American Society of Human Genetics 66th Annual Meeting
October 18–22, 2016 VANCOUVER, CANADA

PLATFORM ABSTRACTS

Tuesday, October 18, 5:00-6:20 pm:

Wednesday, October 19, 9:00-10:30 am, Concurrent Platform Session A:

Wednesday, October 19, 4:30-5:50 pm:

Thursday, October 20, 9:00-10:30 am, Concurrent Platform Session B:

Thursday, October 20, 11:00 am-1:00 pm, Concurrent Platform Session C:

Friday, October 21, 9:00-10:30 am, Concurrent Platform Session D:

Friday, October 21, 4:30-5:50 pm:

Saturday, October 22, 9:00-10:00 am, Concurrent Platform Session E:

Saturday, October 22, 10:15-11:30 am, Concurrent Platform Session F:

Saturday, October 22, 12:00-1:00 pm:

1

Disease heritability estimates using the electronic health records of 9 million patients. *N. Tatonetti[1], F. Polubriaginof[1], K. Quinnies[1], R. Vanguri[1], A. Yahi[1], M. Simmerling[2], I. Ionita-Laza[3], H. Salmasian[4], S. Bakken[1], K. Kiryluk[5], D. Goldstein[6], D. Vawdrey[4].* 1) Biomedical Informatics, Columbia University, New York, NY; 2) Division of Medical Ethics, Weill-Cornell Medical Center, New York, NY; 3) Mailman School of Public Health, Columbia University, New York, NY; 4) Value Institute, NewYork-Presbyterian Hospital, New York, NY; 5) Department of Medicine, Columbia University, New York, NY; 6) Institute for Genomic Medicine, Columbia University, New York, NY.

The heritability of human disease is essential for diagnosis, treatment, and prognosis. Twin studies remain the gold standard for estimating the heritability of disease. However, even the largest study can evaluate only a small number of phenotypes. Electronic health records (EHRs) capture a wide range of clinically relevant variables including continuous data (e.g. clinical laboratory test results) and dichotomous traits (e.g. disease diagnosis), representing a novel resource for heritability studies. Using the emergency contact data provided by 9 million patients at two large academic medical centers, we inferred 4.7 million familial relationships and validated these relationships using both clinical and genetic data sources. We then computed heritability and familial recurrence rates for 663 clinical phenotypes from all major disease categories. Heritability estimates showed high concordance across the two independent centers for both quantitative (rho = 0.85, p = 2.24e-08) and dichotomous traits (rho=0.75, p=8.98e-25). Heritability estimates from the EHR were significantly correlated with those reported in the literature with rho = 0.48, p = 0.024 and rho = 0.44 and p = 0.011 at sites 1 and 2, respectively. When looking at sibling and familial recurrence, perinatal conditions are the most concordant (r^2 = 0.94 for sibling and 0.98 for familial) and the category of miscellaneous and unclassified conditions was the least concordant (r^2 = 0.04 for sibling and < 0.01 for familial). Sibling recurrence and familial recurrence are highly correlated at both medical centers (rho = 0.56, p = 1.00e-50 and rho = 0.58, p = 2.43e-60, respectively). We found eight diseases with heritability estimates greater than 40% that replicated at both medical centers, including obesity with h2=47% (95CI: 40-52, N = 63,840) and h2=69% (95CI: 62-71, N = 9,908), and rhinitis with h2=57% (95CI: 46-55, N = 32,495) and h2=100% (95CI: 99-100, N = 6,173) at the two sites, respectively. We present the first systematic study of disease heritability using EHR data. The results provide insight into the genetic etiology of human disease, especially as diagnosed in a large active medical center, and could be used to improve both detection and treatment.

2

A genome-wide compendium and functional assessment of *in vivo* heart enhancers. *D.E. Dickel[1], C.H. Spurrell[1], I. Barozzi[1], Y. Zhu[1], Y. Fukuda-Yuzawa[1], M. Osterwalder[1], B.J. Mannion[1], I. Plajzer-Frick[1], C.S. Pickle[1], E. Lee[1], T.H. Garvin[1], M. Kato[1], J.A. Akiyama[1], V. Afzal[1], A. Visel[1,2,3], L.A. Pennacchio[1,2].* 1) Functional Genomics Department, Lawrence Berkeley National Lab, Berkeley, CA; 2) U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA; 3) School of Natural Sciences, University of California, Merced, Merced, CA.

Whole genome sequencing is identifying growing numbers of noncoding variants in human disease studies. However, the lack of accurate annotations and an unclear understanding of the functional impact of noncoding sequence variants largely prevent their interpretation. Focusing on the developing and adult heart, an organ whose functional impairment is a predominant cause of mortality and morbidity, we delineate the genome-wide landscape of distant-acting enhancers and highlight their functional importance through mouse knock-out studies. We used integrative analysis of >35 epigenomic datasets from mouse and human pre- and postnatal hearts to create a comprehensive reference of >80,000 putative human heart enhancers. This compendium includes confidence scores for each candidate enhancer and can be easily intersected with human disease data for downstream functional characterization of human WGS or GWAS variants. To understand the role of enhancers in the progression of human heart disease, we also profiled epigenomic and gene expression differences between 18 normal human heart samples and 18 with dilated cardiomyopathy (DCM). Our analysis reveals thousands of enhancers with disease-correlated activity changes, indicating and defining widespread gene regulatory alterations associated with heart disease. To facilitate the use of the genome-wide heart enhancer annotations in human disease studies, we have made them available as a public resource on the VISTA Cardiac Enhancer Browser (heart.lbl.gov). Finally, to demonstrate the importance of enhancers to heart function and their impact on heart disease-associated etiology, we deleted the mouse orthologs of two human enhancers near cardiac myosin genes. In both cases, we observed *in vivo* expression changes and cardiac phenotypes consistent with human heart disease. Taken together, our study provides a comprehensive catalog of human heart enhancers for use in clinical whole genome sequencing studies and highlights the importance of enhancers for cardiac function.

3

Application of a conditional allelic series of the Sloan-Kettering Institute proto-oncogene (SKI) to mechanistically dissect the TGFβ vasculopathies. *B.E. Kang[1,4], D. Bedja[2], H.C. Dietz[1,3,4].* 1) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 2) Department of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 3) Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 4) Howard Hughes Medical Institute, Bethesda, MD, USA.

Marfan (MFS), Loeys-Dietz (LDS) and Shprintzen-Goldberg (SGS) syndromes show substantial phenotypic overlap including aortic root aneurysm. While all of these conditions show an aortic wall signature for high TGFβ signaling, LDS is caused by heterozygous loss-of-function (LOF) alleles in genes encoding positive effectors of TGFβ including ligands (TGFB2 or TGFB3), receptor subunits (TBFBR1 or TGFBR2) or signaling mediators (SMAD2 or SMAD3). In contrast, SGS, which shows complete phenotypic concordance with LDS with the added feature of intellectual disability, is caused by heterozygous LOF mutations in SKI encoding a prototypical TGFβ repressor that displaces positive and recruits negative transcriptional regulators to TGFβ target genes. Such paradoxical results have engendered controversy regarding the precise role of TGFβ in aneurysm progression. We propose a reconciling model in which different vascular smooth muscle cell (VSMC) lineages have a variable vulnerability to heterozygous LOF mutations in TGFβ signaling effectors, resulting in compensatory upregulation of TGFβ ligand production by one lineage that achieves paracrine overdrive of signaling in its less vulnerable neighbor. In keeping with this hypothesis, we showed that LDS second heart field (SHF)-derived VSMCs in the aortic root show signaling collapse and increased TGFβ expression, while neighboring LDS cardiac neural crest (CNC)-derived cells remain signaling competent. We developed a series of conditional Ski alleles that allows robust interrogation of the model. First, VSMC-specific expression of a knock-in Ski mutation known to cause severe SGS in people (p.G34D) results in progressive aortic root aneurysm in association with increased output of TGFβ target genes. Second, we generated transgenic mice that overexpress wild-type SKI in a spatially- and temporally-conditional manner. Postnatal overexpression of SKI specifically in CNC-derived VSMCs of LDS mice suppresses TGFβ target gene expression and rescues the aneurysm phenotype. Finally, we applied a novel conditional null allele of Ski to show that LOF in the CNC of MFS mice exacerbates both TGFβ signaling and aneurysm growth. These data show that aneurysm severity in multiple conditions specifically titrates TGFβ signaling status in the CNC, support therapeutic strategies aimed at TGFβ antagonism, and highlight the importance of consideration of the microenvironments within which genetic alterations exert their phenotypic influence.

4

200 loci complete the genetic puzzle of multiple sclerosis. *N. Patsopoulos[1,2,3,4] on behalf of the International Multiple Sclerosis Genetics Consortium.* 1) Neurology Department, Brigham & Women's Hospital, Boston, MA; 2) Division of Genetics, Brigham & Women's Hospital, Boston, MA; 3) Harvard Medical School, Boston, MA; 4) Broad Institute, Cambridge, MA.

Background Multiple sclerosis (MS) is a common autoimmune disease with complex genetic background that affects the central nervous system (CNS). We present results of the International Multiple Sclerosis Genetics Consortium's (IMSGC) analysis of genome-wide association studies (GWASs), followed by 2 large-scale replication data sets, with overall 47,351 MS subjects and 68,284 controls analyzed. **Methods** We analyzed GWASs totaling 14,802 MS cases and 26,703 controls and ~8 million SNPs post-imputation with the European panel of 1000 Genomes. Next, we applied a stepwise strategy within blocks of 2Mbps (n=1,961) to prioritize 4,842 seemingly statistically independent effects (p<0.01) for replication in autosomal non-MHC genome. Then, we used 2 large-scale datasets: i) a targeted microarray, MS Chip, that we genotyped in 20,282 MS subjects and 18,956 controls, (ii) ImmunoChip platform on an additional 12,267 MS subjects and 22,625 controls. Finally, we used several approaches to prioritized genes and identify significantly enriched pathways. **Results** The joint analysis of 47,351 MS subjects and 68,284 controls resulted in 200 statistically independent effects with $p<5x10^{-8}$ and another 119 with high probability to be associated (joint $p<10^{-5}$ and p<0.05 in both replication sets). The average MAF was 29.7% with a minimum of 1.6%, whereas the odds ratios ranged from 1.05 to 2.06. These 200 effects accounted for ~18% of the narrow-sense heritability. Tissue and cell-specific enrichment, using expression and epigenetic data, consistently highlighted the importance of the immune system with no evidence of enrichment in the CNS. Immune cells, CD4+ T cells and monocytes, and brain *cis*-eQTL analyses identified several genes whose expression was affected by the majority of the 200 GW effects, with few of them being potentially brain-specific. More than 300 genes could be prioritized via multiple integrated layers of data that were statistically enriched (false discovery rate<5%) for several pathways of both the innate and adaptive branches of immunity. More interestingly the enriched pathways ranged from immune system maturation and development to training, and antigen presentation and recognition. **Discussion** We report an analysis of more than 110K samples and report 200 genome-wide MS effects. We illustrate that the majority of these mediate their effect via different immune cells, with a small number of these to be potentially active in a brain-specific manner.

5

Levering gene family information in gene discovery, risk assessment and missense variant interpretation in more than 9,000 trios with neurodevelopmental disorders. *D. Lal[1,2,3], P. May[5], E.B. Robinson[1,2], H. Yuan[4], K.E. Samocha[1,2], J.A. Kosmicki[1,2], R. Krause[5], P. Nuernberg[3], S. Weckhuysen[6], G. Guerrini[7], C. Marini[7], P. De Jonghe[6], S. Biskup[9], A. Poduri[10], B.A. Neubauer[11], B.P. Koeleman[12], K. Helbig[8], Y.G. Weber[9], I. Helbig[13], S. Traynelis[4], A. Palotie[1,2], M.J. Daly[1,2].* 1) The Broad Institute of Harvard and M.I.T, Cambridge, MA; 2) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 3) Cologne Center for Genomics, University of Cologne, Cologne, Germany; 4) Department of Pharmacology, Emory University School of Medicine, Atlanta, Georgia; 5) Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg; 6) VIB-Department of Molecular Genetics, Neurogenetics Group, University of Antwerp, Antwerp, Belgium; 7) Pediatric neurology unit, A. Meyer Pediatric Hospital, Florence, Italy; 8) Division of Clinical Genomics, Ambry Genetics, Aliso Viejo, California, USA; 9) Dpt. of Neurology and Epileptology, Hertie Insitute for Clinical Brain Research, University of Tübingen, Tübingen, Germany; 10) Epilepsy Genetics Program, Boston Children's Hospital; 11) Department of Neuropediatrics UKGM, University of Giessen, Germany; 12) Department of Genetics, University Medical Center Utrecht, Utrecht, The Netherlands; 13) Division of Neurology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, U.S.A.

