit. They identified potential harms to patients, including anxiety, false reassurance, and clinical disutility. Clinicians expressed resentment about anticipated workflow issues including time burden, detracting from other patient care responsibilities, and unreimbursed time. They disagreed about who was responsible for responding to UGR. Ethical concerns included patient privacy and insurance eligibility. **Conclusion:** This study suggests that the idea of receiving UGR for otherwise healthy patients raises important concerns for physicians with regard to utility and efficiency, patient well-being, and medical and ethical responsibility. Findings highlight the importance of strategic workflow integration of UGR, including well-developed clinical decision support as well as resources and safeguards for patients.

224

Non-inferiority of a web platform compared to in-person counselor return of carrier results: A randomized controlled trial. *B.B. Biesecker[1], K.L. Lewis[2], K.L. Umstead[1], J.J. Johnston[2], E. Turbitt[1], K.P. Fishler[2], J.H. Patton[2], I.M. Miller[2], A.R. Heidlebough[1], L.G. Biesecker[2].* 1) Social and Behavioral Research Branch, National Human Genome Research Institute, NIH, Bethesda, MD; 2) Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD.

Importance: A critical bottleneck in clinical genomics is the mismatch between large volumes of results and the capacity of standard results return by a genetic counselor. **Objective:** To test whether a web-based platform is as effective as a genetic counselor for educating patients and returning carrier results from exome sequencing. **Design:** A randomized non-inferiority trial from 2014-2016 was used to compare this education platform to a genetic counselor. Surveys were administered at baseline (T1), immediately following disclosure (T2), and six months later (T4). **Setting:** Longitudinal sequencing cohort at the National Institutes of Health. **Participants:** Of the eligible participants, 571 were heterozygous for a variant in at least one gene that causes a phenotype inherited in an autosomal recessive pattern. After 462 participants (81%) provided consent and were randomized, all but three participants (*N*=459) completed T2 surveys following education and counseling; 85% completed T4 surveys. **Interventions:** We designed a web-based platform that integrated education on carrier results with personal test results to parallel disclosure education by a genetic counselor. The sessions took 21 and 27 minutes, respectively. **Main Outcomes and Measures:** Our main outcomes were knowledge, test-specific distress, decisional conflict, and communication of results to family members. **Results:** Participants were on average 63 years old, and the majority of the sample were parents (76%). The web platform was also non-inferior to the genetic counselor on outcomes assessed at T4: knowledge (d_{mean}=-0.18; lower limit of 97.5% CI=-0.63; non-inferiority margin (d_{NI})=-1), test-specific distress (d_{median}=0; upper limit of 97.5% CI=0; d_{NI}=1), and decisional conflict (d_{mean}=1.18; upper limit of 97.5% CI=2.66; d_{NI}=6). There were no significant differences between modes of education delivery in disclosure rates to spouses, children, or siblings. **Conclusions and Relevance:** This trial demonstrates non-inferiority of web-based communication of carrier results and should spur efforts to shift the communication of lesser-impact genomic test results from the clinic to the Internet to improve efficiency and reduce healthcare costs.

225

Neanderthal introgression reintroduced thousands of ancestral alleles lost in the out of Africa bottleneck. *J. Capra, C. Simonti.* Vanderbilt University, Nashville, TN.

Anatomically modern humans (AMHs) interbred with Neanderthals approximately 50,000 years ago, and as a result, ~1–3% of the genomes of modern Eurasians are derived from of DNA introgressed from Neanderthals. Recent studies have focused on identifying and testing the effects of Neanderthal-derived alleles in AMHs, and these introgressed haplotypes have been shown to influence diverse phenotypes in AMHs including risk for many immune, skin, and neuropsychiatric diseases. However, recent analysis of an introgressed Neanderthal haplotype at the OAS locus revealed that the variant responsible for changes in gene expression is a reintroduced ancestral human allele, rather than a Neanderthal-derived allele. Motivated by this observation of the reintroduction of a functional ancestral allele that was lost in Eurasian populations in the out of Africa bottleneck, we performed a genome-wide search for other lost alleles on introgressed Neanderthal haplotypes in 1000 Genomes Phase 3 European (EUR), East Asian (EAS), and South Asian (SAS) individuals. In each super-population, we identified between ~47,000 and ~57,000 ancestral introgressed alleles. Consistent with the greater levels of Neanderthal DNA in Asian populations, these groups had more reintroduced ancestral alleles. Additionally, we found that nearly 66% of the reintroduced ancestral variants identified in EUR individuals are polymorphic in at least one non-admixed African sub-population (Yoruba, Esan, or Mende). These variants represent an extreme scenario where the ancestral allele was completely lost either before or during the out of Africa transition, and reintroduced by Neanderthal introgression. Many introgressed haplotypes carry more lost ancestral alleles than Neanderthal-derived alleles. Thus, we hypothesized that these alleles might influence phenotypes in modern Eurasian populations. To explore the potential function of these reintroduced ancestral alleles, we performed a phenome-wide association analysis on the introgressed alleles over more than 20,000 European-ancestry individuals with electronic health record data in Vanderbilt's BioVU databank. We identified and replicated several novel associations with clinical phenotypes that are likely driven by these reintroduced alleles.

226

Quantifying the impact of Neanderthal gene flow on human phenotypes. *C.R. Robles[1], A. Ganna[3,4,5,6], A. Gusev[4,7,8], D. Reich[4,9,10], S. Sankararaman[1,2].* 1) Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California Los Angeles, Los Angeles, CA 90095, USA; 2) partment of Computer Science, University of California, Los Angeles, CA 90095, USA; 3) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA; 4) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 5) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 6) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; 7) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; 8) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; 9) Department of Genetics, Harvard Medical School, Boston, MA 02115; 10) Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115.

Genomic analyses have revealed that all present-day non-African populations inherit 1-4% of their genetic ancestry from a population related to the Neanderthals. Given the high divergence of Neanderthals and modern humans, it has been hypothesized that admixture with Neanderthals may have had a major impact on human biology. Previous analyses of the locations of Neanderthal segments within non-African genomes indicate that some of the Neanderthal variants were adaptively beneficial while the bulk of Neanderthal variants were deleterious in the modern human genetic background. Nevertheless, little is known about the impact of Neanderthal ancestry on specific phenotypes. While it is possible to use GWAS to associate Neanderthal alleles with specific phenotypes, the low frequency of Neanderthal mutations (average frequency of 2.5%) substantially reduces the power of such analyses. In order to discover how Neanderthal ancestry affects specific phenotypes, we analyzed the interim release of the UK Biobank dataset (UKBB) consisting of approximately 150,000 genotyped individuals. We identified a set of Neanderthal tag SNPs that captures all confident Neanderthal informative mutations (NIMs) from a previously generated map of Neanderthal ancestry in 1000 Genomes Phase 1 data at r^2>=0.8. These tag SNPs were added onto the UK Biobank Axiom array enabling accurate imputation of the full set of NIMs as well as improved power to detect associations. Genotype imputation yielded 80,159 Neanderthal informative mutations (NIMs) that also passed QC in the UK Biobank data. We performed an association analysis of the NIMs to examine 87 quantitative phenotypes in a subset of 78,801 individuals while using the initial UKBiLeve cohort of 41,397 individuals for replication. We identified several novel NIMs significantly associated with modern phenotypes. We found a total of 35 NIMs that were associated with five phenotypes (Standing height, ankle spacing width, Heel Broadband ultrasound attenuation, Heel quantitative ultrasound index (QUI), Heel bone mineral density (BMD)) at a Bonferroni-corrected significance level of p=5e-10 (correcting for the total number of phenotypes as well as SNPs tested). Associations for four of these five phenotypes were replicated in the validation cohort. These results add to a growing picture of the influence of Neanderthal ancestry on phenotypes in present-day humans.

227

Interpreting human genomic regions depleted of archaic hominin ancestry. *A.B. Wolf[1,2], J.M. Akey[2].* 1) Genome Sciences, University of Washington, Seattle, WA; 2) Ecology and Evolutionary Biology, Princeton University, Princeton, NJ.

Statement of Purpose: Recent studies have identified archaic sequences in modern human genomes that were inherited from Neandertals and Denisovans. Strikingly, the distribution of archaic sequence in the modern human genome is heterogeneous, with some large regions depleted of it. Regions depleted of archaic sequence may represent loci where archaic sequence was strongly deleterious and rapidly purged from modern human populations. Alternatively, the stochastic loss of archaic sequences due to drift could also contribute to these “archaic deserts”. Understanding the formation and characteristics of archaic deserts in the modern human genome will help interpret how archaic admixture influenced human evolution and, possibly, what genes play a role in unique human behaviors. **Methods Used:** We identified introgressed archaic hominin sequence in 503 European, 504 East Asian, and 27 Melanesian individuals using the S* pipeline and the Altai Neandertal and Denisovan reference genomes. We compared empirical data to data from extensive coalescent simulations of a wide variety of neutral demographic models. As well, we used a range of foward-time simlations that included selection to test the effect of deleterious archaic mutations on the distribution of introgressed sequence following admixture. Finally, we leveraged large-scale functional genomics data sets to characterize archaic deserts and to map putatively deleterious sites carried by Neandertals that may have contributed to the generation of deserts. **Summary of Results:** While introgressed archaic sequence appears throughout the modern human genome, several large regions are significantly depleted of it. The overlap of regions depleted of Neandertal and Denisovan sequence is significantly greater than expected due to chance. Modern humans are significantly more enriched for large depletions than expected under neutral models. In simulations that include selection, even weakly deleterious archaic alleles at low frequency cause rapid loss of introgressed archaic sequence in modern humans. The largest regions depleted of archaic sequence differ from the rest of the genome in several key characteristics, such as being significantly enriched for genes expressed in regions of the brain and differing in their levels of sequence diversity. The largest region depleted of archaic sequence contains the *FOXP2* gene, which is associated with speech and language and carries a regulatory change unique to modern humans.

228

Imputing ancient gene expression reveals significant differential expression in Neanderthal tissues. *L.L. Colbran[1], P. Evans[1], E.R. Gamazon[1,2], N.J. Cox[1,2], J.A. Capra[1-3].* 1) Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN; 2) Division of Genetic Medicine, Vanderbilt University, Nashville, TN; 3) Departments of Biological Sciences, Biomedical Informatics, and Computer Science, Center for Structural Biology, Vanderbilt University, Nashville, TN.

The sequencing of DNA extracted from fossils of extinct hominins has yielded significant insights into their phenotypes and evolution, but much about their biology remains inaccessible to study. For example, since RNA is not preserved in these fossils, we cannot directly evaluate differences in gene expression across tissues between modern and archaic humans. To overcome this challenge, we combined the PrediXcan approach with ancient DNA to study unobservable gene expression patterns in ancient human forms. PrediXcan (Gamazon et al. 2015) is a tool for imputing tissue-specific gene expression profiles from patterns of genetic variation in an individual. These imputed profiles of the genetically regulated component of expression have proven sufficiently accurate to identify known and novel associations between genes and phenotypes in modern populations. We constructed an atlas of predicted archaic hominin gene expression for 17,743 genes across 43 tissues by applying PrediXcan models trained to predict the genetically regulated component of gene expression in GTEx Consortium samples to the high quality whole genome Neanderthal and Denisovan sequences. In both hominins, preliminary analyses indicate that brain regions (particularly the hippocampus, anterior cingulate cortex, and basal ganglia), reproductive tissues (vagina, uterus), and spleen have highest proportion of differentially expressed genes compared to modern humans. Many of the most differentially expressed genes have been previously implicated in neurological and immune diseases, platelet count, and height. To gain insight into potential functional effects of genes differentially expressed in archaic hominins, we identified associations of these genes' expression with phenotypes derived from electronic health records in ~10,000 individuals of European descent. Expression levels of genes predicted to be differentially expressed in Neanderthal were associated with a wide range of phenotypes, including autoimmune diseases like lupus and rheumatoid arthritis, neurological phenotypes such as schizophrenia and mood disorders, and phenotypes related to type 1 diabetes, and menopause. Further study of phenotypes in modern humans with similar predicted expression to ancient hominins will enable the study of aspects of diverse ancient human forms that can never be directly observed, and inform how introgressed regulatory SNPs influence expression of particular genes and downstream phenotypes.

229

Resolving variant interpretation differences in ClinVar between 43 clinical laboratories. *S. Harrison[1,2], J. Dolinsky[3], H. Rehm[1,2,4], ClinGen's Sequence Variant Inter-Laboratory Discrepancy Resolution Task Team.* 1) Laboratory for Molecular Medicine, Partners HealthCare Personalized Medicine, Cambridge, MA; 2) Harvard Medical School, Boston, MA; 3) Ambry Genetics, Aliso Viejo, CA; 4) The Broad Institute of MIT and Harvard, Cambridge, MA.

Sharing data in ClinVar provides open access to variant classifications from many clinical laboratories. While the majority of classifications agree, ClinVar has shed light on the important issue of interpretation differences between laboratories, providing a valuable opportunity to resolve differences and positively impact patient care. Recent work with four clinical laboratories found that 53% of interpretation differences were resolved by either updating ClinVar with current internal classifications or reassessment of an older interpretation with current classification criteria (PMID: 28301460). With this finding in mind, ClinGen's Sequence Variant Inter-Laboratory Discrepancy Resolution team will encourage clinical laboratories with outlier interpretations to update ClinVar with current classifications and reassess remaining conflicts using current guidelines. To identify variants that could be resolved by this outlier strategy, interpretations from 43 clinical laboratories in ClinVar were compared, identifying 26,421 variants interpreted by ≥2 clinical laboratories. The majority of classifications were concordant (85.7%; 22,637 variants). Only 2.5% (667 variants) of all shared variants were medically significant differences (MSDs) with potential to impact medical management [pathogenic (P/LP) versus other (VUS/LB/B]. These differences were investigated to determine if submitted interpretations could reach a majority consensus (agreement in classification of at least 2/3 of clinical laboratory submitters). Of the MSDs with ≥3 interpretations (249 variants), 87.6% (218 variants) reached a majority consensus, thus allowing for identification of outlier submissions most in need of reassessment. Outlier submitters on variants with majority consensus will be contacted with a custom report and be encouraged to update ClinVar, if the classification has already changed internally, and reassess remaining outlier interpretations. If the discrepancy remains, other clinical laboratories will be encourage to share internal evidence to facilitate resolution. Based on our initial study results, it is anticipated that this process will resolve at least 79% of MSDs, reducing total MSDs to 0.5%. This process adds to the value of ClinVar and will help the community move toward more consistent variant interpretations which will improve the care of patients with, or at risk for, genetic disorders.

230

An interlaboratory study of complex variant detection in clinical testing. *S. Lincoln[1], J. Zook[2], R. Truty[1], S. Chowhurdy[3], A. Fellowes[4], S. Mahamdallie[5], M. Ferber[6], M. Cleveland[2], C. Huang[7], F. Tomson[7], E. Klee[6], W. DeSilva[4], S. Seal[5], S. Aradhya[1], R. Garlick[7], R. Nussbaum[1], N. Rahman[5], S. Kingsmore[3], M. Salit[2], B. Shirts[8].* 1) Invitae, San Francisco, CA, USA; 2) National Institute of Standards and Technology, Gaithersburg, MD, USA; 3) Rady Childrens Institute for Genomic Medicine, San Diego, CA, USA; 4) Peter MacCallum Cancer Centre, Melbourne, Australia; 5) The Institute of Cancer Research, London, UK; 6) The Mayo Clinic, Rochester, MN, USA; 7) SeraCare Life Sciences, Gaithersburg, MD, USA; 8) The University of Washington, Seattle, WA, USA.

NGS is a capable technique for detecting SNVs and small indels in tractable parts of the genome. However NGS can have limitations for other variant types. We examined one clinical laboratory's cohort of over 80,000 tested patients, and found that pathogenic, medically important variants of technically challenging types are prevalent: They comprise over 9% of pathogenic findings in hereditary cancer genes, 10% in cardiology, 12% in neurology, and 19% in metabolic syndromes. The technical challenges presented are diverse: 32% of the technically challenging pathogenic mutations are single or sub-exon CNVs, 18% are large indels or structural variants, 35% are in homopolymer, low complexity or unmappable regions, and 15% are poorly captured by standard exome kits. Despite their prevalence, developing and evaluating methods that detect such variants is difficult, in part because of the scarcity of positive controls. We selected 23 of these challenging variants and followed a methodology (previously demonstrated for simpler cases) where large synthetic constructs containing these variants were spiked into a known genomic background, creating a single control sample with all 23 alterations. This DNA was blinded and provided to 7 laboratories who sequenced it using a total of 9 NGS workflows, including 5 validated clinical tests with custom bioinformatics and two vendor (Illumina, Ion Torrent) default pipelines. Multiple target enrichment methods were used, as was whole genome sequencing. Twelve of 23 variants were detected by all 9 workflows, but just 2 workflows detected all 23. Importantly, evidence of each variant was present in the raw data files, suggesting that this control strategy is compatible with diverse biochemical methods. Raw data for the synthetic variants mimicked that of the endogenous ones (including presenting similar artifacts) demonstrating that controls such as these may be useful in the development of methods with improved sensitivity. The vendor-supplied bioinformatics pipelines fared the worst, reinforcing the importance of carefully selecting algorithms and parameters. In summary, medically important but technically challenging variants are prevalent across a range of inherited conditions. Although actual patient specimens are also critical, multiplexed synthetic controls may help efficiently assess the analytic range of a clinical test and can help laboratories develop new methods for these challenging variants.

231

The impact of sharing patient-derived data in ClinVar via GenomeConnect. *J.M. Savatt[1], D.R. Azzariti[2], W.A. Faucett[1], M. Landrum[3], D.H. Ledbetter[1], C. Lese Martin[1], V. Rangel Miller[4], E. Palen[1], H. Rehm[2,5,6,7], J. Rhode[4], S. Turner[1], E. Rooney Riggs[1].* 1) Genomic Medicine Institute, Geisinger Health System, Danville, PA; 2) Laboratory for Molecular Medicine, Partners Personalized Medicine, Boston, MA; 3) National Center for Biotechnology Information, Bethesda, MD; 4) Invitae, San Francisco, CA; 5) The Broad Institute of Harvard and MIT, Cambridge, MA; 6) Harvard Medical School, Boston, MA; 7) Department of Pathology, Brigham & Women's Hospital, Boston, MA.

Participants in GenomeConnect (GC), the Clinical Genome Resource patient registry, consent to have their self-reported health information and genomic variants from uploaded testing reports de-identified and shared with approved databases, such as NCBI's ClinVar. The goal of this type of patient data sharing is to increase the availability of genomic information to inform variant interpretation ultimately leading to improvements in patient care. To begin to assess the impact of patient-derived data sharing, we conducted a detailed examination of GC's first ClinVar submission. ClinVar records for each variant submitted by GC were reviewed for previous variant entries and conflicting interpretations. Of the 440 sequence variants submitted or ready for submission by GC, 55.5% have not previously been reported to ClinVar, demonstrating the importance of patients as a data source. The remaining 44.5% of variants have previously been submitted by a clinical laboratory; 59.7% (117/196) of these are from the same reporting laboratory as the GC participant's report. In 10.3% of these cases (12/117), we identified discrepancies in variant interpretation between the result reported to the participant and the laboratory's current entry in ClinVar. Although many laboratories attempt to inform clinicians about updated classifications, this information may not always reach the patient. Realizing that GenomeConnect could serve as a liaison to relay this potentially medically relevant information, we surveyed participants to determine their preferences for receiving such information. Of the 137 consented participants that completed the survey (response rate 19.6%), 99% indicated that they want to receive information about updated variant interpretations from GenomeConnect, and a process for providing these updates is now under development. Of previously submitted variants, 35.2% conflicted with other laboratory submissions. While we will not relay this information back to participants, GC will encourage laboratories to address these discrepancies. Providing patients with updated variant information and collaborating to resolve discrepancies will lead to improved clinical care. By engaging participants in genomic data-sharing efforts, GC contributes information to the public knowledge base that may not have otherwise been available, benefiting both patients and the genetics community.

232

Efficacy of reanalyzing negative clinical WES data to identify new genes in intellectual disability/congenital anomalies. *A. Bruel[1,2], S. Nambot[1,2], V. Quéré[1,2], M. Assoum[1,2], A. Vittobello[1,2,3], S. Moutton[1,2], N. Houcinat[1,2], D. Lehalle[1,2], N. Jean-Marçais[1,2], J. Thevenon[1,2], M. Chevarin[1,2], C. Poë[1,2], T. Jouan[1,2], P. Callier[1,2,3], A. Mosca-Boidron[1,2,3], E. Tisserand[1,2], C. Philippe[1,2,3], F. Tran Mau-Them[1,2,3], Y. Duffourd[1,2], L. Faivre[1,2], C. Thauvin-Robinet[1,2,3].* 1) Inserm UMR 1231, Genetics of Developmental disorders, Université de Bourgogne, Dijon, Bourgogne, France; 2) FHU-TRANSLAD, Université de Bourgogne/CHU Dijon Bourgogne, France; 3) UF diagnostic innovation des maladies rares Laboratoire de génétique moléculaire et de Cytogénétique, CHU Dijon Bourgogne, Dijon, France.

The global diagnostic yield of clinical whole-exome sequencing (WES) in intellectual disability (ID) and/or congenital anomalies (CA) is now about 30%, which means that 70% of patients remain without a molecular etiology after clinical WES. To go beyond the stringent criteria of the ACMG recommendations in variant interpretation, which are limited to established human disease genes, a further analysis extended to the mutated genes in a research environment appears essential to identify novel human disease genes. We performed a systematic research analysis of negative WES data from a clinical setting to create a screening strategy in a cohort of 500 patients with ID/CA and to identify novel molecular bases. In patients negative for causal variants in OMIM genes, the solo WES interpretation was extended through a translational research approach to all variants in order to identify candidate variants, by studying conservation, expression patterns, protein functions and interactions, model organisms, and a literature review. Intensive international data-sharing through public variant databases was used to identify additional patients carrying variants in the same gene, in order to draw definitive conclusions on their implication in the patient's phenotype. With this strategy, we identified or contributed to international collaborations for the identification of 45 disease-causing genes or candidates for ID/CA disorders in 250 cases. The genes were classed in five categories: 1) new genes, unknown in human diseases (17 genes); 2) new genotype-phenotype correlations for known genes supported by data-sharing (5 genes); 3) ultra-rare disorders with low recurrence (8 genes); 4) non-OMIM genes recently published or presented in international congresses (5 genes); 5) candidate genes (10 genes). Most of these resolved cases led to publications by our team or in collaboration with others. The second part of the cohort is still being analyzed. This study demonstrates the power of research reanalysis after negative clinical WES and shows that screening results can rapidly lead to diagnosis. This rapid and cost-effectiveness strategy increases the yield of cases resolved by WES to almost 50%. Despite using omics technologies, such as whole-genome sequencing, many new genes implicated in rare human diseases remain to be identified. The performance of WES will certainly improve in the future.

233

The genetic architecture of pediatric cardiomyopathy. *S. Ware[1], P. Dexheimer[2], S. Bhatnagar[2], A. Sridhar[1], J. Wilkinson[3], M. Tariq[4], J. Schubert[1], S. Colan[5], L. Shi[6], C. Canter[7], D. Hsu[8], S. Webber[9], D. Dodd[9], M. Everitt[10], P. Kantor[11], L. Addonizio[12], J. Jefferies[2], J. Rossano[13], E. Pahl[14], P. Rusconi[15], W. Chung[12], T. Lee[12], J. Towbin[16], A. Lal[10], E. Miller[2], H. Razoky[3], J. Czachor[3], L. Martin[2], B. Aronow[2], S. Lipshultz[3], Pediatric Cardiomyopathy Registry Study Group.* 1) Indiana University School of Medicine, Indianapolis, IN; 2) Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 3) Wayne State University School of Medicine, Detroit, MI; 4) University of Tabuk, Tabuk, Saudi Arabia; 5) Boston Children's Hospital, Boston, MA; 6) New England Research Institutes, Watertown, MA; 7) Washington University, St. Louis, MO; 8) The Children's Hospital at Montefiore, Bronx, NY; 9) Monroe Carell Jr. Children's Hospital at Vanderbilt, Nashville, TN; 10) Children's Hospital Colorado, Denver, CO; 11) Stollery Children's Hospital, Edmonton, AB, Canada; 12) Columbia University Medical Center, New York, NY; 13) Children's Hospital of Philadelphia, Philadelphia, PA; 14) Ann and Robert H. Lurie Children's Hospital, Chicago, IL; 15) University of Miami Miller School of Medicine, Miami, FL; 16) Le Bonheur Children's Hospital, Memphis, TN.

Pediatric cardiomyopathy (CM) is a rare disease that carries substantial morbidity and mortality. While adults and children have shared genetic causes of CM, the genetic basis of early onset CM is not well understood. To investigate the genetic architecture of pediatric CM, we performed exome sequencing on 528 unrelated individuals with familial or idiopathic CM recruited from 12 centers. Phenotypic subtypes were 53% dilated (DCM), 30% hypertrophic (HCM), 11% left ventricular noncompaction/mixed (LVNC/mixed) and 6% restrictive CM (RCM) patients, 55.5% of which were non-Hispanic Caucasian. Median age at diagnosis was 4.5 years (IQR: 0.5-13.2). Exome sequencing identified 793,725 total variants of which 139,386 were rare (MAF < 0.01; 264 rare variants/subject) and predicted to impact protein function. ACMG guideline-based manual interpretation of rare variants among 36 known CM genes identified 99 pathogenic/likely pathogenic variants in 145 subjects. This diagnostic yield of 15% in DCM, 44% in HCM, 50% in RCM, and 31% in LVNC/mixed phenotypes is slightly higher than yield in adult CM. Because exome variant interpretation is particularly challenging for autosomal dominant diseases such as CM, we developed a support vector machine learning approach for variant effect classification. The classifier was trained using 1415 ClinVar CM pathogenic/likely pathogenic variants of the larger list of 62 ClinVar CM genes as a positive control versus gene matched missense variants from the Exome Aggregation Consortium as a negative control. The Recursive Feature Elimination algorithm identified and ranked variant impact features contributing to classifier accuracy. Top-ranked features dealt with conservation, three dimensional, domain, chemical, fold, surface and relative positional properties of protein structure. The classifier performed well, agreeing with manual interpretation 84.2% of the time. Our classifier identified possible disease-causing variants in 46.6% of our cohort, compared with 7.4% of the 1000 Genomes cohort. Strikingly, at least 2 or more potential disease-causing variants were identified in 14.6% of our cohort versus 0.96% in 1000 Genomes ($p < 0.0001$). The frequent occurrence of individuals with multiple candidate causal alleles suggests that a strategy to identify interaction effects from multiple damaging alleles including those of additional non-classic CM genes may be essential to explain early onset Pediatric CM.

234

Congenital heart malformations in Sub-Saharan Africa and Asia: An exome sequencing study. *P. Kruszka[1], S.I. Berger[1], S.K. Hong[1], A.A. Adeyemo[2], P. Tanpaiboon[3], E.N. Ekure[4], M. Muenke[1].* 1) Medical Genetics Branch, NHGRI/NIH, Bethesda, MD; 2) Center for Research on Genomics and Global Health, NHGRI/NIH, Bethesda, MD; 3) Division of Genetics and Metabolism, Children's National Health System, Washington, D.C; 4) Department of Paediatrics College of Medicine, University of Lagos/Lagos University Teaching Hospital Idi-Araba, Lagos, Nigeria.

Background: Congenital heart disease (CHD) is the most common birth defect, affecting approximately 1% of all newborns. There have been multiple large studies genotyping humans with structural CHD in resource rich countries and very little study in developing nations. In this study, we focus on the genomic analysis of individuals with CHD in resource poor countries in Sub-Saharan African and Asia. **Methods:** Clinical examination and echocardiography were used to diagnose patients with both syndromic and isolated CHD. Exome sequencing, assembly, genotyping, and annotation were performed on probands and their parents (trios) by the National Intramural Sequencing Center (NISC). DNA variant list manipulation was performed using perl scripts developed by our group and significant findings were confirmed with Sanger sequencing. Copy number variations were evaluated using both the Illumina HumanExome BeadChip-12v1_A (Illumina Inc. San Diego, CA). Selected variants are being evaluated functionally in the mouse model and zebrafish model. **Results:** 124 probands have completed the pipeline including 111 parent-offspring trios and 13 probands with one parent. Probands included 83 (67%) from Sub-Saharan Africa and 41 (33%) from Asia. Tetralogy of fallot (TOF) was the most common CHD in our cohort occurring in 29 probands (23%), followed by ventricular septal defects in 27 (22%), and then pulmonary stenosis in 13 (10%). Large copy number variations were found in 19 probands (15%) with 22q11.2 deletion syndrome being most common (6%). Syndromic single gene disorders were found 11 probands (9%). Three probands had pathogenic variants in genes known to cause cardiomyopathies (*ACTC1, ACTN2, DSP*). Novel candidate genes were chosen using a strict criterion to minimize false positive. Sixteen genes met this criteria and zebrafish evaluation is currently underway to validate pathogenicity. **Conclusions:** We have initiated the largest CHD study in a non-European cohort using next generation sequencing project to search for novel genetic associations with CHD. We are now finding novel genes that explain the CHD phenotype. Similar to European populations, we found that CHD in diverse populations is enriched with genes associated with genetic syndromes in which CHD comprises a major part of the phenotype. More interesting is our new gene discovery that will increase our understanding of the genetic basis of CHD. .

235

***De novo* noncoding mutations in congenital heart disease.** *F. Richter[1,2], S. Morton[3], J. Homsy[3], A. Kitaygorodsky[4], H. Qi[4], N. Patel[5], K. Manheimer[1,2], D.E. Dickel[6], A. Visel[6,7,8], I. Barozzi[6], M. Linderman[9], G. Hoffman[10], E.E. Schadt[10], D. Jordan[10], R. Do[10], D. McKean[3], J. Priest[11], J.R. Kaltman[12], D. Srivastava[13], J. Yost[14], M. Tristani-Firouzi[14], M. Brueckner[15], E. Goldmuntz[16,17], Y. Shen[4], W.K. Chung[18], J.G. Seidman[3], C.E. Seidman[3], B.D. Gelb[2,8,19], Pediatric Cardiac Genomics Consortium.* 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, Harvard Medical School, Boston, MA; 4) Departments of Systems Biology and Biomedical Informatics, Columbia University Medical Center, New York, NY; 5) Icahn School of Medicine at Mount Sinai, New York, NY; 6) Functional Genomics Department, Lawrence Berkeley National Lab, Berkeley, CA; 7) U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA; 8) School of Natural Sciences, University of California, Merced, Merced, CA; 9) Middlebury College, Middlebury, VT; 10) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 11) Department of Pediatrics (Cardiology), Stanford University, CA; 12) Heart Development and Structural Diseases Branch, Division of Cardiovascular Sciences, NHLBI/NIH, Bethesda, MD; 13) Gladstone Institute of Cardiovascular Disease, San Francisco, CA; 14) University of Utah School of Medicine, Salt Lake City, UT; 15) Department of Genetics; Yale University School of Medicine, New Haven, CT; 16) Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 17) Children's Hospital of Philadelphia, Philadelphia, PA; 18) Departments of Pediatrics and Medicine, Columbia University Medical Center, New York, NY; 19) Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY.

Congenital heart disease (CHD), the commonest birth defect, is attributed primarily to genetic variation, but only 30% is explained, even after whole exome sequencing (WES). As causal variants result in low reproductive fitness, we hypothesized that *de novo* noncoding variants (DNVs) contribute to CHD pathogenesis. We performed whole genome sequencing (WGS) for 350 parent/affected child trios unexplained after WES, comparing DNVs to 514 Simons simplex control trios. We overlapped DNMs with 186 cardiac gene regulatory annotations, assigning DNVs to the closest downstream transcription start site within 10 kb. Burden of cardiac-associated DNVs was tested in 3 gene sets: human CHD, mouse CHD, and genes highly expressed during heart development. A p-value cut-off of 1.8e-4 was used as 92 independent hypothesis tests explained ≥99% of the variance of the highly correlated annotations. We averaged 72 and 75 *de novo* SNVs and indels/trio in cases and controls, respectively, with high PCR-confirmation rates (SNVs 98%, indels 85%). No annotation was enriched in controls. In cases, we observed DNV enrichment in mouse CHD gene H3K27ac/p300 enhancers (n=15, OR=11.4, p=3.7e-6) and their union with ChromHMM active heart enhancers (n=17, OR=5.1, p=5.8e-5). Affected probands constitute a 4% ascertainment differential compared to controls. Phenotype analysis showed noncoding DNVs were implicated more than exonic DNVs among those with isolated CHD (p=0.02). Sensitivity analysis demonstrated significance for all distance cut-offs 5-36 kb upstream of mouse CHD genes, so the 10-kb cut-off was not the driver. To consider more distant H3K27ac/p300 enhancers, we used joint membership of topologically associating domains, observing nominal DNV enrichment 140-190 kb upstream of mouse CHD genes in cases (OR>1.5, p<0.05). Two DNVs altered a *GATA4* cardiac enhancer previously validated in mice. Assessment of cardiac heart RNAseq from 2/17 probands with proximate H3K27ac/p300 enhancer DNVs show significantly lower expression of the mutated gene than in comparable tissues from CHD probands without variants (p=0.039, 10,000 permutations). Of 16 genes with DNVs that altered regulatory annotations defined in mice, 3 cause human CHD. Genes associated with DNVS were enriched in the human phenotype ontology term Autosomal Dominant (p=1.6e-5). These data are the first to implicate *de novo* variants in noncoding cardiac gene regulatory sequences in CHD.

236

Congenital heart defects in Bainbridge Ropers syndrome. *A. Srivastava[1], B. McGarth[2], R. K.C.[1], Y.C. Tsan[1], C.E. Keegan[3], A. Helms[4], S.L. Bielas[1,2].* 1) Human Genetics, University of Michigan Medical School, Ann Arbor, MI, USA; 2) Cellular and Molecular Biology Program, University of Michigan, Ann Arbor, MI, USA; 3) Department of Pediatrics, University of Michigan Medical School, Ann Arbor, MI, USA; 4) Internal Medicine of Cardiology, University of Michigan Medical School, Ann Arbor, MI, USA.

Covalent histone modifications play an essential role in gene regulation and cellular specification required for development. Mono-ubiquitination of histone H2A (H2AUb1) is a reversible transcriptionally repressive mark. Exchange of histone H2A mono-ubiquitination and deubiquitination reflects the succession of transcriptional profiles during development required to produce cellular diversity from pluripotent cells. *De novo* dominant truncating variants in *ASXL3* (*Additional-sex combs like 3*) have been identified as the genetic basis of Bainbridge Ropers Syndrome (BRS), characterized by characterized by persistent severe feeding difficulties, partially penetrant microcephaly, highly penetrant severe hypotonia, failure to meet developmental milestones including walking and highly penetrant nonverbal outcomes. ASXL3 is a component of the polycomb repressive deubiquitinase (PR-DUB) complex, which deubiquitinates H2AUb1. Dysregulation of H2AUb1 was identified as key molecular pathology in primary cells derived from individuals with BRS. The *Asxl3* null animal was generated to investigate the molecular and developmental features of this clinically relevant gene. Constitutive loss of *Asxl3* results in highly penetrant heart defects and perinatal lethality. *Asxl3*$^{-/-}$ mice display severe cardiac hypertrophy and septal defects not previously observed in individuals with BRS. A retrospective echocardiogram analysis of individuals with BRS revealed a partially penetrant bicuspid valve phenotype in >30% of those analyzed. To understand the role of ASXL3-dependent chromatin modifications we analyzed cardiomyocytes differentiated induced pluripotent stem cell (iPSC) lines reprogrammed from individuals with BRS and controls. We analyzed the developmental transcriptomes of cardiomyocytes, enriched for cell-specific surface markers which indicates differential expression of genes important for cardiac lineage specification. This is the first study to evaluate congenital heart phenotypes in the clinical synopsis of BRS. These findings also underscore the important chromatin modifying roles of ASXL3 in polycomb transcriptional repression during organ development.

237

Multi-tissue transcriptome analysis reveals genetic mechanisms of neuropsychiatric traits. *E.R. Gamazon[1], A. Zwinderman[2], N. Cox[1], D. Denys[2], E. Derks[3].* 1) Vanderbilt University, Nashville, TN; 2) AMC, University of Amsterdam; 3) QIMR Berghofer, Brisbane, Australia.

The genetic architecture of psychiatric disorders is characterized by a large number of small-effect variants located primarily in non-coding regions, suggesting that the underlying causal effects may influence disease risk by modulating gene expression. We provide comprehensive analyses using transcriptome data from an unprecedented collection of tissues (889 brain RNA-Seq samples from 10 brain regions and 633 non-brain samples) to gain pathophysiological insights into the role of the brain, neuroendocrine factors (adrenal gland) and gastrointestinal systems (colon). Among significant (FDR<0.05) expression Quantitative Trait Loci (eQTLs), we identify functional target genes for Attention Deficit Hyperactivity Disorder (N_{genes}=4), Bipolar Disorder (N_{genes}=9), and Schizophrenia (N_{genes}=874). Colon-expressed *PLEK2* is a depression-associated gene; this finding is replicated in two independent datasets. Our multi-tissue analysis, which captures effect size heterogeneity and adjusts for differences in eQTL discovery power, demonstrates that 34% of best eQTLs per eGene in colon that are not also eQTLs in brain are likely to be true associations with schizophrenia. We develop a novel systems genetics approach, using the genetically determined component of gene expression, to identify a limited number of genetically defined clusters for disease-associated genes. Finally, we propose gene mechanisms for the well-known 108 schizophrenia loci, including multiple replicated genes (beyond the Complement Component 4A [*C4A*]) in the MHC region. Our analyses highlight the importance of multi-tissue approaches, including most notably the use of non-brain transcriptomes, for interrogating the genetic architecture of psychiatric traits, as 70% of the genes are detected only in inaccessible tissues.

238

Genetically driven gene expression associated with nicotine dependence across ten brain regions reveals novel genes and brain regions. *C.A. Markunas[1], D.B. Hancock[1], Y. Guo[1], G.W. Reginsson[2], R. Sherva[3], A. Loukola[4], C. Minica[5], N.C. Gaddis[1], S.M. Lutz[6], D.F. Gudbjartsson[2,7], K.A. Young[6], D.W. McNeil[8], B. Qaiser[4], P.A.F. Madden[9], L.A. Farrer[3], J. Vink[5,10], N.L. Saccone[9], M.C. Neale[11], H.R. Kranzler[12], M.L. Marazita[13], D.I. Boomsma[5], J. Gelernter[14,15], J. Kaprio[4], N. Caporaso[16], T.E. Thorgeirsson[2], J.E. Hokanson[6], L.J. Bierut[9], N. Cox[17], K. Stefansson[2], E.O. Johnson[1].* 1) RTI International, Research Triangle Park, NC, USA; 2) deCODE Genetics / Amgen, Reykjavik, Iceland; 3) Boston University, Boston, MA, USA; 4) University of Helsinki, Helsinki, Finland; 5) Vrije Universiteit, Amsterdam, The Netherlands; 6) University of Colorado Anschutz Medical Campus, Aurora, CO, USA; 7) University of Iceland, Iceland; 8) West Virginia University, Morgantown, WV, USA; 9) Washington University, St. Louis, MO, USA; 10) Radboud University, Nijmegen, The Netherlands; 11) Virginia Commonwealth University, Richmond, VA, USA; 12) University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA; 13) University of Pittsburgh, Pittsburgh, PA, USA; 14) Yale University School of Medicine, New Haven, CT, USA; 15) VA CT Healthcare Center, Department of Psychiatry, West Haven, CT, USA; 16) National Cancer Institute, National Institutes of Health, United States Department of Health and Human Services, Bethesda, MD, USA; 17) Vanderbilt University, Nashville, TN, USA.

Cigarette smoking is a leading cause of preventable death worldwide. Nicotine dependence (ND) is a highly heritable trait that reduces the likelihood of quitting smoking. Several ND susceptibility genes are known. However, gene regulation across adult human brain regions, as related to ND biology, is poorly understood, with no large-scale genome-wide expression studies. Fortunately, it is now possible to impute genetically driven gene expression using recently developed methods and GTEx RNA-seq as reference data. To identify novel genes and expression patterns associated with ND, we performed a genome-wide gene-based association study using GTEx RNA-Seq data across 10 brain regions, the MetaXcan method, and results from our previous ND GWAS meta-analysis (N=28,677 Europeans/European Americans [EA]). For replication, we applied MetaXcan to a GWAS of heavy vs never smokers from the UK Biobank (N=48,931 EA). We used FDR<0.10 for discovery and a Bonferroni-corrected threshold for replication, by tissue. We identified significant and replicable differential expression of genes on 15q25.1, including nicotinic receptor genes previously established for ND, *CHRNA5* (cingulate cortex, caudate basal ganglia, cortex; $P_{min}=2.6\times10^{-8}$, β = -0.05 [↑expression ↓ND]) and *CHRNA3* (cerebellum, caudate and putamen basal ganglia; $P_{min}=9.3\times10^{-10}$, β= -0.17), as well as two genes not previously associated with ND: *ADAMTS7* (cerebellum, cingulate cortex; $P_{min}=3.6\times10^{-6}$, β= 0.05) and *TMED3* (nucleus accumbens; $P= 4.2\times10^{-7}$, β= -0.08). Based on linkage disequilibrium and tests of overlap of gene models predicting expression, the gene expression signal for *ADAMTS7* appears distinct from the primary ND GWAS signal on 15q25.1 around *CHRNA5/A3*. *ADAMTS7* has been implicated in cardiovascular disease, for which smoking is an established risk factor, as well as in a gene-smoking interaction study for coronary heart disease. We also identified a novel ND-associated gene, *SPDYE3* on 7q22.1 (putamen basal ganglia; $P= 1.5\times10^{-5}$, β= 0.04), although little is known about its function. These results represent the first replicable, genetically driven differential gene expression in human brain associated with a smoking phenotype, indicate the potential importance of brain regions typically overlooked in addiction studies, highlight additional genes of interest in the established 15q25.1 region, and identify a novel ND-associated gene.

239

Using large scale brain eQTL meta-analysis from multiple RNA-sequencing cohorts to identify neurodegenerative and neuropsychiatric risk candidates. *S.K. Sieberts[1], T. Perumal[1], M. Carrasquillo[2], M. Allen[2], J. Reddy[2], K. Dang[1], J. Calley[3], P.J. Ebert[3], A. Dobbyn[4], E. Stahl[4], N. Taner[2], L.M. Mangravite[1], AMP-AD Consortium eQTL Working Group and the CommonMind Consortium (CMC).* 1) Sage Bionetworks, Seattle, WA; 2) Department of Neuroscience, Mayo Clinic Florida, Jacksonville, FL; 3) Lilly Research Labs, Eli Lilly and Company, Indianapolis, IN; 4) Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY.

To date, hundreds of risk loci have been associated with neurodevelopmental, neuropsychiatric and neurodegenerative diseases. When risks are mediated through gene expression, eQTL can implicate the specific genes involved through colocalizing algorithms or global tests of case-control differences in the predicted gene expression. In each case, eQTL from the relevant tissue are necessary to understand the specific expression patterns contributing to disease, which may not be captured in readily available tissue such as blood. Recently, several large initiatives, including the CommonMind Consortium (CMC), the Accelerating Medicines Partnership for Alzheimer's Disease (AMP-AD) and GTEx, have made high-quality data available from the RNA-sequencing and genotyping of post-mortem brain collections representing neurodegenerative and neuropsychiatric cases, and undiseased individuals. Here we generate the best-powered brain eQTL resource to date which spans multiple brain regions and diseases. In total, we have collected 6 cohorts, comprised of tissue from approximately 2400 individuals and representing 4 brain regions, the largest collection of which is from dorsolateral prefrontal cortex comprised of 1800 samples. A common analysis pipeline was applied to each cohort. Samples were imputed to the Haplotype Reference Consortium panel. For each of the AMP-AD cohorts, RNA-seq data was realigned and quantitated using parameters appropriate to sequencing protocol. Following quality control, quantified expression from each of the 6 cohorts was then normalized via *voom*, adjusting for available known clinical and technical covariates, and hidden confounders, prior to applying a linear model to detect eQTL, adjusting for inferred genetic structure and diagnosis. Meta-analysis across cohorts was then performed. Preliminary meta-analysis of the Caucasian samples from the CMC MSSM-Penn-Pitt cohort (n=467) and ROSMAP cohort (n=537) identified almost 3.5 million proximal (distance ≤1 Mb) eQTL at FDR ≤ 5%, 678,144 of which were not identified in either cohort alone. Additionally, we identify eQTL for more than 19,000 genes/lncRNAs which have no significant eQTL in GTEx brain regions. We then use these eQTL to identify candidate genes underlying the GWAS signals in Alzheimer's Disease in the IGAP cohort. Preliminary analysis using *COLOC2* identified 2 GWAS peaks in which a single gene is consistent with colocalization. We apply similar methods to the analysis of schizophrenia.

240

Transcriptome analysis implicates joint dysfunction of GABAergic interneurons and metabolism in schizophrenia. *A. Norris[1,2], M.A. Kondo[3,4], A.E. Jaffe[5,6], X. Chen[7], A. Sawa[4], J. Pevsner[1,2,3].* 1) Neurology, Kennedy Krieger Institute, Baltimore, MD; 2) Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD; 3) Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD; 4) School of Psychiatry, Faculty of Medicine, University of New South Wales, Sydney NSW, Australia; 5) Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD; 6) Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; 7) Nevada Institute of Personalized Medicine and Department of Psychology, University of Nevada, Las Vegas, NV.

Schizophrenia (SZ) is a severe mental illness thought to arise from complex genetic-environmental interactions. RNA-seq of postmortem human brain offers an unbiased approach to identify shared downstream molecular and cellular consequences of divergent underlying genetic and/or environmental liability. Recently several methodological innovations have improved the ability to interpret the biological context of transcriptional changes confounded by RNA degradation in postmortem tissue and celltype heterogeneity in bulk tissue. In this study, we first apply quality surrogate variable analysis (qSVA) to directly estimate postmortem degradation in RNA-seq fastq data from the anterior cingulate cortex (ACC) of patients diagnosed with SZ and matched controls. The vast majority of differentially expressed genes were no longer statistically significant when qSVA estimates were used instead of the standard covariates used to adjust for degradation (postmortem interval and brain pH). To deconvolute the celltype context of the disease-associated transcriptional changes, we subjected the SZ geneset to expression weighted celltype enrichment analysis (EWCE). EWCE analysis revealed significant neuronal enrichment for genes down-regulated in SZ (FDR < 0.05). Sub-celltype analysis, using reported human neuronal subpopulations that were identified by single cell transcriptome data, revealed that these dysregulated genes marked parvalbumin (PV+) GABAergic interneurons. We then characterized the dysregulation of GABAergic interneurons, using pathway analysis and posthoc antipsychotics differential expression tests. Pathway analysis implicated mitochondrial dysfunction (FDR < 0.001), which was linked to PV+ interneurons by a shared upstream transcriptional regulator, peroxisome proliferator-activated receptor gamma coactivator 1alpha (PPARGCIA). Through the application of novel, rigorous methodology that aims to discern transcriptional changes driven by the underlying disease from postmortem artifact, we report the dysfunction of two distinct populations of GABAergic interneurons in SZ. These results extend our understanding of the role of PV+ interneurons in SZ, and reveal possible origins for their dysfunction in mitochondrial deficits.

241

Multi-ancestry genome-wide association study incorporating gene-alcohol intake interactions identifies 18 new lipid loci. *P.S. de Vries[1], M.R. Brown[1], A.R. Bentley[2], T.W. Winkler[3], A.T. Kraja[4], A.C. Morrison[1], CHARGE Gene-Lifestyle Interactions Working Group.* 1) Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA; 2) Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA; 3) Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany; 4) Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA.

Alcohol intake may interact with genetic variants to influence lipid levels. To identify new genetic loci that influence levels of high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, and triglycerides, we conducted a multi-ancestry genome-wide association study incorporating gene-alcohol intake interactions. We included 45 studies in Stage 1, followed by an additional 66 cohorts in Stage 2, altogether encompassing 394,914 individuals of European, African, Asian, Hispanic, or Brazilian ancestry. Dosages of genetic variants were imputed using the 1000 Genomes reference panel. A two degree of freedom test was used to jointly assess genetic main and interaction effects on lipid levels. Inverse variance fixed effect meta-analysis was performed within each ancestry group, and across the combined ancestry groups. Variants with suggestive evidence (P<1E-6) in Stage 1 were tested in Stage 2. The combined analysis of Stage 1 and Stage 2 was then evaluated using the genome-wide significance threshold of P<5E-8. 25,115 genetic variants were suggestively associated with lipid levels in Stage 1 and taken forward to Stage 2. In the combined analysis of Stage 1 and Stage 2, 147 independent loci were significantly associated with lipid levels, of which 18 were completely novel. Eight of the new loci were associated with LDL cholesterol, eight loci were associated with HDL cholesterol, and seven loci were associated with triglycerides. Novel loci included *PCSK5, VEGFB*, *A1CF*, and *DGKQ*, for which independent evidence exists from cell line studies and rodent models to support a role in the regulation of lipid levels. All significant associations were identified using joint tests of main and interaction effects. We successfully identified 18 novel lipid loci in this large multi-ancestry association study using gene-alcohol intake interactions to drive discovery through a joint model of main and interaction effects. Several of the new loci harbor promising candidate genes.

242

Gene x environment interactions in the UK Biobank study: Evidence that both physical inactivity and sleep inefficiency accentuate the genetic risk of obesity. *A.R. Wood[1], S.E. Jones[1], Z. Kutalik[2], H. Yaghootkar[1], R. Beaumont[1], M.A. Tuke[1], K.S. Ruth[1], R.M. Freathy[1], A. Murray[1], M.N. Weedon[1], J. Tyrrell[1], T.M. Frayling[1].* 1) Genetics of Complex Traits, University of Exeter Medical School, Exeter, Devon, United Kingdom; 2) Institute of Social and Preventive Medicine, University Hospital of Lausanne, Switzerland.

Introduction In contrast to main gene-phenotype effects, gene-environment interactions in humans have been difficult to identify and replicate. Meta analyses have identified putative interactions between physical activity and genetic variants associated with obesity, but these findings need replicating and validating with objective measures. We tested the hypothesis that lower levels of physical activity and abnormal sleep patterns accentuate genetic susceptibility to obesity using the largest single resource for testing gene-environment interactions - the UK Biobank - with objective measures of activity derived from accelerometers worn by 103,000 participants. *Methods* We used 120,000 individuals from the first genetic data release of the UK Biobank study, 19,229 of whom had accelerometer data. From accelerometer data, we derived a variety of measures for physical activity and sleep, including total physical activity, bouted activity, sleep duration, and sleep efficiency. We used BMI as the outcome and tested associations with genetics and accelerometer derived measures of activity and sleep, as well as self-reported measures of activity and sleep (N=109,142). We analysed individual BMI variants and a genetic risk score (GRS) for obesity (76 variants). We also performed several negative control experiments mimicking environmental factors with similar properties as activity. *Results* We found evidence of gene-activity interactions in the UK Biobank. For example, the effect of the BMI GRS on BMI was larger in the 50% of people reporting less physical activity, with 10 additional BMI-raising alleles associated with a 3.6kg increase in weight for someone 1.73m tall in the least active 50% of individuals versus 2.8kg in the most active 50% ($P_{interaction}=5x10^{-6}$). This observation was consistent within individuals with objective measures of activity. However, we also observed (weaker) evidence of interaction in the negative control experiments, suggesting residual confounding was present. We also identified a nominal interaction effect between the GRS and sleep efficiency (P=0.023), with stronger BMI genetic effects in individuals sleeping least efficiently. *Conclusions* Our results are consistent with previous studies suggesting that low levels of physical activity and sleep accentuate the genetic risk of obesity, but our results emphasize the importance of using objective measures and negative control phenotypes to test the specificity of gene-activity interactions.

243

Multi-ancestry genome-wide association study of gene × smoking interactions identifies novel lipid loci. *A.R. Bentley[1], Y.J. Sung[2], M.R. Brown[3], C.N. Rotimi[1], L.A. Cupples[4], CHARGE Gene-Lifestyle Interactions Working Group.* 1) Center for Research in Genomics and Global Health, Natl Human Genome Research Institute, Bethesda, MD; 2) Division of Biostatistics, Washington University School of Medicine, St. Louis, MO; 3) 3.Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX; 4) 4.Department of Biostatistics, Boston University, Boston, MA and Framingham Heart Study, Framingham, MA.

The concentrations of high- and low-density lipoprotein cholesterol and triglycerides are associated with smoking, but it is unknown to what extent smoking modifies genetic associations with these traits. We investigated whether accounting for the interaction of smoking status with genetic variants helps identify lipid loci in 133,802 individuals of European (n=90,266), African (n=23,745), Asian (n=13,171), and Hispanic ancestry (n=6,620). We used both a 1 degree of freedom (df) test of interaction as well as a joint (2df) test of main and interaction effects. Meta-analyses were conducted within ancestries and these results were meta-analyzed for trans-ancestry analysis. 17,921 variants from 519 loci (+/- 1 MB) were suggestively associated ($p\leq10^{-6}$), and were evaluated in a stage 2 analysis of 253,436 individuals. 82% of variants replicated in stage 2 ($p<0.05/17,921$). Here we present combined stage 1 and 2 meta-analysis, with $p<5\times10^{-8}$ declared statistically significant. Stage 1 analyses stratified by smoking status were used to better characterize findings. We identified 30 loci that were independent of known lipid loci (novelty defined by physical distance > 1MB). Several notable patterns emerged. There were loci for which the observed association was much stronger among smokers (*BMP6, EXOC6B, MARCH1, G3BP1*, *TP53I11, REEP1*). In contrast, we found loci for which an association among the unexposed was not seen in smokers (*PTPRZ1*, *MAGI2,* intergenic variant rs143396479). For some of our 2df results, associations were similar across strata with no evidence of interaction from the 1df test, suggesting a novel main effect (*EYA3, ETV5*, *TMEM175, B3GNT4, CREB3L2*). Many novel loci were statistically significant only in African ancestry individuals. Allele frequency differences likely explain much of this observation; however, at some of these loci no association was observed in other ancestries, despite similar or higher minor allele frequencies (*MAGI2*, intergenic variant rs12740061). In conclusion, this large, multi-ancestry genome-wide study of gene-smoking interactions on serum lipids identified 30 novel loci. We find evidence for loci that could have been detected only in a stratified or interaction model. Additionally, we demonstrate the importance of including diverse populations, particularly in studies of interaction with lifestyle factors, where genomic and lifestyle differences by ancestry may contribute to novel findings.

244

Machine learning approaches to identify genetic and context-based differences in heart disease prediction between males and females. *S. Raj[1], A.G. Clark[1], M. Sabuncu[2].* 1) Dept. Molecular Biology and Genetics, Cornell University, Ithaca, NY; 2) Dept. Electrical and Computer Engineering, Cornell University, Ithaca, NY.

Males and females show strong sex-based differences in incidence, prevalence, morbidity, mortality and clinical manifestation of cardiovascular disease. Despite these differences, genome-wide association studies rarely incorporate them into studies of cardiovascular disease risk. Meanwhile, machine learning approaches have become powerful tools to predict disease state from genomic data and other risk factors. Here we implement both quantitative genetic and machine learning methods to establish sex-specific predictions of cardiovascular disease state, and quantitatively contrast the prediction accuracy in males vs. females. Toward this goal, we used genetic data from > 10,000 adult individuals of European ancestry from multiple cardiovascular disease cohorts available through dbGAP. We separated these individuals into four categories: male and female cases and controls. First, we used the program GCTA to build a polygenic risk model for the genetic basis of heart disease risk separately in males and females. Second, we calculated a polygenic risk score for each individual, as implemented in PLINK, and in combination with non-genetic risk factor information, used a random forest to predict disease status, again, doing this separately in males and females. Finally, we examined combinations of risk variants, haplotypes, and non-genetic risk factors to understand the joint impact of genes and environment on cardiovascular disease risk separately in males and females. After constructing models to predict disease states in males and females, we tested the accuracy of these models on the opposite sex by exchanging the models and predicting disease state in the opposite sex. Each of the models showed contrasting levels of prediction accuracy for males and females. The most significant difference was obtained through the second model incorporating polygenic risk scores and contextual information. This model showed significant differences in the ability to predict cardiovascular disease between males and females. Models based on genetic information alone, however, showed no strong differences between males and females in risk prediction. Therefore evaluating context-based associations in complex disease may uncover new genetic pathways that contribute to disease pathogenesis depending on other traits, and that differ between the sexes. This type of approach has powerful implications for using clinical and genetic data to predict one's risk to disease.

245

An expanded view of complex traits: From polygenic to omnigenic. *Y. Li[1], E. Boyle[1], J. Pritchard[1,2,3].* 1) Department of Genetics, Stanford University, Stanford, CA; 2) Department of Biology, Stanford University, Stanford, CA; 3) Howard Hughes Medical Institute, Stanford University, Stanford, CA.

A central goal of genetics is to understand the links between genetic variation and disease. Intuitively, one might expect disease-causing variants to cluster in or near core genes and pathways that drive disease etiology. But for complex traits, association signals tend to be spread across most of the genome–including near many genes without an obvious connection to disease. Using ashR to analyze the distribution of regression coefficients from the GWAS of a stereotypical complex trait, we estimated that 62% of all SNPs are associated with nonzero effects on height, and that around 3% of SNPs have causal effects on height. We also analyzed three disease GWASs including rheumatoid arthritis, Crohn's disease, and schizophrenia. Associated variants are enriched in regions that are transcriptionally active in relevant cell types, and absent from regions that are inactive in those cell types; but there is generally little functional enrichment beyond this. These observations imply a need for rethinking conceptual models of complex traits. We propose that gene regulatory networks are sufficiently interconnected that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most heritability can be explained by effects on genes outside core pathways. We refer to this hypothesis as an "omnigenic" model.

246

Joint effects of regulatory and coding variants shape human genetic variation and disease risk. *S.E. Castel[1,2], A. Cervera[1,3], F. Reverter[4], R. Guigo[4], I. Iossifov[1,5], A. Vasileva[1,2], T. Lappalainen[1,2].* 1) New York Genome Center, New York, NY, USA; 2) Department of Systems Biology, Columbia University, NY, USA; 3) Systems Biology Laboratory, Institute of Biomedicine and Genome-Scale Biology Research Program, University of Helsinki, Finland; 4) Centre for Genomic Regulation, Barcelona, Spain; 5) Cold Spring Harbor Laboratory, NY, USA.

In this study, we characterized how common regulatory variants affect the penetrance of rare damaging coding variants via haplotype epistasis, where the functional output of a gene depends on the haplotype combinations of its coding and regulatory variants. We first analyzed this in the general population, using GTEx v6 with exome data from 455 individuals, regulatory variants (eQTLs and sQTLs) from 44 tissues, and improved haplotype phasing. Analysis of eQTLs and allele-specific expression showed that putatively deleterious coding variants (CADD>15, DAF<1%) are depleted from the higher expressed eQTL haplotypes, compared to a matched null (p=2.3e-7). Furthermore, damaging variants were enriched in haplotypes where the corresponding exon is spliced out (p=2.8e-7). This suggests that purifying selection is eliminating haplotypes where regulatory variants increase the penetrance of deleterious variants. Next, we analyzed haplotype epistasis in disease cohorts. Using the Simons Simplex Collection, we showed that in autism-associated genes, individuals with autism had haplotypes with a rare disrupting coding variant paired with a haplotype with lower expressed eQTL allele more often than expected by chance (p=3.2e-4), suggesting joint coding and regulatory effects increasing disease risk via lowered total gene dosage. A similar effect was seen in The Cancer Genome Atlas data of cancer patients, with an enrichment of predicted high penetrance combinations of regulatory and coding variants in tumor suppressor genes (p=1.6e-3). These results indicate that regulatory haplotype configuration of disease-causing rare coding variants affects disease risk. Finally, we used CRISPR to create HEK293 cell lines with different haplotype combinations of a heterozygous eQTL and coding SNP rs199643834 in *FLCN* that causes Birt–Hogg–Dube syndrome. Using transcriptome-wide gene expression as a cellular phenotype readout, we first showed that genes affected by rs199643834 were consistent with the syndrome. In cell lines heterozygous for rs199643834, lower expression of the wild type allele appeared to lead to a stronger transcriptome effect (p=1.1e-23), indicating that the eQTL haplotype of the coding variant affected its penetrance. Altogether, our results demonstrate that joint haplotypic effects of regulatory and coding variants are an important part of the genetic architecture of human traits, and contribute to modified penetrance of disease-causing variants.

247

Evidence of compensatory variations on the remaining allele of rare deletions. *K. Popadin[1], E. Porcu[1], M. Lepamets[2], K. Mannik[1], M. Garieri[3], R. Magi[2], Z. Kutalik[4], A. Reymond[1].* 1) Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; 2) Estonian Genome Center, University of Tartu; 3) Department of Medical Genetics and Development, University of Geneva; 4) Institute of Social and Preventive Medicine, Lausanne.

Every human genome harbors dozens of rare copy number variants (CNVs). To uncover the impact of these variants on human health we need to estimate their individual deleteriousness. While several properties of CNVs e.g. length, ploidy, population frequency, number and nature of disrupted genes could be informative to predict the effect on fitness of a specific CNV, we assessed selection signature as a new and more direct approach to estimate the effect of a CNV. Assuming that dosage alteration of a given gene(s) is the most direct consequence of a deletion, we consider encompassed regulatory SNVs alleles, as potential modifiers of the perturbed expression levels. If the remaining haplotype carries multiple Gain of Expression (GOE) rather than Loss of Expression (LOE) variants we can hypothesize that the expression of the hemizygous gene(s) is partially compensated. We challenged our hypothesis using rare deletions called in 120,000 individuals from the UK Biobank. We determined a test set of 200,000 GTEx cis-eQTLs that are concordant across 44 tissues and called in at least three tissues (blood, brain and any other) that represent the main coexpression clusters in GTEx data. We observed that genome-wide the probability of being a deletion carrier and a hemizygote for GOE allele is shaped by the frequency of the GOE allele in control population (i.e. Null hypothesis). However, genes with strong selection against heterozygous loss of function (Shet>0.04) had a significant increase in frequency of GOE alleles in deletions versus euploid controls (15% of such deletions are compensated by excess of GOE, p< 0.0001 permutation test). This finding is consistent with the expectation that damaging deletions segregating in the human population are partially compensated by regulatory variants. Genes with high Shet (Shet > 0.04) were associated with (i) increased phenotypic severity; (ii) autosomal dominant disorders and (iii) higher load of de novo variants, suggesting that the majority of compensations we observe in our study are of very recent origin. Thus deletions with excess compensatory GOE are expected to be eliminated in non-compensated state, allowing us to estimate individual severity of each deletion and improve our understanding of the integral human genetic load.

248

Pleiotropic noncoding regulatory elements are under purifying natural selection. *D. Radke[1,2,3,4], D.J. Balick[2,3,4], J. Sul[5], S. Akle[2,3,6], M. Maurano[7], R. Green[3,4], J. Stamatoyannopoulos[8], S. Sunyaev[2,3,4].* 1) Program in Genetics and Genomics, Harvard Medical School, Boston, MA; 2) Department of Biomedical Informatics, Harvard Medical School, Boston, MA; 3) Division of Genetics, Brigham and Women's Hospital, Boston, MA; 4) Broad Institute of Harvard and MIT, Cambridge, MA; 5) Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, CA; 6) Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA; 7) Institute for Systems Genetics, New York University, New York City, NY; 8) Department of Genome Sciences, University of Washington, Seattle, WA.

Assessing the role of natural selection on genetic variation in noncoding regions has been previously difficult, because most work has only utilized SNP variation, of which there are no clearly defined loss-of-function (LoF) noncoding variants. Genomic deletions, however, provide a powerful LoF noncoding model by removing the nucleotides altogether. Using three regulatory annotations (DNase1 hypersensitivity, enhancer (H3K4me1) and polycomb-repression (H3K27me3) histone modifications) from primary tissues and cell-types characterized as part of the Roadmap Epigenomics Project, along with deletions from large-scale population projects including 1000GP and our own deletion callset on participants in the Alzheimer's Disease Neuroimaging Initiative (ADNI), we examine differences in deletion allele frequency as it relates to their regulatory element overlap. Regulatory annotations across tissues and cell-types that share the same genomic position can serve as a proxy for the cellular pleiotropy of that position. We hypothesize that regulatory loci exhibiting highly-pleiotropic effects (i.e., loci with regulation across many diverse tissues) should be under stronger purifying selection than cell-type-specific or nonfunctional loci. We develop a statistical method to account for covariance between the tissues/cell-types, enabling a per-base-pair normalized count of co-localized regulatory activity. Analyzing the locus pleiotropy (calculated by our statistic) overlapped by each of the noncoding deletions, we find a statistically significant shift in the allele frequency spectrum (AFS) towards rare alleles not only for deletions overlapping regulatory loci versus nonfunctional loci, but additionally for deletions overlapping highly-pleiotropic loci versus cell-type-specific loci, confirming our hypothesis. We interpret these results as evidence of the widespread action of purifying selection on noncoding regulatory elements, the strength of which is determined by the corresponding amount of pleiotropy. These findings allow for more rigorous noncoding functional interpretation for use in medical or experimental studies, including application of the pleiotropy statistic for use in case/control CNV burden analyses on either segregating or de-novo deletion events.

249

Parent-of-origin effects on gene expression in type 2 diabetes offspring trios. *R.B. Prasad[1], G. Kovacs[2], A. Lindqvist[3], G. Hatem[1], A. Lessmark[1], P. Almgren[1], M. Vitai[2], N. Oskolkov[1], T. Singh[4], P. Vikman[1], M. Åkerlund[1], N. Wierup[3], I. Artner[4], L. Koranyi[2], L. Groop[1,5].* 1) Diabetes and Endocrinology, Lund University Diabetes Centre, Malmö, Sweden; 2) Heart Center Foundation, DRC, H, Balatonfured, Hungary; 3) Neuroendocrine Cell biology, Lund University Diabetes Centre, Malmö, Sweden; 4) Endocrine Cell Differentiation and Function, BMC, Lund University, Lund, Sweden; 5) Finnish Institute of Molecular Medicine (FIMM), Helsinki University, Helsinki, Finland.

Background and aims Type 2 diabetes (T2D) is seen more often in offspring of T2D mothers rather than of T2D fathers. Further, parental sex – specific effects on insulin concentrations have been reported with lowest values seen in sons of diabetic mothers. Parent-of-origin effects (POE), wherein the phenotypic effect of an allele depends on whether it is inherited from the mother or the father, could explain these observations. **Materials and methods** We performed RNAsequencing of 80 trios enriched for type 2 diabetic offspring (77 offspring with T2D) and assessed for differences in correlation of expression between each parent-offspring pair on a background of parental expression correlations. Expression patterns of genes showing parent-of-origin effects in trios were studied in adult and fetal pancreas. Knockdown studies are ongoing to assess their influence on beta cell mass (proliferation) and function (insulin secretion). **Results** 4662 autosomal genes expressed in blood were assessed for parental biases in gene expression. 24 protein-coding genes showed parental biases in gene expression of which 6 genes showed the lowest paternal-maternal correlations including *BMP8A*, *CAMK2G*, *CLCF1*, *PA2G4*, *PSD4*, and *RPS23*. The Bone Morphogenic Protein 8A coding *BMP8A* showed significantly higher gene expression correlation between father-offspring (ρ_{BMP8A}=0.62, P_{BMP8A}=1.81x10^{-09}) than that of mother-offspring (ρ_{BMP8A}=0.11, P_{BMP8A} =3.42x10^{-01}, $Pdiff_{BMP8A}$=1.97x10^{-05}). These data were consistent even when assessed separately for sons and daughters. Differential expression analysis between genders was not significant either in blood or in pancreas (p>0.05), showing clearly that these differences in correlations were not driven by gender. *BMP8A* showed significantly high expression in the fetal pancreas whereas almost no expression was observed in the adult pancreatic islets. The Beta-Actin encoding *ACTB* showed significantly higher father-offspring (ρ_{ACTB}=0.86, P_{ACTB}=3.42x10^{-23}) compared to mother-offspring (ρ_{ACTB}=0.60, P_{ACTB}=6.88x10^{-09}, $Pdiff_{ACTB}$=7.66x10^{-07}). The only gene showing higher mother-offspring correlations was the cardiotrophin-like cytokine factor 1 coding CLCF1 ((ρ_{CLCF1}=0.23, P_{CLCF1}=4.24x10^{-02}, $Pdiff_{CLCF1}$ = 2.06x10^{-04}). **Conclusion** This study demonstrates that parental biases in gene expression exist beyond imprinted genes and can have strong spatial and temporal effects independent of gender. Some of these genes could have significant roles in fetal development.

250

Identification of expression regulation mediating association loci for T2D through characterization of human pancreatic islets and β-cells.

A. Viñuela[1], M. van de Bunt[2], A. Varshney[3], N. Oskolkov[4], P.E. MacDonald[5], L. Scott[6], M.L. Stitzel[7], C.J. Parker[3] for the InsPIRE consortium. 1) Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Geneva, Switzerland; 2) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, Oxford; 3) Department of Computational Medicine & Bioinformatics, University of Michigan, USA; 4) Lund University Diabetes Centre, Department of Clinical Sciences, Skåne University Hospital Malmö, Lund University, Malmö, Sweden; 5) Alberta Diabetes Institute, University of Alberta, Edmonton, Canada; 6) Department of Biostatistics, University of Michigan, USA; 7) The Jackson Laboratory for Genomic Medicine, Farmington, CT.

Understanding the molecular consequences of GWAS loci associated with common disorders requires the study of relevant cell types and tissues to identify the genes through which these variants act. To identify effector genes for T2D loci we performed RNA-sequencing and genotyping in human islets from 420 cadaveric donors to uncover genetic regulatory variants (eQTL) acting in the tissue. Through eQTL mapping we identified variants significantly affecting expression levels in cis for 6,039 genes (FDR<1%). Conditional analysis identified an additional 1,705 significant secondary eQTL signals at 1,288 genes. Integration with high-depth islet ATAC-seq and transcription factor (TF) motif data revealed differential enrichment of high- and low-effect size eQTLs across footprint motifs and allowed us to predict TF motif regulatory directionality (activator or repressor) at the footprint locus. To identify cell-specific regulatory effects that may be mediating the action of GWAS loci, we investigated regulatory differences between islet samples and RNA-seq from bulk FACS sorted β-cells from 26 individuals. By examining strength of association of significant islet eQTLs in beta-cells, we identified 227 islet eQTLs likely to also be β-cell eQTLs and estimated that 46% of islet eQTLs act in β-cells. Using estimates of β-cell proportion in the islet samples, we identified 220 of the 227 eQTLs to be significant GxCell-composition interaction eQTLs when considered alone, supporting the notion that they are β-cell specific regulatory signals. Integrating islet eQTL data with genetic information on 82 known T2D-associated loci identified 11 genes for T2D with at least two lines of evidence. This included *TCF7L2* and two independent signals at *DGKB*. A comparison with the GTEx v6p data, using all tissues, identified 9 genes linked to T2D associated loci, meaning islet samples have suggested significantly more genes to explain T2D association signals. We have shown how islet regulatory state information in combination with expression markedly increases the resolution for the molecular genetic signatures underlying islet gene regulation. In addition, our results demonstrate the power of transcriptomic analysis in a disease-appropriate tissue such pancreatic islets to deliver molecular insights in T2D pathophysiology, and to identify cell specific regulatory signals.

251

Transcriptome sequence analysis at single-cell resolution reveals depot-specific signatures in adipose tissue-derived stromal vascular fraction cells linked to metabolic diseases.

J. Vijay[1], X. Shao[1], M-M. Simon[1], M-C. Vohl[2], A. Tchernof[3], E. Grundberg[1]. 1) Human Genetics, McGill University, Montreal, Quebec, Canada; 2) Institute of Nutrition and Functional Food, Université Laval, Québec, QC, Canada; 3) Québec Heart and Lung Institute, Université Laval, Quebec, Canada.

Adipose tissue and isolated adipocytes are studied extensively to understand underlying mechanisms of obesity and associated metabolic disorders. Likewise, the stromal vascular fraction (SVF) of adipose tissue is studied at length for its regeneration potential. However, involvement of SVF's heterogeneous cellular components in metabolic disorders is not fully elucidated. To this end, we aimed to characterize global gene expression patterns in SVF of subcutaneous (SC) and omental (OM) adipose tissue obtained from obese individuals undergoing bariatric surgery. First, we performed transcriptome sequencing (RNA-Seq) of SVF derived from up to 20 individuals as well as on matching isolated adipocytes. We also utilized RNA-Seq data from the publicly available GTEx resource. We performed differential gene expression (DGE) analysis of SVF and adipocytes derived from OM and SC depots of obese individuals, respectively. In total, we detected 6,664 differentially expressed genes (>2 fold) between SVF and adipocytes independent of depot (i.e. observed in both SC and OM analysis). Ingenuity Pathway Analysis on these genes shows Th1 and Th2 activation (p=2.73E-22) as the most significant pathway with IFNγ (p=6.34E-53) and TNFα (p=2.07E-52) as key regulators. In addition, we found 969 genes enriched for glutamate receptor signaling (p=3.08E-06) to be specific to OM SVF, when comparing expression pattern in adipocytes from the matched depot. DGE analysis between individuals with low (<=24 kg/m^2) and high BMI (>=30 kg/m^2) from GTEx identified 59 obese-associated genes (p-value <=0.05) from either OM or SC. We show that of these genes, SVF has increased expression pattern for 27 genes in OM and 21 in SC. We found that *IGFBP1* – a predictor of diabetes development – exhibits SVF OM-specific expression. Next, we used single cell 3'mRNA-seq to characterize cellular subpopulations of SVF from the same study population of obese individuals. We sequenced ~6300 SVF cells obtaining on average ~67000 reads corresponding to ~1000 genes per cell. In a preliminary analysis, we were able to differentiate multiple cell clusters showing enrichment for stromal stem cells, fibroblasts and various blood cells including macrophages, T cells, B cells, monocytes and dendritic cells. We are currently expanding our single-cell RNA sequencing of SVF from obese individuals in an attempt to disentangle the cellular basis of obesity-related metabolic complications including type 2 diabetes. .

252

Transcriptomic profiling of the developing human islet and mechanisms of type 2 diabetes predisposition. *M. Perez-Alcantara[1], M. van de Bunt[1,2], N. Beer[2], C. Honoré[3], M. Hansson[4], A. Gloyn[2], M. McCarthy[1,2,5].* 1) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; 2) Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Oxford, United Kingdom; 3) Department of Islet & Stem Cell Biology, Novo Nordisk A/S, 2760 Maaloev, Denmark; 4) Global Research External Affairs, Novo Nordisk A/S, Maaloev, Denmark; 5) Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom.

Most variants associated with type 2 diabetes (T2D) predisposition in genome-wide association studies (GWAS) act through defects in insulin secretion. In principle, these could result from deficiencies in islet development and/or mature islet function: most functional studies have focused on the latter. To explore the contribution of disturbed islet development to T2D pathogenesis, we characterised the transcriptome of three fibroblast-derived human iPSC lines (from three different donors) differentiated along the pancreatic endocrine lineage (using a protocol modified from that of Rezania et al 2014). We obtained RNA-Seq profiles of 15,220 protein-coding genes and lincRNAs at seven stages (from definitive endoderm (DE) to beta-like cells (BLC)). Differentially-expressed (ΔExp) genes (q<0.01) with absolute $\log_2$ fold change ($\log_2$FC)>1 were assigned to the stage in which they were most upregulated (vs. the baseline iPSC profile). We identified 9,408 ΔExp genes: these included known markers of islet development (e.g. *PDX1*, *MAFA*, *INS*) which displayed ΔExp at expected stages, confirming the biological relevance of this cellular model. To assess the role of T2D-associated loci in islet development, we considered transcripts mapping near credible set intervals derived from T2D GWAS data (DIAGRAM: ~150,000 European subjects imputed to 1000 Genomes) varying the size of the flanking "window" from 0-500kb to account for cis-regulatory effects. We detected significant gene set enrichment (GSEA q<0.01) for genes mapping close (0-200kb) to GWAS credible sets for ΔExp genes at the most-differentiated (BLC) stage, but no significant excess was detected for the sets of ΔExp genes defined at earlier points in differentiation. Notwithstanding, >70% of this set of GWAS genes displayed ΔExp at pre-BLC stages, indicating the potential relevance of several key T2D GWAS genes to islet developmental processes. For example, *TCF7L2* expression peaked during the foregut stage of pancreatic endocrine development ($\log_2$FC=1.2; q=8.5x10^{-10}). Thus, genomic data from this iPSC-derived differentiation model allows exploration of the contribution made by disturbed human islet development to T2D pathogenesis: these same models will also allow putative developmental regulators to be interrogated by modulating their expression using CRISPR-Cas9-based approaches.

253

The effect of genetic variation on promoter usage and enhancer activity. *M. Garieri[1,2,3], O. Delaneau[1,2,3], F. Santoni[1,4], D. Mull[1], P. Carninci[5], E.T. Dermitzakis[1,2,3], S.E. Antonarakis[1,2,4], A. Fort[1].* 1) Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland; 2) Institute of Genetics and Genomics in Geneva, iGe3, Geneva, Switzerland; 3) Swiss Institute of Bioinformatics, SIB, Lausanne, Switzerland; 4) University Hospitals of Geneva, Service of Genetic Medicine, Geneva, Switzerland; 5) Division of Genomics Technologies, RIKEN Center for Life Science Technologies, Yokohama, Japan.

The identification of genetic variants affecting gene expression, namely expression quantitative trait loci (eQTLs), has contributed to our understanding of mechanisms underlying human traits and diseases. The majority of these variants map in non-coding regulatory regions of the genome and understanding of their mechanism of action remains challenging. We hypothesized that a fraction of eQTLs act at the level of promoter and enhancer functions. We used natural genetic variations and CAGE transcriptomes from 154 unrelated European individuals to map regulatory variants associated with promoter usage (puQTLs) and enhancer activity (eaQTLs). Transcriptome profiles were produced for nucleus-enriched RNAs extracted from lymphoblastoid cell lines. CAGE-tags were mapped to promoter regions of the FANTOM atlas, yielding to the quantification of 38,759 promoters/transcripts. Testing associations between promoters and genetic variants that map in the same topologically associating domains, we identified 5,376 puQTLs in *cis* (FDR<0.05). We found 2,289 puQTLs associated genes display more than one CAGE-peak and thus likely several promoters, suggesting that puQTLs are also implicated in the regulation of differential promoter usage. We characterized five categories of puQTLs associated genes, distinguishing single from multi-promoter genes. Among the multi-promoter genes, we found puQTLs effects that are either specific to one promoter or to multiple alternative promoters with variable effect orientations. Regulatory variants associated with opposite effects on different mRNA isoforms suggest compensatory mechanisms occurring between alternative promoters that do not lead necessary to different transcriptional output. The integration of puQTLs and eQTLs provides insights into the mechanisms underlying eQTLs affecting promoter usage. In a second analysis, we used the quantification of enhancer-derived RNAs (eRNAs) as proxy for enhancer activity, and mapped 110 enhancer-activity QTLs (eaQTL, FDR<0.05). Integrating promoter usage, mRNA quantification and enhancer activity analyses, we illustrate the potential of using complementary molecular phenotypes to dissect the mechanisms underlying enhancer related eQTLs.

254

Massively parallel reporter assays at the population scale: Quantifying the regulatory effect of non-coding variation in a large human cohort. *S.J. Cunningham[1,2], G.D. Johnson[2,3], W.H. Majoros[4,2], C.M. Vockley[2,3], C. Guo[1,2], M.G. Hayes[5], W.L. Lowe Jr.[5], T.E. Reddy[2,3].* 1) University Program in Genetics and Genomics, Duke University, Durham, NC; 2) Center for Genomics and Computational Biology, Duke University, Durham, NC; 3) Biostatistics and Bioinformatics Department, Duke University, Durham, NC; 4) Program in Computational Biology and Bioinformatics, Duke University, Durham, NC; 5) Division of Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL.

The majority of lead genetic variants identified in genetic association studies are in non-coding regions of the human genome. Identifying the causal variants underlying these results remains a major bottleneck for understanding the genetic basis of complex diseases. One new hypothesis is that those genetic association signals result from combinations of non-coding variants with coordinated effects on regulatory element activity. Testing that hypothesis will require comprehensive quantifications of the effects of regulatory variants across entire genetic association loci. To do so, we captured 5 Mb of type 2 diabetes- and gestational hyperglycemia-associated regions from 835 women in the third trimester of pregnancy. The women represented four ancestry groups, and were at the extremes of the oral glucose tolerance test for their ancestry group. We then used a high-throughput reporter assay, STARR-seq, to comprehensively quantify allele-specific regulatory activity in the captured loci across metabolism-relevant cell models. With those results, we are able to evaluate the genetic contributions to regulatory activity in adult-onset diabetes and maternal hyperglycemia. Further, by integrating with QTL studies and genomic data from the ENCODE project, we are now able to evaluate how broadly the model of locally coordinated regulatory variation generalizes to different types of associations in diverse cell systems. We have also used the approach to identify novel regulatory elements in the associated loci that will form the basis for target gene identification. More generally, the population-STARR-seq (POP-STARR) strategy greatly increases the ability of the field to quantify the effects of regulatory variants across entire genetic association loci and across large cohorts of patients.

255

Principles of gene regulation and noncoding variant function from hundreds of enhancer perturbations. *J. Engreitz[1], C. Fulco[1,2], T. Jones[1], R. Anyoha[1], E. Perez[1], M. Kane[1], G. Munson[1], S. Grossman[1,3], E. Lander[1,2,3].* 1) Broad Institute of MIT and Harvard, Cambridge, MA; 2) Department of Systems Biology, Harvard Medical School, Boston, MA; 3) Department of Biology, MIT, Cambridge, MA.

Gene expression in mammals is controlled by millions of noncoding regulatory elements including enhancers, creating what is presumed to be vast network of enhancer-gene connections. Genome-wide association studies (GWAS) have identified thousands of loci associated with one or more common diseases, and in most loci the functional variants are thought to affect enhancer function. Yet, interpreting the functions of noncoding variants is challenging because we do not understand the principles that specify gene-enhancer connections. Here, we develop a high-throughput approach based on CRISPR interference (CRISPRi tiling) to systematically discover the regulatory elements that control any gene in a given cell type. Using this method, we tile >2 megabases of sequence with >100,000 guide RNAs to generate a collection of hundreds of experimentally tested gene-enhancer connections in multiple cell types. We find that a simple model can predict gene-enhancer connections in the genome: the effect of an enhancer correlates with the product of its Activity (as estimated by the quantitative measurements of chromatin accessibility and histone modifications) and its Contact frequency to a target gene (as measured by Hi-C). This Activity by Contact model can classify functional enhancers and predict their quantitative effects on gene expression, supporting a physical contact model for enhancer function. Emergent properties of this model can explain the complex connectivity of gene-enhancer networks, including the phenomenon of enhancers appearing to skip over neighboring genes to regulate more distal ones. We apply this Activity by Contact model to predict gene-enhancer connections across cell types in the immune system, and show that this facilitates interpreting the functions of noncoding variants associated with autoimmune diseases. Our results reveal insights into the mechanisms that specify gene-enhancer connectivity and provide a systematic strategy for linking genetic variants in noncoding regulatory elements with their target genes.

256

Identification of enhancer elements at multiple renal cancer susceptibility loci using a massively parallel reporter assay (MPRA). *L. Machado Colli, L. Jessop, M.J. Machiela, J. Choi, T. Myers, M. Purdue, K. Brown, S.J. Chanock.* Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD.

To date, renal cell carcinoma (RCC) genome-wide associations studies (GWAS) have identified 13 regions in the genome associated with RCC susceptibility. The characterization of genetic susceptibility alleles represents a critical opportunity to understand how genetic variation can influence risk for cancers, providing new insights into the biology. To our knowledge, only two RCC loci have had their functional basis explained (11q13 and 12p12). We have deployed a new high-throughput method to investigate multiple regions in parallel, with the aim to accelerate the identification of promising regulatory regions under the peaks of RCC cancer GWAS signals. We investigate enhancer activity in 774 SNPs that were selected from 20 RCC GWAS regions, which included the 13 achieving genome-wide significance ($P < 5\times10^{-8}$) and 7 promising loci ($P < 10^{-6}$) in RCC GWAS. SNPs were chosen based on correlation (LD with RCC GWAS markers-based on the following LD parameters: (1) R^2>0.4 or D'>0.5 and MAF<0.05 and (2) evidence for enhancer activity based on available ENCODE transcription factor or histone ChIP-seq, FAIRE, and DNase data. The MPRA library was composed of 47,461 oligonucleotides 201 bp in length, compelled of 145 bps contained the SNP in the forward or reverse orientation followed by 10 bp of barcode sequence to allow for identifıcation. Each of the 774 SNPs tested were tagged 10 separate times for both orientations. The MPRA library was cloned into a luciferase reporter vector and transfected into HEK293T and ACHN cell lines. NGS was performed on 5 replicates each, under normoxic and hypoxic conditions, using the Illumina HiSeq 2500. Multivariate analysis included effects of alleles, direction, cell line and condition (hypoxia or normoxia). After correcting for multiple comparisons, 50 SNPs from 16 regions showed signifıcant p-values. From these 50 SNPs, 29 showed enhancer activity, independent of the forward/reverse orientation and had the same effect in both cell lines. As a proof-of-principle, for the 12p12 region, the leading MPRA SNP was rs7132434, for which previously published EMSA and luciferase assay supporting the importance of this SNP in regulating *BHLHE41*. Validations using luciferase assay and EMSA for 14 selected SNPs from 5 different regions are ongoing. This MPRA analysis has identifıed possible functional variants at multiple RCC risk loci and opens opportunities to discover new molecular mechanisms of genetic susceptibility to sporadic RCC.

257

Discovery of unique disease- and gene-specific peripheral blood DNA methylation signatures allows molecular diagnosis and VUS classification in hereditary genetic syndromes. *B. Sadikovic[1], L. Schenkel[1], C. Schwartz[2], K. Boycott[3], P. Ainsworth[1], E. Aref-Eshghi[1].* 1) London Health Sciences Centre, Western University, London, ON, Canada; 2) Center for Molecular Studies, J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, SC USA; 3) Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, ON Canada.

Introduction: Hereditary genetic conditions resulting from mutations in genes involved in epigenetic machinery and chromatin remodeling present with systemic, complex and often overlapping clinical features. Genetic diagnosis is complicated by a large proportion of variants of unknown clinical significance (VUS) in these genes. Recently, we have demonstrated existence of unique, gene/disorder-specific DNA methylation “epi-signatures” associated with such conditions and have begun systematic epi-signature mapping in these disorders. **Methods**: Peripheral blood samples from disease-specific cohorts of patients were assessed for genome-wide methylation changes relative to controls using Illumina Infinium 450k and EPIC genome-wide DNA methylation arrays (>1,500 patents tested). We applied the identified epi-signatures to supervised and non-supervised machine learning techniques to develop unique predictive models for each disorder. **Results**: We identified highly sensitive and specific peripheral blood DNA methylation epi-signatures in: Floating-Harbor Syndrome (*SRCAP*); autosomal dominant cerebellar ataxia, deafness, and narcolepsy (*DNMT1*); alpha thalassemia/mental retardation X-linked syndrome (*ATRX*); Kabuki syndrome (*KMT2D*); Sotos syndrome (*NSD1*); CHARGE syndrome (*CHD7*); and syndromic X-linked intelectual disability, Claes-Jensen type (*KDM5C*). These unique epi-signatures, along with our large reference database, enable near 100% sensitive and specific diagnosis for these disorders. These episignatures enable accurately re-classifycation VUSs in these genes as either pathogenic or benign. We show evidence of partial overlaps between specific episignatures in some of the above conditions including Floating-Harbor, Kabuki, and CHARGE syndromes. **Conclusion**: We demonstrate presence of unique and highly-specific DNA methylation signatures in patients suffering from hereditary genetic conditions involving genes that regulate epigenetic machinery and chromatin remodeling. These epi-signatures can be used for molecular diagnostics in this patient population, and enable interpretation and clinical classification of VUSs in the associated genes. Furthermore, these genomic DNA methylation defects provide insights in the molecular etiologies of these disorders. Ongoing work in our clinical laboratory focuses on systematic mapping of epi-signatures across other epi/genetic syndromes, and transition of this technology into routine clinical use.

258

Variation in mitochondrial DNA copy number influences nuclear DNA methylation. *C.A. Castellani[1], R.J. Longchamps[1], J.A. Sumpter[1], A. Tin[2], J.A. Lane[3], M.L. Grove[4], J. Bressler[4], J. Coresh[2], J.S. Pankow[5], M. Fornage[4], N. Pankratz[3], E. Boerwinkle[4], D.E. Arking[1], CHARGE Aging and Longevity Working Group.* 1) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) Department of Epidemiology and the Welch Center for Prevention, Epidemiology and Clinical Research, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; 3) Department of Laboratory Medicine and Pathology, University of Minnesota School of Medicine, Minneapolis, MN; 4) Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX; 5) Division of Epidemiology & Community Health, School of Public Health, University of Minnesota, Minneapolis, MN.

Variation in mitochondrial DNA copy number (mtDNA-CN) has been proposed to have a role in regulating the nuclear epigenome as evidenced by altered DNA methylation in nuclear genes of mitochondria-depleted cell lines. Given that variation in mtDNA-CN has been associated with several diseases and overall mortality, increased understanding of the cross-talk between nuclear and mitochondrial DNA is critical to elucidating the impact of mtDNA-CN on human health. DNA was extracted from buffy coat for 1,567 African Americans from the Atherosclerosis Risk in Communities (ARIC) study and nuclear CpG methylation was assessed using the Illumina Infinium 450K microarray. The Genvisis software package was used to determine mtDNA-CN from Affymetrix Human SNP 6.0 microarrays. Linear mixed-model regression analyses were performed to determine the association between nuclear DNA methylation (dependent variable) and mtDNA-CN, adjusted for age, sex, site, cell composition, technical artifacts and surrogate variables (Bioconductor package: SVA). Twenty-five independent CpGs reached epigenome-wide significance (residual bootstrapping: p<6.22e-08) and increased mtDNA-CN was associated with increased global methylation (p<2.2e-16, ß=0.149). Weighted Gene Co-expression Network Analysis (WGCNA) identified a module of CpGs associated with mtDNA-CN (p=7e-06, r^2=0.1) and overrepresented by genes related to 'regulation of transcription from RNA polymerase II promoter' (p=1.9e-12). Mendelian randomization was used to establish the direction of causality between mtDNA-CN and nuclear methylation by exploring 1) The relationship between methylation associated SNPs and mtDNA-CN, and 2) The relationship between mtDNA-CN associated SNPs and methylation. To test if methylation is causal of mtDNA-CN, we identified 14 independent methylation quantitative trait loci (meQTLs) (p<5e-08) with signals in 11/25 of the CpGs associated with mtDNA-CN. meQTL SNPs were not associated with mtDNA-CN (p<0.05) and there was no enrichment of p-values, suggesting that nuclear methylation does not cause altered mtDNA-CN. To test if mtDNA-CN is causal of methylation, we leveraged preliminary data from an in-house GWAS of mtDNA-CN (n=7851) and identified an association between mtDNA-CN associated SNPs and mtDNA-CN associated CpGs, indicating that mtDNA-CN may drive nuclear methylation. The results suggest that changes in mtDNA-CN are influencing nuclear DNA methylation which may lead to changes in gene expression.

259

RNAseq in 302 phased trios provides a high resolution map of genomic imprinting. *A.J. Sharp[1], B. Jadhav[1], R. Monajemi[2], K.K. Gagalova[3,4], H.H.M. Draisma[3], T. Lappalainen[5,6], S. Castel[5,6], L. Franke[7], P.A.C. 't Hoen[3], S.M. Kielbasa[2], BIOS Consortium.* 1) Dept of Genetics & Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Dept of Medical Statistics and Bioinformatics, Bioinformatics Center of Expertise, Leiden University Medical Center, Leiden, The Netherlands; 3) Dept of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; 4) GenomeScan B.V., Plesmanlaan 1D, 2333 BZ Leiden, the Netherlands; 5) New York Genome Center, NY, USA; 6) Dept of Systems Biology, Columbia University, NY, USA; 7) Dept of Genetics, University Medical Center Groningen, Groningen, The Netherlands.

Combining allelic analysis of RNAseq data with phased genotypes in family trios provides a powerful method to detect parent-of-origin biases in gene expression genome-wide. We report findings in 302 family trios from two large studies: 165 lymphoblastoid cell lines (LCL) from three ethnicities in the 1000 Genomes Project, and 137 blood samples from the Genome of the Netherlands (GoNL). Based on parental haplotypes, we identified >2.5 million transcribed heterozygous SNPs phased for parental origin in 31,955 Gencode genes, and employed a robust pipeline for measuring allelic expression incorporating stringent correction for potential mapping bias to the reference genome. We performed simulation experiments, assessing several statistical approaches for the detection imprinting. We identified a total of 59 imprinted genes (10% FDR), with 23 (39%) observed in both populations. Our analysis identified 31 genes that have been previously reported as imprinted, while 28 represent putative novel imprinted genes. We identified multiple gene clusters where novel imprinted transcripts showing weak parental expression bias were located adjacent to known strongly imprinted genes. For example, we identified *PXDC1*, a gene which lies adjacent to the paternally-expressed gene *FAM50B*, as showing a 2:1 paternal expression bias. Similarly *ADAM23* lies ~130kb distal to *ZDBF2*, and also exhibits ~2-fold over-expression from the paternal allele. Other novel imprinted genes had promoter regions that coincide with sites of parentally-biased DNA methylation identified in uniparental disomy samples, thus providing independent validation of our results. Using the stranded nature of the RNAseq data in LCLs we identified multiple loci with overlapping sense/antisense transcripts showing opposing imprinting patterns, eg. *RB1*/*LPAR6* and *KCNQ1OT1*/*KCNQ1*. We also identified examples of isoform-specific imprinting, and at many loci also observed clear evidence of read through of imprinted transcription beyond gene annotations. Based on this we applied a sliding window approach to analyze parental expression bias across the entire genome, identifying multiple regions outside of annotated transcripts with evidence of imprinted transcription, suggesting putative novel imprinted lncRNAs. Our data provide a robust map of imprinted gene expression in the genome, identifying many novel imprinted genes, and providing new insights into the nature of genomic imprinting.

260

Genome-wide survey of parent-of-origin effects on DNA methylation identifies candidate imprinted loci in humans. *G. Cuellar-Partida[1], C. Laurin[3], T. Gaunt[2], S. Ring[2], C. Relton[2], G.D. Smith[2], D.M. Evans[1,2].* 1) Genomic Medicine, Diamantina Institute, The University of Queensland, Brisbane, Queensland, Australia; 2) Medical Research Council Integrative Epidemiology Unit, School of Social & Community Medicine, University of Bristol, United Kingdom.

Genomic imprinting is an epigenetic mechanism leading to parent-of-origin dependent expression of alleles. The extent to which imprinting is spread across the human genome is unknown. So far, the number of imprinted genes in humans is estimated to be around 100. In this study, we leveraged genome-wide DNA methylation in whole blood measured longitudinally at 3 time points (birth, childhood and adolescence) and GWAS data in 740 Mother-Child duos from the Avon Longitudinal Study of Parents and Children (ALSPAC) to systematically identify imprinted loci. We reasoned that *cis*-meQTLs at genomic regions that were imprinted would show strong evidence of parent of origin associations with DNA methylation, enabling the detection of both known and novel imprinted regions. Using this approach, we identified genome-wide significant *cis*-meQTLs that exhibited parent of origin effects (POEs) at 37 novel and 49 known imprinted regions ($10^{-10} < P < 10^{-300}$). Among the 37 novel loci, we observed signals near genes implicated in cardiovascular disease (*PCSK9*), and Alzheimer's disease (*CR1*), amongst others. Most of the significant regions exhibited patterns of imprinting consistent with uniparental expression, with the exception of twelve loci (including the *IGF2, IGF1R,* and *IGF2R* genes), where we observed a bipolar-dominance pattern (i.e. the two heterozygous genotypes having greater and lesser average methylation than the homozygotes). Parent-of-origin effects were remarkably consistent across the different time points and were so strong at some loci that methylation levels in heterozygous individuals enabled good discrimination of parental transmissions at these and surrounding genomic regions. The implication is that parental allelic transmissions could in theory be assigned at many imprinted (and linked) loci and hence parent of origin effects detected in GWAS of *unrelated* individuals given a combination of genetic and methylation data. Our results indicate that modelling POEs on DNA methylation is an effective way of identifying genetic loci that may be affected by imprinting.

261

Allele-specific expression improves local expression analysis of GWAS loci: Splicing of GSDMB identified for asthma in the human lungs. *Z. Miao[1,2], A. Ko[1,3], P. Pajukanta[1,2,3].* 1) Dept. of Human Genetics, UCLA, Los Angeles, CA; 2) Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA; 3) Molecular Biology Institute at UCLA, Los Angeles, CA.

Genome-wide association studies (GWAS) have successfully identified variants for complex diseases and traits, however, the current challenge is to elucidate the biological mechanisms underlying these GWAS loci. Limited sample sizes of expression quantitative trait locus (eQTL) studies often restrict the power to identify associations between the regulatory GWAS SNPs and their target genes. Here we aimed to discover local cis regulation of expression not identified by cis-eQTL studies due to the lack of power. To more comprehensively utilize the RNA-seq data from eQTL studies, we performed an allele specific expression (ASE) analysis with our newly developed method, ASElux. To minimize the reference alignment bias and fast and accurately count the allelic reads, ASElux uses an individual's SNPs as input and builds an SNP-aware index system. Using the imputed genotypes from the GTEx study, we counted allelic reads in 273 GTEx lung RNA-seq samples with ASElux and obtained significantly less reference mapping bias when compared to the allelic read counts reported by GTEx. Then we performed a paired t-test and identified 2,598 ASE SNPs ($p<2.32x10^{-6}$). Notably, 1,339 and 1,580 of these SNPs were missed by the GTEx cis-eQTL and ASE analyses, respectively, implying that ASE analysis performed by ASElux is more sensitive in identifying cis regulation of gene expression than the regular cis-eQTL analysis. Among the 1,339 ASE SNPs, a missense variant rs2305480, located in the Gasdermin B (GSDMB) gene is associated with asthma in GWAS and in full LD (R^2=1.00) with another intronic asthma GWAS SNP rs11078927. Here, we identified variant rs11078928 to be in LD (R^2=0.99) with both GWAS SNPs, rs2305480 and rs11078927. Rs11078928 is a splice donor site variant, previously reported to affect splicing of GSDMB in whole blood. To further investigate and verify the splicing effect in a human tissue relevant for asthma, we performed a splice-QTL analysis in 273 GTEx lung RNA-seq samples using LeafCutter and identified rs11078928 as a significant splice-QTL of GSDMB in the lungs ($p=4.63x10^{-35}$). These data indicate that an ASE analysis provides substantially more power for detection of local gene expression; and reveals how rs11078928 affects the risk of asthma through regulating the splicing events in the GSDMB gene, providing thus the biological mechanism underlying the asthma GWAS SNPs rs2305480 and rs11078927 in the human lungs.

262

Global transcriptomic analysis of human hematopoietic stem cells identifies alternative splicing of *HMGA2* in mediating stem cell properties. *M.H. Guo[1,2,3,4], M. Cesana[5,6,7,8], D. Cacchiarelli[1,8,9,10], L. Wahlster[5,6,7,8], S. Doulatov[11], L.T. Vo[5,6,7,8], B. Salvatori[12], C. Trapnell[13], K. Clement[1,8,9], P. Cahan[14], K.M. Tsanov[5,6,7,8], P.M. Sousa[5,6,7,8], J. Barragan[5,6,7,8], B. Tazon-Vega[1,8,9], F.M. Giorgi[15], A. Bolondi[5], A. Califano[11,16], J.L. Rinn[1,8,9], A. Meissner[1,8,9,17], J.N. Hirschhorn[1,2,3], G.Q. Daley[5,6,7,8].* 1) Broad Institute of MIT and Harvard, Cambridge, MA; 2) Division of Endocrinology, Boston Children's Hospital, Boston, MA; 3) Department of Genetics, Harvard Medical School, Boston, MA; 4) University of Florida College of Medicine, Gainesville, FL; 5) Stem Cell Transplantation Program, Division of Hematology/Oncology, Manton Center for Orphan Disease Research, Boston Children's Hospital and Dana-Farber Cancer Institute, Boston, MA; 6) Department of Biological Chemistry and Molecular Pharmacology, Harvard Stem Cell Institute, Harvard Medical School, Boston, MA; 7) Howard Hughes Medical Institute, Boston, MA; 8) Harvard Stem Cell Institute, Cambridge, MA; 9) Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA; 10) Telethon Institute of Genetics and Medicine, Pozzuoli, Italy; 11) Division of Hematology, University of Washington School of Medicine, WA; 12) Department of Systems Biology, Columbia University, New York, NY; 13) Department of Genome Sciences, University of Washington, Seattle, WA; 14) Department of Biomedical Engineering, Institute for Cell Engineering, Johns Hopkins University School of Medicine, Baltimore, MD; 15) Cancer Research UK, Cambridge Institute, Cambridge, UK; 16) Departments of Biomedical Informatics, Biochemistry and Molecular Biophysics, JP Sulzberger Columbia Genome Center, Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY; 17) Max Planck Institute for Molecular Genetics, Berlin, Germany.