Patients with neurodevelopmental disorders (NDDs) show great phenotypic and genetic variability with regard to recently identified genes enriched for *de novo* variants in patients. Investigating disorder specific risk is challenging due to rareness of the disorders and ultra low frequency of variants. Several of the recently identified NDD associated genes have originated from the same ancestral gene, and thus share the same gene family. Here, we demonstrate that gene family information can be leveraged in genetic studies to facilitate gene discovery and interpretation of *de novo* variants in NDDs. We performed a joint analysis of 3982 autism, 4471 developmental delay, and 822 severe seizure disorder exome-sequenced parent-offspring trios. We developed novel approaches to discover and interpret *de novo* variant enriched gene families. In the 9275 NDD trios, we identified 29 gene families significantly enriched for *de novo* variants, including several where no single gene in the family had previously been statistically associated with a NDD. The strongest enrichment was found in the voltage-gated sodium channel gene family (P = 2.03E-39). Given the increased power afforded by aggregating variants across all members of the gene family, we were able to investigate whether the 29 enriched gene families confer risk for a specific subset of the NDDs (sub-NDD). In total, we identified eight gene families for which at least one sub-NDD showed significant risk in comparison to the other NDDs. Seven gene families conferred risk specific for developmental delay. The most significant sub-disorder association was observed for the voltage-gated sodium channel gene family, which confers specific risk for severe seizure disorders (OR=4.23, P=3.00E-10). Next, we investigated the utility of gene family information in missense variant interpretation. We observed that disease associated missense variants fall into sequences identical within all members of the gene family and not in gene specific sites (P < 10E-50). Using publically available non-neuronal disease and control data sets, we show this observation in generalizable. Finally, we provide the functional analysis of a large list of patient and population variants in *GRIN2A* and *GRIN2B* as an illustrative example of how gene family conservation can guide variant interpretation and functional domain identification.

6

Genome sequencing in a “healthy” population: Challenges of variant interpretation. *S. Punj[1], A. Creason[1], J. Huang[1], A. Potter[1], M.O. Dorschner[2,3], D.A. Nickerson[3], G.P. Jarvik[3,4], L.M. Amendola[4], D. Kostiner Simpson[5], A. Rope[5], J. Reiss[5,6], T. Kauffman[7], M. Gilmore[5], P. Himes[5], B. Wilfond[8,9], K.A.B. Goddard[7], C.S. Richards[1] on behalf of the NextGen Project Team.* 1) Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR; 2) Pathology, Division of Bioethics, University of Washington, Seattle, WA; 3) Genome Sciences, Division of Bioethics, University of Washington, Seattle, WA; 4) Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA; 5) Department of Medical Genetics, Kaiser Permanente Northwest, Portland, OR; 6) Obstetrics and Gynecology, Kaiser Permanente Northwest, Portland, OR; 7) Center for Health Research, Kaiser Permanente Northwest, Portland, OR; 8) Department of Pediatrics, Division of Bioethics, University of Washington, Seattle, WA; 9) Truman Katz Center for Pediatric Bioethics, Seattle Children’s Hospital, Seattle, WA.

In the Clinical Sequencing Exploratory Research (CSER) consortium’s NextGen study, preconception carrier screening is performed by whole genome sequencing (WGS). Variants are interpreted in pre-selected genes (728) associated with recessive and X-linked disorders. All participants receive results for lifespan-limiting conditions (177 genes), but the optional categories are: serious (406), mild (93), unpredictable (41), and adult onset disorders (11), as well as additional findings (123) for medically actionable conditions. The majority of participants (92%) requested all categories, and limited information about personal or family history of a genetic disorder and ethnicity was provided to the laboratory. Using the 2015 ACMG guidelines for interpretation of sequence variants, we classified variants in 175 participants, of which ~76% had a pathogenic (P) or likely pathogenic (LP) variant. Twenty-one percent of the variants disclosed to participants were novel, defined as not previously observed in affected individuals. These novel variants were classified as LP because they predicted a loss-of-function in genes where this is known mechanism for disease. For novel splice-site variants (4% of disclosed variants), RNA analysis was performed when possible to assess impact on splicing to aid in variant classification. For novel Copy Number Variants (~2.5% reported variants), an algorithm was developed for classification. The majority of variants returned were missense variants (67%) that were previously reported in an affected individual. None of the novel missense variants, representing the bulk of variants identified in this cohort, met the minimal ACMG criteria for a P or LP classification, suggesting this category of variants is problematic and less likely to be reported in a carrier setting. In terms of overall variant classification, the most frequently used line of evidence was variant frequency in the population databases (80%) followed by evidence for functional effects (53%) and in silico evidence (50%), while several lines of evidence were never used. We will discuss challenges of classifying and returning rare variants in well-known genes, such as *CFTR*, and variants in genes associated with ultra-rare conditions such as Lethal Multiple Pterygium Syndrome in the carrier setting. In our experience, caution must be used when interpreting variants, particularly novel variants, in genomic-based expanded carrier testing of healthy individuals.

7

Identifying Mendelian disease-causing mutations by leveraging splice-affecting variant predictors provides improved guidelines for the interpretation of nonsynonymous, synonymous, and intronic variants of uncertain significance. *Z.T. Soens[1,2], J. Branch[1,2], Y. Li[1,2], K. Wang[1,2], M. Xu[1,2], D. Birch[3], F.B. Porto[4], J. Sallum[5], P. Zhao[6], R. Sui[7], R.K. Koenekoop[8,9,10], R. Chen[1,2,11,12,13].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA; 3) Retina Foundation of the Southwest and Department of Ophthalmology, University of Texas Southwestern Medical Center, Dallas, Texas, USA; 4) Department of Retina and Vitreous, Ophthalmologic Center of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; 5) Department of Ophthalmology and Visual Sciences, Paulista School of Medicine, Federal University of São Paulo, São Paulo, Brazil; 6) Department of Ophthalmology, Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China; 7) Department of Ophthalmology, Peking Union Medical College Hospital, Peking Union Medical College, Dongcheng, Beijing, China; 8) McGill Ocular Genetics Laboratory and Centre, Department of Ophthalmology, McGill University Health Centre, Montreal, Quebec, Canada; 9) Department of Paediatric Surgery, McGill University Health Centre, Montreal, Quebec, Canada; 10) Department of Human Genetics, McGill University Health Centre, Montreal, Quebec, Canada; 11) Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas, USA; 12) Program of Developmental Biology, Baylor College of Medicine, Houston, Texas, USA; 13) Department of Structural and Computational Biology & Molecular Biophysics, Baylor College of Medicine, Houston, Texas, USA.

Many Mendelian disorders are notoriously genetically heterogeneous having a variety of mutations in a multitude of genes that are able to cause disease. This genetic heterogeneity often results in significant portions of affected patients not being able to be confidently assigned a molecular diagnosis after traditional sequencing of the protein-coding exons of the genes associated with the disorder. A confident molecular diagnosis is necessary for a patient to be eligible for personalized treatment and counseling. Leber congenital amaurosis (LCA) is a severe inherited retinal disease which also exhibits a similar trend in that ~25% of all patients lack an actionable molecular diagnosis after sequencing genes known to be associated with the phenotype. There are two primary avenues of investigation for uncovering the genetic etiology of the unsolved ~25%: searching for mutations in genes not yet linked with LCA, and searching for mutations in known disease genes that we are missing or failing to interpret as pathogenic. Seeking to pursue the latter option we chose to assess the prevalence of splice-disrupting variants amongst variants not traditionally annotated as affecting splicing. Applying splice-affecting prediction scores from multiple *in silico* tools to our entire LCA cohort of 714 patients revealed 15 unique variants in 18 patients that were prioritized for splicing functional validation. Five nonsynonymous variants were identified in seven patients previously solved with low confidence, four synonymous variants were identified in four unsolved patients, and six intronic variants were identified in seven unsolved patients. Since not every candidate mutation's gene is expressed in blood we decided to utilize a minigene *in vitro* system to assay each mutation's effect on splicing. Nine variants have been successfully tested in the minigene system to date and all nine have results supporting the variant's disruption of wildtype splicing. By comparing the distribution of scores from the leveraged splice-affecting variant predictors we can recommend a decision tree for other groups to consult when seeking to annotate and prioritize candidate splicing variants in their own cohorts. Splice-disrupting mutations are particularly deleterious to protein function generally resulting in exon loss, and a truncated or absent protein; patients solved with splicing mutations can therefore be considered assigned a molecular diagnosis with a higher degree of confidence.

8

Reinterpreting variant pathogenicity: Lessons from over 60,000 human exomes. *A. O'Donnell-Luria[1,2], E.V. Minikel[1,2], J.S. Ware[2,3,4], N. Whiffin[3,4], M. Lek[1,2], K.J. Karczewski[1,2], K.E. Samocha[1,2], M.J. Daly[1,2], D.G. MacArthur[1,2], Exome Aggregation Consortium.* 1) Analytic and Translational Genetics Unit (ATGU), Massachusetts General Hospital, Boston, MA, USA; 2) Broad Institute of Harvard and MIT, Medical and Population Genetics, Cambridge, MA, USA; 3) National Heart and Lung Institute, Imperial College London, London, UK; 4) MRC Clinical Sciences Centre, Imperial College London, London, UK.

Our ability to sequence human genetic variation has far outpaced our ability to interpret the functional impact of genetic variants. In the diagnostic evaluation of rare disease, this poses a critical challenge as exome sequencing is employed with increasing frequency. While pathogenic variants are identified in 20-70% of exome-sequenced cases depending on the presenting features, many cases remain unsolved. The frequency of a variant is a powerful discriminator of variant pathogenicity. The publicly available Exome Aggregation Consortium (ExAC) dataset, with jointly analyzed exome sequencing data from a collection of 60,706 individuals, provides an unprecedented view of the spectrum of human functional genetic variation extending down to extremely low population frequencies across diverse ancestries. Over 20% of ClinVar and HGMD pathogenic variants are present in ExAC, providing an opportunity to assess the value of deep frequency data for variant assessment. For pathogenic variants with a relatively high allele frequency (greater than 1 in 1000), we identified a number of false positive variants polluting clinical variant databases. Common modes of error include ancestry-specific variants (e.g. *CIRH1A*), overinterpretation of functional data, and misinterpretation of a variant in *cis* with a pathogenic variant. At the very rare end of the frequency spectrum, we unexpectedly found dozens of rare pathogenic variants for severe pediatric-onset dominant conditions in ExAC. For a subset of these, there is evidence of somatic mosaicism in the ExAC sample (e.g. *BRAF*, *ASXL1*) or, very uncommonly, sequencing or alignment errors (e.g. *PQBP1*), multinucleotide polymorphisms (e.g. *ARID2*), but others appear to be examples of genuine incomplete penetrance (e.g. *FGFR2*, *TBX5*). We describe current work exploring potential explanations for the lack of a phenotype in these cases. Our results underscore the tremendous contribution of large population resources such as ExAC in deciding whether variants are pathogenic or benign. The accuracy of variant interpretation has profound implications for patients, as the genetic diagnosis often leads to specific diagnostic and/or treatment recommendations. The power of ExAC as a public resource will continue to grow as sample sizes increase into the hundreds of thousands of individuals.

9

Clinical application of a critical exon mapper to aid in determining pathogenicity of copy number variants of unknown significance. *K.S. Ho[1,2], M. Uddin[3,4], C.H. Hensel[1], M.M. Martin[1], P. Mowery-Rushton[1], S. Page[1], M.A. Serrano[1], S. Venkatasubramanian[5], S. Scherer[3,4,6,7], E.R. Wassman[1].* 1) Lineagen, Inc., Salt Lake City, UT; 2) Department of Pediatrics, University of Utah; 3) The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada; 4) Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada; 5) School of Computing, University of Utah; 6) McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada; 7) Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

Chromosomal microarray (CMA) is recommended as the first-tier diagnostic test for the evaluation of individuals with autism spectrum disorder (ASD), developmental delay (DD), and/or multiple congenital anomalies (MCA). In addition to the detection of copy number variants (CNVs) clinically documented as pathogenic for these disorders, CMA frequently detects rare variants of unknown significance (VOUS). Interpretation of these VOUS to determine their role, if any, in the manifestation of disease remains a challenge. The development of new tools that contribute to VOUS interpretation could have broad impact on the diagnosis and clinical management of a wide range of disorders, and becomes increasingly important as CMA becomes more widely used and accepted in the clinical setting. Uddin et al. developed a relational database tool to identify exons within CNVs that are both highly expressed in the brain and highly conserved, which may help refine the potential pathogenicity of VOUS results. This method, known as the “critical exon mapper” (CEM), was previously shown to identify critical exon containing genes (CEGs) that are potentially pathogenic in an ASD cohort versus sibling controls ($p=10^{-38}$). We report here the application of CEM to a real world population previously referred to a clinical laboratory for diagnostic genetic testing for neurodevelopmental conditions. A total of 5518 individuals were tested between Jun 2012 and Dec 2015, using a CMA optimized for the detection of neurodevelopmental disorders by the addition of 88,435 custom probes to an Affymetrix array. This population manifested pathogenic CNVs in 9.2% and VOUS in 20.3% of patients. Among 1602 patients we evaluated 742 pathogenic and 1363 VOUS CNVs using CEM, and found a significant number of CEGs in both pathogenic (mean=12.4/CNV) and VOUS (mean=1.0/CNV) CNVs, impacting 70% of pathogenic and 40% of VOUS CNVs. A total of 2492 unique CEGs were observed in the pathogenic group and 953 in the VOUS CNVs. By comparison, a small number of “benign” CNVs assessed by CEM had a very low prevalence of CEGs. We cross-referenced the unique VOUS CEGs with OMIM and found that 21% are not represented therein, highlighting the potential limitations of OMIM and other databases for analysis of potentially impactful genes in CNVs. We present several examples demonstrating how CEM might impact the clinical interpretation of VOUS and potentially help identify critical regions that are likely pathogenic.

10

Towards precise genetic diagnosis of human diseases: Experience with *POLG*-related disorders. *R. Bai[1], D. McKnight[1], J. Juusola[1], J. Yang[1], Y. Xie[1], R. Heredia[1], Y. Chen[1], H. Yang[1], D. Arjona[1], A. Balog[1], J. Higgs[1], L. Carey[1], E. Butler[1], H. Cui[1], W.C. Copeland[2], G. Richard[1], S.F. Suchy[1].* 1) GeneDx Inc. Gaithersburg, MD; 2) NIEHS, Research Triangle Park, NC.

Background: Pathogenic variants in the *POLG* gene are the most common cause of single nuclear gene mitochondrial diseases (MtD). To date, 253 published *POLG* variants, mostly identified by Sanger sequencing, have been collected into the Human Gene Mutation Database (www.hgmd.com) and/or the Human *POLG* database (http://tools.niehs.nih.gov/polg/). This study evaluated and curated the published and novel *POLG* variants identified in the past few years. **Methods:** 20136 patients with suspected MtDs or neuro-developmental/muscular /metabolic disorders were tested by whole exome sequencing (WES, 8125 cases) or multi-gene panels (sequencing and deletion/duplication analysis, 12011 cases) with the *POLG* gene included for MtD or Epilepsy; For 9104 patients, both parents or/and other family members were tested for segregation analysis. For 6231 WES cases and 2032 multi-gene panel cases, whole mitochondrial genome analysis (WMGA) was also performed. Variants in the *POLG* gene and other genes were analyzed according to the joint ACMGG and AMP guideline (PMID: 25741868), in which the nature of the variants, the frequency of the variant in affected or unaffected individuals, clinical information, family study results, and whether the patient had diagnostic pathogenic variants in other genes, were all considered. **Results:** For the 253 published variants possibly associated with *POLG* related disorders, 65 were observed 1 to 778 times at our clinical diagnostic laboratory, seven of which were re-classified as benign or likely benign variants (BEN/LBEN). 39 variants were classified pathogenic or likely pathogenic (PATH/LPATH), and 19 variants were classified as variants of uncertain significance (VUS). In addition, 213 novel variants were identified: 43 were classified as PATH/LPATH, 45 were classified as BEN/LBEN, and the remaining 124 variants were of uncertain significance. **Conclusion:** “Pathogenic” *POLG* variants in public databases or individual publications that lack sufficient supporting evidence should be reassessed, as more than half of these would be considered VUS or BEN/LBEN based on current standards. This study updates and expands the *POLG* variant database and will aid in the reliable molecular diagnosis of *POLG*-related disorders. These data also emphasize the importance of combining WES or large multi-gene panels, clinical information, and family studies for the precise evaluation of genetic variants and clinical molecular diagnosis of human diseases.

11

Distribution and clinical impact of functional variants in 50,726 whole exome sequences from the DiscovEHR study. *F. Dewey[1], M. Murray[2], J. Overton[1], L. Habegger[1], J. Leader[2], S. Fetterolf[2], C. O'Dushlaine[1], C. Van Hout[1], J. Staples[1], R. Metpally[2], H.L. Kirchner[2], S. Pendergrass[2], C. Gonzaga-Jauregui[1], S. Balasubramanian[1], A. Lopez[1], J. Penn[1], S. Mukherjee[1], N. Gosalia[1], A. Li[1], S. Bruse[1], K. Praveen[1], I. Borecki[1], G. Yancopoulos[3], O. Gottesman[1], M. Ritchie[1], A. Shuldiner[1], J. Reid[1], D. Ledbetter[2], A. Baras[1], D. Carey[2], DiscovEHR Collaboration.* 1) Regeneron Genetics Center, NY, USA; 2) Geisinger Health System, PA, USA; 3) Regeneron Pharmaceuticals, NY, USA.

The DiscovEHR collaboration between the Regeneron Genetics Center and Geisinger Health System MyCode community health initiative aims to catalyze genomic discovery and precision medicine by coupling high throughput sequencing to a large, integrated healthcare population utilizing longitudinal electronic health records (EHR). Here we describe initial insights from deep whole exome sequencing of 50,726 adult participants of predominantly European ancestry with clinical phenotypes described in EHRs and other clinical data streams. The median duration of EHR data in these participants was 14 years, during which a median of 87 clinical encounters, 687 laboratory tests and 7 procedures were captured per participant. We found ~4.2 million single nucleotide variants and insertion deletion events, of which ~176,000 are predicted to result in loss of gene function (LoF). The overwhelming majority of these genetic variants occurred at minor allele frequency ≤ 1%, and over half were singletons. Each participant harbored a median of 21 rare predicted LoFs. At this sample size, over 90% of genes, including genes encoding existing drug targets and a list of 76 genes (56 designated by ACMG plus 20 additional) conferring risk for actionable, highly penetrant genetic diseases, harbor rare heterozygous predicted LoF variants. Approximately 7% of genes contain rare homozygous predicted LoF variants in at least one individual. Linking these data to EHR-derived clinical phenotypes, we find clinical associations supporting therapeutic targets, including genes encoding drug targets for lipid lowering. We also highlight examples of novel gene discovery, through use of EHR derived lipoprotein laboratory values and exome wide association analyses, identifying novel rare alleles associated with HDL-C (*LIPG, LIPC, LCAT*, *CD36, SCARB1*), LDL-C (*ABCA6, APOH*), and triglycerides (*ANGPTL3, G6PC*). We find that 4.4% of individuals harbor deleterious variants in 76 clinically actionable genes, highlight examples of clinical phenotypes derived from the electronic health record that reflect this genetic disease predisposition, and outline plans to return and act clinically on these findings. DiscovEHR provides a blueprint for large-scale precision medicine initiatives and genomics-guided target discovery and drug development.

12

Extending variant reference databases to over 100,000 samples. *D. MacArthur[1,2], Exome Aggregation Consortium.* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA.

The discovery of genetic variation has been empowered by the growing availability of DNA sequencing data from large studies of common and rare diseases, but these data are typically inconsistently processed and largely inaccessible to most genetics researchers. Since 2014 the Exome Aggregation Consortium (ExAC) has provided a public resource of genetic variants spanning harmonized data from over 60,000 human exomes, which has now become a default reference database for clinical genetics.   Here we describe the expansion of ExAC along two dimensions. Firstly, we announce the release of exome variant frequency data from over 120,000 humans, doubling the depth of frequency information available in protein-coding regions. Secondly, we announce the release of the Genome Aggregation Database (gnomAD), an aggregate frequency database that provides information on variants in both coding and non-coding regions drawn from over 20,000 high-coverage whole human genomes.   We also describe new scientific findings emerging from this increased sample size, including refined maps of genic tolerance for missense and protein-truncating variants; more accurate thresholds for filtering during pathogenic variant discovery; and pilot projects in the targeted re-phenotyping of individuals with “extreme genotypes” identified from within ExAC cohorts.

13

Quantifying aging signals in whole genome sequencing data.

A. Bernal, M. Zhu, C. Lippert, C. Maher, R. Sabatini, T. Wong, P. Garst, E. Kostem, H. Tang, L. Pierce, K. Yocum, F. Och, C. Venter. Human Longevity, Inc., 191 Castro Street FL 2. Mountain View, CA.

Aging is one of the most important risk factor for human diseases. In this study, we analyzed and measured cellular DNA damage signals using whole-genome DNA sequencing (WGS) data and used these to predict a person's chronological age. Over time, genomic DNA in cells gradually acquires damage as a side effect of normal metabolic and molecular processes, such as oxidative stress and cell division. DNA damage can accumulate at specific genomic locations and affects all cells, effectively acting as a genomic clock for age. Damage can also randomly spread throughout the genome of each cell resulting in a state of somatic mosaicism that is thought to be an important cause for biological aging, as well as other diseases. In this work, we analyzed 5587 WGS data samples produced using HiSeq X Ten systems, and observed and measured four distinct signals derived from DNA damage that correlate with the subject's age: gradual shortening of telomeres and mosaic copy number loss of chromosomes X, Y and M. Telomeres are stretches of repetitive DNA that cap chromosome ends that effectively act as protective buffers against DNA loss effects of chromosome replication. To estimate telomere lengths we counted the number of sequence reads that are telomeric and corrected these counts for sequencing reagent chemistry effects. To our knowledge, this is the first study in which these effects have been established on WGS data for estimating telomere lengths. Mosaic loss of chromosomes (MLC) refers to a state in which some cells lose chromosomal segments, but others do not. It has been previously reported that the proportion of sex chromosome aneuploidy, including mosaic loss of X or Y chromosome is associated with aging. The conventional approach for computing MLC has been using SNP arrays. To our knowledge, no attempt of computing MLC using WGS data has been reported. For biological age prediction, we regressed the estimates of telomere length and copy number loss against the chronological age of each sample. Ten-fold crossvalidation experiments on our 5587 WGS samples resulted in a hold-out R^2(MAE) for our estimates of MCL and telomere length were 0.46 (11 years) and 0.24 (13 years), respectively. A combined model obtained a hold-out R^2 of 0.57, equivalent to a mean absolute error (MAE) of 9 years.