While gene expression dynamics have been extensively catalogued during hematopoietic differentiation in the adult, less is known about transcriptome diversity of human hematopoietic stem cells (HSCs) at various stages of development. Here, we performed RNA-seq in human tissues along the developmental progression from fetal liver (FL) to cord blood (CB) to bone marrow (BM) HSCs. Our analyses included expression analyses at the gene and isoform level and were integrated with miRNA profiling in the same tissues. The analyses indicate that HSCs from different developmental stages have distinct expression profiles, including differential expression of key hematopoietic regulators such as *GATA2* and *RBPMS*. Interestingly, *HMGA2,* a transcription factor with a known role in stem cell biology and an important downstream effector of the LIN28-*let-7* axis, did not demonstrate the expected decrease in expression along the HSC developmental progression. Closer examination revealed that this discordant expression pattern is due to expression of a stage-specific shorter isoform of *HMGA2*. This shorter isoform is devoid of the canonical isoform 3'UTR, allowing it to escape targeting by *let-7* miRNAs. This short *HMGA2* isoform has similar chromatin binding properties as that of the canonical isoform, and is similarly capable of stimulating clonogenic capacity and engraftment potential in BM-HSCs. A cell line-based screen identified several candidate splicing factors capable of inducing the *HMGA2* splicing switch. Preliminary experiments suggest that these candidate splicing factors can promote clonogenic capacity and engraftment potential in a manner that is dependent on expression of the short *HMGA2* isoform. In summary, our work identifies a previously uncharacterized isoform of *HMGA2* that helps promote stem cell properties in later developmental stages by allowing it to escape the effects of increasing expression of repressive *let-7* miRNAs. Our work revises the canonical LIN28-*let-7*-HMGA2 axis to include alternative splicing as a means for continued HMGA2 expression and function at later stages of development.

263

Splice quantitative trait loci in human peripheral blood provide novel insight into the molecular determinants of COPD. *A. Saferali[1], A. Lamb[1], M. Parker[1], J.H. Yun[1], R.P. Chase[1], B.D. Hobbs[1], H.M. Boezen[2], K. de Jong[2], E.K. Silverman[1], M.H. Cho[1], P.J. Castaldi[1], C.P. Hersh[1].* 1) Brigham and Women's Hospital, Boston, MA; 2) University of Groningen, Groningen, the Netherlands.

Rationale: Expression quantitative trait locus (eQTL) studies identify single nucleotide polymorphisms (SNPs) that contribute to gene expression and may provide insight into biological mechanisms. While many disease-associated SNPs are eQTLs, a large proportion of genome-wide association study (GWAS) variants are of unknown function. Here, we investigate SNPs that are associated with alternative splicing (sQTL) to identify novel functions for GWAS variants in a common complex disease, chronic obstructive pulmonary disease (COPD). **Methods:** RNA sequencing was performed on whole blood from COPD cases (n=238) and controls (n=226). Associations between all SNP within 1000 kb of a gene (cis-) and gene expression or alternative splicing were tested using linear models, adjusting for age, gender, smoking status, white blood cell differential counts, PEER factors of expression data and principal components of genetic ancestry. A total of 5,815,008 SNPs were tested for association with 27,277 genes and 99,716 splice sites. COPD-associated SNPs were ascertained from a published GWAS (Hobbs et al., Nat Genet 2017;49:426). **Results** We identified a total of 1,230,063 cis-sQTLs (FDR<10%), comprising 514,662 unique SNPs. These sQTLs are associated with 14,741 splice sites corresponding to 18,042 unique genes. Similarly, 1,062,518 cis-eQTLs (688,148 unique SNPs) involving 16,329 genes were detected at 10% FDR. While there is overlap between sQTLs and eQTLs, 58% of sQTLs were not eQTLs. The genomic distribution of sQTLs and eQTLs was similar, with the majority located in intergenic (37% of sQTLs and 41% of eQTLs) and intronic (51% of sQTLs and 47% of eQTLs) regions. To determine what proportion of COPD-associated SNPs are sQTLs, 1746 GWAS SNPs ($p<10^{-5}$) were interrogated in the sQTL and eQTL results. We found that 364 GWAS SNPs were sQTLs (21%) in 20 genes compared to 285 SNPs that were eQTLs (16%) for 13 genes. Furthermore, 160 of the top GWAS SNPs were associated with only splicing and not gene expression. We identified 29 genes which contained at least one alternative splice site between COPD cases and controls. In contrast, only eight genes were differentially expressed between cases and controls. **Conclusions** Many SNPs were associated with alternative splicing in peripheral blood. More COPD-associated variants were sQTLs than eQTLs, suggesting that analysis of alternatively spliced genes can provide novel insights into disease mechanisms.

264

Multiple trait, gene expression, and splice junction associations at the cardiometabolic *MADD-NR1H3* GWAS locus. *C.K. Raulerson[1], A. Ko[2], M. Alvarez[2], T.S. Furey[1,3], M. Laakso[4], P. Pajukanta[2,5], K.L. Mohlke[1].* 1) Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA; 3) Department of Biology, University of North Carolina, Chapel Hill, NC; 4) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland; 5) Molecular Biology Institute at UCLA, Los Angeles, CA.

The variants, genes, and mechanisms that underlie genome-wide association studies (GWAS) loci are not well understood, especially at loci consisting of more than one signal. In a region of ~500 kb near *MADD* and *NR1H3*, we investigated the three proinsulin and three HDL-C GWAS signals for which none of the lead variants exhibit high pairwise linkage disequilibrium (LD) (all r^2<0.7). Previous studies have suggested that *MADD*, encoding MAP Kinase Activating Death Domain, may play a role in proinsulin because it interacts with MAPK, which has been implicated in proliferation of β cells. To identify genetic variants that influence the expression of genes (eQTLs) and the use of particular splice junctions (sQTLs) at this locus, we used 8.9M genotypes (MAF>0.01) and subcutaneous adipose tissue RNA-seq data (50-bp paired-end, mean ~45M reads/sample) from 387 Finns from the METabolic Syndrome in Men (METSIM) study. After STAR-2pass read alignment, we quantified transcripts using salmon, collapsed to the gene level, adjusted for 70 factors using PEER, and performed eQTL association tests using FastQTL. We identified known and novel splice junctions using Leafcutter, quantile-normalized, adjusted for 5 principal components, and performed sQTL association tests using FastQTL. At the *MADD-NR1H3* locus, we further used stepwise conditional analysis and identified two distinct gene-level eQTL signals for *NR1H3* (FDR<1%), and no eQTLs for *MADD*. *NR1H3* encodes LXRα, a transcriptional regulatory that has been implicated in regulation of cholesterol homeostasis and inflammation. The first *NR1H3* eQTL, rs75393320 (P=1.38E-31, β=0.82), exhibits moderate LD (r^2=0.76) with the 2nd proinsulin GWAS signal rs10838687. The second eQTL signal, rs11039149 (P=2.54E-11), exhibits high LD (r^2=1.0) with the 1st proinsulin GWAS signal rs10501320 and appears coincident based on reciprocal conditional analysis. We further identified an *NR1H3* sQTL, rs1449626-C (P= 9.88e-57), associated with increased expression of transcripts using a putative novel splice junction (chr11:47281530-47282792). This splicing event deletes exon 4 in frame, removing the zinc-finger binding domain, suggesting that the isoform that includes this splice variant encodes a less efficient or non-functional transcription factor. Taken together, these results indicate that *NR1H3* has a complex pattern of regulation and may be a target gene for one or more of the nearby GWAS signals.

265

RERE is required for normal eye development in humans and mice. *B. Kim[1], B. Freqeau[2], A. Hernandez-Garcia[1], V. Jordan[3], D. Stockton[4], M. Justice[5], E. Sherr[2], D. Scott[1,3].* 1) Molecular & Human Genetics, Baylor College Med, Houston, TX; 2) Department of Neurology, University of California, San Francisco, CA; 3) Department of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, TX; 4) Department of Pediatrics and Internal Medicine, Wayne State University, Detroit, MI; 5) Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON Canada.

Congenital eye defects occur in approximately 5 in 10,000 newborns. In an effort to identify the genes causing congenital eye anomalies, we performed an ENU screen for recessive eye defects. In this screen, we identified a novel mouse strain (*eyes3*) with a homozygous c.578T>C change in *Rere*. Complementation studies using an RERE null allele revealed that the *eyes3* allele was hypomorphic with *Rere*$^{-/eyes3}$ embryos having severe microphthalmia, coloboma, and optic nerve hypoplasia. We subsequently identified individuals with heterozygous, de novo loss-of-function variants in *RERE* who have microphthalmia, coloboma, optic atrophy and/or hypoplasia, Peter anomaly, iris anomalies, and blepharophimosis. Although RERE clearly plays a role in eye development, the morphogenetic and molecular mechanisms underlying the eye defects caused by RERE deficiency have not been fully elucidated. Using immunohistochemistry, we showed that RERE was expressed in the lens placode, the optic vesicle, the optic stalk and the perioptic mesenchyme of the early embryo. RERE was also expressed in the ganglionic cell layer (GCL), the lens epithelial cells and the cornea at E18.5. Further evaluations of *Rere*$^{-/eyes3}$ embryos on a C57BL/6 background revealed that the number of apoptotic cells was increased in the perioptic mesenchyme secreting the molecules that contribute to specification of the retinal pigment epithelium. In *Rere*$^{-/eye3}$ embryos, the lens vesicle is not developed due to incomplete invagination of the lens placode which, consequently, leads to abnormal expression of PAX6 in the optic cup at E10.5. The lens was not identified in *Rere*$^{-/eyes3}$ embryos at E15.5. Interestingly, the lens development in *Rere*$^{-/eye3}$ embryos on a mixed B6/129S6 background was not affected. On this mixed background, the number of apoptotic cells was increased in the GCL of *Rere*$^{-/eyes3}$ embryos starting at E16.5 and the ganglion cell number was reduced starting at E17.5. Further examinations revealed that each layer of the retina was normally developed in *Rere*$^{-/eye3}$ mice at P21. However, in adult *Rere*$^{-/eye3}$ mice, the retina was significantly degenerated and the GCL disappeared. The optic nerves of *Rere*$^{-/eyes3}$ mice were also shown to be hypoplasic. We conclude that RERE plays a critical role in lens development, and deficiency of RERE leads to ganglion cell apoptosis and optic nerve hypoplasia. These finding provide insight into the eye defects seen in patients with RERE deficiency.

266

Loss of ABCB5 leads to progressive visual loss due to sphingolipid accumulation in retinal pigment epithelium. *G. Gonzalez[1,2], Y. Sasamoto[2], P. Banerjee[3], J. Akula[3], V. Poulaki[3], G. Berg[1,3], M.H. Frank[3], B.R. Ksander[4], N.Y. Frank[1,2].* 1) VA Boston Healthcare, Harvard Medical Sch; 2) Brigham and Women's Hospital, Harvard Medical School; 3) Boston Children's Hospital, Harvard Medical School; 4) Massachusetts Eye and Ear Infirmary, Boston, MA.

Purpose: Retinal pigment epithelium (RPE) has a central role in eye development and is critical for the maintenance of the blood-retinal barrier, and homeostasis and survival of photoreceptors during adult life. Progressive loss of RPE isthought to be a major contributor to age-related macular degeneration (AMD). Currently, the general consensus is that RPE is maintained during aging via proliferation and enlargement of mature adult cells. Recently, however, this idea was challenged when a small population of cells was identified within the human RPE layer that displayed stem cell characteristics in vitro. We hypothesize that the stem cell gene ABCB5 is expressed by a subpopulation of RPE cells, which retain the ability to self-renew and differentiate and are essential for RPE homeostasis. **Methods:** Genetically engineered ABCB5 KO mice and ABCB5 GFP/Cre reporter mice were used to analyze ABCB5 function in retinal maintenance and to monitor murine ABCB5+ RPE cells in situ. Mice were evaluated by immunofluorescence, RNA in situ hybridization, FACS, and liquid chromatography/mass spectrometry (LC/MS). In addition, in vivo evaluation of visual function was performed by comparative full-field electroretinography (ERG) on young and aged ABCB5 KO and ABCB5 WT mice. Two-tailed Student's t-test was used for statistical analysis. **Results:** In the adult eye, we found specific GFP (ABCB5) expression in the RPE of ABCB5 GFP/Cre mice, which was further corroborated by ABCB5 mRNA in situ hybridization and FACS studies. Moreover, up to 50% of cells within the ABCB5 (+) RPE subpopulation did not express the differentiation marker RPE65. ERG analyses revealed reduction in both the amplitude and the sensitivity of the flash responses, as well as longer implicit times in the one-year-old KO animals compared to age-matched WT mice, indicating that loss of ABCB5 results in progressive functional visual abnormalities. There were associated alterations in lipid metabolism revealed by LC/MS analyses, such as accumulation of pro-apoptotic ceramide and its metabolite GM3 ganglioside. **Conclusions**: Our results demonstrate that ABCB5 identifies an RPE65-negative progenitor cell subpopulation within adult RPE. In addition, we show that intact ABCB5 function is important for normal RPE homeostasis and maturation as demonstrated by its role in sphingolipid metabolism.

267

***CNIH4* and *PMEL* rare variants cause ocular pigment dispersion syndrome predisposing to pigmentary glaucoma.** *J.L. Wiggs[1], B.J. Fan[1], K. Allen[1], Q. Zhang[1], M.H. Kang[2], D.J. Rhee[2], D.S. Greenfield[3], R.K. Parrish[4], K. Linkroum[1], L.R. Pasquale[1], E.A. Pierce[1], C.J. Hammond[5], P.G. Hysi[5], N. Weisschuh[6], M.J. Simcoe[5], R.M. Leonhardt[7], R. Ritch[8].* 1) Dept Ophthalmology, Harvard Med Sch, MEEI, Boston, MA; 2) Dept Ophthalmology, Case Western Reserve University School of Medicine, Cleveland, OH; 3) Department of ophthalmology, Bascom Palmer Eye Institute, University of Miami Miller School of Medicine, Palm Beach Gardens, FL; 4) Anne Bates Leach Eye Hospital, University of Miami, Bascom palmer Eye Institute, Miller School of Medicine, Miami, FL; 5) Department of Twin Research and Genetic Epidemiology, King's College London, London UK; 6) Institute for Ophthalmic Research, Centre for Ophthalmology, University of Tubingen, Tubingen, Germany; 7) Department of Immunobiology, Yale University School of Medicine, New Haven, CT; 8) Einhorn Clinical Research Center, New York Eye and Ear Infirmary of Mount Sinai, New York, NY.

Pigment dispersion syndrome (PDS) is an ocular condition that predisposes to a specific type of glaucoma (pigmentary glaucoma, PG), characterized by release of pigment granules from the iris that accumulate in the fluid outflow pathways causing elevation of intraocular pressure (IOP) and damage to the optic nerve. PDS/PG typically develops during the 3rd to 5th decades and is the most common type of glaucoma in young adults. The disease exhibits dominant inheritance with variable penetrance, however causative genes have not yet been identified. Using whole exome and targeted Sanger sequencing and a cohort of 147 PDS/PG index cases from the United States, we identified rare missense mutations in two genes, *CNIH4* (human cornichon homologue 4) (A3G, V5E and G54S) and *PMEL* (premelanosome protein) (G175S, G325V and p.Ser641_Ser642del) in 12 cases. The *PMEL* and *CNIH4* mutations segregated in an autosomal-dominant fashion in 3 families while 9 individuals were isolated case subjects. Four members of one family with earlier disease onset had mutations in both genes (*CNIH4* A3G and *PMEL* G175S). The *CNIH4* G54S mutation was also identified in 2 of 227 cases from an independent German cohort and in none of a total of 791 controls. *PMEL* is known to be important for melanosome formation and ocular and skin pigmentation, while *CNIH4*, recently shown to be involved in endoplasmic reticulum (ER) transport of newly formed G protein-coupled receptors (GPCRs), has not previously been shown to be involved in pigmentation or ocular disease. Knockdown of *CNIH4* in zebrafish revealed decreased ocular pigmentation that was rescued by wild type *CNIH4* but not *CNIH4* mRNA coding for the V5E or G54S missense alleles. Using immunohistochemistry we also identified CNIH4 in human pigmented ocular structures including the iris and the pigmented retinal epithelium. In conclusion we have identified rare variants in *CNIH4* and *PMEL* in PDS/PG cases, identifying the first genes known to be responsible for this blinding disease. Overall our study suggests that *CNIH4* and *PMEL* mutations account for up to 8% of PDS/PG cases, and that abnormalities in genes involved in ocular pigmentation is one pathway for PDS/PG pathogenicity. Supported by: Bright Focus Foundation, March of Dimes Foundation, NIH/NEI P30EY014104, NIH/NIAMS R21-AR068518.

268

Electronic health records elucidate complex relationships between genetics, the anatomy of the eye, and disease. *C.R. Bauer[1], E.D.K. Cha[1], A.B. Paaby[2], D. Lavage[1], S.A. Pendergrass[1] on behalf of the DiscovEHR collaboration.* 1) Biomedical and Translational Informatics Institute, Geisinger Health System, Danville, PA; 2) School of Biological Sciences, Georgia Tech, Atlanta, GA.

In 2016, more than 25% of the US population over the age of 50 suffered from one of four vision impairing diseases of the eye: macular degeneration, glaucoma, diabetic retinopathy, and cataracts. The incidence of these diseases, and visual impairment, is also on the rise. It is important both to diagnose these conditions earlier and to gain a better understanding of the molecular genetic factors that contribute to them. Previous studies have identified many genetic risk factors for these diseases, but have been limited in considering the complex relationships between the structural and functional components of the eye and how they interact with each other and with our genes to affect disease outcomes. Using electronic health record (EHR) data linked to genotypes from the Geisinger Health System MyCode Community Health Initiative cohort, through the DiscovEHR initiative, we have conducted a comprehensive analysis to assess the relationships between ~600,000 common frequency array based genetic variants and multiple ocular traits in ~60,000 individuals. We used 10 quantitative ocular clinical measures of the eye for individuals that underwent an eye exam at GHS, as well ICD-9 derived case/control diagnoses for 45 ocular diseases within the EHR of GHS, requiring 3 or more instances of an ICD-9 code for cases and zero instances of a code to be considered a control. Our results replicate many previous associations, such as for macular degeneration and Fuch's corneal dystrophy. We also found novel associations, such as cup-to-disc ratio (CDR) associated with a locus on chromosome 12 (rs4882462, p=7.04E-09), and two loci (rs7530954, p=1.88E-09 and rs2877651 p= 1.51E-08) that are associated with corneal opacity and macular puckering, respectively. We also performed a PheWAS using the top genetic associations from our quantitative and binary ocular traits, to identify associations between these loci and non-ocular trait diagnoses. Finally, we investigated the association between our quantitative ocular measures and all ICD-9 diagnoses. We found evidence that CDR is linked to other conditions, such as cataracts, in addition to its better characterized association with glaucoma, supporting the notion that diseases of the eye tend to involve complex relationships between multiple structures. Our future work includes further investigation of the interplay between genes, anatomical structure, and biological function for ocular traits.

269

Powerful genome-wide association screening of serum metabolites in 10K individuals. *A. Gallois[1], J. Mefford[2], A. Ko[3], M. Laakso[4], N. Zaitlen[2], P. Pajukanta[3], H. Aschard[1,5].* 1) Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris, France; 2) Department of Medicine, University of California, San Francisco, CA, USA; 3) University of California, Los Angeles, CA, USA; 4) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland; 5) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.

The precision medicine initiative (PMI) and other large-scale cohort programs have initiated the collection of thousands of phenotypes and omics data across millions of individuals in order to improve the quality and efficacy of health care. One of the primary objectives of these initiatives is to map the genetic basis of complex disease phenotypes in the context of high dimensional data. Future success in the analysis of these large-scale high dimensional datasets relies on the development of new methodological approaches. In particular, association mining faces a major multiple comparisons challenge, with true signals buried inside the noise of all associations queried. This is especially true in human genetic association studies where a substantial proportion of the variation is driven by numerous genetic variants of small effect, and where current analysis methods are incapable of integrating thousands of correlated clinical and genetic variables. Here we present CMS (Covariates for Multi-phenotype Studies), an innovative and computationally efficient approach geared toward the analysis of PMI and similarly rich phenotypic data. Our approach keeps the univariate properties of determining association between a single outcome and a single predictor, but as in multivariate approaches, leverages other available variables producing increases in power equivalent to a two or even three fold increase in sample size. We applied CMS in the Finnish Metabolic Syndrome In Men (METSIM) cohort, performing systematic association screening between 600K common genetic variants and more than 100 metabolites in 10,000 individuals. Preliminary analyses show that CMS identified dozens of associated variants that are missed by the standard univariate association screening. On average, we observe a 50% power increase for CMS when compared to the standard approach. All new associations identified by CMS replicated at a nominal significance (P<0.05) in at least one of the previous large-scale metabolite GWAS (Kettunen et al 2012, Shin et al 2013, Rhee et al 2013), providing proof-of-principle evidence of the performance and power gain of CMS for GWAS analyses. Many CMS-identified association signals, such as the one observed with the variants in the *APOA5* region for branched-chain amino acids (BCAA) leucine and isoleucine, are strong candidates and of high interest in the study of metabolic and obesogenic disorders.

270

Genetic determinants of the human plasma proteome and their role in biology and disease. *B.B. Sun[1], J.C. Maranville[2], J.E. Peters[1], C.S. Fox[2], R.M. Plenge[2], J. Danesh[1], H. Runz[2], A.S. Butterworth[1].* 1) MRC BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; 2) MRL, Merck & Co., Inc., Kenilworth, New Jersey, USA.

Proteins are the primary functional units in biology and targets of most drugs. So far, there has been limited knowledge of the genetic factors determining variation in protein levels. Using an expanded high-throughput multiplex aptamer-based proteomic assay with more than twice the proteome coverage of previous studies, we performed genome-wide association studies (GWAS) for 2,994 proteins against 10.6 million genetic variants in 3,301 healthy individuals from the INTERVAL study. We identify 1,927 genetic associations with 1,478 proteins (protein quantitative trait loci [pQTLs]) with additional secondary associations identified through conditional analysis. Our findings represent approximately a 4-fold increase on existing knowledge, enabling us to gain novel insights into the genetic architecture of the human plasma proteome. We examine the extent to which the genetic loci associated with protein levels overlaps with and are enriched for expression quantitative trait loci. We use several approaches to highlight the application of pQTLs to biology and disease. First, we demonstrate how distant pQTLs can be linked to biologically plausible genes and the mediation of distant pQTLs by local protein levels, highlighting the role of protein-protein interactions. In addition, we find epistatic effects of genetically determined phenotypes (blood group and secretor status) on protein levels. Through linking previous disease associations, we use pQTLs to provide insights into possible mechanisms underpinning some of the disease loci. Lastly, we show how our pQTLs can be used to inform causal roles for protein biomarkers in disease through Mendelian randomisation analysis, leveraging the simultaneous measurement of multiple functionally related proteins to account for potential pleiotropic effects. Our data provide a valuable resource for the scientific community, to enhance understanding of biological and disease pathways across a range of domains.

271

***Trans*-eQTL meta-analysis in over 30,000 blood samples identifies genes and pathways affected by disease-related genetic variants.** *U. Võsa[1], A. Claringbould[1], T. Esko[2], L. Franke[1], BIOS Consortium, eQTLGen Consortium.* 1) Department of Genetics, University Medical Center Groningen, Groningen, Groningen, Netherlands; 2) Estonian Genome Center, University of Tartu, Tartu, Estonia.

During the last decade, many disease- and trait-associated genetic risk factors have been identified through genome-wide association studies (GWAS). However, the genes and pathways through which these variants exert their effect to the phenotype are still largely unknown. To identify trait-associated molecular pathways, we have combined the genetic and transcriptomic data from more than 30,000 blood samples within the eQTLGen Consortium, profiled by different expression profiling platforms. Using a standardized analysis pipeline, we performed *cis*-eQTL mapping on ~7 million genetic variants and conducted *trans*-eQTL mapping on 10,598 genetic risk factors, known to be associated with complex diseases and traits. Additionally, we have calculated polygenic risk scores for 1,267 complex traits and correlated those with gene expression levels. We observed a saturation of *cis*-eQTL effects: 88% of 14,575 blood-expressed protein-coding genes show a significant *cis*-eQTL effect, and 67% of these genes show independent secondary signals. We identified *trans*-eQTLs effects for over 30% of all established genetic risk factors for disease, impacting expression of 4,707 unique genes, and observed 5 examples of individual “hub” SNPs having downstream effects on over 200 genes. *Trans*-eQTLs show strong enrichment for overlapping with *trans*-meQTLs ($P<2.2\times10^{-16}$) and Hi-C interchromosomal contacts ($P=1.2\times10^{-14}$). For 2,585 risk SNPs we identified both local and distal effects to the gene expression, enabling us to investigate the relationship between *cis*-eQTL and *trans*-eQTL genes: these pairs of genes more often have protein-protein interactions, and mediation analysis in a subset of samples prioritized several associations where the *trans*-eQTL effect is mediated by the corresponding *cis*-eQTL gene. For individual diseases, the risk SNPs often converge on a shared cluster of genes: for example, the risk SNPs for inflammatory bowel disease converge on three clusters of genes representing B-cell signaling, type I and type II interferon signaling. We observed that the polygenic risk score for HDL cholesterol levels was associated with the expression of genes known to play a role in lipid metabolism (e.g. *ABCA1*, *ABCG1*) and familial hypercholesterolemia (e.g. *LDLR*).

272

An imputation-based approach for matching unknown signals across untargeted metabolomics datasets. *Y.H. Hsu[1,2,3], C. Churchhouse[4,5], T. Esko[6], A. Metspalu[6], J.M. Mercader[3,7], C. Gonzalez[8,9], M.E. Gonzalez[8,9], J.N. Hirschhorn[1,2,3].* 1) Department of Genetics, Harvard Medical School, Boston, MA; 2) Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA; 3) Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA; 4) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA; 5) Analytical and Translational Genomics Unit, Massachusetts General Hospital, Boston, MA; 6) Estonian Genome Center, University of Tartu, Tartu, Estonia; 7) Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 8) Instituto Nacional de Salud Publica, Cuernavaca, Morelos, Mexico; 9) Centro de Estudios en Diabetes, Mexico City, Mexico.

Metabolomics is a powerful approach for discovering biomarkers and metabolic quantitative trait loci. While untargeted profiling methods can measure thousands of metabolite signals, many signals cannot be readily identified as known metabolites or compared across datasets, making it difficult to perform well-powered meta-analyses across studies. To deal with this challenge, we developed a method that uses imputation to match up signals likely to be the same metabolites across mass spectrometry-based profiling datasets. In brief, we imputed the predicted abundance of unknown signals from one dataset to another using shared known metabolites as predictors. The imputation then allowed us to calculate correlation between signals measured in different datasets across a common set of samples and to match signals based on agreement in mass-to-charge ratio (m/z) and this correlation. We used LC-MS (liquid chromatography-mass spectrometry) profiling data from two cohorts to test our method: (1) Obesity Extremes (OE): 13,613 signals (322 known) were measured in 298 samples drawn from the body-mass index extremes of Estonian Biobank and (2) Mexico City Diabetes Study (MCDS): 7,136 signals (242 known) were measured in 824 samples from a prospective type 2 diabetes study. As a proof of principle, we treated shared known metabolites as unmatched signals and applied our method to show that it outperformed a retention time and m/z-based matching approach. Our method correctly matched 58.4% of the shared knowns; furthermore, 91.0% of all matches were strongly correlated ($r^2 > 0.8$) with the true known matches. Next, we used genetic data to validate all OE-MCDS matches, reasoning that useful matches would show similar genetic associations across the two cohorts. We performed genome-wide association analyses for 4,432 matched signal pairs and 207 shared known pairs, which serve as positive controls. We identified potential associated variants for all pairs (best variant with $p < 5e\text{-}8$ in meta-analysis that ignored direction of effect) and then checked for directional consistency of their association. Matched pairs that included unknown signals showed only slightly less consistency than shared known pairs (63.1% vs 70.2%). At more stringent *p*-value thresholds, as many as 81.0% of matched pairs had directional consistency. These results showcase our method as a useful tool for meta-analyzing unknown signals, which can greatly improve the power of untargeted metabolomics studies.

273

First genome wide association study of Internet addiction revealed strong shared risk factors with psychosis. *A. Haghighatfard[1], A. Ghaderi[2].* 1) Department of Biology, Tehran North Branch, Islamic Azad University, Tehran, Ira, Tehran, Iran; 2) Cognitive Neuroscience Lab, Department of Psychology, University of Tabriz, Tabriz, Iran.

Internet addiction disorder (IAD) is listed in section III, as a disorder requiring further studies in the latest diagnostic and statistical manual of mental disorders (DSM-V). Psychological studies were showed significant co-morbidity of Internet addiction disorder (IAD) with depression, alcohol abuse and anxiety disorder. Etiology and genetic bases of tendency to excessive usage of Internet are not clarified. Present study aimed to investigate the genetic bases of internet addiction tendency and psychological and cognitive mechanisms which are involved in internet addiction. DNA was extracted from blood samples of Internet addicted subjects (N= 13890) and 18000 matched non-psychiatric subjects. Diagnoses of IADs were approved by two senior psychiatrists based on Internet Addiction Test Questionnaires (IAT). Genotyping for the subjects were performed using the Affymetrix Genome-Wide Human SNP Array 6.0. In addition a comprehensive assessment on psychological, neuropsychological and neurological characteristics of Internet addiction conducted. Seventy two SNPs in 24 genes have been detected significantly associated with IAD. Most of these SNPs were risk factors of psychiatric disorders. Hyper-geometric analysis was showed most similarity between IAD with autism spectrum disorder, bipolar disorder, schizophrenia and attention deficits hyperactivity disorder. IAD subjects were showed higher anxiety, stress and neuroticism and deficits in working memory, attention, planning and processing speed. These psychological patterns were correlated with severity of Internet addiction symptoms. This is the first genome wide association study of IAD. Results were showed IAD may have genetic bases and strong shared genetic risk factors with neurodevelopmental psychiatric disorders. Genetic risk factor in IAD subjects may cause several cognitive and brain functions abnormalities which lead to excessive Internet usage. It may suggest that IAD should take more seriously because Internet excessive usage could be a marker for vulnerability to sever psychiatric disorders such as Autism and Schizophrenia.

274

Genome-wide analyses of smoking behaviors in schizophrenia: Findings from the Psychiatric Genomics Consortium. *R. Peterson[1], T. Bigdeli[1,2], K. Kendler[1], A. Fanous[1,2,3], Schizophrenia Working Group of the Psychiatric Genomics Consortium.* 1) Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA; 2) Department of Psychiatry and Behavioral Sciences, State University of New York Downstate Medical Center, Brooklyn, NY; 3) Mental Health Service Line, Washington VA Medical Center, Washington DC.

Objective: While 17% of US adults use tobacco regularly, smoking rates among persons with schizophrenia (SCZ) spectrum disorders are upwards of 60%. Several lines of evidence support a shared etiological basis for smoking and SCZ, including recent findings from genome-wide association studies (GWAS). However, few studies have considered whether genetic variants also influence smoking behavior among SCZ cases. *Methods*: We combined single nucleotide polymorphism (SNP) genotypes with self-reported nicotine-use and symptom data from the Psychiatric Genomics Consortium (PGC) study of SCZ. We evaluated whether polygenic risk scores (PRS) constructed from the results of the Tobacco and Genetics (TAG) meta-analyses of smoking behaviors are associated with SCZ risk, or predict these same outcomes in cases. Using genome-wide summary statistics for SCZ, and results from exploratory case-only GWAS of smoking initiation (SI) and cigarettes-smoked-per-day (CPD), we estimated the genetic correlation with TAG phenotypes. *Results*: We demonstrate significant genetic correlations of SCZ with SI (ρ_g=0.159; P=5.05×10^{-10}), CPD (ρ_g=0.094; P=0.006), and age-of-onset of smoking (ρ_g=0.1; P=0.009) in the general population, and successfully replicate findings for SI and age-of-onset in an independent East-Asian cohort. Comparing SCZ-cases to the general population, we observe a significant positive genetic correlation for SI (ρ_g=0.624, P=0.002). Similarly, TAG-based PRS for SI and CPD were significantly associated with SI (P=3.49×10^{-5}) and CPD (P=0.007) among cases. We also successfully replicated a novel SNP association with CPD among cases upstream of *TMEM106B* on chromosome 7 (rs148253479; P=3.18×10^{-8}; N=3,520). *Conclusion*: We provide evidence of a partially shared genetic basis for SCZ and smoking behaviors, and for smoking behaviors among SCZ patients and the general population. Future research needs to address mechanisms underlying associations between these traits to aid both SCZ and smoking treatment and prevention efforts.

275

Genetic associations of maximum regular alcohol intake in the Million Veteran Program. *J. Concato[1,2], N. Sun[1,3], Q. Lu[3], Y. Hu[3], B. Li[3], Q. Chen[1,3], M. Aslan[1,2], K. Radhakrishnan[1], K.H. Cheung[1,4], Y. Li[1,5], R. Pietrzak[6,7], N. Rajeevan[1,5], F. Sayward[1,5], K. Cho[8,9], K. Harrington[8,10], J. Honerlaw[8], S. Pyarajan[8,9], R. Quaden[8], J.M. Gaziano[8,9], H. Zhao[1,3], M.B. Stein[11,12], J. Gelernter[6,7,13] on behalf of the VA Million Veteran Program.* 1) VA Clinical Epidemiology Research Center, VA Connecticut Healthcare System, West Haven, CT, USA; 2) Department of Medicine, Yale University School of Medicine, New Haven, CT; 3) Department of Biostatistics, Yale University School of Public Health, New Haven CT; 4) Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT; 5) Yale Center for Medical Informatics, Yale University School of Medicine, New Haven, CT; 6) Psychiatry Service, VA Connecticut Healthcare System, West Haven, CT; 7) Department of Psychiatry, Yale University School of Medicine, New Haven, CT; 8) Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA; 9) Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 10) Department of Psychiatry, Boston University School of Medicine, Boston, MA; 11) Psychiatry Service, VA San Diego Healthcare System, San Diego, CA; 12) Department of Psychiatry, University of California San Diego, San Diego, CA; 13) Departments of Genetics and Neuroscience, Yale University School of Medicine, New Haven, CT.

The Veterans Affairs (VA) Million Veteran Program (MVP) is one of the world's largest databases of medical and genomic information, with currently >570,000 consented participants and links to the robust VA clinical data as well as questionnaire responses; approximately 350,000 enrollees have genotype information available. In this report, we present results from a VA Cooperative Studies Program initiative (#575B) that seeks to identify genetic risk factors relevant to PTSD and related traits (such as alcohol use phenotypes) in the US veteran population. We conducted a GWAS in 146,660 European-American subjects for a quantitative phenotype based on alcohol consumption: the response to "In a typical month, what is/was the largest number of drinks of alcohol (beer, wine, and/or liquor) you may have had in one day?" Four distinct—and mostly novel—common-variant genomewide-significant (GWS) regions were identified: chromosome 4, lead SNP rs283413 (2.5E-09), gene *ADH1C*; chromosome 8, lead SNP rs7821592 (2.5E-08), closest gene *XPO7* (exportin 7); chromosome 10, lead SNP rs1577857 (3.0E-08), closest gene *LOC105378478*; and chromosome 17, lead SNP rs77804065 (4.8E-12), which maps to an extended LD region identified in prior GWS associations. The lead ADH-region variant was at *ADH1C* rather than *ADH1B*; the latter is the locus most strongly associated previously with *lifetime* maximum drinks measures. The strongly significant region on chromosome 17 includes numerous gene loci including *CRHR1* (corticotropin releasing hormone receptor 1), a strong candidate for affecting drinking behavior. Applying the summary statistics-based version of PrediXcan to calculate gene-level test statistics across 44 tissues, we identified 15 significant genes after Bonferroni correction, including *LRRC37A2* (8.5E-13), previously implicated in cognitive function. Using LD score regression and tissue and cell type-specific annotations (GenoSkyline-Plus), many tissues/cell types showed significant enrichment in functionally annotated regions, with skeletal muscle being most significant (2.3E-05). Based on a comparison between summary statistics from this study and those from published GWASs of 55 traits, 9 of the traits showed statistically significant genetic correlations, including smoking (4.8E-29). These results are described for a novel and clinically relevant measure of alcohol use, and identify significant signals at loci not previously associated with alcohol use.

276

GWAS meta-analysis identifies > 200 novel loci for smoking and drinking addiction. *Y. Jiang, the GWAS and Sequencing Consortia of Alcohol and Nicotine addiction.* Penn State College of Medicine, Hershey, PA.

Smoking and drinking are leading preventable risk factors of human diseases. Smoking and drinking addiction are heritable traits. Yet, very few genetic loci were consistently implicated in previous GWAS. To fill in this knowledge gap, we conducted one of the largest-scale meta-analyses of smoking and drinking phenotypes, comprising of 30 cohorts and ~ 1 million research participants. Datasets include cross sectional studies, electronic medical record based biobanks as well as samples from direct-to-consumer genetic testing companies. The smoking phenotypes include smoking initiation (SI), the age onset of smoking initiation (AI), smoking cessation (SC), cigarettes-per-day (CPD). The drinking phenotypes include drinks per week (DPW), and drinker vs. non-drinker status (DND). We performed genotype imputation for all participating studies using the haplotype reference consortium panel, with the exception of a few studies imputed with the 1000 Genome Project Phase 3 panel. We performed fixed-effect imputation-aware meta-analyses by combining optimally weighted score statistics. After quality control, ~78.5 million SNPs were analyzed in meta-analysis. Genomic control values for common variants and rare variants were both well calibrated (<1.1) and quantile-quantile plot for p-values were well-behaved. At genome-wide significance threshold (p<5e-8), we identified 15 loci associated with CPD, 159 loci with SI, 17 loci with SC, 35 loci with DND and 36 with DPW. Among these loci, 239 are detected at genome-wide significance for the first time. A majority (78%) of the associated SNPs have MAF>5% with average MAF being 34%. There are considerable overlaps between identified drinking and smoking loci. Among the identified, 22 are simultaneously associated with at least one smoking trait and one drinking trait at genome-wide significance. We further gained mechanistic insights on the identified loci via functional enrichment analysis. Loci associated with smoking related traits are enriched with brain-specific enhancers and weak-enhancers (p=2e-6), and loci associated with DPW are enriched with active enhancers in adult liver (p=1e-4) and brain (p=.0005). More detailed functional genomic analyses on the identified loci are underway. Together, the results substantially expanded our knowledge on the genetics of smoking and drinking addictions. The study demonstrates the power of using large genetic datasets for understanding self-reported addiction phenotypes.

277

Multi-omics profiling for individualized precision wellness using blood and saliva. *G.I. Mias.* Biochemistry and Molecular Biology, Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI.

The emerging field of precision individualized wellness is aided by rapid advancements in sequencing and mass spectrometry technologies that allow the monitoring of thousands of molecular components. The next steps involve expansion beyond static genomic information, supplementing such information with the integrated dynamics of other omics (such as transcriptomes, proteomes and small molecules) to reveal individualized responses to disease over time, compared to one's own individualized wellness baseline. We will present findings from our clinical trial, monitoring individualized response to pneumococcal vaccination, where we have carried out integrative profiling on peripheral blood mononuclear cells (PBMCs) and saliva pre and post vaccination. This is to our knowledge the most extensive saliva-based omics dataset on an individual, covering 100 timepoints over the course of one year. The time span covers a healthy period as well as comprehensive monitoring of innate and adaptive immune responses following pneumococcal vaccination. Protein and RNA from saliva and PBMCs were produced at each timepoint (100 timepoints for saliva, 25 for PBMCs), and mass spectrometry proteomics and RNA-sequencing were carried out for all samples in non-targeted comprehensive profiling. The time series for all molecular components were analyzed and integrated using our MathIOmica software, to reveal clear temporal trends across the different components. The trends show distinct categorization of collective molecular response that reflects the activation of the immune system in saliva, both during the innate as well as the adaptive time frames. This research is to our knowledge the first systematic implementation of integrative personalized wellness monitoring, in the context of both including a medically relevant and controlled immune activation, pneumococcal vaccination, as well as displaying the response in non-invasive components (saliva). Our study has the potential to act as an extensible prototype for true universal precision wellness, and future studies for any other diseases that may be monitored in blood and/or saliva, including immunizations. G.I.M. and research reported are supported by grants from MSU and the National Human Genome Research Institute of the National Institutes of Health under Award Number R00 HG007065. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

278

Integrative personal omics profiling during periods of environmental stress. *M. Snyder[1], W. Zhou[1], B. Piening[1], R. Sailani[1], K. Contrepois[1], H. Roest[1], S. Ahadi[1], S. Leopold[2], B. Hansen[2], M. Avina[1], V. Rao[1], T. Mishra[1], S. Rose[1], G. Gu[1], B. Lee[1], S. Chen[1], S. Rego[1], D. Perelman[3], B. Leopold[2], T. McLaughlin[3], E. Sodergren[2], G. Weinstock[2].* 1) Genetics, Stanford University, Stanford, CA; 2) The Jackson Laboratory for Genomic Medicine, Farmington, CT; 3) Endocrinology, Stanford University School of Medicine, Stanford, CA.

Type 2 diabetes mellitus (T2D) is a significant health problem facing our nation, it's showed that early lifestyle or medical intervention in prediabetics can prevent conversion to T2D nearly by half, however, overall our ability to predict which individuals will develop T2D and when this will occur by which mechanism is strikingly inadequate. To better understand these factors for more effective early intervention, in particular to understand changes in response to various physiological stresses in prediabetes and whether those are associated with the risk to convert to T2D, we profiled 107 subjects with more than 900 longitudinal visits total that span over three years. Among these, 98 subjects were prediabetic when enrolled while 7 were healthy and 2 were diabetic controls. Stresses we sampled included respiratory viral infection, antibiotics intakes and others that are linked to the development of diabetes previously. We measured millions of molecular analytes in host blood by multi-omics profiling, including on transcriptome, metabolome and proteome, and also tracked microbial taxonomic changes of four body sites (nares, skin, tongue and gut) over those visits, to fully understand the complex molecular dynamics and regulations in the host-microbiome interactions. A subset of participants were placed on a short-term high caloric diet, followed by additional multi-omic profiling. The dietary perturbation was associated with a wealth of biomolecular expression changes concomitant with weight gain and spanning multiple ‘omes including the microbiome, and the omic response to weight gain differed between prediabetics and healthy controls. For another subset of participants who went through respiratory viral infections, their multi-omic profiling, including the microbiome, responded distinctly to different illness stages during the infection. Overall, the multi-omic profiles of individuals were unique to themselves compared to others regardless diet or illness perturbations. We found significantly higher rate of actionable findings (14%) identified on exome sequencing, and characterized personal health profiles that were most discriminated from their own physiological changes. In total, these large-scale longitudinal data offer a novel and comprehensive view of the dysfunction in cellular networks associated with the progression to T2D and may offer new strategies for managing personal health, predicting and preventing diseases.

279

Sex-specific inbreeding depression in humans. *D.W. Clark[1], P.K Joshi[1], T. Esko[4,5,6], J.F. Wilson[1,2,3] on behalf of the ROHgen Consortium.* 1) Usher Institute of Population Health Sciences, University of Edinburgh, UK; 2) MRC HGU, University of Edinburgh, UK; 3) IGMM, University of Edinburgh, UK; 4) Estonian Genome Centre, University of Tartu, Estonia; 5) Broad Institute, Cambridge, Massachusetts; 6) Department of Genetics, Harvard Medical School, Boston, Massachusetts.

In many plant and animal species the offspring of related parents suffer reduced reproductive success – a phenomenon known to evolutionary geneticists as inbreeding depression. In humans, the importance of this effect remains unclear, partly because reproduction between close relatives is both rare, and frequently associated with confounding cultural factors. To address these difficulties, we have performed a large-scale meta-analysis of more than one million individuals, from over 100 culturally diverse populations. In each individual, dense genotype data was used to identify autozygous genomic segments caused by both recent, and more distant, parental relatedness. We find that increased autozygosity is associated with apparently deleterious changes in 14 of 44 complex traits analysed, including components of reproductive success (age at first sex, age at first birth, number of opposite-sex partners, parity) and medically important traits associated with survival (birth weight, total cholesterol, lymphocyte percentage, haemoglobin concentration). As well as providing insight into the genetic architecture of these traits, our results have direct relevance to highly autozygous individuals. For example, we find that the offspring of first cousins are 1.6 [95% CI 1.4-1.9] times more likely to be childless than their outbred peers, apparently due to reduced fertility. Intriguingly, we also find that, for many traits, the effect of autozygosity is significantly greater in men than in women, suggesting a contribution of sexual selection to human evolution. We present evidence that the observed effects are caused by rare genetic variants and not by unknown environmental confounders. In particular, the effects of autozygosity are consistent across diverse demographic origins, and linear relationships are observed between autozygosity and trait means.

280

Whole-exome sequencing identifies sex-specific risk variation for Alzheimer's disease. *B.W. Kunkle[1], K.L. Hamilton-Nelson[1], A.C. Naj[2], A.B. Kuzma[2], D. Lancour[3], M. Butkiewicz[4], J. Malamon[3], Y. Ma[3], G.W. Beecham[1], W.S. Bush[4], L.S. Wang[2], R. Mayeux[5], J.L. Haines[4], L.A. Farrer[3], G.D. Schellenberg[2], M.A. Pericak-Vance[1], E.R. Martin[1], The Consortium for Alzheimer's Sequencing Analysis (CASA).* 1) University of Miami, Miami, FL; 2) University of Pennsylvania, Philadelphia, PA; 3) Boston University, Boston, MA; 4) Case-Western Reserve University, Cleveland, OH; 5) Columbia University, New York, NY.

Background: Women comprise nearly two-thirds of all Alzheimer disease (AD) cases, suggesting gender-specific risk and protective factors. For example, a number of studies have established that apolipoprotein E (*APOE*) genotype contributes to risk of AD differently in men and women, and several other candidate gene-based studies implicate sex-specific effects on AD from variants within *BDNF*, *LDLR* and *ABCA1*. To identify additional sex-specific genetic associations with AD, we analyzed the Alzheimer's Disease Sequencing Project (ADSP) case-control whole-exome sequencing dataset using a sex-stratified approach. **Methods:** Sex-interaction and sex-stratified analysis were performed on non-Hispanic white subjects in the ADSP case-control WES study (N=5,522 cases, 56.4% female; 4,919 controls, 59.1% female) to investigate genetic risk differences between males and females. Both single-variant (logistic regression) and gene-based tests (SKAT-O) were performed separately in males and females. Model 1 included adjustment for sequencing center and population structure. Additional models added adjustment for age (Model 2), and age and *APOE* genotype (Model 3). **Results:** A variant in the gene *DNAH5* was genome-wide significant ($P<3.5 \times 10^{-7}$, Model 1) for SNP-by-sex interaction ($P=3.22 \times 10^{-7}$). Several other variants show marked differences in strengths of association between males and females, including a synonymous variant in the GWAS-identified risk gene *PICALM* that increases risk for males ($P=9.26 \times 10^{-6}$, Model 0) but shows no association in females ($P=0.495$) (interaction-$P=0.005$). Additionally, gene-based testing (Models 2 and 3) discovered two significant sex-specific associations for AD in the genes *ZNF471* (male $P=0.986$; female $P=1.26 \times 10^{-6}$) and *AP4S1* (male $P=2.31 \times 10^{-7}$; female $P=0.183$). Replication of our top results is ongoing. **Conclusions:** We identified several sex-specific associations with AD. *DNAH5* is an important regulator of primary cilia, organelles important in neuronal signaling and regeneration whose dysfunction has been implicated in AD. *AP4S1*, for which recessive mutations cause the neurological disease hereditary spastic paraplegia, is part of the endocytic pathway, a known AD pathway. *ZNF471* is hypermethylated in nonagenarians. Understanding the nature of these associations could help explain differential risk and progression for AD between the sexes. .

281

SPARK: A large-scale genomic resource of over 20,000 individuals with autism spectrum disorder. *P. Feliciano, The SPARK Consortium.* Simons Foundation, New York, NY.

Although ~100 autism spectrum disorder (ASD) risk genes have been identified from studies of thousands of individuals, larger studies are required to understand the genomic architecture of ASD and to identify additional monogenic, polygenic and environmental risk factors. To accelerate clinical research in ASD, we created SPARK (Simons Foundation Powering Autism Research for Knowledge) with the goal of recruiting 50,000 individuals with ASD and their family members into a longitudinal, recontactable research cohort in which individual genetic causes of autism are returned to participants. In the first year of recruitment, we have enrolled over 20,000 individuals with ASD through a national network of 25 clinical sites and social and digital media campaigns. We performed exome sequencing and SNP genotyping on the first 500 parent-offspring trios and identified two de novo loss of function variants in *BRSK2*, supporting *BRSK2* as a candidate gene for ASD. Overall, we identified pathogenic variants in known ASD risk genes in 3.2% of individuals with ASD and returned those results to individuals using an innovative, centralized genetic counseling service. We also identified de novo loss of function or damaging missense variants in 27 genes, including *DPP6, DOCK8, MEIS2, HRAS* and *SETD1A*, providing additional evidence for a role for these genes in neurodevelopmental disorders. Exome sequencing and genome-wide genotyping for the entire SPARK cohort are underway. De-identified genetic and phenotypic data from the SPARK cohort are available to the research community at https://sfari.org/resources/autism-cohorts/spark. Access to this cohort is also available to qualified researchers at https://sfari.org/resources/sfari-base. Using an engaging, user-friendly online platform, recruitment through clinical sites and social media, active partnership with participants, and collection of saliva for DNA analysis, we developed a large-scale research platform that is efficient, cost-effective and can be applied to other human diseases. SPARK will enable research on groups of individuals sharing specific genetic causes of ASD to provide deeper insight into molecular pathogenesis and development of ASD and related neurodevelopmental disorders.

282

The MSSNG Autism Spectrum Disorder Whole Genome Sequencing Resource. *S. Walker[1], R.K.C Yuen[1,2], D. Merico[1], M. Bookman[3], J.L. Howe[1], B. Thiruvahindrapuram[1], R. Patel[1], J. Whitney[1], N. Deflaux[3], J. Bingham[3], Z. Wang[1], G. Pellecchia[1], J.A. Buchanan[1], C.R. Marshall[1,4], N. Hoang[5], S.L. Pereira[1], T. Paton[1], W. Van Etten[6], M. Szego[1,7], L.J. Strug[1,8], B.A. Fernandez[9,10], L. Zwaigenbaum[11], B.A. Knoppers[12], E. Anagnostou[13], P. Szatmari[14,15,16], W. Roberts[14], R.H. Ring[17], D. Glazer[3], M.T. Pletcher[18], S.W. Scherer[1,2,19,20].* 1) The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada; 2) Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada; 3) Google, Mountain View, California, USA; 4) Department of Molecular Genetics, Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada; 5) Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, Toronto, ON, Canada; 6) BioTeam Inc., Middleton, Massachusetts, USA; 7) Dalla Lana School of Public Health and the Department of Family and Community Medicine, University of Toronto, Toronto, Ontario, Canada; 8) Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Canada; 9) Disciplines of Genetics and Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada; 10) Provincial Medical Genetic Program, Eastern Health, St. John's, Newfoundland, Canada; 11) Department of Pediatrics, University of Alberta, Edmonton, Alberta, Canada; 12) Public Population Project in Genomics and Society, McGill University, Montreal, QC, Canada; 13) Bloorview Research Institute, University of Toronto, Toronto, Ontario, Canada; 14) Autism Research Unit, The Hospital for Sick Children, Toronto, Ontario, Canada; 15) Child Youth and Family Services, Centre for Addiction and Mental Health, Toronto, Ontario, Canada; 16) Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada; 17) Department of Pharmacology & Physiology, Drexel University College of Medicine, Philadelphia, Pennsylvania, USA; 18) Autism Speaks, Princeton, New Jersey, USA; 19) Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada; 20) McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada.