14

Integrating eQTLs and tissue-specificity across 44 normal human tissues with genome-wide association data helps uncover causal genes and pathways for common diseases.

A.V. Segre[1], F. Aguet[1], G. Getz[1,2], K. Ardlie[3], GTEx Consortium. 1) Cancer Program, Broad Institute of Harvard and MIT, Cambridge, MA; 2) Pathology Department, Massachusetts General Hospital, Boston, MA; 3) Medical and Population Genetics Program, Broad Institute of Harvard and MIT, Cambridge, MA.

Since the majority of common variant associations with complex diseases lie in noncoding regions, they are likely to affect disease risk via alterations in gene regulation. A major challenge lies in identifying the causal genes and biological processes through which these genomic loci exert their effect on disease in relevant tissue contexts. The Genotype-Tissue Expression (GTEx) project, whose primary goal is to detect DNA variants associated with gene expression changes (eQTLs) in healthy human tissues, provides an invaluable resource for such questions. Building upon the notion that hundreds of modest effect associations have yet to be detected for complex diseases, we developed a two-step statistical method that tests whether eQTLs significant in a given tissue are enriched for multiple modest to genome-wide significant disease associations, using a rank and permutation-based test that corrects for confounders (distance to TSS, MAF, linkage disequilibrium). If enrichment is found, top ranked eQTL target genes are tested for over-representation in biological pathways and gene or phenotype ontologies. We applied our method to GWAS meta-analysis summary statistics for >20 metabolic, cardiovascular, neurodegenerative and psychiatric diseases/traits, using eQTLs from 44 tissues in <450 GTEx donors (~800-10,000 eGenes/tissue, FDR<5%). Significant GWAS-eQTL enrichment was found for all traits in a range of tissues (*p*<1E-05), and known and new genes and pathways are proposed (e.g. cell cycle, fatty acid metabolism for type 2 diabetes, T2D; axon degeneration, neuron number changes for Alzheimer's disease, AD). While significant enrichment was found in relevant tissues, e.g. visceral adipose and liver, top ranked for T2D, or Brain frontal cortex for AD, eQTL enrichment was also found in less obvious tissues for most traits. To help distinguish between pathogenic tissues and pleiotropic effects, we developed and integrated into our method an eQTL tissue-specificity measure across tissues, based on eQTL effect size (not affected by sample size), correcting for gene expression similarity between tissues. Tissue-specificity analysis will be presented for all traits. Initial results suggest that the observed widespread enrichment is largely driven by shared eQTLs between known pathogenic tissues and the other tissues. Considering the tissue-specificity of genetic regulation may help distill the pathogenic tissue/s and uncover biological changes unique to the disease.

15

Mendelian disease genes provide the ultimate confirmation of PrediXcan validity and suggest novel opportunities for translation of genetic discoveries.

J.E.H. Brown[1], E. Gamazon[1], J.C. Denny[1], L. Basterache[1], H. Im[2], N.J. Cox[1]. 1) Vanderbilt University, Nashville, TN; 2) University of Chicago, Chicago, IL.

We have applied PrediXcan (Gamazon et al, Nat Genet 2015) to more than 18,000 samples (post-QC) in BioVU to test the genetically predicted expression of genes with the medical phenome as represented in ~1600 PheWAS codes. Results of these studies identify novel gene-phenotype relationships that are genome-wide significant with strict Bonferroni correction, and have been validated through model system studies. We argue here that the ultimate validation of PrediXcan associations may be provided by the correspondence of the phenotypes already characterized for “human knock outs” – Mendelian diseases – and the phenotypes we observe in these PrediXcan studies to be associated with reduced genetically predicted expression of Mendelian disease genes. One such disease is Acrodermatitis enteropathica, which includes phenotypes such as dermatitis, gastritis, serious behavioral problems, and anemia. Subjects with reduced predicted expression of *SLC39A4*, the underlying Mendelian gene, are at greater risk for similar phenotypes including behavioral disorders (p<3.4E-9), skin disease (p<1.3E-11), anemias (p<1.3E-13) and digestive system symptoms (p<6.8E-6). Another example is Homocystinuria (due to mutations at *CBS*), most commonly affecting the eyes, central nervous system, skeleton, and circulatory system. Subjects with reduced predicted expression of *CBS* are at a greater risk for ocular disease/defects (p<1.4E-7), intellectual disabilities (p<1.0E-10), spine disorders (p<4.9E-7), and cardiomyopathies (p<3.3E-7). The identification of the genes underlying Mendelian diseases has sometimes suggested relatively innocuous therapies for the disease, such as vitamin or mineral supplementation, or removal of particular foods from the diet. Zinc supplementation is an effective therapy for Acrodermatitis enteropathica as is vitamin B6 supplementation for Homocystinuria. While Mendelian diseases are quite rare, the number of individuals who would benefit from these same innocuous therapies is large when we consider those at highly increased risk for the same serious phenotypes that comprise the disease spectrum for a given Mendelian disease due to the reduced genetically regulated expression of those genes. Collectively, there will be more individuals who would enjoy improved health through treatment with these innocuous therapies for the small number of Mendelian genes where such therapies have been developed than there are people living today with any Mendelian disease.

16

The public sharing of genomic data from the DiscovEHR Collaboration.

M.D. Ritchie[1], J.B Leader[1], T.N. Person[1], F.E Dewey[2], M.F. Murray[1], J.D. Overton[2], H.L. Kirchner[1], A.E. Lopez[2], J. Penn[2], I.B. Borecki[2], G.D. Yancopoulos[2], F.D. Davis[1], W.A. Faucett[1], O. Gottesman[2], J.G. Reid[2], A. Baras[2], D.J. Carey[1], A.R. Shuldiner[2], D.H. Ledbetter[1], Geisinger-Regeneron DiscovEHR Collaboration. 1) Geisinger Health System, Danville, PA; 2) Regeneron Genetics Center.

The DiscovEHR collaboration, between the Regeneron Genetics Center and Geisinger Health System brings together high-throughput sequencing with a large Healthcare Provider Organization that captures a median of 14 years of longitudinal electronic health record (EHR) data in the patient population. This powerful combination serves as a blueprint for large-scale precision medicine research and genomic medicine implementation. Through the sequencing of exomes from more than 50,000 MyCode® participants to date, we have identified more than 4 million rare single nucleotide variants and insertion-deletion events, of which over 176,000 are predicted to result in loss of gene function. An open-access DiscovEHR web browser will be launched at http://www.geisinger.org/for-researchers/. Through this portal, variant frequency data will be made publicly available to enable allele frequency comparisons with other population-based and biobank resources. The community will be able to download a .vcf with allele frequencies (exact for MAF > 0.001 and binned for <0.001) and search by gene, rs#, and position through the browser. Through DiscovEHR, we will also deposit CLIA-confirmed rare variants with clinical relevance into ClinVar based on the 56 genes nominated by the ACMG as clinically actionable, along with an additional 20 genes selected by the Geisinger Institute for Genomic Medicine. We find that 4.4% of MyCode participants harbor deleterious variants in the 76 clinically actionable genes, and have initiated the process to act clinically on this information. For example, we have identified 35 variants (in *LDLR*, *APOB*, and *PCSK9*) associated with Familial Hypercholesterolemia in 229 individuals who have been evaluated using clinical associations from the EHR and will be deposited into ClinVar. We plan to implement this variant identification and clinical validation process for all 76 genes; continuously populating ClinVar with this knowledge. The DiscovEHR portal will serve as a valuable resource for the genetics community for both discovery research and clinical applications. While very rare variants are not available at their precise frequency (to protect participant privacy), DiscovEHR is open to future collaborations and look-ups for specific scientific questions. The dissemination of this resource to the public through a web-based portal, as well as ClinVar deposition of rare clinically relevant variants, will enhance discovery in the scientific community.

17

Cancer risks associated with predisposition gene mutations identified by hereditary cancer panel testing of 85,000 patients.

F.J. Couch[1,2], D.E. Goldgar[3], H. Shimelis[1], J. Lilyquist[2], C. Hu[1], M. Akinhanmi[1], J. Na[2], E.C. Polley[1,2], S.N. Hart[2], R. Huether[4], C. Espenschied[4], R. McFarland[4], T. Pesaran[4], H. LaDuca[4], J.S. Dolinsky[4]. 1) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN; 2) Department of Health Sciences Research, Mayo Clinic, Rochester, MN; 3) University of Utah, Salt Lake City, UT; 4) Ambry Genetics, Aliso Viejo, CA.

Clinical genetic testing of individuals with a personal or family history of breast and ovarian cancer using panels for *BRCA1/2* and other candidate cancer predisposition genes have become routine clinical practice. While the cumulative lifetime and age specific risks associated with mutations in *BRCA1/2* in high-risk families and in the general population are well understood, the risks of breast cancer associated with mutations in many of the other panel genes are not well defined. To estimate risks of breast, ovarian, and other cancers associated with inherited deleterious mutations in cancer predisposition genes we utilized results from 85,379 hereditary cancer panel tests performed by Ambry Genetics. Among these 63,368 reported a personal history of cancer other than basal skin cancer, including 46,320 breast cancers, 6194 ovarian cancers, 4896 colorectal cancers, 2883 endometrial cancers, 1252 pancreatic cancers, 312 gastric cancers, along with smaller numbers of other cancers. Of those with breast and/or ovarian cancer, greater than 90% met National Comprehensive Cancer Network HBOC testing criteria. To estimate gene-specific risks for individual cancers, case-control analyses were performed comparing the frequencies of pathogenic mutations from Caucasian cancer cases with frequencies from Caucasian, non-Finnish, non-TCGA controls from the Exome Aggregation Consortium (ExAC) database. Using breast and ovarian cancer as examples, mutations in ATM and CHEK2 genes were associated with moderate risks (Relative Risk (RR)>2) of breast cancer and limited risk of ovarian cancer, as expected. Pathogenic mutations in CDH1, NF1, and RAD51D were also associated with moderate risks of breast cancer, whereas PALB2 was associated with high risks (RR>5.0) of breast cancer. In contrast, RAD51C, RAD51D, BRIP1, and PALB2 mutations were associated with high risks of ovarian cancer. Risks for several cancers associated with mutations in the various panel-testing genes will be presented. This large clinical testing dataset in combination with public controls provides useful data for many predisposition genes previously lacking risk estimates, and should prove useful for clinical risk management of patients with inherited mutations in these genes.

18

Rare variant cancer association studies with heterogeneous sequencing datasets.

C.D. Huff[1], Y. Yu[1], F. Hu[1], J. Chen[1], S. Chen[1], H. Hu[1], A.S. Deshpande[1], S. Sivakumar[1], Y. Liu[1], J. Fowler[1], S. Shankaracharya[1], B. Moore[2], C. Holt[2], Y. Ye[1], M. Hildebrandt[1], H. Zhao[1], P. Scheet[1], X. Wu[1], M. Yandell[2]. 1) Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX; 2) Department of Human Genetics and USTAR Center for Genetic Discovery, University of Utah School of Medicine, Salt Lake City, UT.

Whole-exome sequencing data is increasingly becoming available to the research community for secondary analyses, providing new opportunities for studies that directly test rare variant, common disease hypotheses. However, the heterogeneous nature of these datasets is a major barrier to large-scale sequencing association studies that incorporate data from multiple sources. Minor differences in sample preparation and sequencing protocols often result in strong technological stratification biases that overwhelm subtle signals of disease association. These biases can be reduced through the use of joint calling and standard quality control procedures. However, as we demonstrate, these approaches alone typically result in poor signal-to-noise ratios and unacceptably high levels of Type I error inflation in exome sequencing association studies with heterogeneous data sources. To address this problem, we developed XQC, a new toolkit to support high-throughput sequencing association studies that greatly mitigates and in some cases eliminates Type I error inflation resulting from technological stratification. XQC optimizes joint variant calling and recalibration procedures based on the target region of each platform, detects and filters variants influenced by technological stratification biases, and assesses population stratification and residual technological stratification from well-behaved markers in each platform. We applied XQC to evaluate the contribution of rare, protein-coding variation to cancer risk by conducting a series of whole-exome case-control studies using VAAST 2.1. The cases consisted of individuals of European ancestry from TCGA involving the following cancer types: breast (783 cases), colorectal (362 cases), melanoma (314 cases), ovarian (272 cases), pancreatic (156 cases), soft tissue sarcoma (219 cases), and pheochromocytoma and paraganglioma (144 cases). The controls consisted of 1726 females and 1781 males of European ancestry. Our results replicate many established susceptibility genes at $p < 0.05$ and provide odds ratio estimates for rare missense and nonsense variation in each gene. Our results also provide support for promising new candidate genes while providing an effective toolkit for combining heterogeneous datasets in large-scale rare variant association studies.

19

Whole genome sequencing identifies germline coding and non-coding variants in familial glioma. *M.N. Bainbridge[1,2,3], G. Armstrong[4], M.C. Wood[3], S. White[1], D.M. Muzny[1], S. Jhangiani[1], R. Gibbs[1,2], M. Bondy[4,5].* 1) HGSC, Baylor College Med, Houston, TX; 2) Molecular and Human Genetics, Baylor College Med, Houston, TX; 3) Codified Genomics, LLC, Houston, TX; 4) Duncan Cancer Center, Baylor College Med, Houston, TX; 5) Department of Pediatrics, Baylor College Med, Houston, TX.

Glioma is a rare cancer associated with poor clinical outcome. Genetic susceptibility plays a significant role in glioma development. There are a few known associations with familial glioma including germline mutations in TP53, RTEL1, NF1/2 and telomere maintenance gene POT1, however, the vast majority of patients' families do not have mutations in these genes. Previous work using whole *exome* sequencing identified only a small number of likely causative alleles. Thus, to decipher the coding and non-coding genetic components of this disease, we whole *genome* sequenced (WGS) germline DNA from 256 familial glioma samples, the largest such cohort ever assembled. Using a high-throughput pipeline we identified structural, small indel and single nucleotide variants as well as the telomere and viral DNA content of the sample. Further, we identified coding and non-coding variants which affected transcription factor binding sites (TFBS), 3’ UTR miRNA binding sites, 5’ UTR start-gains, or created cryptic splice sites. We prioritized the ~1 billion discovered variants using an automated system which scores variants based on MAF, conservation of the affected residue, predicted effect on the gene or TFBS, and previous association of the gene to cancer, including known glioma GWAS peaks and important pathways in gliomagenesis. We further refined these results by prioritizing genes/pathways that are predicted loss of function intolerant in a large population dataset as well as in a control cohort of 1037 ethnically matched WGS samples. Using this approach, we identified numerous truncating variants in known hereditary cancer genes (*MLH1, POLD1, RASSF1, BRIP1, PMS2, MLH3*) and truncating or highly deleterious mutations in known glioma genes (*POT1, RTEL1, NF2, TP53*). In addition, we found rare, highly deleterious or cryptic splicing variants in the conserved mTOR-pathway genes (*MTOR, MLST8, DEPDC5*) and NF1-downstream genes (*PAK1, RAF1, ADCY5*). Furthermore, based solely on extreme population and evolutionary conservation, we prioritized two coding and one cryptic splicing variant in *SLIT3*, a gene which is epigenetically inactivated in cancer and has once been described as associated with pediatric glioma. WGS is a powerful tool for discovery of coding and non-coding variants but requires extensive analytic pipelines, databases, phenotypic association and intelligent filtering methodologies to identify candidate variants from background noise.

20

Genes involved in base excision and direct DNA repair contribute to hereditary breast cancer. *I. Campbell[1], N. Li[1], S. Rowley[1], L. Devereux[1], A. Trainer[2], P. James[2], Lifepool.* 1) Research Div, Peter MacCallum Cancer Ctr, East Melbourne, Victoria, Australia; 2) Familial Cacner Centre, Peter MacCallum Cancer Ctr, East Melbourne, Victoria, Australia.

Identifying the missing hereditary factors of familial breast cancer could have a major and immediate impact on reducing breast cancer risk in these family members. Up to 1,325 candidate breast cancer predisposition genes, identified through exome sequencing of BRCAx families, were sequenced in index cases of up to 4,000 BRCAx families and 4,000 cancer free women from the LifePool study in Australia. Interrogation of the data to refine the highest priority candidates is ongoing, but it is noteworthy that known (PALB2) or suspected (MRE11A) moderately penetrant breast cancer genes showed enrichment of loss of function (LoF) mutations in this dataset. Conversely, some other recently proposed breast cancer genes (BRIP1 and RINT1) did not show a significantly higher LoF mutation frequency in the cases compared to controls. Based on the number of LoF mutations leading candidates include NTHL1 (12 cases versus 4 controls) and ALKBH1 (7 cases versus 2 controls) which are each important members of the base excision repair and direct nucleotide repair pathways. We examined other genes in the base excision and direct repair pathways that were on our sequencing capture design and observed a significant enrichment of potentially deleterious mutations in 12 genes (NTHL1, OGG1, APEX1, APEX2, NEIL1, NEIL2, NEIL3, MUTYH, MPG, ALKBH1, ALKBH2, ALKBH3): Among the 1,638 cases and 1,654 controls analysed to date, 76 LoF variants were detected in these genes among the cases versus 47 LoF variants among the controls (p=0.007). Based on the overall distribution of variants between cases and controls the probability of selecting 12 genes with such enrichment from the 1,325 genes screened was less than 1 in 200. Our data implicates rare mutations in base excision and direct DNA repair pathways genes as moderate-penetrance breast cancer susceptibility alleles.

21

Inherited deleterious germline variants in men with prostate cancer identified by whole exome sequencing. *N. De Sarkar[1], C. C. Pritchard[2], P. Nelson[1,2].* 1) Human Biology, Fred Hutchinson Cancer Research Center., Seattle, WA; 2) Department of Laboratory Medicine, University of Washington, Seattle, WA.

Introduction: Whole exome sequencing (WES) is a platform for the detection of clinically relevant alterations in tumors, making its way from research to the practice of precision oncology. Through efforts designed to define the landscape of somatic exome mutations in cancers, large databases of tumor and matched germline WES have been assembled. Such data has created an opportunity to identify cancer-linked deleterious germline variants that may be prognostic or predictive of outcomes to specific therapeutics. In the present study, we have analyzed germline WES from 499 men comprising the TCGA prostate cancer (PC) project and 142 men comprising the SU2C/PCF PC project to identify clinically relevant deleterious mutations.**Methods:** We applied a germline variant calling and filtering pipeline on germline WES data for 124 genes involved in DNA repair (DRG). We focused on variants classified as deleterious and likely deleterious. We compared the frequency of variant calls in men with localized PC to those with metastatic PC and with frequencies in individuals without cancer in the ExAC database.**Results:** Several deleterious variants in *BRCA1* and *BRCA2* were observed in men with PC and enrichment of deleterious *BRCA2* variants were observed in men with metastatic PC compared to localized PC ($p<0.0001$). Variants in *ATM* and several DNA mismatch repair genes such as *PMS1* and *MSH6* have similar mutation frequencies in men with localized or advanced cancers. *ATM* deleterious variants were enriched in PC cases compared to frequencies in the ExAC population ($p=0.013$). Several deleterious and predicted deleterious mutations were identified in other Fanconi complex genes in the TCGA (2.6%) and SU2C/PCF (0.7%) series. We did not identify any mutation stratification between NCCN defined high-risk cases vs low-risk PC. We observed the presence of deleterious heterozygous variants in several genes including *PMS1, POLE* and *DCLRE1C.* Confirmation of second hits might have implication for disease associations.**Conclusions:** The deleterious and likely deleterious variants evaluated in this study are rare in the normal population and not represented in most genome-wide association studies. In light of Knudson's 2 hit hypothesis deleterious loss of function germline mutations in DNA repair genes might be conferring risks for developing PC and to adverse outcomes. Prospective studies are needed to determine if the germline DRG mutations are prognostic or predictive of clinical outcomes.

22

Ascertainment bias in predicting genetic disease risks. *J. Lachance, K. Patel, A. Teng, A.J. Berens.* School of Biology, Georgia Institute of Technology, Atlanta, GA.

As highlighted by the Precision Medicine Initiative, there is a clear need to use personal genomic information to predict health and disease. An accurate assessment of health disparities requires knowledge of genetic risks in different populations. However, the vast majority of genome-wide association studies (GWAS) have used samples of European ancestry, and it is not cost effective to repeat every study across dozens of global populations. Because of this, an outstanding public health challenge is the prediction of genetic disease risks in non-GWAS populations. Here, we integrate whole genome sequence data from the 1000 Genomes Project with the NHGRI-EBI GWAS Catalog to predict hereditary disease risks across the globe. We then combine empirical data with mechanistic models of population genetics to infer properties of disease risk alleles and populations that are relevant to the generalization of GWAS results. Evolutionary history influences how well results can be generalized across populations (including whether genomic regions are evolving neutrally or under selection and whether populations have diverged recently or in the deep past). We find substantial bias in genetic risk score predictions for ancestral and derived alleles: risks are underestimated in study populations if ancestral alleles tag diseases and risks are overestimated in study populations if derived alleles tag diseases. We find that statistical considerations also contribute to bias in predicting genetic risk including the amount of linkage-disequilibrium in study populations, sample sizes, and choice of genotyping array. Furthermore, genetic architecture affects predicted disease risks. For example, admixed genomes have increased risk if diseases are dominant and decreased risk if diseases are recessive. Focusing on the genetics of cancer, we find that some health disparities (e.g. elevated prostate cancer risks in men of African descent) persist after correcting for ascertainment bias. In sum, we find that simply adding up the number of known disease alleles in an individual's genome is not sufficient, as biased subsets of disease-causing alleles have been missed. By correcting for ascertainment bias, major improvements can be made in the prediction of hereditary disease risks in diverse human populations.

23

The utility of whole exome sequencing in prenatal diagnosis.

M. Walkiewicz[1,2], A. Braxton[2], P. Liu[1,2], F. Xia[1,2], W. Bi[1,2], R. Xiao[1,2], M. Leduc[1], J. Zhang[1], X. Wang[1], L. Meng[1], W. He[2], F. Vetrini[1], P. Ward[1,2], S. Narayanan[1], S. Nassef[1], S. Plon[1], D. Muzny[1,3], J. Lupski[1], R. Gibbs[1,3], I. Van den Veyver[1], Y. Yang[1,2], C. Eng[1,2]. 1) Molecular and Human Genetics, BCM, Houston, TX., Select a Country; 2) Baylor Miraca Genetics Laboratory, Houston,TX, USA; 3) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX.

During the past several years clinical whole-exome sequencing (WES) has been utilized for patients with complex clinical presentations. With a molecular diagnosis rate of about 30%, clinical WES is an effective way of addressing the expensive and time-consuming diagnostic odyssey. With the continuous improvement of WES methodology including updated sequencing instrumentation, shorter turnaround time (TAT), and improved analysis, clinical WES is now being used for prenatal diagnosis. Here, we report the first 61 consecutive prenatal cases from a single laboratory. Prenatal cases are defined as a fetal sample obtained through either a diagnostic procedure or a product of conception (POC). The testing performed includes 40 cases with proband WES followed by Sanger sequencing studies of parental samples and 21 trio WES. Sequencing was performed on the Illumina HiSeq2500 with an average of ~11.4 Gb of data per exome and 97% of the targeted exome regions are sequenced at a depth of 20X. Exome sequencing data were analyzed for small nucleotide changes and large CNVs using the Illumina HumanExome-12v1 array. The proband WES yielded a molecular diagnosis in 31% (12/39) cases whereas trio WES provided molecular diagnosis in 38% (8/21) cases. Diagnoses were reported for 50% (3/6) individuals with a brain anomaly on ultrasound, 37% (10/27) individuals with a brain anomaly and other organ system involvement, and 29% (8/28) individuals with an anomaly not including the brain. For 37 individuals this was the first presentation of the disorder in the family and molecular diagnosis was reported for 41% (15/37) including 8 *de novo* variants, 5 compound heterozygous variants, 1 homozygous variant, and 1 autosomal dominant variant with a mosaic parent. For 24 individuals with a previously undiagnosed family history of a similar phenotype molecular diagnosis was made for 25% (6/24) including 5 compound heterozygous variants and 1 maternally inherited hemizygous variant. With rapid turnaround time and comprehensive analysis, the prenatal trio WES is a valuable tool in prenatal diagnosis revealing a wide variety of underlying Mendelian mechanisms with and without a family history of genetic disorders.