We are performing whole genome sequencing (WGS) of families with Autism Spectrum Disorder (ASD) to build a resource, named MSSNG, to enable the sub-categorization of phenotypes and underlying genetic factors involved. We have created a cloud database containing WGS data and clinical information which is accessible through an internet portal with controlled access. Currently, data from 5205 samples are available including variant calls for single nucleotide variants (SNVs) and small insertion/deletions (indels) with read alignments. From these data, we detected on average 73.8 *de novo* SNVs and 12.6 *de novo* indels per ASD subject and identified 18 new candidate ASD-risk genes, such as *MED13* and *PHF3*. In total, by including *de novo* SNVs and indels and large copy number variants (CNVs), a molecular diagnosis could be determined for 11.2% of ASD cases (Nature Neurosciences, 2017). We are currently working towards the next data release, expected to be autumn 2017, which would bring the total number of genomes available to >7500 with 3602 affected individuals, 3966 parental samples and 103 unaffected individuals. We are also anticipating an expansion of the variants directly available for each sample to include CNVs. Analysis of the CNV data from the first 5205 samples found an average of 22.7 rare (<1% frequency) CNVs >1kb in size per individual sequenced on Illumina platforms, and 7.87 rare CNVs >2kb per individual sequenced by Complete Genomics. Of these, an average of 9.89 and 4.78, respectively, impacted protein-coding regions of genes. We are also analysing structural variant calls using multiple different tools; CREST, LUMPY, Manta, Pindel and DELLY. The MSSNG phenotype database is also being expanded, and dozens of families are being added with multigenerational pedigrees, multiple affected siblings, and participants from clinical trials. Moreover, epigenetic analysis of DNA samples with data from methylation microarrays is adding additional functional data to the genomic information, and all of this is made available to the research community through a simple MSSNG user interface (portal).

283

Rare variants conferring risk for autism and ADHD identified by whole exome sequencing of dried bloodspots. *F.K. Satterstrom[1,2], F. Lescai[3,4,5], D. Demontis[3,4,5], R.K. Walters[1,2], C. Stevens[1], J. Grauholm[3,6], J.B. Maller[1,2], D.M. Hougaard[3,6], T.M. Werge[3,7,8], P.B. Mortensen[3,4,9], B.M. Neale[1,2,10], A.D. Børglum[3,4,5], M.J. Daly[1,2,10], iPSYCH-Broad Consortium.* 1) Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 2) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 3) iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus, Denmark; 4) Department of Biomedicine, Aarhus University, Aarhus, Denmark; 5) iSEQ, Centre for Integrative Sequencing, Aarhus University, Aarhus, Denmark; 6) Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark; 7) Mental Health Centre Sct. Hans, Institute for Biological Psychiatry, Capital Region of Denmark, Roskilde, Denmark; 8) Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; 9) National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark; 10) Department of Medicine, Harvard Medical School, Boston, MA, USA.

To uncover rare genetic variants conferring risk for psychiatric disorders, the iPSYCH-Broad Consortium is sequencing DNA from dried blood samples stored at the Danish Neonatal Screening Biobank and matching them to phenotypes from the Danish Psychiatric Central Registry. After quality control performed using Hail, our current dataset includes over 16,000 exomes, including 4,084 autism cases, 3,536 ADHD cases, 727 cases with both diagnoses, and 5,214 controls. Prior studies have shown that autism cases have a greater burden than controls of protein-truncating variants (PTVs) not found in the "non-psychiatric" subset of the Exome Aggregation Consortium v1 database and in genes recognizably intolerant of such mutations (pLI > 0.9). As these mutational excesses have been more pronounced in ASD cases with comorbid intellectual disability (ID) or global developmental delay (DD), and even more in cohorts ascertained for ID/DD, we focused here on cases and controls with no record of ID/DD. Considering rare "high pLI" PTVs, we observed a significant excess in autism cases (0.30/person, p=6E-14), ADHD cases (0.29/person, p=1E-10), and cases with both diagnoses (0.29/person, p=9E-04) compared to controls (0.21/person). Struck by the similar rates in autism and ADHD, we used the c-alpha test to determine if similar or distinct sets of genes were hit by rare high-pLI PTVs in the two disorders and found no significant difference (p = 0.8) between the sets of genes hit in autism and ADHD. In contrast, we found significant differences between the sets of genes hit in each disorder compared to controls (p = 2E-9 for autism; p = 7E-7 for ADHD). Furthermore, even in the set of genes previously reported as carrying a *de novo* PTV in autism trio studies, a similar and significant excess of PTVs was observed in ASD and ADHD compared to controls (p = 0.01 for autism; p = 0.009 for ADHD). These findings give the strongest evidence to date for a shared rare variant contribution to both autism and ADHD. Pooling case categories to explore single-gene burden identified genes with as many as 13 hits in cases with none in controls. Additionally, several genes identified as potential but unconfirmed hits in earlier studies (e.g., *WDFY3*, *TRIO*, *KAT2B*) have multiple case hits and none in controls, strengthening their association with autism and suggesting they may be related to ADHD as well.

284

Clinical exome sequencing as a first tier clinical diagnostic test for individuals with developmental delay and autism spectrum disorder (ASD/DD). *S.V. Mullegama[1,2], S.D. Klein[3], A.R. Lipson[3], H. Lee[1,2], S.P. Strom[1,2], E.D. Douine[3], W.W Grody[1,2,3,4], E. Vilain[2,3,4], S.F. Nelson[1,2,3,4], J.A. Martinez-Agosto[2,3,4].* 1) Pathology & Laboratory Medicine, UCLA, Los Angeles, CA; 2) UCLA Clinical Genomics Center, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA; 3) Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA; 4) Department of Pediatrics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA.

Purpose: The advent of genome-wide diagnostic tests such as chromosomal microarray (CMA) and clinical exome sequencing (CES) has vastly changed the clinical molecular diagnostic field. CMA is currently an established first-tier test for individuals with developmental disabilities (developmental delay (DD), and autism spectrum disorders (ASD)). Our study aims to establish the molecular diagnostic yield of CES in ASD/DD patients, and the role of reassessment and endophenotyping in enhancing yield. **Methods:** We performed a review and reassessment of 208 patients with ASD/DD that underwent CES at the UCLA Clinical Genomics Center during a 4-year period (January 2012-December 2016). **Results:** A genetic variant was originally reported in 111/208 (53.4%) of cases. Stratification of reported variants into variant calls revealed 8.7% pathogenic (P,18/208), 13.9% likely pathogenic (LP, 29/208), and 30.8% variants of uncertain significance (VOUS, 64/208). Due to the rapidly changing sources of genomic information, such as literature and population genetic variant databases, variant reassessment was conducted on all cases. Variant reassessment increased the molecular diagnostic yield by 8% (64/208, 31.0%). Six cases had original VOUS classifications in nonclinical genes (*AFF4, KAT6A, SLC6A1, HNRNPU, SETD5,* and *NACC1*) that were later associated with newly identified disorders and were reclassified as P/LP (6/208, 2.9%). Furthermore, we sought to identify additional phenotypes associated with the highest molecular diagnostic yields and determined that macrocephaly (38%, $P<0.05$) and microcephaly (50%, $P<0.0001$) increased the molecular yield. Finally, 20% (41/208) of our cases had variants of unknown clinical significance. Functional ranking of these genes by their connectivity to established pathogenic networks in autism enhanced their potential contribution to the observed phenotypes. **Conclusions:** CES offers a higher molecular diagnostic yield (31.0%) for individuals with ASD/DD than CMA (10-20%). Reanalysis over time that incorporates the evolving phenotype of the patient, in addition to new data from literature and variant databases, can aid in enhancing exome molecular diagnostic yield. Overall, our findings demonstrate the diagnostic utility of CES as a first tier test for ASD/DD. .

285

Methylation accurately predicts age of cancer onset in patients with Li Fraumeni syndrome. *B. Brew[1], L. Erdman[1,2], T. Guha[1,2], A. Novokmet[1], A. Dorea[1,2], J. Berman[3], A. Shlien[1,2], D. Malkin[1,2], A. Goldenberg[1,2].* 1) The Hospital for Sick Children, 555 University Ave M5G 1X8 Toronto, Ontario, Canada; 2) University of Toronto, 27 King's College Cir, Toronto, ON M5S; 3) IWK Health Centre 5980, University Ave, Halifax, NS B3K 6R8.

Introduction Li Fraumeni Syndrome (LFS) is a rare hereditary genetic cancer predisposition syndrome. LFS is characterized by germline mutations of the TP53 tumor suppressor gene and is associated with an increased risk of second tumors and a spectrum early onset cancers. We have previously developed and implemented a comprehensive life-long clinical surveillance protocol for individuals with a germline TP53 mutation. We set out to make this screening process more targeted by building a predictive model of age of onset. We accomplished this goal by implementing machine learning methods on germline methylation data. Methods We made use of the Toronto Hospital for Sick Children (SickKids) LFS family cohort in our predictive model of age of onset. In all, we have 80 patients with germline methylation data, consisting of ~450,000 probe sites. We subset this data by identifying regions of the genome that are most differentially methylated between LFS cancer patients and LFS patients without cancer. Results Our machine learning model was able to achieve 86% correlation between true and predicted values of the age of onset. Additionally, we have tested the ability of our models to predict whether an individual will be diagnosed before or after the age of 4. Our classification machine learning model on average achieved 85% accuracy. We validated our results by obtaining additional methylation samples on on 35 LFS patients with cancer and testing our model on that group. We achieve 82% correlation which is inline with our original model. Finally, We attempt to verify that our model does not simply predict the age of sample collection by using our cohort of LFS patients that do not have cancer yet (n = 33). The model has less predictive power on the age of sample collection (65% correlation), indicating that our model does a significantly better job at predicting age of onset and not simply the age of the patient. Conclusion We identify two predictive models for age of cancer onset in LFS patients that achieve high accuracy both when predicting the age of onset as a continuous variable (86% correlation) and whether cancer onset will occur before or after the age of 4, 5, and 6 years (85% accuracy on average). Our model will assist clinicians in targeting high risk patients for screening, lowering the cost of treatment, helping to avoid unnecessary screening in younger patients and raising the likelihood of survival among LFS patients. .

286

Pediatric Cancer Variant Pathogenicity Information Exchange (PeCan-PIE): A cloud-based platform for curating and classifying germline mutations in cancer-related genes. *M.N. Edmonson, A. Patel, D. Hedges, Z. Wang, E. Rampersaud, S. Newman, X. Zhou, M.C. Rusch, C. McLeod, M.R. Wilkinson, C. Pepper, S.V. Rice, J. Becksfort, K.E. Nichols, L.L. Robison, J.R. Downing, J. Zhang.* Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN.

The growing popularity of genome-wide germline genetic testing highlights a need for improved variant annotation, triage and pathogenicity classification. We developed “PeCan-PIE” – the Pediatric Cancer Variant Pathogenicity Information Exchange (https://pecan.stjude.org/pie) – a web portal that facilitates i) automated annotation, classification and triage with our “MedalCeremony” pipeline; ii) an interactive variant page for curation and committee review; and iii) a repository of expert-reviewed germline mutations that may predispose individuals to cancer. MedalCeremony provides a 3-level ranking - Gold, Silver or Bronze- of putative pathogenicity of mutations within cancer-related genes. “Medal” assignment is based on matches to 22 mutation databases, mutation type, population frequency, tumor suppressor status and predicted functional impact. The evidence used for medal assignment is imported into an interactive variant review page where an analyst can enter additional curated information such as primary diagnosis, presence of subsequent neoplasm, family history and literature. ACMG/AMP classification tags can be manually assigned to curated data enabling automated calculation of pathogenicity rating based on ACMG/AMP 2015 guidelines. PeCan-PIE is entirely cloud-based and its easy-to-use graphical interface makes the pipeline accessible to researchers regardless of their computational expertise. Users upload their variants in VCF format, run them through MedalCeremony and import the results to the variant page all within the St. Jude Cloud platform. Importantly, PeCan-PIE allows public sharing of pathogenicity classification and associated supporting evidence, thereby adding to a valuable resource for the broader cancer genetic research community. PeCan-PIE was designed based on our experience in classifying germline mutations in 1,120 patients with pediatric cancers[1]. It has been used to classify 3,423 germline mutations discovered in 3,006 childhood cancer survivors from the St. Jude Lifetime Cohort study. Currently pathogenicity classifications performed for 800 germline mutations are publicly available on PeCan-PIE[1]. This, coupled with the easy access to the MedalCeremony pipeline via the St. Jude Cloud platform, provides a powerful system for investigating mutation pathogenicity in cancer-related genes. Further, we are in the process of extending PeCan-PIE to support genetic studies of non-cancer diseases such as Amyotrophic Lateral Sclerosis.

287

***NBN* germline mutations are associated with pan-cancer susceptibility and show *in vitro* DNA damage response defects.** *S. Topka[1,2], M.F. Walsh[1,3,4], A. Maria[1,2], N. Pradhan[1,5], C. Stewart[1,5], D. Mandelker[1,6], L. Zhang[1,6], M. Berger[1,6,7], Z.K. Stadler[1,4,5], J. Petrini[8], M. Robson[1,4,5], J. Vijai[1,2], K. Offit[1,2,4,5].* 1) Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY; 2) Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, NY; 3) Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, NY; 4) Weill Cornell Medical College, New York, NY; 5) Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering, New York, NY; 6) Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY; 7) Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY; 8) Molecular Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY.

Nibrin, the protein encoded by the *NBN* gene forms a complex with Mre11 and Rad50 (MRN complex) that is crucial for DNA damage repair. Mutations in *NBN* are found in patients with Nimjegen breakage syndrome (NBS), an autosomal recessive disorder characterized by growth retardation, microcephaly, radiosensitivity, immunodeficiency and increased cancer risk. Most NBS patients harbor the common founder mutation c.657del5 that leads to expression of a hypomorphic 70kDa C-terminal fragment produced by alternative translation initiation. Cell lines derived from these patients show increased sensitivity to DNA-damaging agents, chromosome instability, and abnormal cell cycle checkpoint function. Several studies have addressed cancer incidence in individuals with germline *NBN* mutations, showing increased cancer risk for individuals harboring the c.657del5 founder mutation and for carriers of the R215W missense mutation across multiple cancer types. For other *NBN* mutations, conflicting reports exist, as to their association with cancer risk. *NBN* germline mutations were detected in an anonymized set of 9000 individuals through multi-gene panel tumor:normal sequencing (MSK-IMPACT). *NBN* mutations classified as pathogenic or likely pathogenic were shown to confer an increased overall cancer risk (OR:10.77, $p= 3.5x10^{-9}$) and a secondary analysis including carriers of variants of uncertain significance (VUS) showed a moderate increase in cancer risk (OR:1.44, $p= 4.3x10^{-7}$). Overexpression of newly identified truncating *NBN* germline mutations in an *NBN*-deficient cellular background revealed a novel C-terminal truncated fragment that can bind to Mre11. Mutant cells show reduced DNA damage repair and hence decreased overall survival. Impaired Chk2 phosphorylation was also observed, indicating cell cycle checkpoint deficiencies. Patient-derived fibroblast cells also showed decreased survival and reduced phosphorylation of NBN following y-irradiation. Further *in vitro* studies modeling these and additional germline mutations occurring in cancer patients are ongoing, in order to better understand the role of this pathway of DNA damage repair in mediating susceptibility to human malignancies.

288

The spectrum of *MYC* translocations and their effect on gene upregulation in a dataset of 527 multiple myeloma patients. *A. Mikulasova[1], C.T. Ashby[1], R.G Tytarenko[1], S. Deshpande[1], O.W. Stephens[1], E. Tian[1], P.H. Patel[1], C.P. Wardell[1], S. Roy Choudhury[1], G.H. Jackson[2], F.E. Davies[1], G.J. Morgan[1], B.A. Walker[1].* 1) Myeloma Institute, University of Arkansas for Medical Sciences, Little Rock, AR, USA; 2) Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne, United Kingdom.

Introduction: Activation of *MYC* is a late progressional event in multiple myeloma, which negatively affects patients' survival. Here, we focused on the characterization of *MYC* rearrangements and their association with *MYC* upregulation. **Materials and Methods**: In total, 527 pairs of tumor (bone marrow plasma cell separated using CD138 marker) and germline control samples from myeloma patients were analyzed using a targeted sequencing panel of 131 genes and 27 chromosome regions (n=67; KAPA HyperPlus and NimbleGen SeqCap EZ) or exome sequencing (n=460; NEBNext and Agilent SureSelect Human All Exon V5 enriched for *IGH*, *IGK*, *IGL* and *MYC* region capture). Captured region surrounding *MYC* was 2.3 Mb and 4.5 Mb, respectively**.** Normalized tumor/germline depth ratio and MANTA were used for detection of somatic copy-number and structural variants. *MYC* expression level, determined by gene expression microarrays, was available in 66/67 of patients analyzed by targeted sequencing. **Results**: *MYC* translocation was found in 25% (131/527) of patients and occurred as inter-chromosomal translocation involving two chromosomes in most cases (18%, 94/527). 5%, 1%, and 1% showed translocation between more than two chromosomes, inversion of chromosome 8, and co-occurrence of inter-chromosomal translocation and inversion, respectively. Importantly, presence of *MYC* inter- and/or intra-chromosomal rearrangement was associated with significantly higher gene expression (P<0.0001). A total of 61 partner loci of *MYC* translocations were recognized in the dataset of 527 cases; 6 partners were present more than five times and involved super-enhancers near *IGL* (5%)*, IGH* (5%)*, FAM46C* (3%)*, IGK* (2%)*, TXNDC5/BMP6* (2%), and *FOXO3* (2%) genes. We found a region of 1.7 Mb surrounding *MYC* as a translocation breakpoint hot-spot. Using copy-number analysis available for targeted-sequenced cases, we found that 83% (19/23) of translocations are unbalanced, nine of which resulted in additional copies of *MYC* and had the highest expression of *MYC* in the tested dataset. In addition, we found tandem duplication in 6% (30/527) of cases and identified a minimally duplicated region in size of 49 kb, located distally from *MYC*, out of any known gene. **Conclusions**: *MYC* region at 8q24.21 is a hot-spot for heterogeneous inter- as well intra-chromosomal rearrangements and presence of chromosomal translocation, unbalanced in most of the cases, is a driving event in *MYC* upregulation.

289

Integrative analysis of exome sequencing and gene expression data identify novel non-HLA mismatched variants associated with antibody mediated rejection in kidney transplant. *S. Pineda[1,2], T. Sigdel[2], A. Jackson[3], M. Sarwal[2], M. Sirota[1].* 1) Institute for Computational Health Sciences, Department of Pediatrics, University of California, San Francisco (UCSF), CA, USA; 2) Division of Transplant Surgery,Department of Surgery,University of California, San Francisco (UCSF), CA, USA; 3) Johns Hopkins University.

The main problem after kidney transplantation is rejection, due to donor-specific antibodies that damaged the organ by a process called antibody mediated rejection (AMR), which bind to HLA and/or non-HLA (nHLA) molecules. HLA matching is currently the only genetic test applied in donor/recipient (D/R) matching, but it does not completely account for the graft rejection as nHLA loci in the genome also influence the process of rejection. We conducted a pilot study sequencing 28 donor/recipient (D/R) pairs to examine the association between the mismatched variants by D/R and clinical endpoint (AMR (14), Cell-mediated rejection (CMR) (7), no-rejection (NoRej (7)). We found that the total number of mismatches per D/R pair increased in AMR when compared to NoRej (ANOVA - p-value = 0.02). Specific mismatched variants associated with clinical endpoint were then interrogated using Fisher's exact test and 123 variants were nominally significant (94 AMR, 25 CMR and 4 NoRej) (p-value < 0.001). We further applied a machine learning technique using variable selection with Random Forest identifying a subset of 65 variants (OOB = 0.03) showing a clean classification of all three clinical endpoints. A repeated permutation test showed that these results were robust and not artifacts with an average OOB = 0.25. We finally leveraged publically available gene expression to perform an enrichment analysis considering the variant list for AMR mapped to their harbor genes and eQTLs genes. These genes were enriched in those expressed in kidney (p-value = 0.0005), blood vessels (p-value = 0.002), immune-related genes (p-value = 0.007), and cell surface genes (p-value = $4.7*10^{-7}$) genes. Our study highlights that current sequencing methodologies can catalogue nHLA genetic differences. Our analysis has identified D/R specific mismatched variants associated with high risk of AMR. We also were able to predict the clinical endpoint with high confidence and very low error rates. We believe that these variants are functionally relevant as they relate to genes enriched in the kidney (the organ undergoing injury) and blood vessels (anatomical site most affected), are involved in immune function and more likely to be displayed on the surface of the kidney cells. While additional independent validation of these findings is needed, selection of a minimal nHLA variant list that can be added to current HLA testing to enhance our ability to predict AMR even before engraftment of the organ.

290

Leveraging molecular QTL to understand the genetic architecture of diseases and complex traits. *F. Hormozdiari[1,2], S. Gazal[1,2], B. Geijn[1,2], H. Finucane[1,2], C. Ju[3], P. Loh[1,2], X. Liu[1,2], L. O'Connor[1,2], A. Gusev[4,5], E. Eskin[3], A. Price[1,2,6].* 1) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 3) Department of Computer Science, University of California, Los Angeles, California 90095, USA; 4) Dana Farber Cancer Institute, Harvard Medical School; 5) Brigham & Women's Hospital Division of Genetics; 6) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston,.

There is increasing evidence that many GWAS risk loci are molecular QTL for gene expression (eQTL), histone modification (hQTL) and/or DNA methylation (meQTL). Our goal is to identify molecular QTL-based functional annotations that are maximally enriched for disease heritability to help us understand the genetic architecture of diseases. We consider 3 annotations: the set of all significant cis variants (AllCis), the set of top significant cis variants per gene/peak (TopCis), and a continuous-valued annotation equal to the causal posterior probability (CPP) of variants in the 95% credible set, maximized across genes/peaks (MaxCPP); CPP are computed using the CAVIAR fine-mapping method (Hormozdiari et al. 2014 Genetics). We evaluate the functional enrichment of each annotation using stratified LD score regression (S-LDSC; Finucane et al. 2015 Nat Genet). We consider two metrics: enrichment = (% of heritability)/(% of SNPs) and = proportionate change in per-SNP heritability associated to a 1 s.d. change in the annotation value conditional on other annotations in the model. In simulations involving UKBiobank genotypes and simulated phenotypes, we observed severe upward bias for TopCis (because variants linked to the top variant contain excess causal signal, violating S-LDSC model assumptions) but conservative estimates for AllCis and MaxCPP. We used molecular QTL data from the GTEx and BLUEPRINT consortia to build these annotations and analyzed summary association statistics from 20 independent diseases and complex traits (average N=98K); we meta-analyzed all results across 20 traits. We reached 4 main conclusions. First, MaxCPP is far more enriched than AllCis, e.g. for GTEx whole blood: enrichment=4.2x (P=6e-13), 2.1 (P=1e-13) vs. 2.8x, 0.55 for AllCis; all results below are for MaxCPP. Second, annotations based on a meta-analysis of all GTEx tissues performed far better: Enrichment=5.8x (P=5e-44), 3.0 (P=3e-47). Third, this novel signal remains highly significant even after conditioning on 73 functional annotations from the baselineLD model (Gazal et al. biorxiv): 0.57 (P=2e-24). Fourth, in a joint analysis of BLUEPRINT eQTL/hQTL/meQTL annotations with the baselineLD model, both eQTL and H3K27ac-hQTL were statically significant: = 0.30 (P=5e-08) and = 0.20 (P=0.003). In conclusion, our results indicate that eQTL and hQTL are both highly informative for disease heritability. Our new QTL annotations can be used to improve power of association testing.

291

Integrating transcriptome sequencing from Mendelian disease patients and healthy controls enhances genetic variant interpretation. *B. Cummings[1,2], J.L. Marshall[1,2], K.J. Karczewski[1,2], F. Zhao[1,2], B. Weisburd[1,2], The. GTEX Consortium[hA], S. Donkervoort[3], L. Waddell[4], S. Sandaradura[4], G. O'Grady[4], E. Oates[4], J. Dowling[5], S.T. Cooper[4], C. Bonnemann[3], D.G. MacArthur[1,2].* 1) Medical and Population Genetics, Broad Institute, Cambdridge, MA; 2) 1.Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA. USA; 3) Neuromuscular and Neurogenetic Disorders of Childhood Section, Neurogenetics Branch, National Institute of Neurological Disorders and Stroke/National Institutes of Health, Bethesda, Maryland, USA; 4) Institute for Neuroscience and Muscle Research, Kids Research Institute, The Children's Hospital at Westmead, Sydney, Australia; 5) Division of Neurology, Hospital for Sick Children, Toronto, Ontario, Canada.

Exome sequencing is a powerful and cost-effective tool that has become increasingly routine in gene discovery for both common and Mendelian disorders. However, the current diagnosis rate for exome analysis across a variety of rare diseases is approximately 25-50% and variant prioritization in common disease studies remains challenging. One promising approach to increase the discovery of disease-causing variants is RNA sequencing (RNA-seq), which provides direct insight into the splicing and expression of disease genes in samples from both population controls and disease patients. Here we describe an integrated approach of patient tissue RNA sequencing in over 150 exome-unsolved patients with a variety of Mendelian disorders including Mendelian neuromuscular, neurodevelopmental, and kidney disorders. We describe an analysis framework focused on the detection of transcript level changes that are unique to the patient compared to a collection of over 150 tissue-matched controls. We demonstrate the power of RNA sequencing to validate candidate splice-disrupting mutations and to identify splice-altering variants in both exonic and deep intronic regions, yielding an increased diagnosis rate in our patients and identifying novel candidate Mendelian disease genes. We describe best practices for applying RNA-seq to patient cohorts, including tissue selection and technical parameter choices, and report a diagnosis rate of over 30% from transcriptome analysis of a range of exome-unsolved cases. Finally, using available transcriptome data from population controls across over 50 tissue types, we report the development of a base-level expression metric that improves the evaluation of candidate variants identified in disease studies, even for patients on whom no direct transcriptome data are available. We show that an important error mode in variant interpretation involves variants on transcripts with no evidence of expression across surveyed adult tissues. Together, this project expands the use of RNA-seq as a tool to interpret human genetic variation at base-level resolution.

292

Transcriptomic analysis of CD4+, CD8+, and CD14+ cells in newly diagnosed multiple sclerosis patients. *K. Kim[1], E.L. Eggers[1], S.J. Caillier[1], S.L. Hauser[1], J.R. Oksenberg[1], S.E. Baranzini[1,2].* 1) Department of Neurology, Weill Institute for Neurosciences, University of California San Francisco (UCSF), San Francisco, CA; 2) Institute for Human Genetics, University of California San Francisco (UCSF), San Francisco, CA.

Multiple Sclerosis (MS) is an autoimmune condition of the central nervous system characterized by demyelination and neurodegeneration. The exact cause of MS remains unknown and there is no cure. Many MS-associated genes are primarily expressed in immune cells such as T-cells and dendritic cells. Gene expression profiles from whole tissue or peripheral blood mononuclear cells have been reported, however, they consist of many different cell types. Therefore, cell-type specific gene expression can be more informative. Total RNA was purified from $CD3^+CD19^-CD4^+$ and $CD3^+CD19^-CD8^+$ T-cells and $CD14^+$ monocytes collected from MS patients (n=75) participating in the UCSF MS EPIC/ORIGINS studies. 3' mRNA-seq was performed on cell subsets to test for differentially expressed genes (DEGs) between disease courses (RR, PP, CIS) and disease-modifying treatments (DMTs: glatiramer acetate, dimethyl fumarate, natalizumab, fingolimod). DEG analysis was performed using DESeq2. There were 5,198 DEGs when comparing T-cells and monocytes of MS patients, and 172 DEGs between $CD4^+$ and $CD8^+$ T-cells (FDR < 0.05, Log2FC >< ±1.5, baseMean > 1). A disease course comparison in treatment naïve patients further identified DEGs (7 genes in $CD4^+$, 15 in $CD8^+$, and 2 in $CD14^+$ between PPMS and RRMS; 8 in $CD4^+$, 14 in $CD8^+$, 2 in $CD14^+$ between RRMS and CIS; 12 in $CD4^+$, 54 in $CD8^+$, 10 in $CD14^+$ between PPMS and CIS). While most of all the DEGs were down-regulated in PPMS with respect to RRMS and CIS, in turn, RRMS patients showed mainly up-regulated transcripts when compared to CIS. Specifically, *EIF2S3L* and *SNORD8* were significantly regulated in both $CD4^+$ and $CD8^+$ T-cells in RRMS when compared to PPMS or CIS. In addition, *CHI3L2* was down-regulated in PPMS compared to CIS in $CD4^+$. *CHI3L2* is in the same family as *CHI3L1* which has been suggested as a biomarker of MS. When comparing treatment-naïve patients to those on DMTs, *HLA-DQB1* and *PRSS21* were significantly decreased in $CD14^+$ in patients treated with dimethyl fumarate. In contrast, 38 transcripts including those for immune-related genes *S100B* and *MS4A1* were differentially expressed in $CD8^+$ of patients treated with fingolimod. We analyzed gene expression in T-cells and monocytes from newly diagnosed MS patients and observed cell-specific transcriptional changes in both untreated patients and in response to different DMTs. These findings provide important insights into cell-specific gene expression changes in subtypes of MS and in response to DMTs.

293

The contribution of rare variants, polygenic risk, and novel candidate genes to the hereditary risk of breast cancer in a large cohort of breast cancer families. *N. Li[1], S. Rowley[1], D. Goode[1], L. Devereux[2], S. McInerny[3], N. Grewal[3], A. Lee[1], M. Wong-Brown[4], R. Scott[4], A. Trainer[3], K. Gorringe[1], P. James[3], I. Campbell[1].* 1) Research Division, Peter MacCallum Cancer Ctr., Melbourne, Australia; 2) Lifepool, Peter MacCallum Cancer Ctr., Melbourne, Australia; 3) Parkville Familial Cancer Ctr., Peter MacCallum Cancer Ctr., Melbourne, Australia; 4) Division of Genetics, Hunter Area Pathology Service, Newcastle, Australia.

Background: Identifying the missing hereditary factors underlying the familial risk of breast cancer could have a major and immediate impact on managing the breast cancer risk for these families. Methods: We identified candidate breast cancer predisposition genes through whole exome sequencing of BRCAx families and subsequently sequenced up to 1325 genes, along with 76 common low penetrance variants associated with breast cancer, in index cases from 6,000 BRCAx families and 6,000 cancer free women (ethnically matched on principal component analysis). Results: The role of recently described (*PALB2*) or suspected (*MRE11A*) moderately penetrant genes was confirmed. Conversely, the size of the cohort means that the absence of enrichment for loss of function (LoF) mutations provides strong evidence against other reported breast cancer genes (*BRIP1, RINT1, RECQL*). For further moderate risk variants (in *CHEK2, ATM, BRCA2*) we observed significant risk modification based on the polygenic risk score (PRS - calculated from the common variant data), with the risk restricted to the co-occurrence of the rare variant and high PRS. Novel candidate genes were identified based on LoF mutations, including *NTHL1* (OR 2.5, p=0.002): a member of the base excision repair (BER) pathway. DNA sequencing of the breast carcinomas from 17 heterozygous *NTHL1* mutation carriers revealed a strong bias towards a C:G>T:A (C>T) transitions, consistent with a BER defect, which confirmed the recent findings in colorectal carcinomas from bi-allelic *NTHL1* mutation carriers. This data extends the cancer predisposition phenotype of *NTHL1* to heterozygous carriers. In addition to *NTHL1,* there are a large number of candidate genes where the ratio of LoF mutations in cases versus controls indicates that they may convey an actionable level of risk; 46 genes (519 families) meet the basic criteria of multiple LoF variants and an OR >2 for cases versus controls – including previously proposed breast cancer genes *MRE11A*, *BLM, MLH1, MYH, FANCD2* and functionally plausible candidates such as *MLH3, PARP2* and *ATR.* Collectively the OR of breast cancer for LoF mutations in this group of genes is 3.3 (95% CI 2.7-3.9, P=3.5x10-41). Conclusion: Our data shows that the effect of rare variation in established and novel breast cancer genes, along with consideration of the background polygenic risk, together explains a substantial component of the heritable risk of breast cancer in our cohort.

294

The Model Organisms Screening Center for the Undiagnosed Diseases Network. *M.F. Wangler[1,2,3], S. Yamamoto[1,2,3,4], M. Westerfield[5], J. Postlethwait[6], H. Bellen[1,2,3,4,6].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX., USA; 2) Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston TX, 77030, USA; 3) Program in Developmental Biology, BCM, Houston TX, 77030, USA; 4) Department of Neuroscience, BCM, Houston TX, 77030, USA; 5) Institute of Neuroscience, University of Oregon, Eugene OR 97403-1254; 6) Howard Hughes Medical Institute, BCM, Houston, TX, 77030, USA.

Efforts to diagnose patients with rare undiagnosed diseases increasingly involve the use of next-generation sequencing methods to identify potential causative variants, genes, and pathways involved. These efforts are limited by a lack of knowledge of gene function and an inability to predict the impact of genetic variation on encoded protein functions reliably. Diagnostic challenges posed by undiagnosed diseases have solutions in model organism research because model organism researchers provide the medical research community with a wealth of detailed biological information. The Undiagnosed Diseases Network (UDN) is a multi-center effort to solve the most challenging medical mysteries with advanced medical technology. As part of this effort, we established a Model Organisms Screening Center (MOSC) in which genes and variants identified by whole-exome sequencing or whole-genome sequencing in patients enrolled in the UDN are studied systematically in Drosophila and/or zebrafish models. Since launching in 2016, the MOSC has analyzed 125 variants in 85 genes from 53 patients. The MOSC developed a publically accessible informatics tool, MARRVEL, to aid in prioritizing variants to be studied experimentally. Based on this information, the MOSC designed 32 Drosophila models and 12 zebrafish mutant models and are characterizing phenotypes. Recent successful variant validations in which model organisms have aided in diagnosis of the UDN case have included cases with global developmental delay and cerebellar degeneration due to a unique gain-of-function variant in *CACNA1A* that leads to Drosophila photoreceptor degeneration. Other validated examples include *EBF3*, *COG4* and *ATP5D*. The models for these specific patients have offered not only diagnosis but novel insights into disease biology, for example the use of calcium channel blockers to modulate the specific patient defect in the *CACNA1A* protein. Systematic variant validation using model organisms is an effective strategy in human genomics. .

295

Modeling and therapeutic testing of leukodystrophies in 3D human cortical spheroids. *Z. Nevin, M. Madhavan, E. Shick, K. Allan, P. Tesar.* Case Western Reserve University School of Medicine, Cleveland Heights, OH.

Pelizaeus-Merzbacher Disease (PMD, MIM 312080) is an X-linked leukodystrophy caused by mutations in proteolipid protein 1 (*PLP1*). Hundreds of mutations have been identified in patients, who range from mild symptoms of motor delay to severe spasticity and early mortality. We have previously generated PMD induced pluripotent stem cells and oligodendrocytes from a panel of 12 patients using a 2D culture system and demonstrated both distinct and convergent cellular phenotypes in patients with various mutations. However, this system does not capture the complex environment of the brain. Human cortical spheroids provide a 3D system in which to examine development, self-organization, and cellular interactions within the cerebral cortex. We have developed the first platform to generate oligodendrocytes in cortical spheroids. Spheroids derived from patients with a deletion, duplication, or missense mutation (c.254T>G) of *PLP1* display nuances in oligodendrocyte development not appreciated in 2D. PLP1 is the most abundant protein in myelin, yet patients with complete deletion present with mild symptoms. Concordantly, deletion spheroids produced abundant PLP1-negative oligodendrocytes, but with normal morphology. In 2D, missense mutation oligodendrocytes demonstrated frank perinuclear retention of PLP1, and this was recapitulated in spheroids ($p < 0.01$). Moreover, treatment with GSK2656157, an inhibitor of the endoplasmic reticulum stress response, significantly improved mobilization of PLP1 away from the nucleus and into oligodendrocyte processes ($p < 0.05$). Conversely, duplication oligodendrocytes had also demonstrated perinuclear retention in 2D, but did not in the 3D system. Rather, duplication spheroids showed increase in total PLP1 signal, but a significant decrease in the mature oligodendrocyte marker MBP ($p < 0.01$), suggesting precocious but incomplete differentiation. Despite the absence of perinuclear retention, treatment of duplication spheroids with GSK2656157 decreased total PLP1 signal and significantly increased MBP ($p < 0.03$). Lastly, CRISPR/Cas9 correction in the duplication line returned both PLP1 and MBP ($p = 0.005$) signals to wild type levels. This 3D, multi-lineage system provides a versatile platform to observe and perturb the complex cellular interactions that occur in the brain and offers new opportunities for modeling myelin diseases and testing therapeutics in human tissue.

296

Direct assessment of unexpectedly abundant paternal sperm mosaicism allows for the quantification of recurrence risk in autism. *M.W. Breuss[1,2], M. Kleiber[4,5,6], R.D. George[1,2,3], D. Antaki[4,5,6], K.N. James[1,2,3], L.L. Ball[1,2,3], O. Hong[4,5,6], D. Musaev[1,2], A. Nguyen[1,2], O. Devinsky[7], J. Sebat[4,5,6], J.G. Gleeson[1,2,3].* 1) Neurosciences, University of California, San Diego, La Jolla, CA; 2) Rady Children's Institute for Genomic Medicine, San Diego, CA; 3) Howard Hughes Medical Institute; 4) Beyster Center for Genomics of Psychiatric Diseases, University of California, San Diego, La Jolla, CA; 5) Department of Psychiatry, University of California, San Diego, La Jolla, CA; 6) Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA; 7) New York University School of Medicine, New York, New York.

Various sporadic human diseases, ranging from autism spectrum disorders to congenital heart disease and muscular dystrophies, are caused by *de novo* mutations. In a classical model, these are assumed to occur at a low rate in the parental germ cells (10^{-7}-10^{-8}). Consequently, *de novo* mutations identified by genetic testing are often assigned a low risk of recurrence in siblings. This idea is increasingly challenged by the detection of mosaicism in the parents. However, previous studies were largely restricted to the analysis of somatic tissues, whose genetic information is, by definition, not transmitted to the next generation. Here, we directly assessed the presence of inherited "*de novo*" mutations in paternal sperm and discovered abundant, germline restricted mosaicism. These samples were collected from a panel of fourteen families with a proband presenting with autism spectrum disorder. For all of these a candidate *de novo* mutation had been identified in our ongoing genetic studies of this disorder. Employing digital droplet PCR, the causative variants were detectable in 4 sperm samples, but virtually absent or drastically reduced in the somatic tissue for 3. The latter mutations were present a high allelic fractions (AF), comprising SNVs in *NR2F1* (AF=8%) and *GRIN2A* (AF=15%), as well as a large deletion of *CACNG2* (AF=10%). As a consequence, the *GRIN2A* variant, despite being undetectable in the father by classical genetic testing, was inherited by three siblings presenting with phenotypes consistent with this mutation. We next used deep whole genome sequencing (90x) of matched sperm and blood samples of four fathers to test for germline mosaicism of all *de novo* variants present in the offspring. 5-10% of these mutations were detectable in the paternal sperm, half of which were absent or at very low levels (<2%) in the matched blood. These data, together with an unbiased analysis employing mosaic variant detection algorithms, suggest that germline-specific or germline-enriched mosaicism is currently underestimated. This information has important potential implications for clinical practice. Based on our results, genetic analysis of sperm has the potential to quantify individualized recurrence risks for affected families, but could also have predictive value for prospective fathers.

297

Expression quantitative trait loci of primary melanocytes facilitate identification of melanoma susceptibility genes. *T. Zhang[1], J. Choi[1], M. Kovacs[1], S. Loftus[2], M. Xu[1], W. Pavan[2], K. Brown[1].* 1) Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Gaithersburg, MD 20877; 2) Genetic Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD,20892.

Trait-associated common variants are often enriched with expression quantitative trait loci (eQTL) suggesting *cis*-regulation as a main mechanism. While there are a number of publicly-available eQTL resources to aid in dissecting the link between common variants and gene expression, the majority are derived from tissues and do not solely reflect gene expression in specific cell types. To facilitate a better understanding of the *cis*-regulatory landscape of human melanocytes, we collected primary cultures of melanocyte from 106 newborn males, measured transcript levels via RNA sequencing, and assessed genotypes by direct genotyping and imputation. Following standard methods adopted by The Genotype-Tissue Expression (GTEx) project, we identified 4,997 eGenes and 597,335 *cis*-eQTLs (FDR<0.05) including key genes in melanin synthesis pathway. Comparison of these data with GTEx data derived from 44 tissue types suggested that melanocyte eQTLs differ considerably from most tissues. Remarkably, despite the comparatively small sample size, melanocytes ranked in the top 30% GTEx tissues for number of eGenes, higher than any tissue type with similar sample size. Coding SNPs displaying allelic imbalance corroborate melanocyte eGenes, and melanocyte *cis*-regulatory signatures from ENCODE are enriched with melanocyte eQTL. We applied these data to interrogate the 20 known GWAS loci associated with melanoma risk and found that GWAS variants are significantly enriched with melanocyte eQTLs. While GTEx skin tissue collections found significant eQTLs in only six loci, eleven of twenty known melanoma loci overlap melanocyte eQTLs, including *MAFF* (chr22q13.1, *P* = 3.2e-20), *MX2* (chr21q22,3, *P* = 3.47e-15), and *CASP8* (chr2q33-34, *P* = 4.43e-9), as well as three loci that are private to melanocytes - *SLC45A2* (chr5p13.2, *P* = 2.62e-6), *TYR* (chr11q14-21, *P* = 1.30e-4), and *OCA2* (chr15q13.1, *P* = 8.72e-8). Lastly, in order to identify novel melanoma risk gene candidates, we used these melanocyte eQTL data to perform a transcriptome-wide association study (TWAS) using summary statistic data from the largest published melanoma GWAS. Beyond genes at known loci, TWAS uncovered four new melanoma-associated genes (*HEBP1, MSC, CBWD1*, and *GPRC5A*) at genome-wide significant *P*-values. Our data highlights the utility of lineage-specific eQTL resources for annotating GWAS results and represent a robust database of significant utility for genomic research of melanoma risk and melanocyte biology.

298

The landscape of chromatin activity in renal cell carcinoma reveals thousands of germline regulatory variants with somatic interactions. *A. Gusev, M. Freedman.* Dana-Farber Cancer Institute, Boston, MA.

Functional changes as part of tumorigenesis can provide key insights into the mechanisms of cancer. In this work we investigated somatic and germline regulatory features in renal cell carcinoma (RCC) using H3K27ac ChIP-seq data in 10 matched tumor/normal samples and RNA-seq data from 496:66 tumor:normal samples. Unsupervised clustering of H3k27ac activity cleanly separated tumor from normal samples, highlighting extensive epigenetic changes that occur during tumorigenesis. The H3K27ac signal was localized to 101,397 peaks (not overlapping promoters), of which 9,491 were significantly more active in tumors (FDR<1%). Consistent with their role in increasing transcription, tumor-specific enhancers overlapped genes with significantly higher tumor-specific expression; likewise, tumor-specific enhancers overlapped methylation array probes (which are typically repressive) with significantly lower tumor-specific activity. We identified 3,747 super-enhancers, of which 6 were recurrent in all tumors and none of the normal samples. These tumor-specific super-enhancers overlapped known and suspected RCC oncogenes such as *EGLN3* and *BHLHE41*, implicating specific cis regulatory elements. We developed a novel method to test each peak (containing a heterozygous SNP) for allelic imbalance in binding (asbQTL) and applied this method to evaluate tumor/normal differences in allelic imbalance (d-asbQTL) while accounting for local structural variation. This represents a novel, functional approach to identify germline variants that interact with the somatic environment. We identified 1,356 unique asbQTL peaks in normal, 2,868 in tumor, and 1,054 d-asbQTLs (primarily imbalanced in tumor) at 5% FDR. This abundance of d-asbQTLs highlights the matched tumor/normal study design as a unique opportunity to identify putative cancer mechanisms that could not be observed in either study alone. d-asbQTL peaks harbored significantly more recurrent somatic mutations, indicative of positive selection and putative driver mechanisms. Intersecting significant H3K27ac asbQTLs with aseQTLs from RCC RNA-seq increased power to detect QTLs and dramatically reduced the number of putative causal variants. We are now integrating these regulatory variants with RCC GWAS data to infer specific cancer risk mechanisms. The overwhelming majority of somatic and germline variation is non-coding, and our findings showcase a powerful approach to understanding their mechanistic relationship to cancer.

299

Prognostic signature from DNA methylation and corresponding gene expression predicts survival of oral squamous cell carcinoma. *S. Shen[1,2,3], G. Wang[4], Q. Shi[5], R. Zhang[2,3], Y. Zhao[2,3], D. Christiani[1,3], Y. Wei[2,3], F. Chen[2,3,5].* 1) Harvard School of Public Health, Boston, MA; 2) Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, China; 3) China International Cooperation Center of Environment and Human Health, Nanjing Medical University, Nanjing, China; 4) National Health and Family Planning Commission Contraceptives Adverse Reaction Surveillance Center, Jiangsu Institute of Planned Parenthood Research, China; 5) Ministry of Education Key Laboratory for Modern Toxicology, School of Public Health, Nanjing Medical University, Nanjing, China.

DNA methylation has started a recent revolution in genomics biology by identifying key biomarkers for multiple cancers, including oral squamous cell carcinoma (OSCC), the most common head and neck squamous cell carcinoma. Further identification of methylation-specific gene expression changes hold significant potential to improve early diagnosis and survival prediction of OSCC. To identify methylation signatures specific for OSCC, we used a multi-stage screening strategy to develop a DNA-methylation-based prognostic score for overall survival. We used The Cancer Genome Atlas (TCGA) as a training set and further validated results in two independent datasets from Gene Expression Omnibus (GEO). We identified 7 CpG sites and used them to calculate a prognostic score that successfully distinguishes overall survival of OSCC patients and has a moderate ability to predict survival [training set: hazard ratio (HR) = 3.23, $P = 5.52\times10^{-10}$, area under the curve (AUC) = 0.76; validation set 1: HR = 2.79, $P = 0.010$, AUC = 0.67; validation set 2: HR = 3.69, $P = 0.011$, AUC = 0.66]. The signature was both significant in HPV+ and HPV- cases. Expression of genes corresponding to candidate CpG sites (*AJAP1*, *SHANK2*, *FOXA2*, *MT1A*, *ZNF570*, *HOXC4,* and *HOXB4*) also was significantly associated with OSCC patient survival. Integration of DNA methylation and expression provided the best prognostic prediction (AUC = 0.78), suggesting that our DNA methylation- and expression-based score is reliable and practical to predict OSCC prognosis.

300

Epigenetic modifications of innate immunity genes impact early-stage non-small cell lung cancer survival: An integrative analysis of epigenome and transcriptome in Caucasian population. *R. Zhang[1,2,3], Y. Wei[2,3], Y. Guo[1,4], E. Loehrer[1], Z. Liu[4], L. Liang[4], A. Shafer[5], Z. Wang[1], P. Tejera[1], S. Shen[1,2,3], S. Salama[6], T. Fleischer[7], M. Bjaanæs[7], A. Karlsson[8], M. Planck[8], F. Chen[2,3], J. Staaf[8], A. Helland[7], M. Esteller[5], X. Lin[4], D. Christiani[1,3,5].* 1) Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, USA; 2) Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, China; 3) China International Cooperation Center for Environment and Human Health, School of Public Health, Nanjing Medical University, Nanjing, China; 4) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, USA; 5) Pulmonary and Critical Care Division, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, USA; 6) Bellvitge Biomedical Research Institute and University of Barcelona and Institucio Catalana de Recerca i Estudis Avançats, Barcelona, Spain; 7) Department of Genetics, Institute for Cancer Research, Oslo University Hospital - The Norwegian Radium Hospital, Oslo, Norway; 8) Division of Oncology and Pathology, Department of Clinical Sciences Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden.

Background: Non-small cell lung cancer (NSCLC) has poor survival and exhibits histological heterogeneity requiring individualized therapy. Epigenome-wide association studies (EWAS) in Caucasians have attempted to identify methylation biomarkers associated with relapse-free survival (RFS), or progression-free survival (PFS), or methylation subgroups associated with RFS and overall survival (OS). However, due to different study outcomes or aims, results are inconsistent across these studies. Moreover, there is no study focusing on early-stage patients exclusively for histology-specific prognostic probes, though histology affects individual treatments in NSCLC. **Methods**: We conducted a two-stage epigenome-wide association study of tumor DNA methylation and OS to investigate potential histology-specific prognostic methylation biomarkers for early-stage NSCLC patients. The discovery phase included 614 lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) patients from seven countries. Histology-specific CpG probes were tested using a multivariate Cox proportional hazards model. For validation, probes with false discover rate (*q*-FDR) ≤ 0.05 were tested in 617 early-stage LUAD and LUSC patients from The Cancer Genome Atlas (TCGA). Significant probes that passed sensitivity and stratified analyses were further integrated with gene expression data to perform genome-wide methylation transcription analysis and casual mediation analysis. **Results**: We identified 31 LUAD- and 27 LUSC-specific prognostic CpG probes. Taken together, these probes significantly discriminated survival time of LUAD ($P = 4.52\times10^{-18}$) and LUSC ($P = 3.80\times10^{-21}$) and increased the prediction accuracy by 10.04% and 34.25%, respectively. Moreover, 19 LUAD- and 13 LUSC-specific probes are involved in 3,896 significant “methylation-gene expression-survival” pathways, regulating multiple genes including: *GSDMD*, *ICA1L*, *CACNA2D2*, *IRX5*, *MTURN*, *CISH*, *DUOXA1*, *MT1L*, *LRRC37A3*, *PDCD1LG2*, *ZNF750* and *NPTXR*. Majority of these genes are involved in innate immunity pathways. The proportion of effect mediated through a hub gene was high up to 55.53%, indicated gene expressions were important mediators for these cis- or trans-acting prognostic CpG probes. **Conclusion**: These histology-specific prognostic cis- and trans-acting CpG probes may be potential targets of epigenetic drug in NSCLC immunotherapy, but functional studies are warranted.

301

An epigenetic switch confers pleiotropic risk for bone mineral density and hyperglycaemia. *N.A. Sinnott-Armstrong[1,2], I.S. Sousa[1,3], E. Rendina-Ruedy[4], R. Sallari[1], X. Chen[5,6], S.E. Nitter Dankel[7], G. Mellgren[7], A. Guntur[4], D. Karasik[5,8], H. Hauner[3], C. Rosen[4], D.P. Kiel[5,6], Y.-H. Hsu[5,6], M. Claussnitzer[1,3,5,6].* 1) Broad Institute, Cambridge, MA; 2) Department of Genetics, Stanford University, Stanford, CA; 3) Else Kröner-Fresenius Center for Nutritional Medicine, Technical University Munich, Munich, Germany; 4) Center for Molecular Medicine, Maine Medical Center Research Institute, Scarborough, ME; 5) Institute for Aging Research, Hebrew SeniorLife and Harvard Medical School, Boston, MA; 6) Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA; 7) University of Bergen, Bergen, Norway; 8) Faculty of Medicine of the Galilee, Bar-Ilan University, Safed, Israel.

Studies suggest shared etiologies of skeletal and glycemic traits, but the underlying genetic factors are unknown. We performed a bivariate genome wide association study (GWAS) using summary statistics from GWAS meta-analyses of bone mineral density (BMD) and serum fasting glucose levels. At the *3q21.1* locus we found the strongest bivariate association for BMD and fasting glucose levels (negative correlation; p=1.9e-9) and the variant, rs56371916, where a T-to-C polymorphism (14% frequency in Europeans) was found to be causal for hyperglycemia (T allele) and low BMD (C allele). The *3q21.1* locus was enriched in allele specific DNaseI hypersensitivity; furthermore, it is repressed in progenitor cells for primary adipocytes and osteoblasts (2.1-fold, p=0.002) but contains an enhancer which activates during differentiation. De-repression is mediated by a conserved SREBF motif only present in individuals with the major T allele. Motif loss results in failed de-repression and downregulation of *ADCY5* in both osteoblasts and adipocytes (2.7-fold, p=0.0007 and 2.6-fold, p=0.0007, respectively). Network analysis revealed genes involved in lipid oxidation and osteoblast differentiation pathways (p=4.8e-6). In patient-derived osteoblasts, *ADCY5* downregulation leads to a cell-autonomous perturbation of fatty acid oxidation during early differentiation events (3.6-fold, p=0.003) and failure to differentiate (2.9-fold change in alkaline phosphatase activity, p=0.0014), which was further supported through murine osteoblast expression profiling. These data demonstrated that lipid oxidation during early differentiation, mediated by *ADCY5*, is an important factor in osteoblast development. In primary adipocytes, we demonstrated a decreased adrenergic lipolysis rate (1.9-fold, p=0.0012) for the minor allele, consistent with a lower serum glucose. CRISPR C-to-T editing of rs56371916 in patient derived cells restored *ADCY5* activation by *SREBF1*, restored osteoblast differentiation, and accelerated lipolysis rate and glycerol release in adipocytes. Overall, our bivariate GWAS analysis identified a pleiotropic risk locus that acts through *ADCY5*. We have shown that rs56371916 alters lineage-specific Polycomb de-repression, forming the basis for the genetic correlation between low BMD and glucose levels. This may explain why individuals with diabetes have higher BMD, as we find that the same fatty acid oxidation pathways contribute to both phenotypes.

302

Rare *GREB1L* mutations contribute to the genetic heterogeneity of congenital kidney malformations. *K. Khan[1], S. Sanna-Cherchi[2], R. Westland[2,3], P. Krithivasan[2], F. Lorraine[1], H.M. Rasouly[2,4], I.L. Pediaditakis[1], I. Ionita-Laza[5], D.A. Fasel[2], K. Kiryluk[2], M. Bodria[2,6,7], E.A. Otto[8], M.G. Sampson[9], C.E. Gillies[9], A. Mitrotti[2], L. Gesualdo[10], V. Tasic[11], A. Latos-Bielenska[12], G.S. Makar[2], F. Hildebrandt[13], J.V. Wijk[3], M. Saraga[14,15], F. Scolari[16], D.B. Goldstein[17], G.M. Ghigger[18], A. Materna-Kiryluk[12], R.P. Lifton[19], N. Katsanis[1], E.E. Davis[1], A.G. Gharavi[2].* 1) Center for Human Disease Modeling, Duke University, Durham, NC; 2) Division of Nephrology, Columbia University, New York, New York, USA; 3) Department of Pediatric Nephrology, VU University Medical Center, Amsterdam, The Netherlands; 4) Renal Section, Department of Medicine, Boston University Medical Center, Boston, Massachusetts, USA; 5) Department of Biostatistics, Columbia University, New York, NY 10032, USA; 6) Division of Nephrology, Dialysis and Transplantation, IRCCS Giannina Gaslini, Genoa, Italy; 7) Department of Clinical and Experimental Medicine, University of Parma, Parma, Italy; 8) Department of Internal Medicine, University of Michigan Medical, School, Ann Arbor, MI, USA; 9) Department of Pediatrics-Nephrology University of Michigan School of Medicine, Ann Arbor, MI; 10) Section of Nephrology, Department of Emergency and Organ Transplantation, University of Bari, Bari, Italy; 11) University Children's Hospital, Medical Faculty of Skopje, Skopje, Macedonia; 12) Department of Medical Genetics, Poznan University of Medical Sciences, and NZOZ Center for Medical Genetics GENESIS, Poznan, Poland; 13) Department of Medicine, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA; 14) Department of Anatomy, Histology, and Embryology, School of Medicine, University of Split, Split, Croatia; 15) Department of Pediatrics, University Hospital of Split, Croatia; 16) 2Cattedra di Nefrologia, Università di Brescia, Seconda Divisione di Nefrologia Azienda Ospedaliera Spedali Civili di Brescia Presidio di Montichiari, Brescia, Italy; 17) Institute for Genomic Medicine, Columbia University Medical Center, United States; 18) Division of Nephrology, Dialysis, Transplantation, and Laboratory on Pathophysiology of Uremia, Istituto G. Gaslini, Genoa, Italy; 19) Department of Genetics, Howard Hughes Medical Institute, and Yale Center for Mendelian Genomics, Yale University, New Haven, CT.

Renal agenesis and hypodysplasia (RHD) are a major cause of pediatric end-stage renal disease. We conducted whole exome sequencing in 203 patients with RHD and identified diagnostic pathogenic mutations in 8/203 patients. In another 6 patients, we found non-recurrent novel loss-of-function (LoF) variants in genes associated with rare syndromes that include kidney defects (*SETBP1*, *WNT5A*), or in genes whose inactivation results in kidney malformations in the mouse (*SLIT3*, *HSPA4L*, *T*, *SCTR*). To define novel genetic drivers in the remaining cohort of 195 patients, we compared their LOF burden with 6,905 controls. We identified rare LoF variants in *GREB1L* ($P=2.04x10^{-5}$), a gene ubiquitously expressed in the developing mouse kidney. Expansion of our model with novel deleterious missense variants resulted in exomewide significance for *GREB1L* ($P=4.08x10^{-6}$). Three mutations (2 LoF and 1 missense) segregated in an autosomal dominant fashion and one predicted deleterious missense was de novo (joint value for burden, inheritance and de novo occurrence: $P<1.0x10^{-8}$). In a replication cohort of 410 RHD cases, we identified 8 more qualifying LoF/missense variants in *GREB1L*. To directly test our genetic findings, we generated a *greb1l* zebrafish model. Knockdown or CRISPR/Cas9 deletion of *greb1l* in zebrafish showed specific pronephric defects that could be rescued by introduction of wild-type human mRNA. Randomized testing of missense alleles by in vivo complementation showed that 4/4 alleles found exclusively in patients were unable to rescue the phenotype.Taken together, our study provides new insight into the genetic landscape of renal malformations and identifies *GREB1L* as a novel susceptibility gene for RHD.

303

Mutations in the mitochondrial ribosomal protein MRPS22 lead to XX gonadal dysgenesis. *A. Chen[1,12], D. Tiosano[2,3,12], H. Baris[3,4,12], Y. Bayram[5,12], T. Guran[6,12], A. Mory[2], L. Kulnane[7], C. Hodges[7,8], Z. Akdermir[6], S. Turan[6], S. Jhangiani[9], C. Hoppel[10], H. Salz[7], J. Lupski[5], D. Buchner[1,7,11].* 1) Department of Biochemistry, Case Western Reserve University, Cleveland, OH 44106, USA; 2) Division of Pediatric Endocrinology, Ruth Children's Hospital, Rambam Medical Center Haifa 30196, Israel; 3) Rappaport Family Faculty of Medicine, Technion - Israel Institute of Technology, Haifa 30196, Israel; 4) The Genetics Institute, Rambam Health Care Campus, Haifa 3109601, Israel; 5) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030, USA; 6) Department of Pediatric Endocrinology and Diabetes, Marmara University Hospital, Istanbul, 34899, Turkey; 7) Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH 44106, USA; 8) Department of Pediatrics, Case Western Reserve University, Cleveland, OH 44106, USA; 9) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, 77030, USA; 10) Center for Mitochondrial Diseases, Department of Pharmacology, Case Western Reserve University, Cleveland, OH 44106, USA; 11) Research Institute for Children's Health, Case Western Reserve University, Cleveland, OH 44106, USA; 12) These authors contributed equally to this work,.

XX female gonadal dysgenesis (XX-GD) is a rare disorder characterized by primary amenorrhea, hypergonadotrophic hypogonadism, and delayed pubertal development. Autosomal recessive mutations have been identified that cause XX-GD including in the genes encoding the follicle-stimulating hormone receptor (*FSHR*), nucleoporin-107 (*NUP107)*, among others. The identification of these genes has contributed much to our understanding of ovarian development, however this process remains poorly understood. To identify novel genetic causes of XX-GD, we focused on consanguineous families of Middle Eastern descent with primary amenorrhea consistent with an autosomal recessive inheritance pattern. One such family had three female patients below 20 years of age that initially presented with primary amenorrhea and delayed puberty and breast development. Further endocrine and histological evaluation revealed hypergonadotrophic hypogonadism and small ovaries without follicles leading to a diagnosis of XX-GD. Linkage analysis combined with whole exome sequencing (WES) identified a 3 Mb interval on Chromosome 3 that segregated with XX-GD, in which a single homozygous missense variant in the mitochondrial ribosomal protein S22 (*MRPS22*, p.R202H) was identified. The GeneMatcher tool facilitated identification of a second family which has an affected female diagnosed with hypergonadotrophic hypogonadism and ovarian dysgenesis. WES revealed a different homozygous missense mutation in *MRPS22* (p.R135Q) in the second family. Both missense mutations identified in *MRPS22* are rare, occurred in highly evolutionarily conserved residues, and predicted to be deleterious to protein function. The identification of different deleterious missense mutations in *MRPS22* in two independent families with hypergonadotrophic hypogonadism and ovarian dysgenesis suggests that mutations in this gene represent a novel genetic cause of XX-GD. To further evaluate the role of MRPS22 in fertility in an *in vivo* setting, we studied the effect of tissue-specific deficiency of the Drosophila ortholog, mRpS22, using RNAi knockdown. Whereas mRpS22 deficiency in somatic cells of the Drosophila ovary had no effect on fertility, flies with mRpS22 deficiency in germ cells were infertile and failed to develop germ cells. Collectively, our data suggests that mutations in *MRPS22* are a novel cause of XX-GD and that MRPS22 is required for female germ cell development.

304

Mutations of *CDC14A* are associated with nonsyndromic deafness *DFNB32* or HIIMS, hearing impairment infertile male syndrome. *A. Imtiaz[1,2], I. Belyantseva[1], A. Berirl[3], C. Fenollar-Ferrer[4], R. Bashir[2], A. Bouzid[5], U. Shaukat[6], I. Bukhari[2], H. Azaiez[7], K. Booth[7,8], K. Kahrizi[9], A. Maqsood[1,2], E. Wilson[1], T. Fitzgerald[10], A. Tlili[3], R. Olszewski[11], A. Rehman[1], M. Starost[12], A. Waryah[6], M. Hoa[11], L. Dong[13], R. Morell[14], R. Smith[7,8], S. Riazuddin[6,15,16], S. Masmoudi[5], K. Kindt[3], S. Naz[2], T. Friedman[1].* 1) Laboratories of Molecular Genetics, NIDCD/NIH, Bethesda, MD; 2) School of Biological Sciences, University of the Punjab, Lahore, Pakistan; 3) Section on Sensory Cell Development and Function, National Institute on Deafness and Other Communication Disorders, NIH, Bethesda, Maryland, USA; 4) Laboratory of Molecular and Cellular Neurobiology, Section on Molecular and Cellular Signaling, National Institute of Mental Health, NIH, Bethesda, Maryland, USA; 5) Laboratoire Procédés de Criblage Moléculaire et Cellulaire, Centre de Biotechnologie de Sfax, Université de Sfax, Sfax, Tunisia; 6) Center of Excellence in Molecular Biology, University of the Punjab, Lahore, Pakistan; 7) Molecular Otolaryngology and Renal Research Laboratories, Department of Otolaryngology- Head and Neck Surgery, University of Iowa, Iowa City, Iowa; 8) Department of Molecular and Cellular Biology, University of Iowa, Iowa City, Iowa; 9) Genetics Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran; 10) Mouse Auditory Testing Core Facility, National Institute on Deafness and Other Communication Disorders, NIH, Bethesda, Maryland, USA; 11) Auditory Development and Restoration Program, National Institute on Deafness and Other Communication Disorders, NIH, Bethesda, Maryland, USA; 12) Division of Veterinary Resources, National Institutes of Health, Bethesda, Maryland, USA; 13) Genetic Engineering Core, National Eye Institute, NIH, Bethesda, Maryland, USA; 14) Genomics and Computational Biology Core, National Institute on Deafness and Other Communication Disorders, NIH, Bethesda, Maryland, USA; 15) Shaheed Zulfiqar Ali Bhutto Medical University, Pakistan Institute of Medical Sciences, Islamabad, Pakistan; 16) Allama Iqbal Medical College, University of Health Sciences, Lahore, Pakistan.

The Cell Division-Cycle-14 gene (*cdc14*) encodes a dual-specificity phosphatase that is necessary in yeast for exit from mitosis. But the roles of mammalian *CDC14A* gene remain unresolved. We used genetic and mechanistic studies of human CDC14A and its mouse and zebrafish orthologues to clarify its functions in the auditory and reproductive systems *in vivo*. Here we report five recessive truncating and three missense alleles of *CDC14A* in human families segregating progressive, moderate-to-profound deafness linked to the chromosome 1p *DFNB32* locus. We used computational modeling and functional assays to predict how these missense variants impair phosphatase activity of CDC14A as well as the folding of this protein. Our findings suggest that hypomorphic alleles of *CDC14A* are associated with nonsyndromic deafness DFNB32 in two families while more damaging variants cause both deafness and male infertility. In wild-type mouse inner ears, endogenous CDC14A and overexpressed EGFP-CDC14A protein are localized to hair cell kinocilia, basal bodies and stereocilia. Homozygous phosphatase-dead cdc14aa mutant zebrafish, engineered using CRISPR/Cas9, show only subtle variation in kinocilia length and have a normal startle response, mechanotransduction and fertility suggesting that *cdc14ab* may be compensating for the loss of *cdc14aa* function. Several mouse *Cdc14a* recessive mutants including a phosphatase-dead p.Cys278Ser allele, engineered using CRISPR/Cas9, cause perinatal lethality but the exceptional survivors are deaf and males also infertile. In mouse testes, CDC14A localizes to various cell types and to sperm flagella. Mutant alleles of *Cdc14a* cause degeneration of the seminiferous tubules and sperm have an abnormal morphology. Auditory hair cells of postnatal bi-allelic *Cdc14a* mutant mice develop normally, but subsequently degenerate, causing deafness. These findings define a new deafness syndrome and firmly establish new roles for mammalian CDC14A for hearing and spermatogenesis.

305

Discovery of cerebral palsy genes through trio-based whole exome sequencing in cryptogenic individuals. *S. Bakhtiari[1,2], S.C. Jin[3], X. Zeng[3], M. Corbett[4], S. Padilla-Lopez[1,2], B. Norton[1,2], H. Magee[1,2], C. van Eyk[4], R.P. Lifton[3,5], K. Bilguvar[3], A.H. MacLennan[4], J. Gecz[4], M.C. Kruer[1,2].* 1) Barrow Neurological Institute, Phoenix Children's Hospital; 2) Department of Child Health, University of Arizona College of Medicine Phoenix; 3) Department of Genetics, Yale University; 4) University of Adelaide; 5) Laboratory of Human Genetics and Genomics, Rockefeller University.

Cerebral palsy (CP) is a major neurodevelopmental disorder that impairs motor function. Traditionally, CP has been attributed to environmental insults such as intrauterine infection, premature birth, thrombosis, and asphyxia. Nevertheless, up to one-third of affected children are not exposed to any of these risk factors. Such 'cryptogenic' forms of CP may instead be attributable to genomic variants that disrupt normal brain development. In this study, we performed whole exome sequencing of a single cohort of 95 parent-offspring trios with cryptogenic CP and identified that the majority (63%) of affected probands carry at least one putatively damaging (splice site, premature truncation, frameshift and MetaSVM-deleterious missense) variant. Among the deleterious mutations, 38% are X-linked, 35% are de novo, and 27% are biallelic variants. Notably, several genes were previously implicated in other neurodevelopmental disorders, including SYNGAP1 (intellectual disability (ID), epilepsy, autism), CTNNB1 (ID, autism), KCNT1 (epilepsy, ID), GABRA1 (autism, ID), GNB1 (ID), ZDHHC9, ZDHHC15 (ID, epilepsy), WDR45 (intellectual disability), and TRIO (intellectual disability). We identified several novel rare de novo CNVs as well as three recurrent CNVs (2p25.3, 2q37.3, 11q21) in our cohort. Pathway analysis identified distinct biological mechanisms significantly enriched in CP patients, which include the Ras superpathway, axon guidance, and the cytoskeleton; inflammation-related and vascular pathways were absent. Using a context-specific de novo probability model, we found a significant enrichment of protein-altering de novo mutations in cases (1.4-fold enrichment, P-value = 2.0x10-3). When restricting the analysis to genes in the Ras superfamily, there was marked enrichment of de novo damaging (109-fold enrichment, P-value = 7.3x10-8) and loss-of-function mutations (158-fold enrichment, P-value = 1.1x10-6), which clearly suggests an important role for the Ras superfamily in CP. We further estimated that damaging de novo mutations in the Ras superfamily can account for ~4.3% of cases. Taken together, our findings indicate a significant role for de novo mutations and CNVs in CP pathogenesis, illustrate both clinical and molecular overlap of CP with other neurodevelopmental disorders, and suggest common molecular pathways may unify genes that contribute to CP neurobiology.

306

Debunking the cerebral palsy (CP)-birth asphyxia myth: Frequent genomic etiologies for CP identified by exome sequencing. *A. Moreno-De-Luca[1], A. Gonzalez-Mantilla[1], S.M. Myers[1], J. Reid[2], J. Overton[2], D.H. Ledbetter[1], C.L. Martin[1] on behalf of the DiscovEHR collaboration.* 1) Autism & Developmental Medicine Institute, Geisinger Health System, Danville, PA; 2) Regeneron Genetics Center, Tarrytown, NY.

Cerebral palsy (CP) is the most common physical disability of childhood, with an incidence of 2–3 per 1,000 live births. CP encompasses a group of developmental brain disorders (DBDs) characterized by motor deficits that are often accompanied by other DBDs and medical disorders, such as intellectual disability and epilepsy. Historically, birth asphyxia, caused by adverse intrapartum events, was assumed to be the leading cause of CP. However, large population-based studies have now shown that birth asphyxia accounts for less than 10% of CP and the specific etiology remains elusive in most cases. A growing body of evidence suggests that a large proportion of CP is caused by many diverse and individually rare genomic abnormalities, as is the case for other DBDs, such as autism and intellectual disability. As part of the DiscovEHR collaboration between Geisinger and the Regeneron Genetics Center we surveyed 47,589 patient-participants with exome sequencing data, and found 87 with a diagnosis of CP in their electronic health record (frequency of 1 in 547). Exome sequencing was used to assess for potentially causative variants. We identified loss-of-function (LOF) variants in 15 cases (17%); 8 variants are in genes previously known to cause CP or other DBDs, including *ADD3* (Spastic quadriplegic CP), *MECP2* (Rett syndrome), *TCF4* (Pitt-Hopkins syndrome), and *SMARCA4* (Coffin-Siris syndrome). In the remaining 7 cases, we identified rare LOF variants in novel CP candidate genes (*SIRT1*, *HECW1*, *NEURL4*, *KDM5A*, *KIF23*, *MAP3K12,* and *GPBP1*), all of which are predicted to be extremely LOF intolerant. We also identified private missense variants predicted to be deleterious in an additional 34 cases (39%); 20 of which occur in genes previously associated with DBDs, and 14 that are in novel CP candidate genes predicted to be haploinsufficient. All LOF and missense variants are absent from population-based exome repositories (ExAC, EVS, 1000G). Our results show that CP, like other neurodevelopmental disorders, is genetically heterogeneous with shared genomic underpinnings. We also provide evidence that a large proportion of CP has a genetic etiology, debunking the long-standing misconception that most CP is caused by birth asphyxia. This paradigm shift has the potential to change the current diagnostic approach for CP and lead to an increase in research efforts to elucidate the neurobiology of this disorder, eventually contributing to the development of novel therapies for CP.

307

Why West: Comparative analysis of age, etiology, genes and molecular pathways in infants who do and do not develop spasms. *S. Chakravorty[1], S. Koh[2], R.P. Saneto[3], Z.M. Grinspan[4], J.E. Sullivan[5], E.C. Wirrell[6], R.A. Shellhaas[7], J.R. Mytinger[8], W.D. Gaillard[9], E.H. Kossoff[10], I. Valencia[11], K.G. Knupp[12], C. Wusthoff[13], C. Keator[14], N. Ryan[15], T. Loddenkemper[16], C.J. Chu[17], E.J. Novotny Jr.[18], J. Coryell[19], M. Hegde[1], A.T. Berg[20].* 1) Human Genetics, Emory University, Atlanta, GA; 2) Pediatrics, Emory University, Atlanta, GA; 3) Neurology, University of Washington, Seattle, WA; 4) Weill Cornell Medical College; 5) UCSF Benioff Children's Hospital; 6) Neurology, Mayo Clinic; 7) University of Michigan CS Mott Children's Hospital; 8) Nationwide Children's Hospital, Columbus, Ohio; 9) Children's National Medical Center; 10) Johns Hopkins University Medicine; 11) Pediatrics, Drexel University; 12) Children's Hospital Colorado; 13) Neurology, Pediatrics, Stanford University; 14) Cook Children's Medical Center; 15) Pediatrics Neruology, Children's Hospital of Philadelphia; 16) Neurology, Boston Children's Hospital; 17) Neurology, Massachusetts General Hospital; 18) Neurology, University of Washington; 19) Oregon Health & Science University Healthcare; 20) Northwestern University Ann & Robert H. Lurie Children's Hospital of Chicago.

Infantile spasms (IS) are the defining seizure type of West syndrome, a form of early life epilepsy (ELE) associated with seizures, developmental consequences and early mortality. This is the first comparative study of infants with and without spasms on some key clinical and molecular differences. 775 children were prospectively recruited at the time of initial diagnosis of epilepsy (onset <3 years). Age of onset (AoO) and gestational age (GA) were compared between children who presented with spasms (N=271), and who never developed spasms over the course of a 1-yr follow-up (N=503). 327/775 (42%) had ≥1 clinical genetic investigations, including chromosomal microarrays, epilepsy gene-panel, exome, and other targeted tests. Of these, pathogenic variants were identified in 142 (43%). We also performed comparative gene ontology and molecular Pathway Analysis on genes harboring pathogenic variants. AoO of spasms had a striking peak near 6 months whereas other seizures were more concentrated at younger ages. Gene-pathway analyses on 60 genes harboring pathogenic variants reveal a much broader spectrum of "biological pathways" affected in the spasms. For spasms, central nervous system (CNS) development, cell cycle regulation, mitochondrial metabolism, immunological processes are more affected than that in non-spasm epilepsies. By contrast, in the non-spasm group, maintenance and neuromuscular motor activity pathways appear preferentially affected. In spasms, neuronal cellular organelles (Golgi, ER, mitochondria, lysosome, centromere / kinetochore, dendrites) were uniquely enriched. By contrast, in the non-spasms group, axonal areas such as synapse, node of Ranvier, and plasma membrane were preferentially enriched. Of note, no children with *SCN1A* or *PRRT2* pathogenic variants developed spasms while most with Trisomy 21, *CDKL5*, *TSC2*, but not *TSC1* developed spasm. Our molecular analyses reveal the quintessential neurodevelopmental component of spasms with distinct peak of AoO associated that in turn associated with dysregulation of large numbers of developmental pathways. There is also a clear genetic and cellular component dichotomy between IS and other ELEs. Such findings can advance the understanding of the unique molecular underpinning of spasms and may play a role in patient stratification for future therapeutic trials. Use of whole genome sequencing could further examine the full range of molecular contributions to these diverse early life epilepsies.

308

Epilepsy: Whole exome sequencing of 6K cases and 14k controls confirms a significant role for ultra-rare deleterious coding variants. *D. Howrigan[1,2] on behalf of the Epi25k Consortium.* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Stanley Center for Psychiatric Genetics, Broad Institute, Cambridge, MA.

Recent advances in genetic sequencing have enabled the discovery of highly penetrant rare genetic variants underlying rare forms of epilepsy, however the broader genetic landscape of rare genetic variation across epilepsy subtypes remained to be discovered. In turn, global consortium efforts are essential for facilitating new genetic discoveries by aggregation of larger patient cohorts, particularly for more common forms of epilepsy. The international Epi25k consortium has been formed combining previous national and multinational research groups in order to unite synergistic efforts to discover new genetic risk factors for epilepsy, marking it as the largest exome sequencing study on epilepsy to date. Within the Epi25k framework, whole exome sequence (WES) data have recently been generated for over 6,011 epilepsy patients: 3,037 individuals diagnosed with genetic generalized epilepsy (GGE), 2,186 individuals diagnosed with familial or sporadic non-acquired focal epilepsy (NAFE), and 788 individuals diagnosed with epileptic encephalopathies (EE). These patients were compared against 14,334 individuals not ascertained for epilepsy drawn from multiple independent collections at the Broad Institute. All controls were screened negative for neurodevelopmental disorders. All WES data were jointly called, with extensive quality control to match for population structure, remove low quality samples and variants, and restrict to shared exome capture targets. Preliminary analyses confirm the significant enrichment of ultra-rare protein truncating (PTV) and predicted damaging missense coding variants among all epilepsy types over controls, and particularly among genes depleted for loss-of-function mutations (PTV OR = 1.34, p = 4e-8; damaging missense OR = 1.07, p = 6e-5). Furthermore, we see a strong enrichment among genes previously identified in EE, GGE, and NAFE (PTV OR = 7.48, p = 2.8e-6; damaging missense OR = 1.78, p = 1.3e-4), demonstrating the strong replicability of extant gene findings to date in an independent cohort. Upcoming downstream analyses will elucidate novel shared and subtype specific genetic discoveries among epilepsy syndromes, and will be facilitated by expected doubling of the patient cohort in the upcoming months. .

309

TOPMed whole genome sequence association analysis of type 2 diabetes. *J. Wessel[1], J. Brody[2], B. Hidalgo[3], A. Manning[4] on behalf of the Trans-Omics for Precision Medicine (TOPMed) Program, Diabetes Working Group.* 1) Indiana University, Indianapolis, IN; 2) University of Washington, Seattle, WA; 3) University of Alabama at Birmingham, Birmingham, AL; 4) Harvard University, Boston, MA.

The majority of genetic variants significantly associated with type 2 diabetes (T2D) reside in the non-coding genome, with many causal variants still unknown. We leveraged Whole Genome Sequence (WGS) phase 1 data from NHLBI's Trans-Omics for Precision Medicine (TOPMed) initiative to perform a T2D WGS association (WGSA) pooled analyses. WGSA analyses included samples with deep (>30x) sequence coverage in n=13,408 (n=2,607 cases and n=10,801 controls) from 12 studies (65% European and 34% African ancestry). We used mixed effects models adjusting for sex, age, ancestry, study, empirical kinship and 10 principal components to adjust for relatedness and population structure. In multi-ethnic analyses, common (minor allele frequency (MAF)>0.05) variant associations (P value<$5x10^{-8}$) were identified at known locus with T2D: *TCF7L2* (rs7903146, P value=$2.0x10^{-12}$ and 8 additional variants). For variants that have not been previously described, we identified 8 rare non-coding variant (MAF<.01) associations in 3 known loci: *GCKR* (chr2:27987368, MAF=.03%, P value =$2.0x10^{-9}$), *ADAMTS9* (chr3:64184329, MAF=.06%, P value =$2.6x10^{-9}$) and *CDKN2A/B* (chr3:64184329, MAF=.02%, P value =$3.8x10^{-8}$ and 5 additional variants). Three rare nonsynonymous variants are in candidate genes that are outside of known loci in *NTN4* (A166T, MAF=.03%, P value =$1.5x10^{-9}$) and *DIABLO* (A60V, MAF=.06%, P value =$1.8x10^{-8}$) are between T2D loci *TSPAN8* and *HNF1A*; and *OIP5* (F20I, MAF=.2%, P value =$1.4x10^{-8}$) downstream of *RASGRP1*. Additional associations included non-coding rare variants in loci not previously described with T2D; including intergenic rs778917988 (MAF=0.03%, P value=$2.0x10^{-8}$) near *SESN3*, a known glucose-homeostasis gene; and 7 variants in *SEMA6D* (chr15:47474740, MAF=.2%, P value =$1.3x10^{-8}$), a gene previously associated with a common variant and BMI. In ancestry specific analyses, an intronic variant in the ncRNA *RP11-99E15.2* (chr14:77210987, MAF=.12%, P value =$1.8x10^{-8}$) was significant in Europeans only. Preliminary results suggest multi-ancestry WGSA can discover novel loci for complex traits. Work is ongoing to refine annotation for non-coding gene-based tests, perform fine-mapping, and extend into phase 2 and 3 data (N~14,000 cases and ~55,000 controls).

310

Eighteen novel loci associated with type-2 diabetes in multi-ethnic analyses involving 484,989 individuals. *J. Lee[1,2], D. Miller[3], J. Huang[4,5], J. Lynch[6,7], S. Damrauer[1,8], Y. Sun[9,10], T. Assimes[11,12], J. Lee[11], K. Cho[4], S. Muralidhar[13], Q. Shao[7], S. DuVall[6], K. Lee[4], D. Rader[1,14], M. Gaziano[4,5], J. Concato[15], P. Tsao[11,12], C. O'Donnel[4,5], K. Chang[1], J. Meigs[16], P. Wilson[17], P. Reaven[18], L. Phillips[17], D. Saleheen[1,2].* 1) Corporal Michael Crescenz VA Medical Center, Philadelphia, PA; 2) Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA; 3) ENR Memorial Veterans Hospital, 200 Springs Road (152), Bedford, MA 01730; 4) Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Health System, Boston, MA; 5) Brigham Women's Hospital, Harvard Medical School, Boston, MA; 6) Department of Veterans Affairs Salt Lake City Health Care System, Salt Lake City, UT; 7) University of Massachusetts College of Nursing and Health Sciences, Boston, MA; 8) Department of Surgery, University of Pennsylvania, Philadelphia, PA; 9) Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA; 10) Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA; 11) VA Palo Alto Health Care System, Palo Alto, CA; 12) Department of Medicine, Stanford University School of Medicine, Stanford, CA; 13) Office of Research and Development, Veterans Health Administration, Department of Veterans Affairs, 810 Vermont Ave. N.W., Washington D.C. 20420; 14) Department of Genetics, University of Pennsylvania, Philadelphia, PA; 15) VA Connecticut Healthcare System, West Haven, CT and Yale University School of Medicine, New Haven, CT; 16) Massachusetts General Hospital, Harvard Medical School, Boston, MA; 17) Atlanta VA Medical Center, Emory University School of Medicine; 18) University of Arizona, Arizona State University, Phoenix, AZ 85012.

Introduction: To evaluate the contribution of both coding and non-coding variation in relation to type-2 diabetes (T2D), we conducted multi-ethnic meta-analyses involving 484,989 individuals of European, Hispanic, South Asian, East Asian and African descent, by leveraging the individual participant data available in the Million Veteran Program (MVP), a mega-biobank established by the United States Department of Veterans Affairs (VA). **Methods**: We first used 286,731 MVP participants of European (n = 210,255), African (n = 56,292) and Hispanic (n = 20,184) descent with available genotype data and applied an ICD-9/10 based algorithm to identify individuals with and without type-2 diabetes (cases =59,501, controls =227,230). Stratified by self-reported and PCA clustered ethnicity, we used genetic data imputed by using the 1000 Genomes reference panel (ver. 3) and conducted discovery studies in association with T2D followed by an overall meta-analysis involving all participants. Analyses were adjusted for the first 10 principal components generated for each ethnicity. For replication studies, we subsequently used unpublished and non-overlapping data from the Penn-T2D meta-analyses comprising of 48,437 individuals of South Asian (n = 28,139) and European (n = 20,298) descent and data for T2D from the DIAGRAM consortium and conducted combined meta-analysis on 484,989 participants (n = 107,866) in relation with T2D. **Results**: In the MVP dataset alone, 3,656 variants were found to be genome-wide significant ($P<5x10^{-8}$). As a positive control, an intronic variant, rs7903146, at the *TCF7L2* locus was found to be most significantly associated with T2D ($P = 5.07x10^{-217}$). In the overall meta-analyses, 206 variants at 18 distinct loci (e.g., KLF12), including exonic variation at 4 novel loci (e.g., DRG2), were found to be genome-wide significant that have not been previously reported in relation with T2D. **Conclusions**: By leveraging the VA MVP dataset, we conducted a large-scale multi-ethnic meta-analysis and identified 18 novel T2D risk loci. These loci implicate several novel pathways including cellular proliferation, cell signaling and inflammation in the etiology of T2D. .

311

Discovery and fine-mapping of type 2 diabetes susceptibility loci across ethnically diverse populations. *A. Mahajan[1], H. Kitajima[1], X. Sim[2], M.CY. Ng[3], W. Zhang[4], J.E. Below[5], D. Taliun[6], K.J. Gaulton[7], A.P. Morris[1,8] on behalf of the DIAMANTE Consortium.* 1) WTCHG, University of Oxford, Oxford, United Kingdom; 2) Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore; 3) Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA; 4) School of Public Health, Imperial College London, London, UK; 5) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA; 6) Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA; 7) Department of Pediatrics, University of California San Diego, La Jolla, CA, USA; 8) Department of Biostatistics, University of Liverpool, Liverpool, UK.

To discover and enhance fine-mapping resolution of type 2 diabetes (T2D) loci, we are conducting meta-analysis of genome-wide association studies (GWAS) representing the most comprehensive view of the genetic contribution to the disease in terms of sample size and ethnic diversity. The interim data freeze included 99,265 T2D cases and 545,212 controls (>40% non-Europeans) from 72 GWAS, each imputed up to reference panels from the 1000 Genomes Project or Haplotype Reference Consortium. Association summary statistics across GWAS were combined using meta-regression accounting for ancestry. Approximate conditional analysis was performed to identify distinct signals within loci attaining genome-wide significance ($p<5x10^{-8}$). For each signal, we defined credible sets of variants that accounted for 99% of the posterior probability of driving the association and evaluated enrichment across a diverse range of functional genomic annotations. We identified 110 loci at genome-wide significance, including 37 mapping outside regions previously implicated in T2D. Conditional analyses revealed 46 additional distinct association signals across the loci, including 11 at *KCNQ1*, 5 at *INS-IGF2*, and 4 each at *CDKN2A-B* and *CCND2*. Whilst allelic effects on T2D risk of index variants were predominantly consistent across populations, for the first time we observed strong evidence of heterogeneity correlated with ancestry at loci such as *LEP* (rs7778167, $p_{HET}=8.2x10^{-16}$, East Asian specific) and *KCNQ1* (rs11819853, $p_{HET}=2x10^{-10}$, varying direction/magnitude of effect between ethnic groups). Compared with previous efforts, we substantially improved fine-mapping resolution, highlighting 17 signals where a single variant accounted for >99% of the posterior probability of driving the association, including at *JAZF1* (rs10226758), *CDC123-CAMK1D* (rs11257655), *TCF7L2* (rs7903146), and 2 signals at *KCNQ1* (rs2237884 and rs2237895). The posterior probability of association was significantly enriched in coding exons ($p=1.4x10^{-5}$) and transcription factor binding sites for PDX1 ($p=2.6x10^{-6}$) and FOXA2 ($p=1.8x10^{-5}$), key regulators of β-cell development and function. The final data freeze will include over a million individuals and anticipated to further improve prioritisation of causal variants and discern their functional consequences, providing a better understanding of the causal mechanisms of T2D risk.

312

Inverting the problem: Environmental signal maximization in type 2 diabetes using genetic correction. *J. Blangero[1], J.M. Peralta[1,2], J.E. Curran[1], S. Williams-Blangero[1].* 1) South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX; 2) Menzies Institute for Medical Research, University of Tasmania, Hobart, TAS.

While genetic studies typically focus on gene discovery, we have developed a novel approach that exploits genetics to facilitate discovery of systematic environmental (i.e., non-genetic) factors in disease risk. In our approach, we treat genetics as a confounder that reduces power to detect environmental signals. Pedigree-based designs are a powerful way to obtain statistical control of genetics to enhance the signal-to-noise ratio of environmental factors. In this study, we control for genetic factors in type 2 diabetes risk (T2D) by estimating individual-level genome-wide additive genetic effects using a Best Linear Unbiased Prediction procedure. The resulting estimated genetic value (EGV) then is subtracted from the phenotypic value to obtain an estimated environmental value (EEV) that is free of the expected additive genetic signal. Conditional on such genetic correction, environmental signals are increased. As an example, we estimated EEVs for T2D risk using a whole genome sequence derived relationship matrix in ~1,200 participants in the San Antonio Family Heart Study from large extended pedigrees. We then searched the transcriptional exposome using genetically corrected PBMC-derived transcriptomic profiles. We observed strong evidence for genes whose environmentally-driven expression values correlate with T2D risk. To characterize systematic environmental features, we performed spatial covariance analyses using geocoded residence information upon the ~1,000 transcripts that were environmentally correlated with T2D risk. These transcripts exhibited an excess ($p=2.0\times10^{-27}$) of significant local spatial structure (i.e., differences between individuals increased with spatial distance). For example, the environmental component of *ARH3* gene expression was negatively correlated with T2D and showed evidence for spatial structure (p=0.0025) that accounted for nearly 18% of its variation. Correlation between individuals was halved at residential distances of 2.5 miles and eliminated by ~30 miles, suggesting an underlying systematic environmental factor. *ARH3* inhibits poly(ADP-ribose) chains that are known to be influenced by air pollutants such as particulate matter and benzene. Thus, a testable hypothesis is that spatial variation in *ARH3* expression is due to spatial variation in air pollutants. Our results suggest that this novel approach to controlling for genetics can be employed to enhance the detection of environmental factors in disease risk.

313

Design and implementation of a sequencing panel for eMERGE Phase III. *L. Witkowski[1,2], M.V. Harden[3], E. Kudalkar[4], M.S. Leduc[5], L.M. Mahanta[1], E. Hynes[1], L.J. Babb[1,6], M.J. Bowser[1], C. Graham[1], C-F. Lin[1], S.B. Baxter[1,3], S.B. Gabriel[3], S.J. Aronson[1], M.S. Lebo[1,2], B.H. Funke[1,7], N.J. Lennon[3], H.L. Rehm[1,2], H. Zouk[1,7].* 1) Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA; 2) Dept of Pathology, Brigham & Women's Hospital, Harvard Medical School, Boston, MA; 3) Clinical Research Sequencing Platform, Broad Institute, Cambridge, MA; 4) Dept of Dermatopathology, Northwestern University, Chicago, IL; 5) Dept of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 6) Geneinsight a Sunquest Company, Boston, MA; 7) Dept of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA.

The electronic MEdical Records and GEnomics (eMERGE) Network is an NHGRI-funded project that combines DNA biorepositories with electronic health records for large scale, high-throughput genetic research and support for implementing genomic medicine. In Phase III of this project, a major aim is to sequence and assess the implications of detecting clinically relevant variants in 25,000 individuals in 109 genes and ~1500 single nucleotide variants (SNVs) submitted by nine participating sites at two CLIA-certified centralized sequencing and genotyping centers (CSGs): Partners Healthcare/Broad Institute and Baylor College of Medicine. A clinical curation effort was performed for all genes and SNVs submitted for inclusion on an NGS panel. A total of 88 genes met strong/definitive association to disease, including cancer susceptibility, cardiac disease, and neurological and connective tissue disorders. A further 68 SNVs were considered clinically relevant. These genes and SNVs were further reviewed for actionability by the eMERGE clinical annotation working group, leading to a consensus list of 68 genes and 14 SNVs suggested for return to participants. To ensure consistency in variant interpretation between CSGs, a harmonization effort was developed where CSGs exchange monthly downloads of interpreted variants in eMERGE panel genes to generate a discrepancy report. Discrepancies in reportable variants are discussed for attempted resolution. Of 24 discrepancies thus far, 16 variants spanning four disease areas affected reporting, with consensus achieved for all. Our eMERGE panel was clinically validated and used to sequence DNA from >4000 individuals thus far. Data has been analyzed for 3917 individuals from three sites, comprising a diagnostic cohort of patients with colorectal cancer or polyps (n = 1191), a cohort unselected for phenotype (n = 1475), and a cohort recruited for suspicious genotype (n = 1251). The positive rates in these cohorts are 5%, 2.8%, and 15.7%, respectively. In the diagnostic cohort, 2.4% of patients had clinically significant variants relating to their disease, while 2.6% of patients had reportable secondary findings. One of these patients had both an indication-based returnable result and a secondary finding. Overall, 7 individuals had more than 1 pathogenic variant, and 4 had pathogenic copy number variants. eMERGE sites are now returning results to participants and studying the impact of this process in the broader medical context.

314

Genomic sequencing results from the first 100 newborns sequenced in the BabySeq Project. *O. Ceyhan-Birsoy[1,2], K. Machini[2,3,4], J. Murry[2,3,4], M.S. Lebo[2,3,4], S. Fayer[4,5], C. Genetti[4,6], T.S. Schwartz[4,6], G.E. VanNoy[4,6], T.W. Yu[4,6,7], P.B. Agrawal[4,6,8], R.B. Parad[4,9], I.A. Holm[4,6], A. McGuire[10], R.C. Green[4,5,11], A.H. Beggs[4,6], H.L. Rehm[2,3,4,11].* 1) Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY; 2) Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA; 3) Department of Pathology, Brigham & Women's Hospital, Boston, MA; 4) Harvard Medical School, Boston, MA; 5) Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA; 6) Division of Genetics and Genomics, The Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA; 7) Department of Neurology, Boston Children's Hospital, Boston, MA; 8) Division of Newborn Medicine, Boston Children's Hospital, Boston, MA; 9) Department of Pediatric Newborn Medicine, Brigham and Women's Hospital, Boston, MA; 10) Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX; 11) The Broad Institute of MIT and Harvard, Cambridge, MA.

Genomic sequencing (GS) in newborns provides the opportunity to detect a wide range of conditions for which early knowledge can improve health outcomes. The BabySeq Project is a randomized controlled trial exploring the use of GS in healthy and ill newborns. Half of the enrolled newborns receive standard care, while the others receive GS in addition to standard care. Parental samples are collected to assist in variant interpretation (e.g. assess phase, de novo status, or segregation). All babies receive a Newborn Genomic Sequencing Report that returns risk and carrier status for highly penetrant diseases with onset or management during childhood as well as pharmacogenomics (PGx) variants relevant to the pediatric population. Additionally, ill newborns in the NICU and any baby who develops symptoms that may warrant a genomic analysis during the course of the study receive an indication-based analysis. Report generation has been completed for the first 100 newborns in the sequencing group, twelve of whom were enrolled from the NICU. Variants associated with monogenic disease risk were identified in four babies in the healthy cohort indicating risk for cardiomyopathy, SVAS, and biotinidase deficiency. In the NICU subjects, seven had congenital heart defects and four had other congenital anomalies. Five babies enrolled in the healthy cohort had presentations that later prompted an indication-based analysis for creatine deficiency, hyperbilirubinemia, bilateral hip dysplasia, or congenital heart defects. While no variants that definitively explain their current condition were detected in the ill newborns, risk for KBG syndrome and cancer were identified in two babies. Carrier status for one or more variants was identified in ninety newborns, suggesting that returning carrier information has a significant impact on the number of newborn GS reports with positive findings. The number of carrier status variants per report ranged from zero to six, with a median of two variants per report. PGx variants relevant to medications that may be used during childhood were identified in five newborns. Eighteen variants identified in the newborns were tested in the parents to guide their interpretation, demonstrating that having parental samples available for testing adds power to the GS results analysis in newborns. We continue to study the impact of our reports on clinical outcomes and their utilization by parents and physicians to evaluate the use of GS in newborns.

315

Precision medicine screening using whole genome sequencing, whole body imaging and noninvasive functional diagnostics. *C.Y.-C. Hou[1], P. Brar[1], D. Karow[1], A. Kahn[2], N. Shah[1], V. Lavrenko[1], H.C. Yu[1], A.M. Dale[3], L. Guo[4], T.J. Jonsson[4], B.M. Wittmann[4], I. Bartha[1], S. Ramakrishnan[1], A. Bernal[1], J. Brewer[5], W.H. Biggs[1], A. Telenti[1], B.A. Perkins[1], C.T. Caskey[1,6], J.C. Venter[1].* 1) Human Longevity Inc., 4570 Executive Dr., San Diego, CA; 2) Division Cardiovascular Medicine, UC San Diego School Medicine, 9434 Medical Center Dr, San Diego, CA; 3) Depts of Neurosciences and Radiology, UC San Diego School Medicine, San Diego, CA; 4) Metabolon, 617 Davis Dr, Ste 400, Morrisville, NC; 5) Department of Neuroscience , UC San Diego School Medicine, 9500 Gilman Dr, San Diego, CA; 6) Molecular and Human Genetics, Baylor College of medicine, One Baylor Plaza, Houston, TX.

Whole genome sequence (WGS) of N of 1 coupled with personal and three generation pedigree information has been the initial approach for precision medicine. One of the significant burdens in this approach is the considerable number of variants with uncertain significance or misclassification of clinical significance observed in each WGS analysis that require further clarification. We utilized a series of functional measurements and clinical tests combined with WGS data to make better disease risk prediction, disease diagnosis and treatment strategies. The technologies include metabolomics, MRI (full body and brain), CT cardiac scan, electrocardiogram, echocardiogram, continuous cardiac monitoring system (iRhythm), dual-energy X-ray absorptiometry, cognitive/gait quantifications, clinical laboratory tests and microbiome analysis. Our cohort size is 625 and recruitment is ongoing. Based on the results of first phase of comprehensive analysis of 209 individuals (median age = 55 yrs, range 20-98 yrs, 34.5% female, 7 families), 4 individuals had early stage neoplasms, 8 individuals had cardiac abnormalities requiring immediate medical care, and 52 individuals had medically significant genomic variants correlating with clinical data, including 12 genotype and metabolomics associations. Through our approach of combining WGS with noninvasive functional diagnostics, we have identified genomic evidence that supports behavior changes or adjustments to clinical care that may prevent premature mortality in some participants. We are working to quantitatively integrate genomics with functional data to discover new genotype and phenotype associations through machine learning as our cohort size continues to grow. These findings demonstrate the value of integrating WGS and noninvasive clinical assessments for a rapid and integrated point-of-care clinical diagnosis of age-related diseases that contribute to premature mortality.

316

Five years of clinical whole genome sequencing in asymptomatic individuals: Insights and future directions. *M. Cremona, V. Gainullin, R. Hagelstrom, W. Li, J. Avecilla, M. Bennett, K. Bluske, C. Brown, N. Burns, A. Chawla, A. Coffey, ICSL Curation Team, B. Juan, A. Malhotra, M. Mcginniss, F. Mullen, M. Rajan, V. Rajan, E. Ramos, A. Scocchia, S. Ajay, M. Eberle, ICSL past and present members, D. Perry, D. Bentley, R. Taft.* Population and Medical Genomics Department, Illumina Inc., San Diego, CA.

The Illumina Clinical Services Laboratory (ICSL) offers a clinical whole genome sequencing (cWGS) test for asymptomatic adults and identifies variants in ~1700 genes associated with Mendelian disorders, including the 59 genes defined as actionable by the ACMGG. This test detects SNVs, small insertions and deletions in exons and within 15 bp of the splice site boundary. Variants are interpreted and classified based on modified ACMGG guidelines with the addition of a category, VUS-suspicious (VUS-S), variants with insufficient evidence to qualify as likely pathogenic but with evidence suggesting they may contribute to disease. Over five years, 1512 individuals received the ICSL cWGS predisposition test. Of these, 744 individuals received a clinically relevant finding. We reported 310 pathogenic/likely pathogenic (P/LP) variants in genes associated with autosomal dominant (AD) conditions for which the majority (54%) were hematological diseases. Variants in genes that confer increased cancer risk and immunological disorders of variable penetrance each accounted for 19% of AD P/LP reported variants. Forty three reportedly unaffected individuals had homozygous or compound heterozygous variants in genes with autosomal recessive presentations, the majority (80%) of which were associated with hereditary hemochromatosis. Ninety-two participants (6%) received a P/LP result in one of the 59 ACMGG actionable genes. The vast majority (86%) received carrier status results. For reported VUS-S, 418 participants (28%) received a finding in one of 181 dominant conditions (eye, neurological, hematological, cancer, and muscular diseases among others). Two hundred and eighteen individuals (52%) carried loss-of-function VUS-S (stop-gained, frameshift and splice site) in 136 AD genes. Considering tolerance to loss-of-function (LoF) variation, for AD genes in which VUS-S were found, there were at least 66 with a pLI<0.1, and 23 with pLI >0.95 (ExAC database). These findings highlight the complexity of interpreting newly characterized LoF variants in asymptomatic individuals. Screening roughly 7% of characterized genes yields findings of medical significance for the vast majority of tested individuals. The number of genes in predisposition tests is likely to expand, increasing both its clinical value and the complexity of the interpretation. We will discuss its continued evolution in the context of refined gene and variant curation, expanded gene sets and additional variant types.

317

Genetic analysis of enhancer RNA (eRNA) variation in human population. *H. Kwak[1], K. Kristjánsdóttir[1], H.M. Kang[2].* 1) Molecular Biology and Genetics, Cornell University, Ithaca, NY; 2) Department of Biostatistics, University of Michigan, Ann Arbor, MI.