24

The genomic autopsy: Using whole exome and whole genome sequencing to solve complex fetal and neonatal presentations.

C.P. Barnett[1,7], A. Byrne[2,3,8], L. Moore[4], Y. Khong[4], N. Manton[4], J. Lipsett[4], S. Yu[2], M. Dinger[5], M. Babic[2,3], P.J. Brautigan[2,3], Q. Schwartz[3], P.Q. Thomas[6], C.N. Hahn[2,3], F. Feng[6,7], A.W. Schreiber[6,7], K. Kassahn[2,3], H.S. Scott[2,3,6,7,8,9]. 1) Pediatric & Reproductive Genetics Unit, Women's and Children's Hospital, North Adelaide, South Australia, Australia; 2) Department of Genetics and Molecular Pathology, SA Pathology; 3) Centre for Cancer Biology, An alliance between SA Pathology and University of South Australia; 4) Department of Anatomical Pathology, SA Pathology at the Women's and Children's Hospital; 5) Kinghorn Centre for Clinical Genomics, the Garvan Institute; 6) School of Biological Sciences, University of Adelaide; 7) School of Medicine, University of Adelaide, Adelaide, South Australia, Australia; 8)) School of Pharmacy and Medical Sciences, Division of Health Sciences, University of South Australia, Adelaide, South Australia, Australia; 9) ACRF Cancer Genomics Facility, Centre for Cancer Biology, SA Pathology.

Background: Congenital abnormalities are the most frequent reason for stillbirth and termination of pregnancy between 18-22 weeks gestation. Late fetal deaths *in utero*, or in the newborn period, affect ~3.5/1000 births and in 60% of cases the precise cause of death is not clinically apparent. Formal post mortem examination is performed in South Australia in ~60% of these mid-term and late losses (200 post-mortems/year). After chromosomal abnormalities have been excluded, no definitive cause of the congenital abnormalities or perinatal death is identified by the post-mortem in ~50% of cases. Aim: To use whole exome sequencing (WES) and whole genome sequencing (WGS) to identify genetic causes of fetal and newborn abnormalities that result in termination of pregnancy, death due to congenital abnormalities, death in utero or death in the newborn period, in view to providing families with answers regarding cause and likelihood of recurrence of these congenital problems. Methods: We are performing WES or WGS on cases of fetal and perinatal death, using either an Illumina HiSeq 2500 (WES) or X Ten System (WGS). Samples are available for ~1,200 cases. High priority cases are fetuses with congenital abnormalities with consanguineous parents (singles); fetuses with multiple malformations; and unexplained fetal/newborn death (trios). Bioinformatic and experimental laboratory techniques are being used to confirm causality of variants. Results: This project is in its first year. Of the 2 WES/ 6 WGS's performed on 8 families to date, 4 have resulted in the identification of causative variants; 2 in known disease genes and 2 in novel disease genes. These novel discoveries are supported by in-depth functional evidence. A further 2 cases likely representing novel discoveries are currently under investigation to prove causality, with a third case expected to represent an allelic disorder of a known gene. Data will be presented from the first WGS's performed, with specific focus on our discovery of a new autosomal recessive polycystic kidney disease gene, confirmed in a CRISPR-Cas9 modified homozygous mutant mouse model. The discovery of this disease gene has led to a successful reproductive outcome for a family. Discussion: Genomic autopsy using WES/WGS offers enormous potential as an adjunct to traditional autopsy in providing accurate genetic counselling for families who have experienced pregnancy loss, death in utero, termination of pregnancy or death in the newborn period.

25

Gene identification in fetal malformation phenotypes. *I. Filges[1], N. Meier[1], O. Lapaire[2], S. Wellmann[3], P. Miny[1], I. Hösli[2], E. Bruder[4], S. Teranli[5].* 1) Medical Genetics, University Hospital Basel, Basel, Switzerland; 2) Obstetrics and Gynecology, University Hospital Basel, Basel, Switzerland; 3) Neonatatology, University Children's Hospital Basel, Basel, Switzerland; 4) Pathology, University Hospital Basel, Switzerland; 5) Center for Ultrasound, Freie Strasse, Basel, Switzerland.

Prenatal ultrasound identifies an increasing number of fetal malformation phenotypes of unknown cause. For a significant number of these, often lethal, phenotypes the underlying causal mutations are not identified yet, but animal data predict that up to 30% of the protein-coding genes of our genome are implicated in embryonic development. Little attention has been paid to gene identification in these disorders which often may present with a specific phenotype during fetal life. We aim at identifying novel disease genes where mutations lead to autosomal recessive forms of rare malformation phenotypes and to describe the particular prenatal presentation. Whole exome sequencing (WES) is used in selected families with phenotype recurrence in sibs who died during pregnancy or after birth because of their malformations. We correlate the malformation pattern, confirmed by autopsy, to developmental pathways in embryogenesis. Genes with homozygous or compound heterozygous variants will be considered candidates. Those harboring truncating variants will be prioritized, since loss of function variants are more likely to be causal compared to other variant classes. The comparison of human and animal morphology is an important means to validate novel potentially causal genes. We present several novel prenatal malformation patterns, identified by ultrasound, and their phenotypic variability. Among the genes we identified are *KIF14*, *CENPF* and others we will present, that are involved in early cell division processes, and are functionally linked to developmental pathways of ciliary disorders. Comparisons to animal model phenotypes further support causality of the mutations. Genes involved in cell division may be considered important candidates in early human development. We illustrate the successful application of WES approaches to identify mutations in recessive disease genes leading to early errors of morphogenesis. It is important to document prenatal findings in order to precisely characterize prenatal phenotypes. Identifying the mutations will improve recurrence risk counseling and allow prenatal diagnosis for future pregnancies in affected families. Continuing delineation of specific fetal phenotype-genotype correlations will be one of the requirements for successful implementation of WES in a prenatal setting in the future. Lethal fetal phenotypes may also represent an important model to study the roles of genes for which little to nothing is known.   .

26

The BabySeq Project: Preliminary findings from a randomized trial of exome sequencing in newborns. *R.C. Green[1,2,3,4], I.A. Holm[2,5,7], H.L. Rehm[1,2,3,4], A.L. McGuire[6], P.B. Agrawal[2,5,7], R.B. Parad[1,2,5], M.H. Helm[1], C.A. Genetti[5,7], A.H. Beggs[2,5,7] for the BabySeq Project.* 1) Brigham and Women's Hospital, Boston, MA; 2) Harvard Medical School, Boston, MA; 3) Partners Healthcare Personalized Medicine, Cambridge, MA; 4) The Broad Institute of MIT and Harvard, Cambridge, MA; 5) Boston Children's Hospital, Boston, MA; 6) Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX; 7) The Manton Center for Orphan Disease Research, Boston, MA.

Background: Newborn genomic sequencing (GS) has the potential to provide comprehensive and clinically useful information on a range of conditions. The BabySeq Project is the first randomized clinical trial exploring the impact of GS on infants and their parents. Methods: Families are enrolled from Boston Children’s Hospital and Brigham and Women’s Hospital (BWH) ICUs and from the BWH well baby nursery. Within each cohort, half of the infants are randomized to receive GS. All families receive results during a session that involves review of state-mandated newborn screening results and family history. Families randomized to GS also receive a report that includes pharmacogenomic (PGx) variants and pathogenic and likely pathogenic variants in 1800 genes associated with dominant and recessive monogenic conditions with childhood age of onset and high estimated penetrance, without regard for actionability. For infants with a phenotype, an indication-based analysis of candidate genes is performed and variants of uncertain significance in these genes are also returned. Medical record reviews and surveys collect medical, behavioral and economic outcomes. Results: To date the study has been offered to the parents of 254 ICU infants and 1193 healthy infants, of whom 14 (6%) and 85 (7%) respectively have enrolled. Parents who declined before meeting with a study genetic counselor most often cited study logistics as their primary reason. Parents who declined after meeting with a study genetic counselor most often indicated one or more of these reasons for declining: potential to receive unfavorable or uncertain results (38%), insurance discrimination (26%), and confidentiality/privacy concerns (22%). Time from DNA extraction to issuance of GS report has averaged 50 days. GS results have been returned to the families of 4 ICU infants and 27 healthy infants. Two dominant monogenic variants have been identified in 2 healthy infants where no medical or family history suggestive of either disease was reported: a pathogenic, paternally inherited *ELN* variant, and a likely pathogenic, maternally inherited *VCL* variant. Recessive carrier variants have been returned for 26 infants, and PGx variants have been returned for 2 infants. Conclusion: The BabySeq Project is a proof-of-concept trial examining the risks and benefits of implementing GS in newborn clinical care. Additional recruitment and outcomes data are being collected and will be reported.

27

Clinical utility of expanded carrier screening: Reproductive behaviors of at-risk couples. *K.K.L. Wong[1], K. Ready[1], C. Lieber[1], J.D. Goldberg[1], I.S. Haque[1], G.A. Lazarin[1], C. Ghiossi[2].* 1) Counsyl, South San Francisco, CA; 2) California State University: Stanislaus, Turlock, CA.

Introduction: Expanded carrier screening (ECS) analyzes dozens or hundreds of recessive disease genes for couples planning to have children. The literature on the clinical utility of screening conditions is scarce. We surveyed at-risk carrier couples, identified through ECS, to learn about their reproductive decisions.**Methods:** Patients underwent ECS via Counsyl laboratory for up to 110 genes. At-risk carrier couples (ARCC) were those in which both partners were carriers for the same autosomal recessive diseases. We invited 537 ARCC who received their results between April 2014 and August 2015 to participate via email, SMS message, and paper survey in an IRB-approved study questionnaire.**Results:** Of 537 eligible participants, 76 completed the questionnaire; 12 who reported a family history were excluded from analysis. 45 (70%) were not pregnant at time of screening. Of the 45 participants that were not pregnant, 62% indicated that they would choose IVF with PGD, prenatal diagnosis, gamete donation, adoption, or no reproduction. 29% indicated that they were not planning to alter reproductive plans, indicating perceived severity as a major reason. The remainder did not indicate clear plans. Of the 19 participants that were pregnant, 42% (8) elected prenatal diagnosis. Of the remainder, 2 reported interest in testing, but miscarried before the procedure could be done while 9 did not consider the condition sufficiently severe to consider pregnancy termination. Of 8 pregnancies that underwent prenatal diagnosis, 5 were unaffected and 3 were affected. 2 of the affected pregnancies was terminated and 1 was continued. Fisher’s exact test revealed the association between the severity of the disease and clinical utility was significant (p=0.000145), whereas the association between pregnancy status and clinical utility was not (p=0.088).**Conclusions:** Most ARCC altered reproductive planning, demonstrating clinical utility of this information. Perceived severity of the condition factored into decision making with milder diseases less likely to change planning.

28

Noninvasive prenatal screening for common aneuploidies in a Canadian province: A cost effectiveness analysis. *F. Rousseau[1], L. Nshimyumukiza[2], J.A. Beaumont[3], J. Duplantie[2], S. Langlois[4], J. Little[5], F. Audibert[6], C. McCabe[7], J. Gekas[1], Y. Giguère[1], C. Gagné[3], D. Reinharz[2].* 1) CHU de Québec-Université Laval, Quebec, Québec, Canada; 2) Département de médecine sociale et préventive, Faculté de médecine, Université Laval, Québec (Qc), Canada; 3) Département de génie informatique et logiciel, Faculté de sciences et de génie, Université Laval, Québec (Qc), Canada; 4) Department of Medical Genetics, University of British Columbia, Vancouver (BC), Canada; 5) School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa (ON), Canada; 6) Département d'obstétrique-gynécologie, Faculté de médecine, Université de Montréal, Montreal (QC), Canada; 7) Department of Emergency Medicine University of Alberta Hospital, Edmonton (AB), Canada.