Enhancer RNAs (eRNA) are non-coding RNAs transcribed bidirectionally from the sequences of enhancer regions. The amount of eRNA generated is directly related to the enhancer activity and proposed to be a better predictor of regulatory targets than chromatin modifications or accessibility. Therefore, population-scale analyses of eRNAs will provide a more comprehensive view of genetic variations that regulate gene expression. Quantifying eRNA has been challenging due to their instability, but recent techniques have enabled a global capture of nascent RNAs at the synthesis stage allowing the measurement of eRNAs under active transcription. Using precision run-on capped nascent RNA sequencing (PRO-cap) in lymphoblastoid cell lines (LCLs) from 69 Yoruba individuals, we quantified eRNA levels to identify actively transcribed enhancers in high resolution, and found genetic variants associated with the eRNA levels. We found 74,988 enhancer candidates in LCLs, 40% of which are variably transcribed between individuals. They have structures composed of central transcription factor binding sites surrounded by divergent bidirectional eRNA initiation sites that are 160 base pairs apart on average. These eRNA regions are enriched with known expression quantitative trait loci (eQTLs) and disease associated variants. Correlation analyses of the transcription levels between enhancers and promoters show that enhancer-promoter interactions are prominent within 200 kilobases, and are disrupted by intervening insulators or strong promoters. Furthermore, we identified 11,394 enhancers where eRNA levels are significantly (FDR < 0.05) associated with genotypes (transcription initiation quantitative trait loci: tiQTLs), 34% of which are also eQTLs. We also found variants affecting the relative ratio between the divergent bidirectional eRNA pairs (directional initiation quantitative trait loci: diQTLs). Interestingly, while tiQTLs are enriched in the central transcription factor binding sites, diQTLs are enriched near the eRNA initiation sites, reflecting a dual-hub model of enhancer architecture. This architecture appears to determine the mode of interaction between clustered enhancers, where there is both eRNA interference and eRNA synergism depending on their orientation and distance. Our results collectively provide evidences that genetic variation of enhancer function relies not only on transcription factor binding, but also on eRNA transcription itself.

318

Intra- and inter-chromosomal chromatin interactions mediate genetic effects on gene expression. *O. Delaneau[1], K. Popadin[2], M. Zazhytska[2], K. Kumar[3], G. Ambrosini[3], A. Gschwind[2], C. Borel[1], D. Marbach[4], D. Lamparter[4], S. Bergmann[4], P. Bucher[3], S. Antonarakis[1], A. Reymond[3], E. Dermitzakis[1].* 1) Department of Genetic Medicine and Development, University of Geneva, Geneva 4, Geneva canton, Switzerland; 2) Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; 3) Swiss Institute for Experimental Cancer Research, EPFL, Lausanne, Switzerland; 4) Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.

Population measurements of gene expression and genetic variation enable the discovery of thousands of expression Quantitative Trait Loci (eQTLs), a useful resource to determine the function of non-coding variants. To understand the effects of eQTLs on regulatory elements such as enhancers and promoters, and to dissect local regulatory networks, we quantified gene expression (mRNA) and three key histone modifications (H3K4me1, H3K4me3 and H3K27ac) across two cell types (Fibroblast and Lymphoblastoid Cell Lines) in 80 and 320 densely genotyped European samples, respectively. First, we find that nearby regulatory elements form 12,583 local chromatin modules spanning up to 1Mb, often comprising multiple sub-compartments and overlapping topologically associating domains (TADs). They bring multiple distal regulatory elements in close proximity, vary substantially across cell types and drive co-expression at multiple genes. Second, we show that this regulation layer is under strong genetic control: we discovered ~44k chromatin QTLs (cQTLs) affecting ~30% of the chromatin peaks, representing to our knowledge the largest cQTL collection so far. In addition, we quantified the activity of chromatin modules by using dimensionality reduction techniques and find QTLs for up to ~70% of them (>5,000 AmodQTLs; module-activity QTLs). Besides module activity, we also find ~50 genetic variants that perturb the structure of modules (SmodQTLs for module-structure QTLs). These act in complex ways leading to a mix of negative and positive correlation among regulatory elements with similar complex effects on gene expression. Third, we show that the module QTLs have downstream effects on higher-order phenotypes. Besides being highly enriched for GWAS hits, they also extensively overlap cis-acting eQTLs (~65%) and we highlight three possible mechanisms of how chromatin modules mediate genetic effects on gene expression: (1) By using Bayesian network modeling, we show that local chromatin interactions within modules mediate the long range effects of some cis-acting eQTLs, (2) We discovered 6 well-replicated trans-eQTLs with strong evidence that they act through direct inter-chromosomal chromatin interactions, (3) We discovered 25 chromatin modules in which the accumulation of rare variants has a significant effect on expression of nearby genes.

319

From mammals to fish and back again: Discovering new regulators of early cardiac development. *M.D. Wilson[1,3,5], X. Yuan[1,2,3,6], M. Song[1,2,3], P. Devine[4], B.G. Bruneau[4], I.C. Scott[2,3,5].* 1) Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada; 2) Developmental Stem Cell Biology, SickKids Research Institute, Toronto, ON, Canada; 3) Department of Molecular Genetics, University of Toronto, Canada; 4) University of California, Developmental and Stem Cell Biology, San Francisco, CA; 5) co-corresponding authors; 6) first author.

Understanding the mechanisms underlying heart formation is crucial for uncovering causes of congenital heart disease. However, we still have an incomplete understanding of the gene regulatory networks that drive early cardiac lineage specification. While the zebrafish is a powerful model for studying heart development *in vivo*, there is a paucity of markers that can identify early cardiac progenitor cells. We first tested whether a mouse cardiac cis regulatory element (CRE) near the *Smarcd3* could serve as a marker of early cardiac lineage in zebrafish. We generated a zebrafish GFP reporter line driven by this *Smarcd3* enhancer. Using *in-situ* hybridization and immunostaining we found that this mouse enhancer expressed GFP in embryonic regions known to give rise to cardiac lineages. To discover cardiac regulatory programs within these GFP +'ve cells, we conducted bulk and single-cell mRNA-seq at the end of gastrulation. We identified a cluster of cells that co-expressed known cardiac markers along with several novel genes. Using *in-situ* hybridization we determined that 14/18 these novel genes are expressed in embryonic cardiac domains. To identify CREs that function in early heart development we profiled open chromatin in this GFP +'ve population using ATAC-seq. We identified ~3800 open chromatin regions that were unique to the GFP +'ve cell population and these regions were enriched for the GATA DNA binding motif and heart development pathways. By using direct and indirect human-zebrafish sequence alignments, we found 176 open chromatin regions that overlap conserved non-coding elements. Remarkably more than two thirds of these conserved regions overlapped human open chromatin regions. These anciently conserved open chromatin regions were enriched for Polycomb repressive complex 2 binding in human embryonic stem cells and developmentally active transcription factors. *In vivo* assays confirmed many of these anciently conserved open chromatin regions were active during zebrafish heart development. Furthermore, several human orthologous CREs drove GATA-dependent cardiac expression following the same spatial-temporal dynamics observed in zebrafish, even though many of them showed modest sequence conservation. Overall, our cross-species functional and comparative epigenomic analysis of zebrafish and human open chromatin regions identified a novel set of genes and ancient enhancers that are active during early heart development.

320

Genomic variation modulates islet regulatory landscape: Application of islet-specific deep learning model. *A. Wesolowska-Andersen[1], M. Thurner[1], J. Torres[1], J. Fernandez[1], M. McCarthy[1], A. Mahajan[1], A.L. Gloyn[2,3], M. McCarthy[1,2].* 1) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; 2) Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK; 3) NIHR Oxford Biomedical Research Centre, Oxford, UK.

Genome-wide association studies point to defects in pancreatic islets as the key mediator of type 2 diabetes (T2D) aetiology. Translational discoveries of these findings have been hampered by challenges in pinpointing the causal variant and interpretation of the biological significance of noncoding variation. Recently, deep learning models have been successfully applied to study the effects of sequence variants on DNA accessibility. Here we collected a comprehensive set of 29 islet specific regulatory features derived from in-house and publicly available datasets, including DNA accessibility through ATAC-seq (whole islets and FACS-sorted alpha, beta and acinar cells), methylation, binding of several histone marks and transcription factors (TFs). We processed all data uniformly and used it for training of a convolutional neural network (CNN). The resulting CNN learned the islet specific regulatory patterns with mean area under curve (AUC) accuracy of 0.847 per feature (range: 0.713-0.976). The best predictive power was achieved for promoter associated features, such as unmethylated regions (AUC=0.976) and H3K4me3 (AUC=0.938), as well as TF binding, in particular for CTCF (AUC=0.956). Features related to heterochromatin or actively transcribed coding regions proved the most difficult to predict ($AUC_{H3K27me3}$=0.7529, $AUC_{H3K79me2}$=0.713). Convolution filters in the first CNN layer recovered binding motifs of several TFs with known function in the islets, including FOXA2, HNF1A, MAFB, NEUROD1 and RFX6. We then explored how variants in T2D GWAS credible sets (DIAGRAM: ~150,000 European subjects imputed to 1000 Genomes) influenced the predicted presence of islet regulatory features. For 394 of the 25,187 variants across 96 GWAS signals we observed a >10% prediction change for at least one of the features, and we found 9 variants with >50% change in prediction probability. Among those strongest predictions, we found that rs4338565 (*HMGA2* locus) resulted in disruption of a FOXA2 binding site, and concurrent gain of H3K9me3 repressive histone mark, while rs4742488 (*PTPRD* locus) led to loss of DNA accessibility, with largest loss predicted for the alpha cells, and mediated through disruption of the Maf recognition element (MARE) motif recognized by multiple TFs. These examples illustrate how deep learning techniques can shed light on the tissue and cell specific biological mechanisms underlying the T2D GWAS variants and aid in prioritization for follow-up studies.

321

Genome-wide association study of chronotype in 355,728 individuals identifies 106 genetic variants influencing circadian rhythms in humans. *S.E. Jones[1], J.M. Lane[2,3,4], V. van Hees[5], J. Tyrrell[1], R.N. Beaumont[1], K.S. Ruth[1], M.A. Tuke[1], H. Yaghootkar[1], Y. Hu[6], R.M. Freathy[1], A. Murray[1], K.V. Allebrandt[7], P.R. Gehrman[8], D.A. Lawlor[9,10], M.K. Rutter[11,12], R. Saxena[2,3,4,13], T.M. Frayling[1], A.R. Wood[1], D.A. Hinds[6], M.N. Weedon[1], the 23andMe Research Team.* 1) Genetics of Complex Traits, University of Exeter Medical School, Exeter, UK; 2) Center for Human Genetic Research Massachusetts General Hospital, Boston, Massachusetts 02114, USA; 3) Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA; 4) Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts 02142, USA; 5) Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam, Netherlands; 6) 23andMe Inc., Mountain View, California, USA; 7) Institute of Medical Psychology, Ludwig-Maximilians-University, Munich, Germany; 8) Perelman School of Medicine of the University of Pennsylvania, Philadelphia, PA, U.S.A; 9) MRC Integrative Epidemiology Unit at the University of Bristol, Bristol BS8 1TH, UK; 10) School of Social and Community Medicine, University of Bristol, Bristol BS8 1TH, UK; 11) Division of Endocrinology, Diabetes & Gastroenterology, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester M13 9PL, UK; 12) Manchester Diabetes Centre, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester M13 9PL, UK; 13) Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA.

Disrupted circadian rhythms are associated with several human diseases, particularly psychiatric disorders and metabolic disease. Chronotype (morning/evening person) is one aspect of an individual's circadian rhythm. Recently, the first genetic variants have been associated with self-report chronotype. In this study, we tripled the size of the self-report GWAS study and performed the first large-scale genetic study of objective measures of chronotype. We first performed a meta-analysis of genome-wide association studies of self-report chronotype in 355,728 research participants from 23andMe (N=248,094) and UK Biobank (N=107,634). We identified 106 independent genome-wide significant loci. To validate these, we derived objective measures of sleep using 7-day actigraphy data from ~26,000 genotyped UK Biobank individuals and calculated midpoint sleep as a chronotype analogue. Eighty-nine of 106 variants (binomial P=2.4x10^{-13}) were associated with midpoint sleep in the expected direction, and a genetic risk score of the 106 chronotype variants was associated with midpoint sleep (P=1.7x10^{-25}). Almost every known circadian rhythm gene was present amongst the 106 associated loci (e.g. *ARNTL, PER1, PER2, PER3, CRY1*). Genes in the photo-transduction cascade, visual processing (e.g. *RGS16, INADL*) and obesity pathways (e.g. *FTO*) were also present. Tissue enrichment analysis demonstrated that genes expressed in the visual sensory organs (particularly the retina), regions of the hindbrain, the hypothalamus and the central nervous system were overrepresented (P<0.05). LD score regression analyses identified genetic correlations between morningness and "subjective well-being" (r_G=0.20, P=7.1x10^{-8}), "depressive symptoms" (r_G=-0.19, P=2.0x10^{-7}) and Schizophrenia (r_G=-0.10, P=2.9x10^{-6}). We investigated the causal associations between chronotype and several health outcomes using Mendelian Randomisation. We found evidence that being a morning person is causal for lower risk of depression (MR-Egger OR=0.80, P=0.011). Conversely, although variants at the *FTO* obesity locus are strongly associated with chronotype (P=1x10^{-14}), we did not find evidence from MR that BMI causally influences whether an individual is a morning or evening person (P=0.15). In conclusion, we identified novel variants associated with self-report chronotype, validated them using objective measures from activity monitor data, and provide new insights into the biology of circadian rhythms and links to disease.

322

Genetic associations of low sleep hours in a sample of >130,000 subjects from the Million Veteran Program. *H. Zhao[1,2,3], N. Sun[2], Q. Lu[2], Y. Hu[2], B. Li[2], Q. Chen[1,2], M. Aslan[1,4], K. Radhakrishnan[1], K.H. Cheung[1,5], Y. Li[1,6], R. Pietrzak[7,8], N. Rajeevan[1,6], F. Sayward[1,6], K. Cho[9,10], K. Harrington[9,11], J. Honerlaw[9], S. Pyarajan[9,10], R. Quaden[9], J.M. Gaziano[9,10], J. Concato[1,4], M.B. Stein[12,13], J. Gelernter[3,7,8,14] on behalf of the VA Million Veteran Program.* 1) VA Clinical Epidemiology Research Center (CERC), VA Connecticut Healthcare System, West Haven, CT; 2) Department of Biostatistics, Yale University School of Public Health, New Haven CT; 3) Department of Genetics, Yale University School of Medicine, New Haven, Connecticut; 4) Department of Medicine, Yale University School of Medicine, New Haven, CT; 5) Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT; 6) Yale Center for Medical Informatics, Yale University School of Medicine, New Haven, CT; 7) Psychiatry Service, VA Connecticut Healthcare System, West Haven, CT; 8) Department of Psychiatry, Yale University School of Medicine, New Haven, CT; 9) Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA; 10) Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 11) Department of Psychiatry, Boston University School of Medicine, Boston, MA; 12) Psychiatry Service, VA San Diego Healthcare System, San Diego, CA; 13) Department of Psychiatry, University of California San Diego, San Diego, CA; 14) Department of Neuroscience, Yale University School of Medicine, New Haven, CT.

The Veterans Affairs (VA) Million Veteran Program (MVP) is in the process of building one of the world's largest medical and genetic information databases, and currently has >570,000 consented participants. Approximately 350,000 enrollees have curated genotype information available from a customized Affymetrix microarray, linked to VA electronic health record data and questionnaire responses—resulting already in the largest sample for studying posttraumatic stress disorder (PTSD)-relevant traits reported to date. Here we present results from a VA Cooperative Studies Program study (CSP #575B) that seeks to identify genetic risk factors relevant to PTSD and related traits in the U.S. veteran population; sleep disturbance frequently occurs in PTSD. We considered self-report data regarding the number of sleep hours, a phenotype based on responses to "How many hours do you usually sleep each day (24 hour period)?" with six options: "5 or less", "6", "7", "8", "9", and "10 or more". Our GWAS classified subjects as low sleepers "5 or less" compared to normal or long sleepers, "more than or equal to 7" (with 6-hour sleepers "unclassified"). After data cleaning, 118,563 European-American (EA) and 14,678 African-American (AA) were retained. In the EA sample, one SNP in the *CENPM* gene achieved genome-wide statistical significance (5.4E-09). We applied MetaXcan to calculate gene-level test statistics across 44 tissues and identified 1 significant gene after Bonferroni correction (*DESI1*, 8.9E-08). Based on the comparison between summary statistics from this study and those from published GWASs of 55 diseases or traits, 18 showed statistically significant evidence of genetic correlations, including education (3.7E-48), cognitive function (3.6E-36), age at first birth (2.2E-28), BMI (5.1E-16), depression (1.2E-13), and neuroticism (2.2E-11). No genome-wide significant associations were found in the much smaller AA sample. Our results suggest broad sharing of genetic factors between sleep hours and cognitive function, development, and psychiatric disorders.

323

Functional consequences of genetic loci associated with IQ in a meta-analysis of 87,740 individuals. *J.R.I. Coleman[1,2], J. Bryois[3], H. Gaspar[1], P. Jansen[4,5], J. Savage[4], N. Skene[6], R. Plomin[1], J. Hjerling-Leffler[6], P.F. Sullivan[3,7], D. Posthuma[4,8], G. Breen[1,2].* 1) Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, London, United Kingdom; 2) NIHR Biomedical Research Centre for Mental Health, South London and Maudsley NHS Trust, London SE5 8AF, UK; 3) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-17177 Stockholm, Sweden; 4) Department of Complex Trait Genetics, VU University, Center for Neurogenomics and Cognitive Research, Amsterdam, 1081 HV, The Netherlands; 5) Department of Child and Adolescent Psychiatry, Erasmus University Medical Center, Rotterdam, the Netherlands; 6) Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-17177 Stockholm, Sweden; 7) Departments of Genetics, University of North Carolina, Chapel Hill, NC, 27599-7264, USA; 8) Department of Clinical Genetics, VU University Medical Center (VUMC), Neuroscience Campus Amsterdam, Amsterdam, 1081 HV, The Netherlands.

Background Intelligence is a fundamental human characteristic, central to social interaction and frequently disrupted in mental illness. A recent GWAS meta-analysis identified 18 loci associated with IQ at genome-wide levels of significance. Adding data from a GWAS of extremely high IQ individuals to this meta-analysis, we integrated localised gene expression data to determine tissue-specific enrichment of genomic signatures in IQ. Method GWAS data from a population study of 78,308 individuals (Sniekers et al, 2017) was meta-analysed with an extreme-case sampling study of 1,247 individuals with IQ > 145 and 8,185 population controls (total Neff = 82,637). The enrichment of genetic associations to genomic annotations (using partitioned LDScore) and tissue-specific gene-expression levels (using MAGMA and LDScore) was assessed. Tissue-specific annotations were assessed body-wide and between brain tissues, using reference data from GTEx. Predicted effects of genetic variation at IQ-associated loci on gene expression in GTEx brain tissues was assessed using MetaXcan. Increased granularity of enrichment was obtained using brain cell-specific analyses from a large single-cell RNA sequencing reference dataset. Results Genomic loci significant in the IQ GWAS were enriched in regions conserved across mammalian species. Tissue-specific analyses demonstrated significant enrichment of brain-expressed genes in GWAS IQ loci using multiple methods. Between brain tissues, significant enrichment was identified in MAGMA analyses for the frontal cortex. Within the frontal cortex, *RNF123* and *RBM6* were significantly down-regulated and up-regulated respectively. Cell-specific analyses identified enrichment in pyramidal neurons of the somatosensory cortex and CA1 region of the hippocampus, as well as embryonic GABAergic neurons. Discussion Integrating results from GWAS of IQ with tissue- and cell-specific measures of gene expression implicates pyramidal neurons, and regions of the prefrontal cortex and hippocampus in IQ. One locus identified by GWAS may function to alter the expression of multiple genes, including *RNF123, RBM6* and *AMT*. Depending on the spatial and temporal effect of this locus, this could have downstream consequences for several processes, including glycine metabolism, dendrite development and the immune response. Such functional examination of GWAS results can generate hypotheses for further neurobiological exploration.

324

Large-scale genetic study of risk tolerance and risky behaviors identifies 29 new loci and reveals shared genetic influences. *J.P. Beauchamp[1], R. Karlsson Linnér[2,3], P. Biroli[4], M.A. Fontana[5], E. Kong[6], F. Meddens[2,3], R. Wedow[7], D.J. Benjamin[5,8], P.D. Koellinger[2,3,9], Social Science Genetic Association Consortium.* 1) University of Toronto, Toronto, Canada; 2) VU University Amsterdam, Amsterdam, the Netherlands; 3) Erasmus University Rotterdam, Rotterdam, the Netherlands; 4) University of Zurich, Zurich, Switzerland; 5) University of Southern California, Los Angeles, CA; 6) Harvard University, Cambridge, MA; 7) University of Colorado Boulder, Boulder, CO; 8) National Bureau of Economic Research, Cambridge, MA; 9) German Institute for Economic Research - DIW Berlin, Berlin, Germany.

Risk tolerance—defined as the willingness to take risks to obtain greater rewards—is an important variable in the behavioral sciences. Twin studies have established that risk tolerance is moderately heritable, and it is one of the most studied phenotypes in social science genetics. To date, however, few genetic variants have so far been found to robustly associate with it or with risky behaviors. We conducted genome-wide association studies (GWAS) of risk tolerance (n = 108,689) and of four risky behaviors: automobile speeding propensity, alcoholic drinks per week, whether one has ever been a smoker, and lifetime number of sexual partners (n = 93,625-186,102). Our GWAS identified three genome-wide significant loci associated with risk tolerance as well as a total of 26 additional new loci associated with the four risky behaviors or their first principal component. The set of the most significant loci associated with risk tolerance replicated in an independent meta-analysis of 59,747 individuals. Gene-based association analyses further identified six genes associated with risk tolerance, all of which are in or near the loci identified by our GWAS of risk tolerance. One of these genes, *CADM2*, has previously been associated with age at first sexual intercourse, BMI, and lifetime cannabis use. We report evidence of substantial pleiotropy between risk tolerance and the four risky behaviors. Risk tolerance is strongly genetically correlated with automobile speeding propensity (r_g = 0.46, $P = 5.6\times10^{-16}$), number of sexual partners (r_g = 0.57, $P = 1.2\times10^{-38}$), lifetime cannabis use (r_g = 0.37, $P = 2\times10^{-4}$), self-employment (r_g = 0.76, P = 0.03), and with neuropsychiatric disorders such as attention-deficit/hyperactivity disorder (r_g = 0.32, $P = 8.1\times10^{-9}$), bipolar disorder (r_g = 0.24, $P = 5.3\times10^{-5}$), and schizophrenia (r_g = 0.23, $P = 1.0\times10^{-10}$). Polygenic scores of risk tolerance predict a wide range of phenotypes, such as real-world risky behaviors, personality traits, and ADHD. Bioinformatics analyses revealed that variants regulating expression in the central nervous system and in adrenal or pancreatic tissues are enriched for association, but yielded no evidence of enrichment for biological pathways and candidate genes that have been the focus of prior genetics research on risk tolerance. Our results help to begin elucidate the genetic and environmental factors that underlie variation in risk tolerance and risky behaviors.

325

Profiling immunoglobulin repertoires by RNA sequencing across 8555 samples from 53 GTEx tissues. *H. Yang[1], S. Mangul[1,2], I. Mandric[3], N. Strauli[4], D. Montoya[5], J. Rotman[1], W. Van Der Wey[1], J. Ronas[6], B. Statz[1], A. Zelikovsky[3], R. Spreafico[2], S. Shifman[7], N. Zaitlen[8], M. Rossetti[9], M. Ansel[10], E. Eskin[1,11].* 1) Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA; 2) Institute for Quantitative and Computational Biosciences, University of California Los Angeles, Los Angeles, CA, USA; 3) Department of Computer Science, Georgia State University, Atlanta, USA; 4) Biomedical Sciences Graduate Program, University of California, San Francisco, CA, USA; 5) Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, Los Angeles, CA, USA; 6) Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, USA; 7) Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel; 8) Department of Medicine, University of California, San Francisco, CA, USA; 9) Immunogenetics Center, Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA; 10) Department of Microbiology and Immunology, Sandler Asthma Basic Research Center, University of California, San Francisco, San Francisco, CA, USA; 11) Department of Human Genetics, University of California Los Angeles, Los Angeles, USA.

Assay-based approaches provide a detailed view of the adaptive immune system by profiling immunoglobulin (Ig) receptor repertoires. However, these methods carry a high cost and lack the scale of standard RNA sequencing (RNA-Seq). Here we report the development of ImReP, a novel computational method for rapid and accurate profiling of the immunoglobulin repertoire from regular RNA-Seq data. ImReP can also accurately assemble the complementary determining regions 3 (CDR3s), the most variable regions of Ig receptors. We applied our novel method to 8,555 samples across 53 tissues from 544 individuals in the Genotype-Tissue Expression (GTEx v6) project. ImReP is able to efficiently extract Ig-derived reads from RNA-Seq data. Using ImReP, we have created a systematic atlas of Ig sequences across a broad range of tissue types, most of which have not been studied for Ig receptor repertoires. We also compared the GTEx tissues to track the flow of Ig clonotypes across immune-related tissues, including secondary lymphoid organs and organs encompassing mucosal, exocrine, and endocrine sites, and we examined the compositional similarities of clonal populations between these tissues. The Atlas of Immune Immunoglobulin repertoires (The AIR), is freely available at https://smangul1.github.io/TheAIR/ , is one of the largest collection of CDR3 sequences and tissue types. We anticipate this recourse will enhance future immunology studies and advance the development of therapies for human diseases. ImReP is freely available at https://sergheimangul.wordpress.com/imrep/ .

326

Genome-wide analysis of transcriptional and cytokine response variability in activated human immune cells. *S. Kim-Hellmuth[1,2], M. Bechheim[3], B. Puetz[4], P. Mohammadi[1,2], J. Schumacher[5], V. Hornung[3,6], T. Lappalainen[1,2].* 1) New York Genome Center, New York, NY; 2) Department of Systems Biology, Columbia University, New York, NY; 3) Institute of Molecular Medicine, University of Bonn, Bonn, Germany; 4) Statistical Genetics, Max Planck Institute of Psychiatry, Munich, Germany; 5) Institute of Human Genetics, University of Bonn, Bonn, Germany; 6) Gene Center and Department of Biochemistry, Ludwig-Maximilians-University Munich, Munich, Germany.

The immune system plays a major role in human health and disease. Understanding variability of immune responses on the population level and how it relates to susceptibility to diseases is vital. In this study, we aimed to characterize the genetic contribution to interindividual variability of immune response using genome-wide association and functional genomics approaches. For this purpose, we studied genetic associations to cellular (gene expression) and molecular (cytokine) phenotypes in primary human cells activated with diverse microbial ligands. We isolated monocytes of 134 individuals and stimulated them with three bacterial and viral components (LPS, MDP, and ppp-dsRNA). We performed transcriptome profiling at three time points (0 min/90 min/6 h) and genome-wide SNP-genotyping. In addition, we profiled five cytokines produced by peripheral blood mononuclear cells activated by five components from the same individuals to perform a genome-wide association study. Comparing expression quantitative trait loci (eQTLs) under baseline and upon immune stimulation revealed 417 immune response specific eQTLs (reQTLs). We characterized the dynamics of genetic regulation on early and late immune response, and observed an enrichment of reQTLs in distal *cis*-regulatory elements. Analysis of signs of recent positive selection and the direction of the effect of the derived allele of reQTLs on immune response suggested an evolutionary trend towards enhanced immune response. Furthermore, multivariate GWAS analysis of cytokine responses to diverse stimuli revealed 159 genome-wide significant loci; however, only a small number of these could be reliably linked to potentially causal eQTLs in monocytes. Finally, given the central role of inflammation in many diseases, we examined reQTLs as a potential mechanism underlying genetic associations to complex diseases. We uncovered novel reQTL effects in multiple GWAS loci, and showed a stronger enrichment of response than constant eQTLs in GWAS signals of several autoimmune diseases. These results indicate a substantial, disease-specific role of environmental interactions with microbial ligands in genetic risk to complex autoimmune diseases. While tissue-specificity of molecular effects of GWAS variants is increasingly appreciated, our results suggest that innate immune stimulation is a key cellular state to consider in future eQTL studies as well as in targeted functional follow-up of GWAS loci.

327

Multi-omics and deep-phenotyping integration predicts cytokine response to pathogens. *Y. Li[1], O.B Bakker[1], R. Aguirre-Gamboa[1], S. Sanne[2], M. Oosting[2], S. Smeekens[2], M. Jaeger[2], S. Withoff[1], R.J. Xavier[3,4], L.A.B. Joosten[2], V. Kumar[1], M.G. Netea[2], C. Wijmenga[1,5].* 1) Genetics, University Medical Centre Groningen, Groningen, Groningen, Netherlands; 2) Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, The Netherlands; 3) Center for Computational and Integrative Biology and Gastrointestinal Unit, Massa- chusetts General Hospital, Harvard School of Medicine, Boston, MA 02114, USA; 4) Broad Institute of MIT and Harvard University, Cambridge, MA 02142, USA; 5) Department of Immunology, University of Oslo, Oslo University Hospital, Rikshospitalet, 0372 Oslo, Norway.

Background. There is substantial inter- and intra-individual variation in immune response to infection. While both genetic and non-genetic factors are known to contribute to this variability, the relative contribution of various factors is poorly known, and it is difficult to predict the strength of the responses to infection in different individuals. Therefore, a comprehensive assessment of the different host and non-host factors contributing to the variation in immune response in humans is needed. **Methodology.** We investigated the immune system response to stimulation in ~500 Western-European healthy volunteers from the Functional Genomics Project (http://www.humanfunctionalgenomics.org/site/) by measuring *ex-vivo* production of 6 cytokines after exposure to 18 human pathogens or stimuli. Immune, molecular and clinical phenotypes were determined, and we assessed their correlation with host factors, including genome, metabolome, transcriptome, and microbiome. We thereafter integrated the overlap between the immune system response to pathogens and the genetic risks for immune- related diseases. Lastly, we applied machine-learning approaches to built predictive models for cytokine responses by integrating multi-omics and baseline immune profiles. **Results.** In-depth analysis of baseline immune parameters, molecular profiling and cytokine production capacity in 500 healthy subjects shows large inter- individual variability in immune system functionality, but also clear patterns of co-regulated components Among 91 different cytokine-pathogen pairs, 10 categories of host factors explain up to 67% (on average 36%) of observed inter-individual variation in cytokine response. Data integration of these results with genome-wide association studies demonstrates that individuals with high genetic risk for autoimmune diseases tend to be high producers in cytokine response to pathogens. Human cytokine response can be accurately predicted using only genetics, with prediction performance up to 0.80. Information on multi- omics and baseline immune parameters increases the prediction only slightly up to 0.82. **Conclusions** We demonstrate that multi-omics and deep-phenotyping data can be integrated to predict the immune response. This prediction model could be used to identify individuals with modified ability to react to pathogens or immune stimuli, hence at high risk of infections or immune-mediated diseases. .

328

Tensor decomposition for myeloid and brain tissue gene expression identifies neurodegenerative disease associated *trans*-eQTLs. *S. Ramdhani, T. Raj.* Ronald M. Loeb Center for Alzheimer's Disease, Departments of Neuroscience, Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai Hospital.

A steadily growing number of studies have identified and characterized *cis* expression quantitative trait loci (eQTLs) in human primary cells and tissues. However, identifying distal regulation on gene expression (*trans*-eQTLs) are far more difficult to detect because of their smaller effect size and large number of tests for thousands of transcripts. Here, we extended a Tensor Decomposition Method (Hore et. al., 2016) to uncover multi-tissue gene networks linked to distal genetic variation (trans-eQTLs). The tensor decomposition method uses the Parallel Factor Analysis (PARAFAC) (Caroll and Chang [1970]) within a Bayesian framework. All estimation is done via Variational Bayes with the sparsity element coming from a select prior. We apply this method to several large-scale gene expression datasets including 367 peripheral monocytes, stimulated with Interferon (IFN) and LPS from 261 individuals. We extended the method to identify multi-omics networks using gene expression, splicing, histone, methylation datasets from prefrontal cortex tissue of 461 subjects. At FDR < 0.10, we identified 11, 16 and 43 unique gene networks or components associated with genetic variants in baseline, LPS and IFN, respectively. We observed several molecular features of the *trans*-QTLs including enrichment of RNA-binding proteins (for *trans*-splicing QTLs) and members of signaling proteins (i.e., *IRF7*). When exploring the impact of disease-associated variation, we identified 1,161 *trans*-eQTL's (with 283 unique components) in primary and stimulated monocytes across 388 complex traits at FDR 10%. We report robust evidence that some disease-associated variants affect expression of multiple genes in *trans:* the AD-associated SNP rs983392 (*cis* effect on *MS4A4A/6A*) increased expression of over 36 genes including several in interferon signaling pathway (*IRF1*, *IRF7*, *IRF8*, etc.), innate immune response (*ABI3*, *CD2AP*) and complement cascade (*C1QB*, *C1QC*, *C2*). We identified a PD-associated SNP rs823118 (*cis* effect on *RAB7L1*) impacts expression of over 26 genes involved in response to innate immune processes (i.e., *CCL5*, *CCR7*, *CD80*, *CX3CR1*, *ICAM3*, *IRF1*). These results suggest the important role for innate immune-mediated mechanisms for Alzheimer's and Parkinson's disease. In summary, our approach represents a powerful framework for understanding the effect of genetic variants on gene networks contributing to disease, and may help in elucidating the underlying biology of complex disease.

329

Predicting splicing directly from sequence using a neural network model of splice site competition. *D.A. Knowles[1], Y.I. Li[1], K.K. Fahr[2], J.K. Pritchard[1,3].* 1) Stanford University, Stanford, CA; 2) Illumina Inc., San Diego, CA; 3) Howard Hughes Medical Institute, Stanford University, Stanford, CA.

RNA splicing is dysregulated in many human diseases across the spectrum from rare (e.g. Duchenne muscular dystrophy) to complex (e.g. multiple sclerosis). Approaches using computational methods and reporter assays have been used to discover splicing regulatory elements, i.e. sequence motifs associated with alternative splicing levels (e.g. percent spliced-in, or "PSI") of known cassette exons. However, predicting usage of splice junctions from DNA sequence alone, i.e. learning a model of intron/exon-definition, remains particularly challenging. We developed Grasshopper, a deep learning method to predict splice site (SS) usage from DNA sequence. We model splicing outcomes as a series of SS choices. For every 5' SS we model the choice of 3' SS out of all AG dinucleotides within 100kbp downstream, quantified using using split reads from RNA-seq. We learn a mapping from sequence to SS usage rate using a hybrid neural network consisting of a convolutional neural network (CNN) on 800bp of sequence context centered at each AG combined with a dense network on a smaller sequence context (80bp). A Dirichlet-multinomial likelihood models competition between available SS while accounting for count overdispersion. An analogous model is learned for 5' SS choice. To test the performance of Grasshopper, we used GTEx RNA-seq data, holding out odd-numbered chromosomes, and could correctly predict 70% of the most frequently used 3' SS (out of thousands of AGs within 100kb). This corresponds to a drastic improvement compared to only the 9% when using MaxEntScan. Of interest, features learned de novo by the model include the canonical branch point, 5' SS and 3' SS motifs, the polypyrimidine tract and the GAAGA exonic splicing enhancer. We also used our model to predict changes in SS usage across mammalian evolution. We predicted the correct direction for 89% of SS with observed changes in usage proportions greater than 50%. Most notably, we were able to use in-silico mutagenesis to identify candidate mutations underlying the observed cross-species differences. Finally, we used Grasshopper to interpret putative rare disease variants: rare or de novo variants may disrupt splicing by creating novel canonical splice junction dinucleotides (GT or AG). Using whole-genome sequencing from GTEx we obtain an AUC of 99% in predicting novel AG that are used as 3' SS at some frequency in corresponding RNA-seq data. This opens a new avenue for prioritizing variants in rare disease settings.

330

Prioritizing functional rare variants: The impact of rare variation on alternative splicing. *B.J. Strober[1], J. Davis[3,4], M. Grove[4], Y. Kim[2], P. Parsana[2], E. Tsang[3,4], J. Merker[4], S. Montgomery[3,4], A. Battle[2].* 1) Biomedical Engineering, Johns Hopkins University, Baltimore, MD; 2) Department of Computer Science, Johns Hopkins University, Baltimore, MD, 21218, USA; 3) Department of Genetics, Stanford University, Stanford, CA, 94305, USA; 4) Department of Pathology, Stanford University, Stanford, CA, 94305, USA.

Rare genetic variation is abundant in the human genome, yet interpreting and predicting its functional effects remains a significant challenge. Previous efforts have shown functional rare variants are enriched among individuals that exhibit abnormal expression of a gene relative to the population, particularly in the genomic region near that gene. In this study, we develop a novel, probabilistic framework to identify individuals that exhibit abnormal alternative splicing, rather than total expression. For this, we utilized whole genome sequencing and multi-tissue RNA-sequencing data across 344 European-ancestry individuals from the Genotype-Tissue Expression (GTEx) project. For each gene, we fit a dirichlet-multinomial distribution directly to split-read counts of alternatively spliced junctions, and then label an individual as a “splicing outlier” if that individual deviates significantly from expectation based on this fitted dirichlet-multinomial. Compared to normal individuals, splicing outliers are greater than 100 times more likely to harbor a rare variant in the regions surrounding splice sites, suggesting rare variation often underlies abnormal splicing events. Total expression outliers, as well as splicing outliers identified from a heuristic-based approach, do not yield enrichment of this magnitude. Functionally, splicing outliers are enriched for splicing variants, evolutionarily conserved variants, and variants nearby splice sites. We then applied RIVER, a probabilistic graphical model developed in previous work, to integrate information from both genomic annotations and splicing outlier status of an individual to predict the probability a rare variant has a functional effect on splicing. We used held out pairs of individuals to evaluate the accuracy of our model. RIVER performed significantly better than using genomic annotations alone. We are now applying RIVER to rare disease individuals from the Clinical Genomics Service (CGS) to establish splicing outlier mediated links between rare variation and rare disease. Additionally, we plan to enhance RIVER to allow it to capture patterns of splicing outlier sharing across tissues, as well as account for confounding effects, such as a nearby structural variant. Overall, we demonstrate that rare variants contribute to aberrant splicing events and we describe a novel framework to prioritize functional rare variants contributing to changes in splicing.

331

Investigating non-coding variants: Characterizing potential disease causing variants on microRNA binding sites. *C. Lau, T. Frisby, N. Balanda, M. Malicdan, B. Pusey, W. Gahl, D. Adams.* NIH Undiagnosed Diseases Program, National Institutes of Health, Bethesda, MD.

The NIH Undiagnosed Diseases Program (UDP) enrolls patients with diseases that remain undiagnosed despite extensive diagnostic evaluation. Exome and genome sequencing are frequently used by the UDP as part of our effort to find a diagnosis for our patients. Approximately 25% of the over 1000 enrolled patients eventually received a diagnosis at the UDP either by clinical acumen or by exome or genome sequence analysis. As part of our on-going effort to arrive at a diagnosis for the undiagnosed, we are studying the previously uncharacterized non-coding variants from exome or genome data to identify potential disease-causing genetic lesions. In one approach, we have detected a set of previously uncharacterized non-coding variants located on potential microRNA (miRNA) binding sites. As a group, miRNAs play critical roles in regulating gene expression by binding to targets commonly found in the 3’UTR of transcripts. Abnormalities in miRNA biogenesis, regulation, and recognition of targets have increasingly been linked to clinical conditions. To characterize these variants, we utilize principles of Mendelian consistency and population genetics to identify variants in 3’UTRs that segregate with diseases and that are rare in the general population. We then adapted miRNA target predicting tools including miRanda, PITA and Targetscan to predict the differences in binding affinities of miRNA to wildtype (WT) and variant binding sites, and selectively prioritize those variants with significant differences in miRNA binding affinities when compared to WT. By studying known OMIM disease-causing genes, we found that disease-causing genes that are dose-sensitive contain the most highly conserved miRNA binding sites. Similarly, we found that genes with low tolerance to variation in non-coding regions contain the most highly conserved miRNA binding sites. We hypothesize that evolutionary conservation of miRNA binding sites is directly correlated to their physiological function. We thus utilize the principles of dosage sensitivity and variant tolerance combined with evolutionary conservation to further prioritize candidate variants. To validate these findings, we have developed a molecular probe method to measure the extent of miRNA induced silencing complex (RISC) related mRNA degradation. The studies described here enable us to define a pool of non-coding variants that constitute new potential targets for functional assessment and correlation with our patients’ illnesses.

332

Annotating pathogenic non-coding variants in genic regions. *S. Gelfman[1,2], Q. Wang[1,2], K.M. McSweeney[1,2], Z. Ren[1,2], F. La-Carpia[3], M. Halvorsen[1,2], K. Schoch[4], F. Ratzon[5], E.L. Heinzen[1,3], M.J. Boland[1,6], S. Petrovski[1,7], D.B. Goldstein[1,2].* 1) Institute for Genomic Medicine, Columbia University Medical Center, New York, NY; 2) Department of Genetics and Development, Columbia University Medical Center, New York, NY; 3) Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY; 4) Department of Pediatrics, Duke University Health System, Durham, NC; 5) Department of Pathology, Lenox Hill Hospital, New York, NY; 6) Department of Neurology, Columbia University, New York, NY; 7) Department of Medicine, Austin Health and Royal Melbourne Hospital, University of Melbourne, Melbourne, Australia.

Identifying the underlying causes of disease requires accurate interpretation of genetic variants. Current methods ineffectively capture pathogenic non-coding variants in genic regions, resulting in overlooking synonymous and intronic variants when searching for disease risk. Here we present the Transcript-inferred Pathogenicity (TraP) score, which uses sequence context alterations to reliably identify non-coding variation that causes disease. High TraP scores single out extremely rare variants with lower minor allele frequencies than missense variants. TraP accurately distinguishes known pathogenic and benign variants in synonymous (AUC=0.88) and intronic (AUC=0.83) public datasets, dismissing benign variants with exceptionally high specificity. TraP analysis of 843 exomes from epilepsy family trios identifies synonymous variants in known epilepsy genes, thus pinpointing risk factors of disease from non-coding sequence data. TraP outperforms leading methods in identifying non-coding variants that are pathogenic and is therefore a valuable tool for use in gene discovery and the interpretation of personal genomes.

333

Joint imputation of gene expression in 44 tissues identifies context-specific associations for complex traits. *Y. Hu[1], M. Li[1], Q. Lu[1], S. Muchnik[2], J. Wang[3], H. Zhao[1,2,3,4].* 1) Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA; 2) Department of Genetics, Yale School of Medicine, New Haven, CT, USA; 3) Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA; 4) VA Cooperative Studies Program Coordinating Center, West Haven, CT, USA.

Gene-level association tests based on GWAS summary statistics and multi-tissue eQTL studies have successfully discovered novel trait-gene associations and replicated known ones in recent years, providing a powerful approach to studying the genetic architecture of complex traits. The key idea behind this approach is building a gene expression prediction model from tissue specific expression-genotype data, which is however challenging mainly due to the limited sample size of expression profile for a single tissue. In this project, we proposed MTIE, a multi-task learning approach to jointly predict the gene expression in 44 tissues using the GTEx data (n=450). Comparing with the single-tissue elastic net model, our model achieved both consistent and significant improvement in prediction accuracy across all tissues, especially for tissues with moderate sample sizes (p=0.04~1.3E-8; n<150) and generated significant prediction models for a considerably higher number of genes (13% ~ 300% increment across tissues). Applied to 50 complex traits with publicly accessible GWAS summary statistics (n=4.5 million), we detected 18,928 significant trait-gene-tissue associations (compared with 12,208 detected by MetaXcan). In addition to the improved total number of associations, MTIE could identify more genes in trait-related tissues with small sample sizes. In Alzheimer's disease (AD), we not only replicated known risk genes in multiple tissues such as *APOE* (p=4.6E-9~3.2E-76), but also identified many novel and significant genes in brain and liver tissues. In particular, *APOC4* (p=2.5E-9) is specifically abundant in liver, which is consistent with results in mouse model and supports the notion that lipid metabolism is critically involved in AD etiology. In order to perform a more comprehensive and powerful gene-level test, we have also developed a joint testing framework to combine single-tissue results while explicitly taking tissue similarity into account. We applied the joint test to AD and schizophrenia and identified 44 and 218 significant genes, respectively, including 8 and 57 genes that were not identified by single tissue test. Through jointly modeling multi-tissue gene expression, cross-tissue expression correlation, and GWAS summary statistics, MTIE substantially increases the power in identifying both tissue-specific and ubiquitous gene-trait associations, and provides systematic and novel insights into the genetic basis of many complex human traits.

334

Transcriptome-wide association studies are vulnerable to false positives due to co-regulation. *M. Wainberg[1], N. Sinnott-Armstrong[2], D. Knowles[2,3], D. Golan[2], H. Tang[2], M. Rivas[4], A. Kundaje[1,2].* 1) Department of Computer Science, Stanford University, Stanford, CA; 2) Department of Genetics, Stanford University, Stanford, CA; 3) Department of Radiology, Stanford University, Stanford, CA; 4) Department of Biomedical Data Science, Stanford University, Stanford, CA.

Imputed transcriptome-wide association studies (TWAS) are a promising new family of methods for integrating GWAS and reference expression data to discover genes associated with complex traits. We show that existing TWAS methods are vulnerable to false positives resulting from cis co-regulation: if two genes are co-regulated by the same variants, or variants in linkage disequilibrium (LD) with each other, then both genes may appear to be associated with the trait even if only one is causal. Gene-trait association testing based on Mendelian Randomization (MR) is also vulnerable to false positives because cis co-regulation, as a form of pleiotropy, violates one of the core assumptions of MR. The more tightly a pair of genes is co-regulated in cis, the more difficult it becomes to distinguish causality based on GWAS and expression data alone. We find that the effect of cis co-regulation in TWAS is analogous to that of LD in GWAS. Rather than treating every TWAS association as causal, we consider groups of hit genes as “co-regulation blocks”, by analogy with LD blocks, and fine-map the causal gene(s) within each co-regulation block. In a GWAS for age-related macular degeneration (AMD), we show that while TWAS identifies 36 gene-trait associations, these correspond to only 8 distinct co-regulation blocks, most with only a single causal gene according to our fine-mapping protocol. We identify several novel candidate genes for AMD with roles in programmed cell death and retinal development, including genes at sub-significant GWAS loci and one gene at a known GWAS locus, *BCAR1*, with stronger literature evidence than the canonical nearest genes to the locus, *CTRB1/2*. TWAS begins by building a predictive model of each gene’s expression from genotype on a reference panel of individuals. Existing TWAS methods train gene expression models on all variants within up to 1Mb of the gene. Yet relatively few genes have strong eQTLs at such a distance, and restricting the feature space to a smaller window often improves predictive accuracy. We partition reference panel individuals into training, validation and test sets and adaptively select the window size on a per-gene basis to maximize validation set accuracy, and show that this improves test set accuracy by over 50%, which translates into a 33% increase in the number of co-regulation blocks. Together, our results put TWAS on a more secure methodological footing and increase their utility for complex trait mapping.

335

Meta-prediction of gene expression levels from genotypes across multiple tissues and datasets. *A. Liu, H.M Kang.* Biostatistics, University of Michigan, ANN ARBOR, MI.

Transcriptome wide association studies (TWAS) can be used as a powerful method to understand the underlying biological mechanisms behind GWAS by mapping gene expression levels with phenotypes. In TWAS, gene expression is often imputed from individual-level genotypes of known cis-regulatory genetic variants, and prediction models trained from external resources, such as GTEx data. In this setting, a most straightforward approach to predict expression levels of a specific tissue is to use the model trained from the corresponding tissue type. When multiple prediction models trained from different tissue types are available, prediction models from multiple tissue types could improve the prediction accuracy because of shared eQTLs between the tissues and increase in effective sample size. Here, we explore the optimal way to combine predicted expression levels across multiple tissues and datasets for a target tissue and population. We first produced 45 tissue-specific sets of predicted expression levels for 465 GEUVADIS lymphoblast cell lines (LCL) samples from their genotypes, trained using prediXcan models from GTEx tissues and DGN whole blood samples. Among these predicted expressions, the best prediction accuracy for the European population (n=344) came from the model based on GTEx LCLs, with 1,858 genes being predicted with significant accuracy (p-value<0.05 compared to null distribution) among 3,611 predicted genes. We produced multi-tissue predicted expression as linear combinations of the 45 tissue-specific predicted expression sets, using predictive R-squared as weights. Our multi-tissue prediction accurately predicted 3,832 genes, 2-fold larger than single tissue only. This was mainly due to the number of predictable genes being increased from 3,611 to 14,126 in the multi-tissue setting. Among the 3,611 GTEx LCL genes, the number of accurately predicted genes were comparable (n=1,799). In the African population (n=77), only 619 genes were accurately predicted from LCLs due to smaller sample size. Here, multi-tissue prediction outperformed LCL prediction by accurately predicting 636 genes among cis-eQTLs in LCL, and 421 additional genes that did not have cis-eQTLs in LCL. We also extend our meta-prediction method to meta-TWAS to leverage multiple tissues in TWAS analysis with summary-level statistics. Our results capitalize on the importance of integrating multiple datasets and tissues to unravel regulatory impacts of genetic variants on complex traits.

336

Variant-sensitive prediction of RNA expression using convolutional neural networks. *M. Abdalla[1,2,3], M. Abdalla[4], C.C. Holmes[1,2], M.I. McCarthy[1,3].* 1) Wellcome Trust Centre for Human Genetics, University of Oxford, United Kingdom; 2) Department of Statistics, University of Oxford, United Kingdom; 3) Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, United Kingdom; 4) Department of Computer Science, University of Toronto, Canada.

Most variation associated with common disease lies in non-coding regions of the genome. Modelling the effect of this variation on gene expression remains an unsolved problem. Here we demonstrate that the *cis*-component of RNA expression in lymphoblastoid cell lines (n=117) can be modelled as a function of 1*Mb* sequences centred on the transcription start sites, without *a priori* specification of expression quantitative trait loci (eQTLs), using a tailored end-to-end trained deep convolutional neural network (CNN). CNNs are feed-forward artificial neural networks that enforce local connectivity patterns between neurons in adjacent layers to leverage spatially-local dependencies in the input sequences. This modified CNN architecture, called pBrain, can predict absolute expression levels (driven by invariant differences between genes; using *withheld* genes as test cases, 10-fold cross-validated r^2=0.23). Incorporating genomic and epigenetic annotations into the input sequence further improves performance (10-fold cross-validated r^2=0.63). The pBrain model can be extended to predict variation in expression (*i.e.* direct measure of *cis*-heritability between *nearly-identical* gene sequences). To interrogate the sensitivity of pBrain, we selected a subset of highly-reproducible eQTLs with high-throughput experimental validation (from publicly available sources). When restricted to 2kb upstream flank (for compute reasons in this pilot study), pBrain achieves comparable performance in predicting eQTL effects to that obtained by experimental methods, such as massively parallel experimental assays (Spearman's rho=0.38; p=0.002); indicating the architecture has the potential to identify functional variants underlying eQTLs. Moreover, by computing the gradient of pBrain's predictions with respect to the input sequences while holding the weights fixed, we can generate saliency maps that visualize which positions of the input sequence contribute to the prediction. Overlaying these maps with genomic annotations reveals how different sequence elements positively and negatively regulate gene expression. We further highlight the utility of the model in transcriptome-wide association studies, in interrogating the consequences of epigenetic modifications by measuring the *in silico* impact of histone (de)acetylation and (de)methylation, and in elucidating clusters motivated by shared transcriptional regulation (using the neural activations of the penultimate layer).

337

Aggregation of population-based genetic variation over protein domain homologues strongly improves diagnostic prediction of missense variants. *L. Wiel[1,2], H. Venselaar[2], J.A. Veltman[3,4], G. Vriend[2], C. Gilissen[3].* 1) Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboudumc, Nijmegen, Noord-brabant, Netherlands; 2) Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, the Netherlands; 3) Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, the Netherlands; 4) Institute of Genetic Medicine, International Centre for Life, Newcastle University, Newcastle upon Tyne, United Kingdom.

We investigated whether combining information from homologous domains across different genes improves variant interpretation. For this purpose we developed a framework that maps population variation from the Exome Aggregation Consortium (ExAC) and pathogenic mutations from the Human Gene Mutation Database (HGMD) onto 5,250 Pfam protein domains that cover 41% of all protein coding sequences. We note that 71% of pathogenic missense variants from HGMD are present in a protein domain. We aggregated population-based and disease-causing genetic variants across 30,853 homologous protein domains into 2,750 meta-domains. We find that genetic tolerance is consistent across 97% of domain homologues in different genes ($p < 0.05$, Bonferroni corrected), and that patterns of genetic tolerance faithfully mimic patterns of evolutionary conservation (Pearson 0.97, p-value < 1e-308). Interestingly, we find that 2,201 of the aggregated domain positions (0.47%) are not evolutionary conserved, but still highly intolerant to normal variation while also containing one or more disease causing missense variants. Via repeated Monte Carlo experiments we find that for 22% of meta-domains high-frequency population variation re-occurs at the same positions within the domain across homologues. Residues that are not evolutionary conserved, but nevertheless depleted of population-based variation are especially strong predictors of sites of pathogenic missense variants. An informative example concerns domain positions 17 and 21 in the "EGF-like domain" (PF00008) that are depleted of population-based variation in 60 genes and cause disease in at least 3 of these genes (*NOTCH3*, *JAG1*, and *CRB2*). Summing across all genes, we find that for 24% of meta-domains, pathogenic missense variants re-occur at the same aligned position in domain homologues significantly more often than could be expected by random chance.

338

De novo missense mutation clustering identifies candidate neurodevelopmental disorder genes. *S.H. Lelieveld[1], L. Wiel[2], H. Venselaar[2], R. Pfund[1], G. Vriend[2], J.A. Veltman[1,3], H.G. Brunner[1,4], L.E.L.M. Vissers[1], C. Gilissen[1].* 1) Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands; 2) Centre for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, the Netherlands; 3) Institute of Genetic Medicine, International Centre for Life, Newcastle University, Newcastle upon Tyne, United Kingdom; 4) Department of Clinical Genetics, Maastricht University Medical Centre, Maastricht, the Netherlands.

Haploinsufficiency (HI) is the most common mechanism through which dominant mutations exert their effect and cause disease. Typically, these pathogenic mutations are spread throughout the gene, and result in absence of protein product. In contrast, non-haploinsufficiency (NHI) mechanisms, such as gain-of-function and dominant-negative mechanisms, are often characterized by the spatial clustering of missense mutations within a gene, thereby affecting only particular regions or base pairs of a gene. Here we exploit this property and developed a method to specifically identify genes with significant spatial clustering patterns of *de novo* mutations in large patient cohorts. We applied our method to a dataset of 4,061 *de novo* missense mutations from published exome studies of patient-parent trios with (neuro)developmental disorders (NDDs) and identified 15 genes with statistical clustering of mutations. Among the 15 genes that we identified there was a strong enrichment for known NDD genes (12 out of 15, p=1.65e-04) thereby supporting our approach. Strikingly, 11 out of these 12 genes are known to act through a disease mechanisms other than HI and for 8 of these known genes we found extensive functional evidence supporting NHI mechanisms. Interestingly, the identified genes are significantly less tolerant to population variation than known HI genes (p=8.59e-03) and are significantly depleted for truncating mutations (p<1.00e-05) in the patient cohort. The 3 genes that were not previously linked to NDDs (*ACTL6B*, *GABBR2* and *PACS2)* are involved processes that are known to be disrupted in NDDs. Through a collaboration we found that multiple patients with the same *PACS2* mutation have been independently identified that all share clinical phenotypes. Finally, we performed 3D-modeling of protein structures to show that, unlike known HI genes, clustering mutations are unlikely to affect protein folding and more likely to disturb protein interactions/complex formation (p=1.26E-03). In conclusion, we developed a method for the identification of disease genes based on the significance of spatial mutation clustering within a gene. We identified three genes with similar clustering patterns that we propose as candidate NDD genes. Our findings support the notion that these mutations mostly exert their pathogenic effect through disease mechanisms other than HI.

339

Missense mutations disrupting the ATPase domain of CHD3 cause a novel neurodevelopmental syndrome with intellectual disability, macrocephaly and impaired speech and language. *L. Snijders Blok[1,2], J. Rousseau[3], J. Twist[4], S. Ehresmann[3], L. Faivre[5], J. Thevenon[5], M. Assoum[5], L. Rodan[6], C. Nowak[6], J. Douglas[6], K.J. Swoboda[7], M.A. Steeves[7], I. Sahai[7], C.T.R.M. Stumpel[8], P. Wheeler[9], M. Willing[10], E. Fiala[10], A. Kochhar[11], W.T. Gibson[12], A.S.A. Cohen[12], R. Agbahovbe[12], J. Rankin[13], I.J. Anderson[14], S. Skinner[15], R. Louie[15], H. Warren[15], A. Afenjar[16], R. Lewandowski[17], J. Propst[17], M. Choi[18], J.H. Chae[18], S. Price[19], M. Cho[20], C. Zweier[21], A. Reis[21], M. Bialer[22], C. Moore[22], M. Swinkels[23], E.H. Brilstra[23], G.R. Monroe[23], G. van Haaften[23], R. Newbury-Ecob[24], the DDD study[25], L .D. Shriberg[26], P. Deriziotis[2], T. Kleefstra[1], H.G. Brunner[1], M. Takaku[4], J.D. Roberts[4], R.M. Petrovich[4], S. Machida[27], H. Kurumizaka[27], P.A. Wade[4], S.E. Fisher[2], P.M. Campeau[28].* 1) Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands; 2) Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands; 3) CHU Sainte-Justine Research Center, Montreal, Canada; 4) National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA; 5) CHU de Dijon, Dijon, France; 6) Boston's Children's Hospital, Boston, USA; 7) Massachusetts General Hospital, Boston, USA; 8) Maastricht University Medical Center, Maastricht, the Netherlands; 9) Nemours Childrens Clinic, Orlando, FL, USA; 10) Washington University School of Medicine, St. Louis, MO, USA; 11) Valley Children's Hospital, Madera, CA, USA; 12) BC Children's Hospital, University of British Columbia, Vancouver, BC, Canada; 13) Royal Devon and Exeter NHS Trust, Exeter, UK; 14) University of Tennessee Medical Center, Knoxville, TN, USA; 15) Greenwood Genetic Center, Greenwood, SC, USA; 16) Armand Trousseau Hospital, Paris, France; 17) VCU Medical Center, Richmond, VA, USA; 18) Seoul National University, Seoul, Korea; 19) Oxford University Hospitals NHS Foundation Trust, Oxford, UK; 20) GeneDx, Gaithersburg, MD, USA; 21) Institute of Human Genetics, FAU Erlangen-Nürnberg, Erlangen, Germany; 22) Northwell Health, Manhasset, NY, USA; 23) University Medical Center Utrecht, Utrecht, the Netherlands; 24) University Hospital Bristol, Bristol, UK; 25) Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; 26) Waisman Center, Madison, WI, USA; 27) Waseda University, Tokyo, Japan; 28) Sainte-Justine Hospital, University.

CHD3 is a member of the CHD subfamily of chromatin remodelers, and chromatin remodeling is of crucial importance during neurodevelopment. Subunit exchange of CHDs occurs in the core ATPase subunits of the NuRD complex and is required for distinct aspects of cortical development. More specifically, CHD4 promotes the early proliferation of progenitors, CHD5 facilitates neuronal migration and CHD3 ensures proper layer specification. Upon whole genome sequencing of a cohort of children with a primary diagnosis of Childhood Apraxia of Speech (CAS), we discovered an individual with a *de novo* missense variant in *CHD3*. Through an international search, we collected a total of 25 patients with *de novo* mutations in this gene. These patients have 18 different mutations in *CHD3*; 16 missense mutations, one canonical splice site mutation and one frameshift mutation. The majority of the missense mutations in *CHD3* cluster within the ATPase/helicase domain. FLAG-CHD3 constructs were overexpressed, purified and used in radiometric ATPase assays with recombinant nucleosomes or naked dsDNA. Some mutations decreased the activity (R1121P, R1172Q, N1159K), while others did not affect the ATPase activity (R1187P) or increased it (L915F) in these assays. Further experiments are underway to assess additional possible mechanisms for the deleterious effect of the mutations on development, such as altered interaction with other proteins or altered ability to remodel chromatin and regulate gene expression in cells. All 25 individuals with *de novo* mutations in *CHD3* show developmental delays, varying from borderline impairment to severe intellectual disability. Many patients have persistent hypotonia. Behavioural problems are frequent and include autism spectrum disorders and ADHD. In some individuals, speech and language impairment exceeded what would be expected based on their level of general cognitive impairment. Macrocephaly with enlarged CSF spaces, and hypertelorism with frontal bossing were often present. In conclusion, *de novo* mutations in *CHD3* cause a novel neurodevelopmental syndrome, characterized by intellectual disability, macrocephaly and impaired speech and language. *De novo* loss of function mutations in *CHD2, CHD7,* and *CHD8* each cause specific forms of neurodevelopmental delay, while *de novo* missense mutations in *CHD4* cause intellectual disability and macrocephaly. Thus, these conditions may reflect impairment of normal NuRD complex activity during cortical development.

340

Findings from the Critical Assessment of Genome Interpretation, a community experiment to evaluate phenotype prediction. *G. Andreoletti[1], R.A. Hoskins[1], J. Moult[2], S.E. Brenner[1], CAGI Participants.* 1) University of California Berkeley, CA; 2) IBBR, University of Maryland, Rockville, MD.

The Critical Assessment of Genome Interpretation (CAGI, 'kā-jē) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. CAGI participants are provided genetic variants and make predictions of resulting phenotype. Independent assessors evaluate the predictions by comparing with experimental and clinical data. CAGI challenges thus far have included predicting the biochemical impact of non-synonymous variants and of the impact of noncoding regulatory variants on gene expression; predicting the impact of mutations in cancer driver genes on cell growth; prediction of individuals' complex trait status based on exome data; matching personal genomes to phenotypic trait profiles; and matching variant data to clinical diagnoses. Results from the CAGI experiments are described in a collection of 20 articles, shortly to appear in a special issue of *Human Mutation*. There have been notable discoveries throughout the CAGI experiments, and general themes have emerged. Some examples: For a number of challenges, independent assessment has found that top missense prediction methods are highly statistically significant, but individual variant accuracy is limited. Missense methods also tend to correlate better with each other than with experiment (for reasons that may reflect biases in the predictive methods but also in the experimental assays). Although overall missense accuracy is limited, there are a subset of variants where methods may be sufficiently reliable to provide strong evidence for clinical use. Protein three-dimensional structure-based missense methods do well in a few cases, while sequence-based methods have more consistent performance. Bespoke approaches often enhance performance. Interpretation of non-coding variants shows promise but is not at the level of missense. In challenges using clinical data predictors have been able to identify causal variants that were overlooked in the initial clinical pipeline analysis. The results have also highlighted possible diagnostic ambiguities. Additionally, the results suggest that running multiple uncalibrated methods and considering their consensus may result in undue confidence in a pathogenic assignment, so we advise against this procedure. Detailed information about CAGI may be found at https://genomeinterpretation.org.

341

Development of a pipeline to support translational research: Incorporating multiple data sources to select genetic pharmacomimetics and implement PheWAS in UK Biobank. *K. Song[1], M. Chiano[2], B.H. Dessailly[2], A. Pandey[1], T. Johnson[2], M.G. Ehm[1], M.R. Nelson[1], J.C. Whittaker[2], R.A. Scott[2], L.M. Yerges-Armstrong[1].* 1) Target Sciences, GlaxoSmithKline, King of Prussia, PA; 2) Target Sciences, GlaxoSmithKline, Stevenage, UK.

The phenome-wide association study (PheWAS) is an important tool to support the use of genomics in therapeutic target validation. It can be used to validate disease associations and identify previously unsuspected associations with other traits, providing a richer understanding of biological and disease mechanisms. Scaling up PheWAS to support large-scale target validation efforts is not without challenge. Selection of informative variants which mimic the pharmacological perturbation of specific drug targets (or tool variants) for study is often a manual process with some targets being better served by single SNP instruments than others. Phenotype diversity is often limited to a single data type (eg. self-report or EHR data) or aggregations of individual, relatively small studies, and approaches are required to standardise diverse phenotypes in high-throughput fashion. Inference of associations across differentially powered and correlated phenotypes and interpretation of associations in their genomic context provides an additional layer of challenge. With this in mind, we developed an integrated PheWAS approach to conduct target validation analysis at scale on participants from UK Biobank. Potentially informative variants for study were selected by collating non-synonymous coding variants and significant sQTL and eQTL variants from GTEx, as well as multi-SNP cis-based predictors of gene expression levels. Disease classifications were defined in 133,397 genotyped participants from UK Biobank by truncated ICD10 codes from linked hospital episode statistic (HES) data, self-report, nurses interview, and where possible, combinations of those data sources to improve power. Rank-normalised quantitative traits were also analysed, resulting in a set of 2,085 phenotypes for analysis by linear or logistic regression. Our pipeline also provides standard output to facilitate inference, including Manhattan plots and forest plots to compare effect sizes across related and distinct phenotypes. Regional plots and approximate conditional analyses also enable investigation of the genomic context of index variants and independence of signals from surrounding associations. We will describe the construction and implementation of the PheWAS approach, as well as extensions incorporating the ability to test genetic scores and perform burden tests to further develop the utility of PheWAS in target validation.

342

Big data distributed system for phenome and genome management and analysis in a large health system. *X. Liu, W.S.W. Wong, P. Kothiyal, W. Zhu, F. Zhou, S. Gao, S. Madhappan, L. Smith, H. Hunter, A. Black, J.F. Deeken, J.E. Niederhuber.* Inova Translational Medicine Institute, Fairfax, VA.

The increasing use of High Throughput Sequencing (HTS) and the scale of data produced along with the huge footprint of various public data resources can quickly overwhelm the bioinformatics analysis paradigm based on traditional clusters and relational databases. Innovative “Big data” solutions built on the open-source Apache Hadoop and Spark cluster computing technology have been employed to address this challenge. ADAM (http://bdgenomics.org/projects/adam/) and Hail (https://hail.is) are two of the cutting-edge projects in the area of big data genomics. Although both projects are much more efficient in processing and storing the vast genomic data, they are not user-friendly for end users who lack a strong bioinformatics background. The steep learning curve of grasping Hadoop and Spark programming is a challenge even for bioinformaticians. To leverage these powerful new tools while considering the practical applications to support Inova Health System’s translational genomic research, we are building an integrated system composed of a Hadoop data warehouse (DW) with Cloudera Impala as the backend, an ETL (Extraction, Transformation, Loading) workflow using ADAM and Spark, an analysis platform middle tier powered by Spark and Hail, and a Web front-end for ad hoc query and interactive data analysis. Currently, our Impala DW stores the variants from over 2200 whole genome sequenced family trios (over 8000 individuals), ~6000 transcriptomes (mRNA and small RNAs from over 4000 individuals at different time points), ~2400 methylomes, public and internal annotation databases, as well as clinical phenotypes extracted and anonymized from Inova’s Electronic Medical Record (EMR). In-house pipelines are used to efficiently extract and filter genotypes, gene expression, DNA methylation, and clinical phenotypes. Phenome-Wide Association Studies (PheWAS) are then carried out for each genomic and epigenomic locus. Furthermore, our system allows standard machine learning algorithms to be applied to the data with simple workflows. Final results are stored in the Impala DW. We are currently building a Web front-end to present researchers and clinicians PheWAS results and their associated annotations. Lastly, we provide an R front-end powered by Spark for interactive and advanced statistical analysis. Examples on several use cases are presented to demonstrate the power of our integrative big data genomic system.

343

A wellness study of 108 individuals using dense, dynamic, personal data clouds. *A. Magis[1], N.D. Price[1,2], J.C. Earls[2], G. Glusman[2], R. Levy[2], C. Lausted[2], D.T. McDonald[2], U. Kusebauch[2], C.L. Moss[2], Y. Zhou[2], S. Qin[2], R.L. Moritz[2], K. Brogaard[1], M. Conomos[1], G.S. Omenn[2,3], J.C. Lovejoy[1,2], L. Hood[1,2].* 1) Arivale, Inc, Seattle, WA; 2) Institute for Systems Biology, Seattle, WA; 3) Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI.

In order to understand the basis of wellness and disease, we and others have pursued a global approach termed ‘systems medicine’. The defining feature of systems medicine is the collection of diverse longitudinal data for ‘healthy’ individuals, to assess both genetic and environmental determinants of health and their interactions. We report the generation and analysis of longitudinal multi-omic data for 108 individuals over the course of a 9-month study called the Pioneer 100 Wellness Project (P100). This study included whole genome sequencing; and at three time points measurements of 218 clinical laboratory tests, 643 metabolites, 262 proteins, and abundances of 4,616 operational taxonomic units in the gut microbiome. Participants also recorded activity using a wearable device (Fitbit). Using these data, we generated a multi-omic correlation network and identified communities of related analytes that were associated with physiology and disease. We demonstrate how connectivity within these multi-omic communities identified known and candidate biomarkers in an unsupervised manner; e.g., the metabolite gamma-glutamyltyrosine and the protein FGF21 were densely interconnected with clinical analytes associated with cardiometabolic disease. We calculated polygenic scores for 127 traits and diseases using effect estimates from published genome-wide association studies (GWAS). By including these polygenic scores in the multi-omic correlation network, we identified molecular correlates of polygenic disease risk in undiagnosed individuals. For example, the polygenic risk score for inflammatory bowel disease (IBD) was negatively correlated with plasma cystine in this unaffected cohort. Abnormally low levels of cystine have previously been associated with IBD in case-control studies. Our results suggest that lower levels of blood cystine may be more common in individuals with higher genetic risk for IBD, even before the disease manifests. We have expanded these analyses to a larger cohort of more than 2000 individuals enrolled in a commercial wellness program (Arivale, Seattle, WA) for whom longitudinal multi-omic data have been collected. Using this larger cohort, we identified additional associations with cumulative genetic risk for disease in a ‘healthy’ population, including Alzheimer’s disease. These results illustrate how multi-omic longitudinal data will improve understanding of health and disease, especially for the detection of early transition states.

344

Statistical methods and computational algorithms to enable robust and efficient cloud-scale joint variant calling of >60,000 deeply sequence genomes. *H. Kang, J. LeFaive, A. Tan, C. Scheller, T. Blackwell, G. Abecasis.* Biostatistics Dept, Univ Michigan, Ann Arbor, Ann Arbor, MI.

The amount data produced by the ultra-high-throughput sequencing has now surpassed 100,000 deeply sequenced genomes, ten quadrillon sequenced bases, 20 petabytes of raw sequence data (uncompressed). Joint variant calling at this scale must be performed on the cloud computing environment to ensure timely delivery of the callset. Existing algorithms and tools designed for joint variant calling are no longer adequate to handle this scale of data. Here we present key statistical methods and computational algorithm that allows us to scale the joint variant calling on the cloud across ~65,000 deeply sequenced genomes for the Trans-Omics Precision Medicine (TOPMed) studies. Our new version of Genomes on the Cloud (GotCloud) variant caller capitalizes on several improvements in the sequence analysis and variant calling procedures. First, we leverage the contamination and ancestry estimates obtained from the QC steps into the genotype calling procedure to increase the genotype accuracy. In deep sequence data, a small amount of contamination (e.g. 1%) can create frequent genotyping errors. Our methods reduce the Mendelian errors by 4-fold compared to the previous methods, particularly on common or population-specific variants. Second, we leverage the genetic ancestry and familial relatedness to robustly filter variant sites. Certain genetic features of variant sites, such as Hardy-Weinberg Equilbrium test statistics are very useful metrics to identify potentially false variants, but it poorly behaves in a large joint variant callset that contains heterogeneous populations. We developed variant filtering methods that leverages the population structure and inferred familial relatedness to modify the variant features so that the filtered variant sites are much more specific to true variant sites while retaining important population-specific variants (such as the Duffy blood group variant; rs2814778) across the common and rare allele frequency spectrum. Finally, we efficiently scaled our methods by the number of sequenced genomes on the cloud. We re-implemented our variant consolidation and joint genotyping steps to be hierarchically structured and perform I/O sequentially, so that minimum amount of memory and disk operations are required while processing >60,000 genomes and >400 million potential variant sites together. Our modified software architecture reduced the computational cost of variant calling and turnaround time by orders of magnitudes.

345

Assessing the effect of a range of modifiable risk factors on overall cancer risk: A Mendelian randomization study. *J. Ong[1,2], J. An[1], G. Chenevix-Trench[1], P. Gharahkhani[1], S. MacGregor[1].* 1) Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Herston, Brisbane, Queensland, Australia; 2) School of Medicine, Herston, University of Queensland.

The identification of modifiable risk factors is crucial in public health efforts to control cancer. Observational and Mendelian randomization (MR) studies have shown that for specific cancers, some modifiable risk factors are important in determining cancer risk. However, the patterns of association can vary considerably across different types of cancer; for example, increased body mass index increases the risk of esophageal adenocarcinoma but decreases the risk of breast cancer. It is hence unclear how modifying risk factors will affect the overall cancer burden. Traditional observational studies have addressed this but issues such as confounding and reverse causality make interpretation difficult. Using 42,000 cancer cases and 188,000 controls from GERA and the UK BioBank Pilot study, we performed MR analyses to investigate the effect of various modifiable risk factors on overall cancer risk. Instrumental variables for each risk factor were extracted from the literature or estimated using UK Biobank data. As proof of principal, we used MR to show that increased height was associated with overall cancer risk (P=2e-6, with larger effect in females), consistent with observational studies. We then tested obesity, fasting glucose, cholesterol levels, coffee and alcohol consumption; all were not associated with cancer risk and for all apart from coffee and alcohol consumption, our power was high enough to allow us to rule out all but small effects of these traits on cancer risk. Lower red blood cell count was associated with cancer risk although this result was not significant after correction for multiple testing. We tested for sex specific effects but found none except for height (Causal OR in females= 1.09, p-value<1e-5 per SD increase in height; Causal OR in males= 1.05, p-value=0.04). For risk factors specific to women, we found a nominally significant causal association between menopausal status and overall cancer risks (not significantly after correction for multiple testing). Our analyses reinforce the connection between height and cancer risks established by previous observational studies. At the meeting we will present results on the full UK Biobank cohort – this will either further strengthen our null results for traits such as coffee and alcohol consumption or allow us to show there is a moderate (and potentially important in public health terms) effect of these traits on overall cancer risk. .

346

Evaluating cancer risk in children with non-chromosomal birth defects: A population-based registry linkage study and family-based genetic analysis. *S. Sisoudiya[1], H.E. Danysh[2,3], M.E. Scheurer[2,3], A. Brown[2,3], A. Sabo[4], S.E. Plon[1,2,3], P.J. Lupo[2,3].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 3) Texas Children's Cancer Center, Texas Children's Hospital, Houston, TX; 4) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX.

Introduction: The association of chromosomal anomalies with increased childhood cancer risk is well-established. However, less is known about the role of non-chromosomal birth defects on cancer risk. In the Genetic Overlap Between Congenital Anomalies and Cancer in Kids (GOBACK) Study we have a two-pronged approach to this question: (1) a population-based registry linkage study to identify novel birth defect-childhood cancer (BD-CC) associations, and (2) a family-based sequencing study to evaluate the underlying genetic mechanisms. Methods: We obtained linked BD-CC registry data from Texas (discovery) and Michigan (replication) for the period 1996-2009 to identify novel associations. Cox proportional hazard models were used to calculate hazard ratios (HRs) and 95% confidence intervals (CIs). Analyses were restricted to those children without chromosomal anomalies. Children with both cancer and birth defects (and their parents) are being recruited from the Texas Children's Cancer Center and using statewide cancer registries. Genetic analyses including whole-genome sequencing (WGS – 40X coverage) is performed and data analyzed using Platypus to identify *de novo* and compound heterozygous variants, and structural rearrangements. Results: We identified a cohort of 5,275,792 live births in Texas and 2,566,771 live births in Michigan. Children with non-chromosomal birth defects had an increased risk for childhood cancer (Texas HR=2.27, 95% CI: 1.98-2.60; Michigan HR=3.24, 95% CI: 3.02-3.47). Eighteen of the 35 (51%) birth defects evaluated in Texas were significantly associated with childhood cancer including: patent ductus arteriosus (HR=3.18, 95% CI: 2.46-4.12) and spina bifida (HR=7.59, 95% CI: 4.20-13.72). Thirteen of these associations (72%) replicated using Michigan data. Children with spina bifida were particularly likely to develop renal tumors (HR=10.33, 95% CI: 1.45-73.55). Cancers strongly associated with birth defects including: germ cell tumors (HR=3.94, 95% CI: 2.07-7.49) and hepatic tumors (HR=9.65, 95% CI: 5.66-16.47). To date, 65 families are enrolled in GOBACK and WGS completed on 12 trios. Conclusion: Our population-based analysis of over seven million live births demonstrates that children with non-chromosomal birth defects have an elevated risk of cancer with novel, recurring birth defect/cancer patterns. Work is underway to complete molecular characterization of these children with combined developmental and cancer susceptibility conditions.

347

Capture-recapture method gives estimate of the number of families in Central Ohio with Lynch syndrome. *B.H. Shirts[1], J.M.O. Ranola[1], R. Pearlman[2], H. Hampel[2].* 1) Laboratory Medicine, University of Washington, Seattle, WA; 2) Division of Human Genetics, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center, OH.

Background: Lynch syndrome (LS) causes dominantly inherited increased risk of colorectal, endometrial, and other cancers. Universal screening of all colorectal cancers (CRC) for LS has been shown to be a cost effective way to prevent colon cancer as this can lead to identification of families with unaffected high-risk individuals and increased surveillance. With increased surveillance colorectal cancer caused by LS can be virtually eliminated. Accurate estimates of the number and size of families with Lynch syndrome could inform policymakers about the relative costs and benefit of universal screening measures for Lynch syndrome. **Capture-Recapture Method:** In this study we use the capture-recapture or mark-recapture method from ecology to estimate the number of families in central Ohio with LS. Briefly, in ecology animals would be caught, marked, and released (m1). At a later time, animals would again be caught and released (n1) noting how many marked animals were recaught (m2). The total population is then calculated as n = m1*n1/m2. **Application to Hereditary Disease:** We screened 1566 CRC cases in central Ohio from 1999-2005. We again screened 2510 CRC cases in central Ohio from 2013-2016. The unit of interest was a family not an individual. A family was considered "marked" in one of two ways, by having a known family relationship with another case or by having an assumed relationship defined by a shared rare genetic variant. **Results:** The first screening identified 44 cases with LS and the second identified 94 with LS. We found 12 LS variants shared between at least one individual identified in the first and second studies. We further found that 3 to 8 individuals in the first study were related to individuals in the second study depending on the degree of relatedness. We estimate there are between 800 and 1500 families in the area that have LS. There are about 250 mutations present in the area that cause LS, only 93 of which were identified by this study. With continued universal screening it would take about 26 years to detect 50% of LS families and about 95 years to detect 90% of LS families. **Conclusions:** This is the first time, to our knowledge, that the capture-recapture method has been applied to human genetics to estimate the burden of specific heritable disease variants in a specific population. This method may be used for genetic epidemiology studies of other diseases.

348

Identification of novel breast cancer risk genes using a gene-based analysis of regulatory variants. *J. Beesley[1], M. Ferreira[1], W. Shi[1], F. Al-Ejeh[1], P. Kraft[2,3], W. Zheng[4], A. Antoniou[5], D.F. Easton[5,6], G. Chenevix-Trench[1] on behalf of BCAC and CIMBA.* 1) QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia; 2) Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 3) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 4) Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee; 5) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 6) Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK.

Genome-wide association studies (GWAS) of breast cancer have identified >150 loci harbouring risk alleles. Most candidate causal risk variants at GWAS signals are non-coding and therefore likely to act through gene regulation which hinders causal gene identification. Our aim was to integrate information from expression quantitative trait loci (eQTLs) across multiple tissues to identify potential target genes of breast cancer risk variants. We hypothesised that breast cancer risk genes are regulated by eQTLs operating not just in breast tissue, but also in the immune system and adipose tissue. Breast cancer risk variants were identified in the largest meta-analysis to date (the OncoArray), totalling 69,501 estrogen receptor positive (ER+) cases, 21,468 estrogen receptor negative (ER-) cases, and 105,974 controls, all of European ancestry. Genes with sentinel eQTLs in linkage disequilibrium with breast cancer risk variants were identified using EUGENE (PMID 27554816), a gene-based test that aggregates information across multiple independent eQTLs, considering both *cis*- and *trans*-effects. In a combined analysis of the three tissues across relevant datasets, we identified 216 genes (188 protein-coding and 28 non-coding) associated with overall breast cancer risk at 63 known loci, and 52 genes at 18 loci associated with risk of ER- breast cancer (gene-based significance level $P < 3.4 \times 10^{-6}$). These included previously validated targets *ABHD8*, *ESR1* and *RMND1*. At loci not yet found by GWAS, we identified eight genes at seven loci (*DNPH1*, *FLOT1*, *GSTM4*, *HSF2*, *METTL10*, *SIK3*, *TMEM205*, and *ULK3*) as *cis*-regulated signals. We knocked down seven genes for which high expression levels were associated with risk by siRNA and assayed proliferation in normal breast epithelial and breast cancer cell lines. Knockdown of *ALS2CR12*, *AP006621.6*, *PIDD1*, *RP11-15A1.7*, *STXBP4* and *ZNF404*, but not *RMND1*, reduced proliferation and/or colony formation. *RMND1* is a known target of the 6q25 risk locus (PMID 26928228) and may operate through another hallmark of cancer. These findings provide evidence for putative target genes at about half of the known breast cancer risk loci, suggest regulatory activity may occur in additional tissues types beyond breast epithelial cells and highlight loci potentially associated with risk as yet undiscovered by GWAS. They also implicate *ALS2CR12*, *AP006621.6*, *PIDD1*, *RP11-15A1.7*, *STXBP4* and *ZNF404* as novel breast cancer oncogenes underlying cancer risk.

349

Germline heterozygous mutations in *KDM3B* cause a syndrome with intellectual disability and short stature. *I.J. Diets[1], K. Baltrunaite[2], E. Waanders[3], M.R.F. Reijnders[1], R. Pfundt[1], S. Bergevoet[4], A. Vulto-van Silfhout[1], G. Beunders[5], J. Thevenon[6], L. Perrin[6], B. Keren[7], A. Afenjar[7], C. Nava[7], S. Bartz[8], B. Peri[8], N. Verbeek[9], K. van Gassen[9], H. Brunner[1], B. van der Reijden[4], V. Hwa[2], T. Kleefstra[1], N. Hoogerbrugge[1], A. Dauber[2], R.P. Kuiper[1,3], M. Jongmans[1,9].* 1) Department of Human Genetics, Radboud University Medical Center and Radboud Institute for Molecular Life Science, Nijmegen, The Netherlands; 2) Cincinnati Center for Growth Disorders, Division of Endocrinology, Cincinnati Children's Hospital Medical Center, Cincinnati, USA; 3) Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands; 4) Department of Laboratory Medicine, Laboratory of Hematology, Radboud University Medical Center, Nijmegen Center for Molecular Life Sciences; 5) Department of Clinical Genetics, University Medical Center Groningen, The Netherlands; 6) Centre de Génétique et Centre de Référence Anomalies du Développement et Syndromes Malformatifs de l'Inter-région Est, Centre Hospitalier Universitaire Dijon, Dijon, France; 7) Service de génétique, CHU Paris Est - Hôpital d'Enfants Armand-Trousseau, Paris, France; 8) Division of Endocrinology, Children's Hospital of Colorado, Aurora, Colorado, USA; 9) Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands.

Whole exome sequencing has facilitated the discovery of many novel genes causing intellectual disability (ID) and an increased risk for hereditary cancer. Using whole exome sequencing in a cohort of patients with childhood cancer and additional features pointing towards genetic predisposition, we identified mutations in *KDM3B* in two patients with ID and a hematological malignancy. By exploration of an in-house database and through GeneMatcher, we identified 11 additional patients with mutations in *KDM3B*. The mutations (2 nonsense and 8 missense) were *de novo* in 9 patients and inherited from a similarly affected parent in 2 patients. The patients share a specific phenotype that includes ID/developmental delay, short stature, feeding difficulties in infancy, joint hypermobility, behavior problems and characteristic facial features such as a broad mouth, pointy chin, large ears, low columnella and downslanted palpebral fissures. Two patients developed cancer in childhood, namely acute myeloid leukemia and Hodgkin lymphoma, respectively. *KDM3B* is located on chromosome 5q31, a region which is frequently deleted in both myeloid leukemia and myelodysplasia, and it has been suggested as the candidate tumor suppressor gene in this region. *KDM3B* contains a Jumonji-C domain and is involved in H3K9 demethylation, a crucial part of chromatin modification required for transcriptional regulation. Mutations in several components of chromatin modification pathways have been found to cause syndromes characterized by ID, and deregulation of chromatin remodeling is also observed in many cancers. We identified both missense and truncating mutations, suggesting haploinsufficiency as the most likely mechanism. The mutations cluster in three regions of the protein and in-silico analysis showed a probable pathogenic effect of the missense mutations. Also, *KDM3B* is a gene which is extremely intolerant for normal variation. To further explore the effect of the mutations on *KDM3B* function, we performed immunoblot analyses for mono-, di- and tri-demethylation status. We show that mutant KDM3B loses its ability to efficiently demethylate H3K9me3, therefore leading to a disturbance of chromatin modification. In conclusion, mutations in *KDM3B* cause a syndrome characterized by ID and short stature, and might increase the risk of malignancies.

350

***De novo* germline variants in *Histone 3 Family 3A (H3F3A)* and *Histone 3 Family 3B* (*H3F3B*) associated with a severe neurodegenerative disorder with unique functional effect different from somatic mutations.** *E.J. Bhoj.* Children's Hospital of Philadelphia, Philadelphia, PA.

Histones are nuclear proteins that associate with DNA packaged it into condensed chromatin. They are dynamically decorated with post-translational modifications (PTMs), which regulate processes like DNA repair, gene expression, and mitosis/meiosis. The specific Histone 3 Family 3 histones (H3.3), encoded by *H3F3A* and *H3F3B*, mark active genes, maintain epigenetic memory, and maintain heterochromatin and telomeric integrity. Specific somatic mutations in *H3F3A* have been associated with pediatric tumors, but no germline mutations have been described. Here we report 23 patients, ages 4 months to 32 years, with *de novo* missense germline mutations in *H3F3A* or *H3F3B* who share a core phenotype of progressive neurologic dysfunction and congenital anomalies, but no malignancies yet. All patients have mild to profound developmental delay, and some also have seizures, developmental regression, congenital heart disease, Craniosynostosis, but no recognizable facial gestalt. These 16 mutations in 23 patients, are all *de novo* and not found in large population datasets. There are three recurrent mutations in our cohort; p.R18G and p.A115G each in two unrelated patients, and p.T46I in four unrelated patients. We hypothesized that these missense mutations contribute to the shared patient phenotype through epigenetic dysregulation of histone PTMs. Histones PTMs within the nucleosome affect chromatin state, mitotic initiation, and gene expression. We analyzed histones from lymphoblasts and fibroblasts from several *H3F3A* and *H3F3B* patients by mass spectroscopy (MS) and demonstrated that the mutant histone proteins are present at a level similar to that of wild-type H3.3. We quantified PTMs on mutant histones and demonstrated strikingly aberrant patterns of local, but not global, dysregulation of histone PTM. These data suggest that the pathogenic mechanism of germline histone mutations is distinct from that of the global dysregulation of cancer-associated somatic histone mutations. RNA-Seq on patient lymphoblast and fibroblasts showed a statistically significant upregulation of genes related to mitosis and cell division. Fibroblast lines derived from multiple unrelated patients showed increased proliferative capacity compared to normal human fibroblast control lines, which may contribute to the phenotype. We anticipate that mechanistic understanding of this novel syndrome will provide insight into new therapeutic targets to prevent the neurologic degeneration in these patients.

351

Haploinsufficiency of the chromatin-remodeling bromodomain PHD finger transcription factor *BPTF* leads to developmental delay, microcephaly, and dysmorphic features. *P. Stankiewicz[1,2,*], T.N. Khan[3,*], P. Szafranski[1], L. Slattery[4], J.A. Bernstein[4], C.W. Brown[5,6], B. Bostwick[1], H. Streff[1], S. Rednam[7,8,9], S. Scollon[1], K.L. Bergstrom[1], D.W. Parsons[1,7,8,9], S.E. Plon[1,7,8,9], M.W. Vieira[10], C.R.D.C. Quaio[11], W.A.R. Baratela[11], J.C. Acosta Guio[12], R. Armstrong[13], S.G. Mehta[13], C.A. Bacino[1,2], R. Xiao[1,2], A.M. Breman[1,2], J.L. Smith[1,2], M.E. Hurles[14], N. Katsanis[3,15], E.E. Davis[3], Y. Yang[1,2].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Baylor Genetics, Houston, TX; 3) Center for Human Disease Modeling, Duke University Medical Center, Durham, NC; 4) Department of Pediatrics, Division of Medical Genetics, Stanford University, Stanford, CA; 5) University of Tennessee Health Science Center, Memphis, TN; 6) Le Bonheur Children's Hospital, Memphis, TN; 7) Texas Children's Hospital, Houston, TX; 8) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 9) Texas Children's Cancer Center, Texas Children's Hospital, Houston, TX; 10) PUCSP, Faculdade de Ciências Médicas e da Saúde, São Paulo, Brazil; 11) Fleury Medicina e Saúde, São Paulo, Brazil; 12) Especialista en Genética Médica. Instituto de Ortopedia Infantil Roosevelt, Bogotá, Cundinamarca, Colombia; 13) East Anglian Medical Genetics Service, Clinical Genetics, Addenbrooke's Treatment Centre, Addenbrooke's Hospital, Cambridge, UK; 14) Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK; 15) Department of Cell Biology, Duke University Medical Center, Durham, NC. *Equal contribution.

Chromatin-remodeling, an essential process regulating DNA accessibility and gene expression, is controlled by four conserved ATP-dependent protein complexes: SWI/SNF, ISWI, CHD, and INO80. The ISWI family member NURF (Nucleosome Remodeling Factor) is a key regulator of development conserved through phylogeny. In humans, NURF consists of SNF2L (ISWI homolog), pRBAP46/48, and the largest subunit BPTF (Bromodomain PHD finger transcription factor). Although dysregulation of chromatin-remodeling has been identified in multiple developmental disorders as well as in cancer, thus far, no germline variants in *BPTF* have been definitively linked to human disease. Here, we report one missense and six loss-of-function variants involving brain-expressed *BPTF* on 17q24.2 in unrelated individuals (three Caucasian and four Latino) aged between 2.1-13 years, who manifest developmental delay (DD)/intellectual disability (ID) (7/7), microcephaly (5/6), and dysmorphic features (6/7). Three frameshift changes, c.2860dup (p.E954fs), c.5216_5217delTGva (p.V1739fs)), and c.3360_3370+1del, one nonsense variant, c.8650A>T (p.K2884X), and two small (88 kb and 196 kb in size) copy-number variant deletions are all predicted to disrupt *BPTF*. In five subjects, the variants arose *de novo*; the origin of the other two changes could not be determined. Using CRISPR-Cas9-genome editing of *bptf* in zebrafish to induce a loss of gene function, we observed a significant reduction in a defined region of the head of 3-day post-fertilization (dpf)bptf CRISPR F0 mutants compared to control larvae. Using TUNEL and phospho-histone H3 (PH3) staining to assess apoptosis and cell proliferation, respectively, we found a significant increase in cell death in F0 mutants compared to control. Finally, we injected bptf single guide (sg)RNA and Cas9 into -1.4col1a1:egfp transgenic embryos and observed a significant increase of the ceratohyal angle of the craniofacial skeleton in *bptf* CRISPR F0 mutants. Importantly, the phenotype of larvae injected with bptf sgRNA alone (without Cas9) was indistinguishable from that of controls, excluding the possibility of sgRNA toxicity-induced phenotypic differences. Our studies demonstrate that haploinsufficiency of *BPTF* results in DD/ID, microcephaly, and dysmorphic features. Interestingly, *BPTF* is also disrupted in microcephalic individuals with 17q24.2 deletions, demonstrating that haploinsufficiency of *BPTF* also contributes to the Stankiewicz-Isidor syndrome (MIM 617516). .

352

Germline mutations on the histone H4 core cause a developmental syndrome by affecting DNA damage response and cell cycle control. *G. van Haaften[1,3], F. Tessadori[2], J. Giltay[1], J. Hurst[4], M. Massink[1,3], K. Duran[1,3], H.R. Vos[3], R.M. van Es[3], D.D.D. Study[5], R. Scott[4], K. van Gassen[1], J. Bakkers[2,6].* 1) Department of Medical Genetics, University Medical Center Utrecht, Utrecht, Utrecht, Netherlands; 2) Hubrecht Institute-KNAW and University Medical Center Utrecht, Utrecht, The Netherlands; 3) Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands; 4) North East Thames Regional Genetics Service, Great Ormond Street Hospital, London, UK; 5) the Deciphering Developmental Disorders Study; 6) Department of Medical Physiology, Division of Heart and Lungs, University Medical Center Utrecht, Utrecht, the Netherlands.

Covalent modifications of histones have an established role as chromatin effectors, as they control processes such as DNA replication, transcription and repair or regulate nucleosomal structure. Loss of modifications on the histone N-tails, whether due to mutations in genes belonging to histone-modifying complexes or directly affecting the histone tails, cause developmental disorders or play a role in tumorigenesis. More recently, modifications affecting the globular histone core are being uncovered as crucial for DNA repair, pluripotency and oncogenesis. Here we report monoallelic missense mutations affecting Lysine 91 in the histone H4 core (H4K91) in three individuals with a syndrome of growth delay, microcephaly and intellectual disability. The human genome contains fifteen histone H4 genes, all differing at the nucleotide level but encoding an invariant H4 protein. RNA sequencing analysis of patient cells showed that ±8% of H4 cDNA molecules carried the mutated allele. We observed differentially expressed histone genes and cell-cycle related genes compared to controls, suggesting an effect on processes such as DNA replication or cell cycle progression. Analysis of the chromatin fraction of patient fibroblasts by mass spectrometry revealed that 1-2% of histone H4 molecules contained the mutated residue. Expression of the H4 mutants in zebrafish embryos recapitulate the developmental anomalies seen in the patients. We show that the H4 mutations cause genomic instability, resulting in increased apoptosis and cell cycle progression anomalies during early development. Mechanistically, our findings indicate an important role for the ubiquitination of H4K91 in genomic stability during embryonic development. Our results highlight the functional importance of the histone core and establish H4K91 and its modifications in the realm of human genetic disorders. Loss of Lysine 91 on histone H4 acts in a genetically dominant manner. On a biological level, our data presented here point at a mechanism involving inherent DNA damage accumulation and early perturbed cell cycle through which missense mutations affecting K91 are causative for an identifiable syndrome consisting of dysmorphic features and intellectual disability.

353

Pathogenic and therapeutic epigenetic modulation of a novel super-enhancer at the *TGFB2* locus in systemic sclerosis. *J.Y. Shin, H.C. Dietz.* Johns Hopkins University, Baltimore, MD.

Systemic Sclerosis (SSc) is a rare and complex disorder characterized by adult-onset predisposition for progressive fibrosis of the skin and viscera with high mortality. Disease associates with an overt inflammatory prodrome and sustained auto-inflammation, but it is unclear if this is a marker or driver of disease. In the absence of a strong genetic signature in twin studies (monozygous=dizygous=4% concordance) or GWAS (only indicative of inflammatory predisposition), there are few pathogenic insights and no specific treatment for SSc. Using RNAseq, we found that primary dermal fibroblasts (PDFs) from SSc patients maintain a strong fibrotic synthetic repertoire (FSR) after many passages in culture (e.g. *COL1A1*, *ACTA2*). This correlated with specific upregulation of TGFβ2 (but not β1 or β3) mRNA and protein expression that was prone to further amplification by TGFβ treatment. Use of a TGFβ receptor kinase inhibitor or a specific TGFβ2-neutralizing antibody fully silenced the FSR in SSc cells. Together, these data suggest a mechanism to “lock in” the FSR in SSc, with particular relevance for TGFβ2. We posited epigenetic regulation of gene expression as an integrator of both genetic and environmental triggers. ATACseq revealed an open chromatin conformation for a sequence-constrained region just distal to the *TGFB2* gene in SSc PDFs, with direct correlation between accessibility and TGFβ2 mRNA levels. This element was enriched for acetylated H3K27 and occupancy by the histone acetyltransferase (HAT) EP300. CRISPR-based targeting of HAT activity to this element was sufficient to induce or accentuate TGFβ2 expression in control or SSc PDFs, respectively, validating functional enhancer status. Treatment of SSc cells with a HAT inhibitor (HATi) was sufficient to normalize TGFβ2 expression and silence the FSR, but full rebound occurred within 24hrs after drug removal. These findings were suggestive of super-enhancer priming that can be initiated by inflammatory effectors (e.g. NF-kB) and enforced by BRD4 recruitment. In support of this hypothesis, we found NF-kB binding sites and high BRD4 occupancy at the implicated TGFβ2 enhancer in SSc PDFs. Treatment with the BRD4 inhibitor JQ1 normalized TGFβ2 expression and the FSR, which was now refractory to drug removal. These data inform the pathogenesis of SSc including an initiating role for inflammation, allow for precise disease modeling, and identify both therapeutic targets and biomarkers for use in clinical trials.

354

Multiple functional variants at the 13q14 risk locus for osteoporosis regulate *RANKL* expression through long-range super-enhancer. *D.L. Zhu, X.F. Chen, N.N. Wang, B.J. Lu, Y. Rong, T.L. Yang.* School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

Receptor activator of nuclear factor-κB ligand (*RANKL*) is a TNF-like cytokine that is necessary for osteoclast formation and survival. Previous genome-wide association studies (GWASs) have identified several susceptibility single nucleotide polymorphisms (SNPs) near *RANKL* at 13q14.11 for osteoporosis. However, these SNPs all located in the intergenic regions over 100 kb upstream of *RANKL*, and their functional roles remain unknown. Therefore, the aim of this study was to identify the target gene(s) and the functional roles of these SNPs at 13q14.11. Here we conducted an integrative analysis combining expression quantitative trait locus (eQTL), genomic chromatin interaction (Hi-C), epigenetic annotation and functional assays. The eQTL analysis identified five potential functional SNPs (rs9533090, rs9594738, r8001611, rs9533095 and rs9594759) exclusively correlated with *RANKL* mRNA expression ($P < 0.001$). Hi-C analysis further validated that these 5 SNPs had long–range interactions with *RANKL* promoter in CD34 and GM12878 cells. The first 4 SNPs were in strong linkage disequilibrium (LD) within the same block, which is located in a super-enhancer region. To validate the functionality for these 5 SNPs, we conducted luciferase assay both in HEK293T and U2-OS cells. We observed the most significant constitutive enhancer activities in PRE (putative regulatory element) 1a surrounding rs9533090 and PRE 1b surrounding rs9594738. We further deleted the region containing PRE1a and PRE 1b using CRISPR/Cas9 genome editing and observed drastic down-regulation of *RANKL* expression in both mRNA and protein levels in U2-OS cells. We noticed that the PRE 1a, 1b regions were enriched in a cluster of multiple transcription factor binding sites (TFBSs), suggesting that multiple transcription factors (TFs) might cooperatively active the super-enhancer activity. In conclusion, we deciphered the functional basis for BMD variants in 13q14.11 in which a super-enhancer containing 4 functional SNPs affect *RANKL* expression to affect BMD via long-range regulation. Our results suggested that the methodology could be widely applied to identify regulatory SNPs and their target genes in other human complex diseases.

355

Protein-mediated 3D genome architecture in human B cells by HiChIP. *Y. Fu[1], R. Pelikan[1], C. Lareau[2], M. Aryee[2], J. Kelly[1], P. Gaffney[1,3].* 1) Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK; 2) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; 3) Department of Pathology, University of Oklahoma School of Medicine, Oklahoma City, OK.

Three-dimensional (3D) chromatin loops are critical for transcriptional regulation by bringing distant promoters and enhancers into close proximity. Protein factors, such as cohesin and CCCTC-binding factor (CTCF) facilitate the formation of the functional 3D chromatin conformation. Therefore, studying the protein-mediated 3D genome architecture can help to understand how specific genes contribute to human diseases. HiChIP is a new method to identify protein-mediated 3D chromatin looping information, and it was shown to have higher efficiency and specificity compared to Hi-C for the purpose of identifying such loops. In this study, we performed HiChIP assays to map the DNA looping patterns in EBV-transformed human B cell lines originally obtained from 3 systemic lupus erythematosus (SLE) patients. We measured 3D DNA contacts mediated by an essential nucleus protein, CTCF, and an epigenetic histone mark, histone 3 lysine 27 acetylation (H3K27ac). We also performed ChIP-seq assays to refine accurate anchor ranges. Using a minimum input of 1 million cells, we obtained approximately 100M paired-end tags (PETs), of which 30% represent intrachromosomal long-range interactions spanning between 5KB and 2MB. Distinct anchor and looping patterns were observed when targeting the two nucleus markers. The H3K27ac HiChIP analyses showed more cell-specific loops in enhancer-promoter regions, compared to CTCF HiChIPs, supporting its known role as an active enhancer marker throughout the genome. We further investigated the chromatin interactions approximately 185kb upstream of the tumor necrosis factor alpha inducible protein 3 (*TNFAIP3*) gene. This non-coding region was previously identified as an SLE and rheumatoid arthritis (RA) susceptibility region, however the function of associated SNPs (e.g. rs10499194, rs6920220) in this region is still not clear. Our H3K27ac HiChIP data showed that this region is enriched with H3K27ac-mediated looping activities. Interestingly, although this region is flanked by the oligodendrocyte transcription factor 3 (*OLIG3*) and *TNFAIP3* gene, we found that the associated SNPs are mainly looping to further upstream genes, interleukin 22 receptor alpha 2 (*IL22RA2*) and interferon gamma receptor 1 (*IFNGR1*). Our data provide new insight into finding direct functional targets associated with GWAS variants that are in non-coding regions. The data can also help to identify potential causal risk SNPs within a large linkage disequilibrium (LD) region.