BACKGROUND: Non-invasive prenatal testing (NIPT) using cell-free fetal DNA in maternal plasma has been developed for prenatal screening for Down syndrome (DS) and other common severe aneuploidies [Trisomy 18(T18), Trisomy 13 (T13)]. Although it has been reported to have a high accuracy, only little evidence about its cost effectiveness (CE) is available. **OBJECTIVE**: To evaluate, using computer simulations, the CE of fetal aneuploidy screening strategies involving NIPT. **METHODS**: A semi-Markov agent-based model was used to simulate the CE of 13 screening strategies for aneuploidies (6 current prenatal testing strategies, 1 for universal NIPT and 6 incorporating NIPT as contingent test) in the context of Quebec public health care. Comparisons were made for a virtual cohort with maternal age distribution similar to that of expected Quebec pregnant women in 2014. Data input parameters were retrieved from a thorough literature search and in government databases. The outcomes considered were the total direct costs borne by the Quebec universal health care system, total number of chromosomal anomalies detected (DS, T18, T13), number of invasive procedures and number of euploid fetal losses. The outcomes measured for cost effectiveness analysis were cost per T21 case detected and the incremental cost per additional T21 case detected (ICER). **RESULTS**: at a baseline risk cut-off of 1:300 (combined FTS, integrated, serum integrated, and QUAD), a risk cut-off of 1:30 (contingent and sequential) and a cost of $795 per NIPT sample, strategies with NIPT as contingent test are less costly and most safer (fewer procedure-related losses) but detect 10-13% fewer aneuploidy cases than their corresponding current options. The universal (first line) NIPT detects more aneuploidy cases (9-32.5%), has fewer procedure-related losses but is much more expensive than other screening strategies. Among the 13 options tested, the Serum Integrated_NIPT is the most cost effective option followed by the Serum Integrated with an ICER of $ 61 623. Results were sensitive to the NIPT unit cost and risk cut-offs used for current screening strategies in one way sensitivity analysis but were robust in probabilistic sensitivity analysis where the Serum Integrated_NIPT remain the most cost effective option at a threshold of $ 50 000 per additional DS case detected. **CONCLUSION**: The Serum Integrated_NIPT is likely to be the most cost-effective option.

29

Identifying genetic ASD-risk factors in over 2,000 whole-genome sequenced familial samples. *E.K. Ruzzo[1,4], L. Perez-Cano[1,4], J. Jung[2], D. Kashef[2], L. Wang[1], S. Sharma[2], M. Duda[2], G.M. McInnes[2], J.K. Lowe[1], D.H. Geschwind[1], D.P. Wall[3].* 1) Center for Neurobehavioral Genetics and Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, California, USA; 2) Division of Systems Medicine, Department of Pediatrics, and Department of Biomedical Data Science, School of Medicine, Stanford University, Stanford, California, USA; 3) Department of Pediatrics, Psychiatry and Biomedical Data Science, Stanford University, Stanford, California, USA; 4) These authors contributed equally to this work.

As part of the Hartwell Autism Research and Technology Initiative (iHART), we generated whole-genome sequencing data for >4,000 individuals from the family-based Autism Genetic Resource Exchange (AGRE) cohort. Recently, exome sequencing studies demonstrated the importance of *de novo* mutation to ASD risk in simplex families. The iHART project seeks to 1) identify novel risk factors in noncoding regions, 2) identify structural variants at high-resolution, and 3) elucidate the role of inherited risk loci. To date, we have analyzed 422 families containing at least two affected children (some monozygotic twins) and both biological parents. Whole-genome sequence data were generated with the Illumina HiSeqX. Single nucleotide variants and indels were identified following GATK's best practices. Raw structural variant (SV) calls were generated by BreakDancer, SMuFin, GenomeSTRiP, and LUMPY. After QC filtering, we categorize the inheritance of each variant in a child (e.g., *de novo*, paternally inherited, unknown phase) and annotate variant/gene properties to facilitate filtering for likely pathogenic variants, such as those transmitted to all affected children. We observe 35 intolerant genes (RVIS) harboring a rare (AF <0.01%) likely protein-disrupting variant transmitted to all affected children in ≥4 families. One such gene, *TTC3*, harbors qualifying variants in six families and lies within the Down syndrome critical region. We also found rare large deletions transmitted to all affected children in ≥2 families with ≥90% overlap with four known syndromic deletion loci. After Bonferroni correction, three large deletion events were significantly enriched in affected vs. unaffected family members, including the 17q11.2 deletion identified in seven families (p=5.1x10^{-7}). In addition to inherited risk analyses, we observe an enrichment of rare *de novo* variants in a coexpression module found in early fetal brain development (p=0.042) and identify *de novo* variants in syndromic autism genes including *SHANK3*, *TSC2*, and *DDX3X*. We will continue to explore the role of noncoding variation by leveraging experimentally defined regulatory regions in human fetal brain. The iHART genome data will be housed in a cloud-computing repository and made available as a community resource. In addition to advancing our understanding of autism genetics, the iHART project represents a landmark for the study of other complex diseases and investigations using whole-genome sequence data.

30

Whole genome sequence-based resource for autism research. *R.K.C. Yuen[1,2], D. Merico[1], J.L. Howe[1], B. Thiruvahindrapuram[1], J. Whitney[1], R. Patel[1], N. Hoang[3], J.R. MacDonald[1], Z. Wang[1], T. Nalpathamkalam[1], W. Sung[1], S. Walker[1], J. Wei[1], C.R. Marshall[1,4], G. Pellecchia[1], A. Chan[1,5], L. D'Abate[1,5], M. Zarrei[1,2], M. Uddin[1,2], M. Bookman[6], N. Deflaux[6], E. Anagnostou[7], L. Zwaigenbaum[8], B.A. Fernandez[9,10], P. Szatmari[11,12,13], B.M. Knoppers[14], M. Elsabbagh[15], D. Glazer[6], M. Pletcher[16], S.W. Scherer[1,2,5,17].* 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) Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, Toronto, ON, Canada; 4) Department of Molecular Genetics, Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada; 5) Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada; 6) Google, Mountain View, California, USA; 7) Bloorview Research Institute, University of Toronto, Toronto, Ontario, Canada; 8) Department of Pediatrics, University of Alberta, Edmonton, Alberta, 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) Autism Research Unit, The Hospital for Sick Children, Toronto, Ontario, Canada; 12) Child Youth and Family Services, Centre for Addiction and Mental Health, Toronto, Ontario, Canada; 13) Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada; 14) Public Population Project in Genomics and Society, McGill University, Montreal, QC, Canada; 15) Department of Psychiatry, McGill University, Montreal, QC, Canada; 16) Autism Speaks, Princeton, New Jersey, USA; 17) McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada.

Research into the genomics of autism has brought us to the recognition that its spectrum has profoundly complex etiologies, which will only be fully unraveled with novel approaches on a massive scale. In an international initiative named *MSSNG* (discovering the *MiSSiNG* information in autism), we are performing whole genome sequencing (WGS) in order to determine the complete genetic architecture of thousands of families having Autism Spectrum Disorder (ASD). Here, we report the WGS (>30X coverage) of 5,200 samples from individuals and family members with ASD, accompanied by detailed phenotype information, creating a database stored and accessible for research in the Google Cloud and made widely available through an internet portal with controlled (but unhindered) access. Applying our newly developed variant detection pipeline to the WGS data, we found an average of 73.8 *de novo* single nucleotide variant (SNVs) and 12.6 *de novo* insertion/deletions (indels) or copy number variations (CNVs) per autism subject. Collectively, 224 *de novo* and putative loss-of-function (LoF) mutations were found. Each sample was also assessed using a high-resolution microarray and the resulting CNV and genotype data was compared with the WGS, as well as other data in the literature, to yield a list of most relevant genes and CNVs for ASD diagnostics. We highlight eight novel ASD risk gene discoveries and discuss how specific genetic diagnoses in other families enable the sub-categorization of potential medical risk factors. Through support of open access research, MSSNG has enabled a better understanding of ASD and helps to set a course for other initiatives of its kind.

31

Uncovering somatic mosaicism in autism. *R.A. Barnard[1], D.R. Krupp[1], Y. Duffourd[2], S.A. Evans[1], S.J. Webb[3], R. Bernier[3], E. Fombonne[4], J.B. Riviére[5], B.J. O'Roak[1].* 1) Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR; 2) Genetique des Anomalies du Developpement, Université de Bourgogne, Dijon, France; 3) Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA; 4) Psychiatry, Oregon Health & Science University, Portland, OR; 5) Human Genetics, McGill University, Montreal, Québec.

Autism spectrum disorder (ASD) has a strong genetic component and complex genetic architecture that has yet to be fully elucidated. Somatic mosaic mutations (SMs) have been firmly implicated in several neurodevelopmental/brain disorders including epilepsy, cortical malformations, and overgrowth syndromes (Poduri *et al.*, 2013). Pathways underlying these syndromes, e.g. PI3K/mTOR are also implicated in syndromic and nonsyndromic ASD. In previous work focusing on germline *de novo* mutations, we were surprised to validate ~5% of new mutations as likely SMs (O'Roak *et al.*, 2012). To systematically evaluate the role of SMs in ASD, we are leveraging exome data of ~2,300 families from the Simons Simplex Collection (SSC), including parents, proband, and an unaffected sibling. We first examined presumed *de novo* germline calls published from the SSC (Iossifov *et al.*, 2014, Krumm *et al.*, 2015). We find strong evidence for 4% of sites being consistent with a somatic versus germline event (binomial p<0.001). Next, we recalled all exome positions in the SSC using a custom somatic SNV calling pipeline that utilized complementary calling approaches (VarScan 2.3.2, LoFreq 2.1.1, in house). Using single molecule molecular inversion probes, we fully validated predicted mutations for 100 families, including 24 sequenced deeply (median ~200x). We used this data to build a logistic regression model to score variants and differentiate true SMs from noise. We applied this model to the quad subset (n=1,781) of the SSC dataset to predict *high-confidence* SMs at various allele frequencies (AF)/coverage thresholds. We do not find strong evidence for global mutation burden in affected versus unaffected children, or fathers versus mothers. We estimate 0.15/child SMs in the unique exome sequence (min 15% AF). Consistent with their age, we see an ~3-fold increased somatic rate in parents versus children. We observe 10% of these parent mutations are transmitted *germline* to a child as occult *de novo* mutations, which requires an early embryonic origin. This finding has important potential implications for recurrence risk for families and may explain some instances of parents with subclinical ASD features. We find somatic missense mutations in previously implicated high-confidence ASD risk genes, including *CHD2*, *CTNNB1*, *KMT2C*, *RELN*, and *SYNGAP1*, further suggesting that this class of mutations, which are generally not detectable using standard methods, are contributing to population risk.

32

Contribution of mosaic variation to autism spectrum disorders. *D. Freed[1,2], J. Pevsner[1,2].* 1) Johns Hopkins School of Medicine, Baltimore, MD; 2) Kennedy Krieger Institute, Baltimore, MD.