356

Integration of chromosomal interactions with local adipose gene expression identifies obesity genes beyond GWAS. *D.Z. Pan[1,2], K.M. Garske[2], A. Ko[3], Y.V. Bhagat[2], M. Alvarez[2], J. Boocock[2], C.K. Raulerson[4], C.A. Glastonbury[5], K.S. Small[5], J.S. Sinsheimer[2,6], K.L. Mohlke[4], M. Laakso[7], P. Pajukanta[1,2,3].* 1) Bioinformatics IDP, UCLA, Los Angeles, CA; 2) Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, USA; 3) Molecular Biology Institute at UCLA, Los Angeles, USA; 4) Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA; 5) Department of Twin Research and Genetic Epidemiology, King's College London, UK; 6) Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, USA; 7) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.

To identify genetic variants influencing obesogenic gene expression in human adipose tissue, we combined expression quantitative trait loci (eQTLs) with chromosomal interaction data in primary human white adipocytes (HWA). We first identified *cis*-eQTLs (+/-1 Mb from the TSS, FDR<5%) using RNA-sequence (RNA-seq) data from 335 subcutaneous adipose biopsies from the Finnish Metabolic Syndrome In Men (METSIM) cohort. Each *cis*-eQTL and its target gene was replicated in subcutaneous adipose RNA-seq data (n=277) from GTEx (79.0% replicated), passing GTEx permutations. Next, we sought to detect distal enhancers and regulatory regions interacting with promoters by generating promoter capture Hi-C (pCHi-C) data in HWA. We searched for enrichments of transcription factor (TF) binding sites using HOMER by comparing the open chromatin regions of the chromosomal interactions in primary HWA to a background of macrophage interactions. We identified 30 TFs (FDR<5%) including PPARG (p-value=0.01) and CEBP (p-value=1.0x10^{-4}), which are critical players in adipogenesis. We also performed LD Score analysis and found that the same open chromatin regions explain ~4.6% of the heritability of local gene expression despite the small number of SNPs in this category (p<0.0002, enrichment=20.3±5.2, proportion of genome-wide SNPs= 0.23%). We then integrated these HWA distal enhancer-promoter interactions with the METSIM adipose *cis*-eQTLs to find variants affecting expression via chromosomal interactions in adipocytes. To focus on genes related to obesity, we correlated the expression of the target genes with BMI, identifying 54 genes with looping *cis*-eQTLs (p<1.2x10^{-5}), 48 of which were replicated in the adipose RNA-seq data of the TwinsUK cohort (n=720; adj. p<9.3x10^{-4}). Of these BMI-correlated eGenes, 4 contained a cardiometabolic GWAS SNP: *MAP2K5* (rs4776984, BMI), *LACTB* (rs3784671, metabolites), *ACADS* (rs12310161, metabolites), and *ORMDL3* (rs8076131, lipids). We also performed Transcriptome Wide Association Study (TWAS), showing that these 4 GWAS SNPs significantly (p<4.0x10^{-5}) affect obesity-related phenotypes, mediated through gene expression of their 4 target genes. Overall, our results uncover four new BMI genes, *MAP2K5*, *LACTB, ACADS, and ORMDL3*. In addition to the 4 genes, we identify a novel set of 44 replicated, BMI-correlated genes that are beyond the reach of GWAS, possibly representing the non-additive genetic component of BMI, such as gene-environment interactions.

357

Asprosin: Using genetics to discover a novel orexigenic hormone. *C. Duerrschmid[1], Y. He[2], C. Wang[2], J. Bournat[1], Y. Xu[2], A. Chopra[1,3].* 1) Department of Molecular & Cellular Biology, Baylor College of Medicine, Houston, TX; 2) Department of Pediatrics and Children's Nutrition Research Center, Baylor College of Medicine, Houston, TX; 3) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.

We recently described two patients with Neonatal Progeroid syndrome, that present with an extremely low body mass index, and partial lipodystrophy. 5 other cases in addition to our patients have been reported, and in all cases, the underlying etiology is a mutation within a 71 basepair region of the roughly 8600 bp *FBN1* gene. In all cases, the transcripts are predicted to escape mRNA nonsense mediated decay, resulting in a truncated protein product. The NPS mutation affects the 140 amino acid C-terminal polypeptide, which is cleaved off profibrillin to generate mature Fibrillin-1. Previously, this polypeptide had not been associated with a function of its own. We showed that the C-terminal polypeptide is present in low nanomolar concentrations in the plasma of unaffected humans, and significantly reduced in NPS patients. We named it asprosin, and demonstrated that it acutely increases hepatic glucose release during fasting to maintain plasma glucose levels in the normal range. Importantly, immunologic loss-of-function using monoclonal antibodies is protective against insulin resistance and type II diabetes via acute reduction in the plasma glucose burden. Here we present evidence, that asprosin, in addition to its glucogenic effect, tilts energy balance towards hyperphagia. Peripheral injection increases food intake in mice, while leaving energy expenditure unaltered. Mice with the NPS mutation are extremely lean, eat less than their wildtype littermates and are immune to obesogenic and diabetogenic effects of a prolonged high fat diet. Immunologic neutralization of asprosin results in hypophagia and weight loss, improving the plasma glucose burden, insulin sensitivity and body weight in genetically and diet-induced obese mice. Asprosin crosses the blood-brain barrier and directly, in a dose-dependent manner, increases firing frequency and membrane potential of hypothalamic orexigenic AgRP neurons, and inhibits anorexigenic POMC neurons through GABAergic signals. These effects are nullified by neutralizing asprosin, protecting from genetic and environmental forms of obesity. Our results demonstrate, that circulating asprosin crosses into the brain, where it stimulates food intake by activating orexigenic AgRP, and inactivating anorexigenic POMC neurons. Immunologic sequestration abolishes this effect, and protects against obesity. In addition to systemic asprosin being a glucogenic hormone, asprosin functions – equally importantly – as an orexigenic hormone.

358

Novel mutations in *ADCY3* cause severe, monogenic obesity in humans.

S. Saeed[1,2], A. Bonnefond[1], F. Tamanini[2], M.U. Mirza[3], J. Manzoor[4], Q.M. Janjua[5], S.M. Din[6], J. Gaitan[7,8], A. Milochau[7,8], E. Durand[1], E. Vaillant[1], A. Haseeb[6], F.D. Graeve[1], I. Rabearivelo[1], O. Sand[1], G. Queniat[1], R. Boutry[1], D.A. Schott[9], H. Ayesha[10], M. Ali[11], T.A. Butt[12], T. Rinne[13], C. Stumpel[14], A. Abderrahmani[1,2], J. Lang[7,8], M. Arslan[5,6], P. Froguel[1,2]. 1) UMR 8199 – EGID, University of Lille, CNRS, Lille, France; 2) Department of Genomics of Common Disease, Imperial College London, London; 3) Department of Pharmaceutical and Pharmacological Sciences, Rega Institute for Medical Research, KU Leuven (University of Leuven), Leuven, Belgium; 4) Department of Paediatric Endocrinology, Children's Hospital, Lahore, Pakistan; 5) Centre for Research in Molecular Medicine, The University of Lahore, Lahore, Pakistan; 6) Department of Biological Sciences, Forman Christian College, Lahore, Pakistan; 7) Laboratory of Membrane Chemistry and Biology (CBMN), CNRS UMR 5248, Université de Bordeaux, France; 8) Université de Bordeaux, 351 Cours de la Libération, Talence, France; 9) Department of Pediatrics, Zuyderland Hospital, Heerlen, the Netherlands; 10) Department of Paediatrics, Punjab Medical College, Faisalabad, Pakistan; 11) Department of Paediatrics, Mayo Hospital, King Edward Medical University, Lahore, Pakistan; 12) Department of Paediatrics, Fatima Memorial Hospital, Lahore, Pakistan;; 13) Department of Human Genetics, Donders Centre for Neuroscience, Radboud University Medical Center, Nijmegen, The Netherlands; 14) Department of Clinical Genetics and GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center, The Netherlands.

Single-gene mutations leading to obesity have provided critical insights into the molecular and physiological mechanisms that control energy homeostasis and body weight. Thus far, loss-of-function mutations in six genes coding for ligands, receptors and enzymes involved in the leptin signaling pathway have been identified in only 4–6% of cases of severe obesity. We previously showed that in a highly consanguineous population in Pakistan, recessive mutations in known obesity genes can explain ~30% cases of severe, early-onset obesity. These unexpected findings suggest that novel monogenic forms of obesity could be identified in this population and prompted us to perform whole exome sequencing (WES) on a cohort of 138 children with severe obesity (BMI SDS >3) and their family members. Our WES data analysis identified novel recessive mutations in adenylate cyclase 3 (*ADCY3;* NM_004036.4) in three probands with severe obesity, hyperphagia and anosmia. These extremely rare variants included a frame-shift (c.3315del; p.Ile1106Serfs*3), a missense (c.191A>T, p.Asn64Ile) and a splice-site (c.2578-1G>A) mutation. In an independent investigation, an affected child of European origin was also identified as a carrier of compound heterozygous *ADCY3* variants (c.1268del; p.Gly-423Alafs*19 and c.3354_3356del; p.Phe1118del). These ADCY3 mutations were functionally characterized by measuring cAMP production in BHK cells in which recombinant Myc-tagged, wild-type or mutant proteins were transiently over-expressed. Stimulation of adenylate cyclase activity with forskolin and concomitant inhibition of endogenous IBMX led to a significantly higher increase in cAMP production (~50-fold) in *ADCY3* wild-type cells compared to *ADCY3* mutant cells (~25-fold increase). These results were validated by luciferase assay, whereby the stimulatory effect of mutant *ADCY3* on the activity of the cAMP responsive element (CRE-Luc) was significantly lower compared to the wild-type (7-9-fold increase versus 13-fold increase, respectively). Homology modeling followed by molecular dynamic simulations indicated that the variants in *ADCY3* would cause structural deformations to the catalytic sites, imparting a significant reduction in the ability of ADCY3 to bind ATP. These data strongly suggest that the aforementioned mutations confer loss of protein function and thus highlight adenylate cyclase signaling as an important mediator of energy homeostasis.

359

Common and rare variants associated with body mass index (BMI) among multi-ethnic veterans: The Million Veteran Program.

J. Huang[1,2], J.J. Lee[3,4], N. Sun[5,6], Y.L. Ho[1], K. Cho[1,2], Y. Sun[7,8], J. Lynch[9,10], T. Assimes[11,12], S. Muraldihar[13], J.M. Gaziano[1,2], J. Concato[5,14], P. Tsao[11,12], P. Wilson[15], K.M. Chang[3,4], C.J. O'Donnell[1,2], D. Saleheen[3,4] for the VA Million Veteran Program. 1) Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA; 2) Brigham Women's Hospital, Harvard Medical School, Boston, MA; 3) Corporal Michael Crescenz VA Medical Center, Philadelphia, PA; 4) Perlman School of Medicine, University of Pennsylvania, Philadelphia, PA; 5) Veterans Affairs (VA) Cooperative Studies Program, VA Connecticut Healthcare System, West Haven, Connecticut; 6) Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut; 7) Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA; 8) Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA; 9) Department of Veterans Affairs Salt Lake City Health Care System, Salt Lake City, UT; 10) University of Massachusetts College of Nursing & Health Sciences, Boston, MA; 11) VA Palo Alto Health Care System, Palo Alto, CA; 12) Department of Medicine, Stanford University School of Medicine, Stanford, CA; 13) Office of Research and Development;Department of Veterans Affairs; 14) Yale University School of Medicine, New Haven, CT; 15) VA Atlanta Health Care System, Atlanta VA.

Background: Body mass index (BMI) provides a widely used estimate of adiposity that is a strong predictor of common diseases, including cardiovascular disease and cancer. Recent large-scale genome-wide association study (GWAS) consisting of multiple meta-analyzed cohorts of nearly 400,000 individuals discovered a total of 97 loci, but explained only 2.7% of BMI variance. **Design and Methods**: In the Million Veteran Program (MVP), a large mega-biobank drawn from the Veterans Health Administration system in the United States, we defined BMI in all participants using an algorithm that includes longitudinally collected measurements of height and weight derived from the electronic health record. Using ~50 million genotyped and imputed common and rare variants from 353,948 enrollees in the MVP, we examined their genetic association with BMI phenotypes within and across three ethnic groups: non-Hispanic European Americans (EUR, N=225,216), non-Hispanic African Americans (AFR, N=56,246), and Hispanics (HIS, N=20,661). For common variants (MAF>=5%), we defined novel loci as those with genome-wide significant variants (*P*<5E-08) in MVP and whose lead variant are >500 kb apart from previously reported variants. For low frequency and rare (MAF<5%) variants, we focus on missense and splice region variants. **Results**: *For common variants***,** we discovered 114 and 9 novel loci for EUR and AFR, respectively. Among the novel loci are SNPs within *CEBPZ, NAT8, TINK*. None of the 9 lead SNPs in AFR are significant in EUR. Through step-wide conditional analysis and further conditioning on previously published BMI variants, we found a total of 168 and 11 novel variants in both known and novel loci for EUR and AFR, respectively. **For rare variants**, we discovered 8 novel functional variants for EUR, including one inframe-deletion within *ZBTB10*. For AFR, we discovered one rare missense variant within *GIPR*. **Conclusion**: In MVP, we identified multiple novel common and rare variants underlying BMI, beyond those identified in recent large GWAS with comparable sample sizes. Collaborative replication and validation efforts are ongoing. Large biobank cohorts with detailed phenotyping derived from electronic health records and high quality genetic profiling can continue to contribute substantial novel information regarding the genetic architecture of complex traits, such as BMI.

360

Promoter or enhancer activation by CRISPRa rescues haploinsufficiency caused obesity. *N. Matharu[1,2], S. Rattanasopha[1,3], L. Maliskova[1,2], Y. Wang[4], A. Hardin[1,2], C. Vaisse[4], N. Ahituv[1,2].* 1) Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, 94158,; 2) Institute for Human Genetics, University of California San Francisco, San Francisco, California, 94158, USA; 3) Doctor of Philosophy Program in Medical Sciences, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand; 4) Diabetes Center, University of California San Francisco, San Francisco, California, 94143, USA.

Haploinsufficiency, having only one functional copy of a gene, leads to a wide range of human disease and has been associated with over 300 genes. Large-scale exome sequencing studies estimated that there could be around 3000 genes that can potentially contribute to disease upon heterozygous loss of function. Here, we tested if CRISPR activation (CRISPRa) could rescue haploinsufficiency in vivo. We targeted the promoter or enhancer of an existing functional copy of haploinsufficient gene using single guide RNA and a nuclease deficient dCas9 fused with an activator domain VP64 in mice. Haploinsufficiency of *SIM1*, a transcription factor expressed in the hypothalamus that is involved in the regulation of food intake through the leptin pathway, results in severe obesity in humans and mice. CRISPRa targeting of either the *Sim1* promoter or its ~270kb distant hypothalamic enhancer using transgenic mice, rescued the obesity phenotype in *Sim1* heterozygous mice. Despite using a ubiquitous promoter for CRISPRa, *Sim1* was upregulated only in tissues where the promoter or enhancer are active, suggesting that *cis*-regulatory elements can determine CRISPRa tissue-specificity. To further translate CRISPRa into a potential post-natal therapeutic strategy, we delivered dCas9-VP64 and sgRNA targeting either the *Sim1* promoter or its enhancer using adeno-associated virus (AAV) to the hypothalamus. Transcriptional upregulation of *Sim1* in the hypothalamus after AAV stereotaxic injections led to reversal of the weight gain phenotype of *Sim1* heterozygous mice in a long lasting manner for both promoter and enhancer targeted CRISPRa. Our results show that CRISPRa could be used to exploit the inherent tissue specificity of *cis*-regulatory elements and thus have a potential to be developed further for therapeutic applications. This novel therapeutic strategy can be used to treat many other disorders resulting from altered gene dosage.

361

Genome-wide analysis of UK Biobank to aid drug discovery in a consortium of pharmaceutical companies. *M.R. Nelson[1], A. Day-Williams[2], R.A. Scott[3], S. Glass[4], N. Smaoui[5], Z. Ding[6], C. Franklin[6], C. Vangjeli[6], M. Weale[6], S. John[7], M.G. Ehm[1], A. Holden[8], G. McVean[6], C. Spencer[6], Genomics Resources Consortium.* 1) Genetics, GSK, King of Prussia, PA; 2) Quantitative Target Sciences, Merck, Boston, MA; 3) Genetics, GSK, Stevenage, UK; 4) Genetics, Teva Pharmaceuticals, Malvern, PA; 5) Pharmacogenomics, Takeda Pharmaceuticals, Deerfield, IL; 6) Genomics plc, Oxford, UK; 7) Genomics and Computational Biology, Biogen, Cambridge, MA; 8) Genomics Resources Consortium, Chicago, IL.

Understanding the genetic influence on human biology provides causal insights that improve how we select drug targets and indications. In spite of the growing catalog of genetic associations for many traits, there remains a need for integrated variant-level summary statistics in order to fully explore genetic effects on disease subsets that could have the greatest impact on drug discovery decisions. The Genomics Resources Consortium (GRC) is a non-profit consortium with 10 current pharmaceutical companies working to form partnerships and develop resources to generate these data from large-scale biobanks. Genomics plc has analyzed UK Biobank (UKB) data across a range of traits, to understand the pleiotropic effects of genetic variants and their relevance to questions about efficacy and safety of drug targets. We conducted GWAS across 47 traits defined from the UKB survey data. We then addressed the effect of combining self-reported, nurse interview, and medication data to progressively refine phenotypes, with a focus on asthma. Our analyses quantify the trade-off between increased power to detect associations with larger sample sizes, and discovery of larger effect size estimates with higher confidence phenotypes. We found a high degree of concordance between the effects estimated in UKB and published meta-analyses for several traits evaluated, providing confidence in the quality of the phenotypic data. Visualizing the effect of genetic variants across these traits (PheWAS) has provided insights into several existing drug targets, including *PCSK9*, *GLP1R* and *IL33*. For example, we found the low frequency splice variant in *IL33* previously associated with lower eosinophils and asthma risk was also associated with lower risk of eczema and hay fever, and possibly elevated risk of type 2 diabetes. By combining the data from the biomarkers, baseline questionnaire, hospital, and registry data, we investigated several traits in more detail, including asthma subtypes and severity, migraine comorbidities, epilepsy treatment effects on cognitive function, and characterizing overall health. We discuss how the insights gained from this collaboration shape the development of a pre-competitive approach to enable analyses of data across a network of biobanks. Bringing these data together effectively will be key to increasing the power to conduct genetic research into questions that matter most in characterizing overall health and improving patients' lives.

362

Pharmacogenetic testing in the Veterans Health Administration: Recommendations from the VA Clinical Pharmacogenetics Subcommittee. *A. Stone[1], J.T. Callaghan[2], M.A. Mendes[3], V.M. Pratt[4], R. Przygodzki[5], M.T. Scheuner[6], S.A. Schichman[1], J.L. Vassy[7] for the VA Clinical Pharmacogenetics Subcommittee.* 1) Central Arkansas Veterans Healthcare System, Little Rock, AR., Select a Country; 2) Richard L Roudebush VAMC; 3) VA San Diego Healthcare System; 4) Department of Medical and Molecular Genetics, Indiana University School of Medicine; 5) Office of Research and Development, US Dept of Veterans Affairs; 6) VA Greater Los Angeles Healthcare System; 7) VA Boston Healthcare System.

Objectives FDA labels for more than 160 drugs include pharmacogenetic (PGx) considerations, and expert groups including PharmGKB and the Clinical Pharmacogenetics Implementation Consortium (CPIC) have graded several drug-gene pairs as having strong evidence of analytic and clinical validity. However, the Veterans Health Administration (VA), the largest U.S. integrated health system, currently lacks policies regarding PGx testing. The VA Clinical Pharmacogenetics Subcommittee, comprised of pharmacists, laboratory medicine physicians, medical geneticists, and health services researchers, was charged with issuing recommendations for PGx testing in patient care. **Methods** A literature review in March 2015 identified 29 drug-gene pairs with high PharmGKB or CPIC levels of evidence. The Subcommittee surveyed VA pathology and laboratory medicine service chiefs about current use of these tests in VA. Members met monthly through April 2017 to discuss the clinical utility of testing for each drug-gene pair. If a quorum was present, members voted by majority rule on whether each test should be required, recommended, or not recommended for use in VA. Veteran-specific contextual factors were considered. **Results** Nineteen service chiefs or designees completed the survey for 23 health systems. Of the 29 drug-gene pairs, 9 were performed at VA facilities and 19 sent to reference laboratories within the prior year. The Subcommittee's review determined that 4 (14%) of 29 tests should be required before prescribing the associated drug, such as *HLA-B* 1502 to avoid carbamazepine-associated Stevens-Johnston syndrome and *G6PD* to avoid rasburicase-associated acute hemolytic anemia. The Subcommittee recommended 12 (41%) tests, including *CYP2D6* to avoid codeine toxicity. Thirteen (45%) tests were not recommended, such as *CYP2C19* for clopidogrel dosing. Although these 13 drug-gene pairs had the highest PharmGKB and CPIC evidence grades (1A and A), the Subcommittee voted against routinely recommending them for lack of studies demonstrating improved patient outcomes or because alternative tests were available for clinical decision-making. The Subcommittee identified specific contexts in which these tests might be recommended. **Conclusions** The Subcommittee used expert consensus to make recommendations for evidence-based use of preemptive PGx testing in VA. Dissemination and implementation strategies promoting appropriate PGx testing and prescribing of associated drugs are needed.

363

Phenome-wide association studies (PheWAS) across large "real-world" population cohorts support drug target selection. *D. Diogo[1], C. Tian[2], C. Vangjeli[3], C. Spencer[3], C. Franklin[3], M. Weale[3], M. Alanne-Kinnunen[4], H. Mattsson[4], E. Kilpeläinen[4], R. Samuli[4], M. March[5], D. Reilly[1], M-P. Reeve[6], J. Hutz[6], N. Bing[7], S. John[8], D. MacArthur[9], M. Daly[9], H. Hakonarson[5], V. Salomaa[10], A. Palotie[4], D. Hinds[2], P. Donnelly[3], R. Plenge[1], A. Day-Williams[1], C. Fox[1], H. Runs[1].* 1) Genetics and Pharmacogenetics (GpGx), Merck & Co. Inc., Boston, MA, USA; 2) 23andMe, Mountain View, CA, USA; 3) Genomics Plc, Oxford, UK; 4) University of Helsinki, Helsinki, Finland; 5) Children's hospital of Philadelphia, Philadelphia, PA, USA; 6) Eisai, Boston, MA, USA; 7) Pfizer, Boston, MA, USA; 8) Biogen, Boston, MA, USA; 9) Broad Institute, Boston, MA, USA; 10) National Institute for Health and Welfare, Helsinki, Finland.

Drug targets supported by human genetics show increased success rates in clinical trials. Phenome-wide association studies (PheWASs), which assess whether a genetic variant is associated with multiple phenotypes across a phenotypic spectrum, have been proposed as a possible aid to therapy development through elucidating mechanisms of action, identifying alternative indications, or predicting adverse drug events (ADEs). Here, we evaluate whether PheWAS can inform decision-making on target programs during drug development. We selected 25 SNPs linked through genome-wide association studies (GWASs) to 19 candidate drug targets for common-disease therapeutic indications. We independently interrogated these SNPs through PheWAS in four large "real-world data" cohorts (23andMe, UK Biobank analysed by Genomics plc, FINRISK, CHOP) for association with a total of 1,891 binary endpoints. We then conducted meta-analyses for 145 harmonized disease endpoints in up to 697,815 individuals and joined results with summary statistics from 57 published GWASs. Our analyses replicated 70% of powered known GWAS associations and identified 72 novel associations with FDR<0.1, of which 9 reached study-wide significance after multiple test correction (P<1.8e-6). By leveraging the directions of effect and the point estimate of the effect sizes, our study identifies novel associations that may predict ADEs (e.g., acne, high cholesterol, gout and gallstones for rs738409 (p.I148M) in *PNPLA3*; asthma for rs1990760 (p.T946A) in *IFIH1*), and allows the comparison of the genetic efficacy and safety profiles of several candidate targets within a single indication (e.g., *F11*, *F12* and *KNG1* for thromboembolism). The results of this investigation reveal that PheWAS, despite limitations, provides powerful additional information to inform drug discovery.

364

Large-scale PheWAS in UK Biobank: A paradigm shift in genetic evaluation of prospective and existing drug targets. *L.M. Yerges-Armstrong[1], K. Song[1], M. Chiano[2], B.H. Dessailly[2], A. Pandey[1], T. Johnson[2], M.G. Ehm[1], M.R. Nelson[1], J.C. Whittaker[2], R.A. Scott[2].* 1) Target Sciences, GlaxoSmithKline, King of Prussia, PA; 2) Target Sciences, GlaxoSmithKline, Stevenage, UK.

Where genetic variants mimic the pharmacological perturbation of a drug target, genetic association results have the potential to identify new target candidates, validate primary therapeutic indications, suggest alternative indications and highlight putative on-target safety concerns. These opportunities are best realised with large-scale, deeply phenotyped collections combined with genetic data. To date, such analyses have largely relied on aggregating results from multiple studies, a labour-intensive and time-consuming strategy not scalable to large numbers of genes. Moreover, evaluation of traits related to disease progression or severity has been limited due to reliance on case-control studies. The advent of large-scale biobank resources offers the opportunity to rapidly perform such studies at hitherto unseen scale and phenotypic resolution. Here we illustrate how this advance dramatically enhances the use of genetics to validate prospective and existing drug targets. In up to 133,398 participants from the UK Biobank, we performed large-scale phenome-wide association studies (PheWAS) of variants likely to mimic pharmacological perturbation of drug targets. To identify associations with phenotypes related to primary indications as well as those which might highlight potential additional indications or on-target safety concerns, we included over 2,000 traits from baseline questionnaires, physical exam, biological sampling and hospital inpatient health records. For example, a recent whole-genome sequencing study in the Icelandic population identified a rare loss-of-function (LoF) variant (rs146597587) in *IL33* which may mimic pharmacological inhibition of IL33 (a therapeutic strategy proposed for asthma). Here, we confirmed the role of IL33 haploinsufficiency in lower eosinophil count and protection against asthma. We also implicate potential additional indications for IL33 inhibition by identifying associations of the *IL33* LoF allele with lung function ($p<2.5x10^{-5}$), protection from COPD (OR=0.63; p=9.7e-05), and “hayfever, allergic rhinitis or eczema” (OR=0.65; $p=2.5x10^{-5}$;). We will discuss multiple examples from this resource for validation of existing and prospective therapeutic targets. The imminent expansion of the UK Biobank resource to include 500,000 genotyped individuals with eventual linkage to further electronic health records promises a paradigm shift in our ability to bring genetic evidence to support therapeutic target validation.

365

Discovering cell-specific regulatory networks using natural genetic variation. *C. Romanoski, L. Stolze, M. Whalen.* University of Arizona, Tucson, AZ.

Regulatory networks are remarkably cell type-specific in that each cell type utilizes less than ten percent of all possible enhancers (promoter-distant, *cis*-regulatory elements) to direct appropriate gene expression for that cell type. The proper complement of enhancers are established and maintained by the actions of a few prominent transcription factors (TFs) that cooperate with one another to access chromatin. To discover the sets of cooperative transcription factors that define enhancer function, we have established a method that integrates high-throughput sequencing data for enhancer features (such as chromatin accessibility or TF binding) with natural genetic variation. This method was first developed for application to genomes of inbred mouse macrophages and has been extended to human endothelial cells. The underlying principle is that non-coding genetic variation sometimes perturbs enhancer function by altering the genomic sequences at TF binding motifs. This will have the consequence of diminished TF binding for the respective factor as well as cooperating TFs where binding motifs are unchanged. Our method searches for enrichment of diminished TF binding (or chromatin accessibility) at the set of genomic loci where a binding motif of interest is ‘mutated’ by one allele compared to their ‘wildtype’ counterparts. Using this method, we found that genome sequence combined with ChIP-seq data for one individual was sufficient to validate that two nominated TFs, ERG and JUN, indeed collaborate to bind endothelial cell-specific enhancers. This approach is now being utilized for analysis of 50 human aortic endothelial cell cultures from genetically diverse individuals as a way to identify functional non-coding genetic variation in human cells.

366

Genetic and epigenetic regulation of gene expression in the human liver. *M. Caliskan[1], J. Segert[1], S. Rao[1], M. Trizzino[1], Y. Park[1], K.M. Olthoff[2], A. Shaked[2], D.J. Rader[1,3], B.E. Engelhardt[4], C.D. Brown[1].* 1) Department of Genetics, Perelman School of Medicine, University of Pennsylvania; 2) Division of Transplant Surgery, Perelman School of Medicine, University of Pennsylvania; 3) Department of Medicine, Perelman School of Medicine, University of Pennsylvania; 4) Computer Science Department, Princeton University.

The liver performs a large number of critical functions, including detoxification of endogenous and exogenous toxins, synthesis of essential proteins, and regulation of carbohydrate, lipid, and drug metabolism. As such, 'liver disease' applies to over 100 types of diseases and disorders that result from liver dysfunction. In fact, a recent GTEx study reported liver as a critical tissue for explaining the mechanisms at genome wide association study (GWAS) loci. To help interpret the functional consequences of the GWAS variants, we generated the largest allele-specific gene expression and histone modification maps of the human liver to date. We generated genome-wide genotype, RNA-seq, and ChIP-seq data on H3K27ac and H3K4me3 histone modifications in up to 241 individuals. At a false discovery rate (FDR) of 5%, we identified local expression quantitative trait loci (eQTL) for 2,554 autosomal genes and 13 X-chromosome genes. Of the 2,567 eQTL loci, 308 (12.0%) have been previously associated with at least one complex phenotype based on NHGRI GWAS catalog and are enriched for genes implicated in cardiovascular disease, blood lipid levels, and HIV infections. Our deep ChIP-Seq data have dramatically expanded the catalog of liver cis-regulatory elements: 84.1% of genes expressed in the liver have significant histone modification within 1kb of their transcription start sites. Extensive functional validation of these predicted regulatory elements supports their ability to regulate gene expression. We then identified genetic variants associated with histone modifications (hQTL) and identified 110 local hQTLs for H3K4me3 and 1,066 hQTLs for H3K27ac at an FDR of 10%. Through integration of hQTL and eQTL signals in our dataset, we were further able to fine-map candidate functional variants at 22 distinct GWAS loci. Overall, our study highlights the benefits of integrating multiple cellular traits in fine mapping of GWAS loci and contributes to basic understanding of genetic and epigenetic regulation of gene expression in the human liver tissue.

367

Interaction between two promoter/enhancer variants in *CYP7A1*: Robust effect on hepatic mRNA expression and association with lipids and risk of CAD and diabetes. *D. Wang[1], K. Hartmann[1], M. Seweryn[2], W. Sadee[1].* 1) Center for Pharmacogenomics, The Ohio State University, Columbus, OH; 2) Center for Medical Genomics OMICRON, UJ CM, Krakow, Poland.

Cholesterol 7alpha-hydroxylase (CYP7A1) catalyzes the first and rate-limiting step in biosynthesis of bile acids from cholesterol, which also serves as a main pathway for cholesterol removal from the body. Association studies including genome wide association study (GWAS) have identified SNPs in *CYP7A1* to be associated with lipid level, CAD and other phenotypes; however, results were inconsistent, and causative SNPs remain to be elucidated. In this study, we used chromatin conformation capture combined with high throughput sequencing (4C assay) to identify regulatory regions for CYP7A1, followed by CRISPR-mediated genome editing to determine interacting regulatory regions. The results revealed a novel enhancer and a repressor region for *CYP7A1*, located >10kb downstream of *CYP7A1* promoter. To identify regulatory variants,, we screened for SNPs located in these regulatory regions and used allelic mRNA expression imbalance to test their effect on CYP7A1 expression in human livers. This approach led to a SNP located at the enhancer region that regulates CYP7A1 expression. Only when considered together, of this enhancer SNP and the previously identified promoter SNPs (rs3808607) robustly determines CYP7A1 expression; moreover, we find the 2-SNP combination to be significantly associated with lipid level, risk of CAD and diabetes in publicly available clinical cohorts.

368

Functional regulatory mechanism of smooth muscle cell-restricted *LMOD1* coronary artery disease locus. *C. Miller[1], V. Nanda[2], T. Wang[3], M. Pjanic[1], B. Liu[4,5], T. Nguyen[1], T. Quertermous[1], N. Leeper[2].* 1) Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA; 2) Vascular Surgery, Stanford University School of Medicine, Stanford, CA; 3) Genetics, Stanford University School of Medicine, Stanford, CA; 4) Biology, Stanford University School of Medicine, Stanford, CA; 5) Pathology, Stanford University School of Medicine, Stanford, CA.

Coronary artery disease (CAD) remains the leading cause of mortality and morbidity worldwide and has an estimated heritability of 40%. Meta-analyses of genome-wide association studies have now identified 73 genome-wide significant loci and the most recent multi-ethnic meta-analysis identified a number of arterial wall-specific loci. One of these loci include the smooth muscle cell-restricted factor, LMOD1 (Leiomodin 1), a member of the actin filament nucleator family that is highly enriched in smooth muscle tissues including main arteries and visceral organs of the gastrointestinal system. We hypothesized that LMOD1 may serve as a potent marker of the phenotypic modulation of smooth muscle cells during atherosclerosis. Using the assay for transposase accessible chromatin (ATAC-seq) we recently identified a non-coding regulatory variant, rs34091558, which is in tight linkage disequilibrium (LD) with the lead CAD GWAS variant, rs2820315 (P=7.7E-10; OR=1.05). LMOD1 mRNA and protein was also shown to be downregulated in human carotid atherosclerotic arteries. However, the causal mechanism for how these variants alter LMOD1 expression/function and CAD risk remains unclear. Expression quantitative trait loci (eQTL) mapping in GTEx and STARNET databases confirmed that rs34091558 is among the top eQTLs for LMOD1 in vascular tissues. Allelic expression imbalance analyses in heterozygous HCASMC donors further demonstrated *cis*-acting effects on LMOD1 gene expression. Position weight matrix (PWM) motif analyses identified the protective allele at rs34091558 to form a consensus forkhead box O3 (FOXO3) binding motif, which is predicted to be disrupted by the risk allele. FOXO3 chromatin immunoprecipitation and reporter assays demonstrated reduced FOXO3 binding and transcriptional activity by the risk allele in HCASMC. Platelet-derived growth factor BB (PDGF-BB) stimulation also significantly reduced LMOD1 expression coincident with FOXO3 knockdown. Finally, both gain and loss-of-function for FOXO3 and LMOD1 in HCASMC delineated a regulatory circuit by which LMOD1 decreased HCASMC proliferation and migration and increased cell contraction. Taken together, these results provide compelling functional evidence that: 1) common genetic variation is associated with dysregulation of LMOD1 expression, and 2) changes in vessel wall processes through LMOD1 dysregulation may partially explain the heritable risk for CAD.

369

Molecular diagnostics of Mendelian disorders via RNA sequencing. *H. Prokisch[1,2], D.M. Bader[3,4], C. Mertes[3], R. Kopajtich[1,2], G. Pichler[5], A. Iuso[1,2], T.B. Haack[1,2], E. Graf[1,2], T. Schwarzmayr[1,2], C. Terrile[1], E. Koňaříková[1,2], B. Repp[1,2], P. Lichtner[1], C. Leonhardt[6], B. Funalot[7], A. Donati[8], V. Tiranti[9], A. Lombes[10,11,12], C. Jardel[10,13], D. Gläser[14], R.W. Taylor[15], D. Ghezzi[9], J.A. Mayr[16], A. Rötig[6], P. Freisinger[17], F. Distelmaier[18], T.M. Strom[1,2], T. Meitinger[1,2], J. Gagneur[3], L.S. Kremer[1,2].* 1) Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany; 2) Institute of Human Genetics, Klinikum rechts der Isar, Technische Universität München, München, Germany; 3) Department of Informatics, Technische Universität München, Garching, Germany; 4) Quantitative Biosciences Munich, Gene Center, Department of Biochemistry, Ludwig Maximilian Universität München, Munich, Germany; 5) Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Martinsried, Germany; 6) Neuropädiatrie, Neonatologie, Villingen-Schwenningen, Germany; 7) INSERM U1163, Université Paris Descartes - Sorbonne Paris Cité, Institut Imagine, Paris, France; 8) Metabolic Unit, A. Meyer Children's Hospital, Florence, Italy; 9) Unit of Molecular Neurogenetics, Foundation IRCCS (Istituto di Ricovero e Cura a Carettere Scientifico) Neurological Institute “Carlo Besta”, Milan, Italy; 10) Inserm UMR 1016, Institut Cochin, Paris, France; 11) CNRS UMR 8104, Institut Cochin, Paris, France; 12) Université Paris V René Descartes, Institut Cochin, Paris, France; 13) AP/HP, GHU Pitié-Salpêtrière, Service de Biochimie Métabolique, Paris, France; 14) Genetikum, Genetic Counseling and Diagnostics, Neu-Ulm, Germany; 15) Wellcome Trust Centre for Mitochondrial Research, Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK; 16) Department of Pediatrics, Paracelsus Medical University, Salzburg, Austria; 17) Department of Pediatrics, Klinikum Reutlingen, Reutlingen, Germany; 18) Department of General Pediatrics, Neonatology and Pediatric Cardiology, University Children's Hospital, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany.

Across a large variety of Mendelian disorders, 50 to 75% of undiagnosed patients do not receive a genetic diagnosis by whole exome sequencing indicative of underlying regulatory variants. In contrast, whole genome sequencing allows the discovery of all genetic variants, but their significant number, coupled with a poor understanding of the non-coding genome, makes their prioritization challenging. Here, we demonstrate the power of directly sequencing transcriptomes by providing a molecular diagnosis for 11% of unsolved patients and strong candidates for others. RNA-seq identified 12,680 transcribed genes across 105 patient fibroblast cell lines, 47 from cases without genetic diagnosis. We systematically searched for three possible causes for the rare disorder: i) genes with aberrant expression level, ii) genes with aberrant splicing and iii) mono-allelic expression of rare variants and found a median of 1 aberrantly expressed gene, 5 aberrant splicing events, and 6 mono-allelically expressed rare variants. To evaluate the consequences of reduced RNA expression on the protein level we performed quantitative proteomics for 31 patients. Follow-up experiments established disease-causing roles for each kind of aberrant expression. Strikingly, the majority (70%) of the 30 private exons we discovered arose from sites that were weakly spliced in other individuals. Our finding shows that weakly spliced cryptic exons are loci more susceptible to turn into strongly spliced sites than other intronic regions. Hence, these results suggest that the prioritization of deep intronic variants of unknown significances gained through WGS could be improved by annotating weak splice sites. One such intronic exon-creating event was found in three unrelated families in the complex I assembly factor TIMMDC1. The new exon resulted in a frameshift and a premature stop codon leading to the absence of the encoded complex I assembly factor. Quantitative proteomic analysis revealed global complex I reduction. The complex I deficiency was rescued upon lentiviral transduction of the wild-type allele and a minigene assay demonstrated that the intronic SNV was responsible for the aberrant splicing. Together, this established *TIMMDC1* as a novel disease-associated gene. In conclusion, our study expands the diagnostic tools for detecting non-exonic variants of Mendelian disorders and provides examples of intronic loss-of-function variants with pathological relevance.

370

Massively parallel reporter assays combined with cell-type specific expression quantitative trait loci profiling identified a functional melanoma risk variant in HIV-1 inhibitor gene, *MX2*. *J. Choi[1], T. Zhang[1], M. Makowski[2], A. Vu[1], M. Kovacs[1], L. Colli[1], M. Xu[1], N. Lam[1], S. Chanock[1], S. Loftus[3], W. Pavan[3], M. Vermeulen[2], K. Brown[1].* 1) Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD; 2) Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University Nijmegen, the Netherlands; 3) Genetic Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD.

Recent melanoma Genome-Wide Association Study (GWAS) has identified 20 common susceptibility loci via meta-analysis of >12,000 cutaneous malignant melanoma cases. Identifying functional variants from these GWAS loci and unraveling their risk-conferring mechanisms presents a significant challenge due to limited resolution of genetic structure and lack of high-throughput assays. To address this challenge, we employed a combined approach of massively-parallel reporter assays (MPRA) and primary cultured human melanocyte expression quantitative trait locus (eQTL, n=106) profiling. From 16 GWAS loci whose signals are not readily explained by protein-coding variants, 835 high-LD variants (r^2 > 0.4 with the lead SNP) were prioritized based on promoter/enhancer-relevant ENCODE annotation in primary melanocytes or melanomas. To simultaneously examine allele-specific transcriptional activities of these variants, MPRA was performed on pooled luciferase constructs of ~64,000 iterations representing 145bp encompassing each variant and coupled with 10 unique sequence tags per allele, direction, and promoter type. The libraries were transfected into melanoma cells and resulting transcribed 10bp tags were quantitated using massively parallel sequencing. We identified ~100 melanoma-specific functional variants from 13 loci displaying allelic transcriptional activity at FDR < 0.05. By overlaying significant MPRA SNPs and top melanoma GWAS SNPs (r^2 > 0.8) with significant melanocyte eQTLs (FDR < 0.05), we further nominated eight variants exhibiting allelic transcription patterns mimicking allelic expression of local genes. Mass-spectrometry of these variants demonstrated allele-preferential binding of multiple nuclear proteins. Among them, rs398206, which is within the first intron of *MX2* gene and correlated with *MX2* levels, demonstrated allele-specific binding by transcription factor YY1, a finding validated by electromobility shift assay/supershift and chromatin immunoprecipitation. Ectopic expression of *MX2*, previously identified as a gene associated with HIV-1 resistance, conferred a growth advantage in both melanoma cells and primary human melanocytes. Our study provides a rapid identification strategy for functional GWAS variants, and elucidates an unexpected role of *MX2* in melanomagenesis, suggesting a pleiotropy of *MX2* and its common functional variant in antiviral immunity and melanoma susceptibility.

371

Integrative analysis of loss-of-function variants in clinical and genomic data reveals novel genes associated with cardiovascular traits. *B.S. Glicksberg[1,2], L. Amadori[1,3], N.K. Akers[1], K. Sukhavasi[4], O. Franzén[1,5], L. Li[1,2], G.M. Belbin[1,6], K. Shameer[1,2], M.A. Badgeley[1,2], K.W. Johnson[1,2], B. Readhead[1,2], B.J. Darrow[3], E.E. Kenny[1,6], C. Betsholtz[7], R. Ermel[8], J. Skogsberg[9], A. Ruusalepp[5,8], E.E. Schadt[1,2,5], J.T. Dudley[1,2,10], H. Ren[11], J.C. Kovacic[3], C. Giannarelli[1,3], S.D. Li[1], J.L.M. Björkegren[1,4,5,9], R. Chen[1].* 1) Department of Genetics and Genomic Sciences, The Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029, USA; 2) The Institute for Next Generation Healthcare, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029, USA; 3) Cardiovascular Research Center and Cardiovascular Institute, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029, USA; 4) Department of Pathophysiology, Institute of Biomedicine and Translation Medicine, University of Tartu, Biomeedikum, Ravila 19, 50411, Tartu, Estonia; 5) Clinical Gene Networks AB, Jungfrugatan 10, 114 44 Stockholm, Sweden; 6) Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029, USA; 7) Department of Immunology, Genetics and Pathology, Uppsala University, 751 85 Uppsala, Sweden; 8) Department of Cardiac Surgery, Tartu University Hospital, 1a Ludwig Puusepa Street, 50406 Tartu, Estonia; 9) Integrated Cardio Metabolic Centre, Department of Medicine, Karolinska Institutet, Karolinska Universitetssjukhuset Huddinge, 141 86 Stockholm Sweden; 10) Department of Health Policy and Research, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029, USA; 11) Department of Pediatrics, Herman B Wells Center for Pediatric Research, Center for Diabetes and Metabolic Diseases, Stark Neurosciences Research Institute, Indiana University, 635 Barnhill Dr., MS2049, Indianapolis, IN 46202, USA.

The delineation and association of loss-of-function variants (LoFs) with human diseases and phenotypic traits continues to play an increasingly important role in the discovery and validation of novel therapeutic targets. In the current study, we integrated the analysis of genetic and clinical data from the Mount Sinai Bio*Me* Biobank (*n*=10,511) and identified LoFs in 433 genes significantly associated with at least one of ten major cardiovascular disease (CVD) traits in the Mount Sinai Hospital's electronic medical record (EMR) system. Next, we used RNA-sequence data from the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) study obtained from seven metabolic and vascular tissues in 600 coronary artery disease patients and validated 115 of the 433 genes whose expression levels were concordantly associated with corresponding CVD traits. *APOC3* and five novel genes predicted for lowering plasma lipids in Bio*Me* and STARNET liver data were successfully validated using gene silencing in HepG2 cells with marked effects on levels of media triglycerides, APOB100 and PCSK9, and cellular LDLR. In addition, we identified LoF in a novel therapeutic target associated with lower plasma cholesterol and glucose levels in Bio*Me* that were also confirmed in STARNET, and showed pharmacological inhibition of the target in C57BL/6 mice not only significantly lowered fasting glucose levels but also affected body weight. In summary, by integrating genetic and EMR data, and then leveraging one of the world's largest human RNA-sequence datasets (STARNET), we identified several potential targets for CVD therapeutics that merit further investigation.

372

Phenomewide association study of life course health events: Analyzing 50 years of hospitalization, prescription drug use and death data. *S. Ripatti[1,2], A. Havulinna[1,3], T. Kiiskinen[1], P. Helkkula[1], H. Hautakangas[1], P. Häppölä[1], S. Ruotsalainen[1], J. Koskela[1], M. Kurki[4,5], I. Surakka[1], M. Perola[3], M. Kallela[6], V. Salomaa[3], B. Neale[4,5], M. Pirinen[1], H. Laivuori[1], E. Widén[1], M. Daly[1,4,5], A. Palotie[1].* 1) Institute for Molecular Medicine Finland FIMM, Helsinki, Finland; 2) Public Health, University of Helsinki; 3) National Institute for Health and Welfare; 4) Broad Institute of MIT and Harvard; 5) ATGU, Massachusetts General Hospital; 6) Helsinki University Hospital.

The long history of systematically coded major health events using electronic nation-wide population-based health registries in the Nordic countries provide special opportunities for genetic studies. In particular, the possibility to link information from various registries including for example prescription drug usage, hospitalization and deaths using unique personal electronic identifiers together with genomic profiles, allow for rich genetic and epidemiological study designs. We illustrate these opportunities by analyzing data on 30,000+ Finnish individuals with over 50 years of registry-based health event data. We present heritability estimates and genetic correlations on hundreds of health events and show how the heritability estimates typically get higher as the level of details in the ICD codes grow. We also provide phenome-wide scans across the major health events for 1,999 loss-of-function (LoF) variants enriched in Finland. As an example, a novel protein truncating variant *rs760351239* in *ANPTL8* gene (MAF=0.0015 in Finns, over 80-times enriched in Finns compared to Non-Finnish Europeans) is associated with lower triglycerides levels (beta=-0.66 on sd scale, p=1.2*10^-8) and protective of T2D (OR=0.29, p=0.036). Finally, we show that two registry-based indicators for healthy ageing, one using YODA (Years of Drugs Applied) score based on prescription drug purchase history and another using cumulative months of hospitalization periods, both predict mortality (RR=1.023 per YODA year, p=2e-16; RR=1.074 per month of hospitalization, p=2e-16, respectively). Furthermore, YODA score can be shown to associate with individual genetic markers linked to ageing, like *APOE e4* (OR=0.941, p=0.006), and polygenic risk score for CAD (RR=1.09, p=3e-15). As the number of genetically profiled individuals grow, the rich life course health event data provide a powerful platform to study health events of carriers of candidate variants and test for predictive tools to disease risk and management across hundreds of diseases.

3022

A multi-hit *de novo* mutation model for idiopathic simplex autism. *T.N. Turner[1], B.P. Coe[1], D.E. Dickel[2], K. Hoekzema[1], B.J. Nelson[1], M.C. Zody[3], Z.N. Kronenberg[1], F. Hormozdiari[4], A. Raja[1,5], L.A. Pennacchio[2,6], R.B. Darnell[3,7,8], E.E. Eichler[1,5].* 1) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA; 2) Functional Genomics Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; 3) New York Genome Center, New York, NY 10013, USA; 4) Department of Biochemistry and Molecular Medicine, University of California, Davis, Davis, CA 95817, USA; 5) Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA; 6) U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA; 7) Laboratory of Molecular Neuro-Oncology, The Rockefeller University, New York, NY 10065, USA; 8) Howard Hughes Medical Institute, The Rockefeller University, New York, NY 10065, USA.

Large copy number variants (CNVs) and *de novo* gene-disruptive mutations account for only a fraction of autism spectrum disorder (ASD). To further our understanding of ASD genetic architecture, we generated and analyzed genome sequence data from 516 idiopathic simplex ASD families (2,064 individuals corresponding to parents-proband-unaffected sibling quads). Families were selected where no loss-of-function or CNV candidate mutation was previously characterized. From this resource of >59 million single-nucleotide variants (SNVs) and 9,212 private CNVs, we focus on *de novo* mutations (DNMs), including 88 smaller CNVs and 133,992 SNVs where we establish an overall validation rate of 95.5%. We identify, on average, 94.3 +/-22 DNMs per child yielding an overall higher mutation rate than previous reports ($1.5x10^{-8}$ SNVs per site per generation). This is due to better access to repetitive DNA where we find the mutation rate to be significantly elevated (chi-squared test $p=1.7x10^{-74}$). Comparing probands and unaffected siblings, we observe several DNM trends. Probands carry more gene-disruptive CNVs (p=0.03) and SNVs resulting in severe missense mutations (p=0.01). Probands also show an excess of noncoding DNMs mapping to predicted fetal brain promoters, embryonic stem cell enhancers and within 3' UTRs. These differences become more pronounced when restricting to autism genes and considering multiple DNMs. We show that probands are about twice as likely to have two or more DNMs in protein-coding or noncoding regulatory DNA of known autism risk genes when compared to their unaffected siblings (overall $p=2.6x10^{-4}$, OR=2.3). Considering all genes genome-wide, this oligogenic DNM signal is conspicuously enriched for genes with elevated expression in striatal neurons ($p=3x10^{-3}$)—a pattern that persists when only considering noncoding events. The oligogenic burden also appears more pronounced in autism females, helping to explain the apparent discrepancy between the female protective effect and the absence of the Carter effect in autism families. We estimate that multi-hit noncoding DNMs contribute to autism in ~6% of individuals in our study and represent the third largest class of DNM for autism after monogenic *de novo* SNVs and CNVs. These data strongly argue that exome sequencing alone will be insufficient to diagnose autism in the population. We are now applying this model to the study of 7,584 additional genomes to understand more genetically complex cases of ASD.