Autism spectrum disorders (ASDs) are a group of highly heritable disorders which are diagnosed in an estimated 1 in 68 children. Prior studies have conservatively estimated *de novo* mutations as contributing to 27% of disease incidence in sporadic cases. Occasionally these experiments have indicated that the identified mutations are present as mosaics, arising post-zygotically. However comprehensive evaluation of the incidence of mosaic mutations and the contribution of these mutations to disease has not yet been performed. We investigated the occurrence of mosaic mutation in 8,938 individuals from 2,388 families in the Simons Simplex Collection (SSC) using previously generated whole-exome sequence data. We developed methods for the detection of mosaic variation directly from whole-exome sequence data of pedigrees. Application of these methods to SSC data resulted in the identification of 221 mosaic variants. We validated both the presence and mosaic status of the identified mosaic variants using Sanger sequencing, pyrosequencing, and read-backed phasing approaches. The results of our validation indicated that the precision of our classification of variants as mosaic was near 84% while all variants were validated in tissues in which they were detected. We found three missense or frameshift mosaic mutations in genes previously implicated in ASD in probands but none in siblings. Comparison of the rates of mosaic mutation in probands and unaffected siblings through measurement of ascertainment bias indicated that 42% of detectable mosaic variants contribute to disease and these variants contribute to 3.8% of ASD incidence. These results demonstrate that mosaic mutations occur frequently and that these mutations may have a role in the etiology of ASD.

33

Autism redefined: Genomic pathway approach to autism spectrum disorder.

S. Smieszek, J.L. Haines. Case Western Reserve University, Cleveland, OH.

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disease primarily characterized by deficits in verbal communication, impaired social interaction and repetitive behaviors. It exemplifies profound clinical heterogeneity, which poses challenges in diagnosis and treatment. Genetic studies have pointed to hundreds of presumptive causative or susceptibility genes in ASD, making it difficult to find common underlying pathogenic mechanisms and suggesting that multiple different genetic etiologies for ASDs influence a continuum of traits**.** Deep phenotyping analysis allowed for re-categorization of genetic variants. Our previous analysis suggested the existence of two significant subgroups within the existing ASD classification. To investigate this hypothesis in greater detail we have performed in-depth analysis using phenotypic and genetic data from Autism Genetic Resource Exchange (AGRE) and Autism Genome Project (AGP). Our initial findings on both phenotypic and genetic data (1,262 cases and 2,521 controls using familial transmission disequilibrium test) suggest existence of two groups that range in severity. Findings were replicated in a validation dataset. Genetic risk scores (GRS) were used to sum up the total effect of several single-nucleotide polymorphisms characteristic of the two clusters. The high discriminatory ability of the genetic risk score to define cluster 1 from cluster 2 case group at different combinations of sensitivity and specificity was assessed and clearly demonstrates strong signal with AUC being 0.74. There is a significant signal differentiating the 2 clusters relying on non-genetic risk factors and even greater signal when using both non-genetic risk factors and GRS. The detection and validation of the two groups allowed us focus on convergence of findings at the pathway level. ASD heterogeneity was leveraged via large scale pathway analysis within those two categories, which led to identification of a driver gene set across significant pathways. The significant pathways in cluster 1 (severe, affected = 300) include autoimmune disease, vitamin B6 metabolism, whereas in cluster 2 (non-severe, affected = 921) included oxytocin signaling pathway, WNT signaling pathway and glutamatergic synapses (all at P < 0.001).We envision that systematic study of all genomic pathways obtained given a set of redefined categories will yield profound findings for ASD even in the absence of strong individual variant information.

34

Maternal and fetal genetic control of mid-gestational and neonatal levels of markers of immune function.

M. Traglia[1], L.S. Heuer[2,3], K.L. Jones[2,3], C.K. Yoshida[4], R. Hansen[3,5], R. Yolken[6], O. Zerbo[4], G.C. Windham[7], M. Kharrazi[7], G.N. DeLorenze[4], P. Ashwood[3,8], L.A. Croen[4], J. Van de Water[2,3], L.A. Weiss[1]. 1) Department of Psychiatry and Institute for Human Genetics, University of California, San Francisco, 401 Parnassus Ave, LangPorter, San Francisco, CA 94143, USA; 2) Department of Internal Medicine, Division of Rheumatology, Allergy, and Clinical Immunology, University of California, Davis, California, USA; 3) MIND Institute, University of California, Davis, California, USA; 4) Divison of Research, Kaiser Permanente Northern California, Oakland, California, USA; 5) Department of Pediatrics, University of California, Davis, California, USA; 6) Stanley Division of Developmental Neurovirology, Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 7) Division of Environmental and Occupational Disease Control, California Department of Public Health, 850 Marina Bay Pkwy, Bldg. P, Richmond, CA 94804, USA; 8) Department of Medical Microbiology and Immunology, University of California, Davis, California, USA.

The immune system plays an important role in neurodevelopment, and increasing evidence suggests a link between immune system dysregulation and autism spectrum disorder (ASD). Animal models show that maternal immune activation during gestation impacts fetal brain development and subsequent behavior, potentially driven by alterations in levels of cytokines and chemokines which serve as Soluble Immune Mediators (SIMs). We therefore aimed to assess whether levels of SIMs during pregnancy or at birth might be determined by maternal and/or fetal genetics, and thus influence ASD risk. We utilized a multi-ethnic genotyped population-based nested case-control study of 790 women and 764 of their newborns (390 ASD cases, 400 controls) in the EMA (Early Markers of Autism) study (Croen, Autism Res 2008; Tsang, PLoS ONE 2013). Mid-gestational levels of 22 SIMs were measured in maternal serum, and 42 in neonatal bloodspots. We first estimated the maternal and neonatal genome-wide SNP-based heritability (h^2_g) for each SIM and then performed GWAS to identify specific loci contributing to individual SIMs. Finally, we assessed the relationship between genetic SIM determinants and ASD outcome. Levels of two maternal SIMs showed > 80% maternal h^2_g (P<0.05, each) and the levels of 4 separate neonatal SIMs showed > 50% neonatal h^2_g in a REML model (Yang, Nat Genet 2010) adjusted for genetic ancestry, maternal sociodemographic confounding factors as well as offspring affection status. Genome-wide association via linear regression revealed 23 independent loci associated with 27 SIMs (P<5x10^{-8}): 4 maternal alleles were associated with maternal SIMs, 5 maternal alleles were associated with neonatal SIMs, and 24 neonatal alleles were associated with neonatal SIMs. These results highlight the pleiotropic contribution of a neonatal locus mapping to *PLCL2*, (P_{min}=2x10^{-22}), associated with autoimmune diseases. An additional neonatal highly significant locus near *CCL15/CCL23* is associated with levels of neonatal MIP and MPIF (P_{min}<10^{-106}). Five maternal SIMs were associated with offspring affection status (P<0.05, each). Of these, only sIL-2Rα showed one suggestively associated locus that maps near the promoter of *IL2RA* (rs41295055; P=2x10^{-7}*)* but the SNP was not associated with ASD outcome. Our results demonstrate strong mutual contribution of both maternal and neonatal genetics to maternal and neonatal SIMs, however further research is required to elucidate roles in the development of ASD.

35

DDRGK1 regulates SOX9 ubiquitination and its loss causes a human skeletal dysplasia. *A.T. Egunsola[1], Y. Bae[1], M. Jiang[1], D.S. Liu[1], Y. Chen-Evenson[1], T. Bertin[1], J.T. Lu[2,3], L. Nevarez[4], N. Magal[5], E.S. Swindell[6], D.H. Cohn[2,7,8], R.A. Gibbs[1,2], P.M. Campeau[9], M. Shohat[10], B.H. Lee[1].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 3) Department of Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX; 4) Department of Molecular Cell and Developmental Biology, University of California, Los Angeles, Los Angeles, CA; 5) Recanati Institute of Medical Genetic, Rabin Medical center, Petach Tikva, Israel; 6) The University of Texas Graduate School of Biomedical Sciences, Houston, TX; 7) Department of Orthopaedic Surgery, University of California, Los Angeles, Los Angeles, CA; 8) International Skeletal Dysplasia Registry, University of California, Los Angeles, Los Angeles, CA; 9) Department of Pediatrics, University of Montreal, Montreal, QC, Canada; 10) 10Maccabi Genetic institute and Bioinformatics unit - Sheba Cancer Research Center, Sackler School of medicine, Tel Aviv University, Tel Aviv, Israel.

Shohat type spondyloepimetaphyseal dysplasia (SEMD) is a skeletal dysplasia primarily affecting cartilage and is characterized by long bone and vertebral defects. Additional features include premature osteoarthritis, disproportionate short stature, hepatosplenomegaly, lordosis, genu varum and joint laxity. Radiographically, patients have delayed bone age, platyspondyly, radiolucency of the femoral metaphyses and fibular overgrowth. The genetic basis of Shohat type SEMD is unknown. By performing whole exome sequencing on three affected individuals from two families, we identified a homozygous c.408+1G>A donor splice site mutation in the *DDRGK1* gene. This mutation causes a frameshift resulting in a premature stop codon and loss of DDRGK1 protein in patient tissues. We used zebrafish and CRISPR/Cas9 generated mouse models to investigate the role of *Ddrgk1* in cartilage development and chondrocyte differentiation. Knockdown of *ddrgk1* causes craniofacial defects in zebrafish embryos, while *Ddrgk1*$^{-/-}$ mice are embryonic lethal between E11.5-12.5 with delayed chondrogenic mesenchymal condensation in the limb buds. To address a potential molecular mechanism, we knocked down *Ddrgk1* in chondrogenic ATDC5 cells and found decreased protein levels of SOX9, the chondrocyte master transcription factor, and of its target gene type II collagen (*Col2a1)*. To validate *Sox9* as an epistatic downstream target of *Ddrgk1*, we rescued the *ddrgk1* craniofacial phenotype by overexpressing *sox9a* mRNA in *ddrgk1* zebrafish morphants. Furthermore, we found that DDRGK1 and SOX9 interact to form a complex by co-immunoprecipitation, while *Ddrgk1* overexpression decreases SOX9 ubiquitination in HEK293T cells. In conclusion, our results demonstrate that the *DDRGK1* loss-of-function mutation causes Shohat type SEMD by disinhibition of SOX9 ubiquitin-dependent proteasomal degradation, and consequently, decreased *Col2a1* mRNA expression. Interestingly, the deletion of *Ddrgk1* in mice delayed mesenchymal/chondrogenic condensation, partially phenocopying *Sox9* loss of function. Consistent with these results, we found decreased SOX9 protein and *Col2a1* mRNA levels, and increased apoptosis in E11.5 *Ddrgk1*$^{-/-}$ mouse limb buds. Finally, *Sox9* overexpression rescued the chondrogenic and craniofacial phenotype in zebrafish *ddrgk1* mutants. Taken together, these data identify a novel mechanism regulating chondrogenesis via modulation of SOX9 ubiquitination, where dysregulation causes a human skeletal dysplasia.

36

Heterozygosity for _MYH3_ mutations produce spondylocarpotarsal syndrome leading to further insights into the pathophysiology and locus heterogeneity associated with progressive vertebral fusions. *J. Zieba[2], W. Zhang[5], K. Forlenza[1], J.H. Martin[1], K. Heard[5], D.K. Grange[7], M.G. Butler[8], T. Kleefstra[9,10], R.S. Lachman[6], D. Nickerson[11,12], D.H. Cohn[1,4,5,6], J.X. Chong[11,12,13,14], J.H. Bamshad[11,12,13,14], D. Krakow[1,2,3,4,6], University of Washington Center for Mendelian Genomics.* 1) Orthopaedic Surgery, University of California at Los Angeles, Los Angeles, CA; 2) Human Genetics, University of California at Los Angeles, Los Angeles, CA; 3) Obstetrics and Gynecology, University of California at Los Angeles, Los Angeles, CA; 4) Orthopaedic Institute for Children, University of California at Los Angeles, Los Angeles, CA; 5) Department of Molecular, Cell and Developmental Biology, University of California at Los Angeles, Los Angeles, CA; 6) International Skeletal Dysplasia Registry, University of California at Los Angeles, Los Angeles, CA; 7) Division of Genetics and Genomic Medicine, Department of Pediatrics, Washington University School of Medicine, Saint Louis, Missouri; 8) University of Kansas Medical Center, Kansas City, KS, USA; 9) Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands; 10) Department of Cognitive Neurosciences, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands; 11) University of Washington Center for Mendelian Genomics, University of Washington, Seattle, Washington, USA, 98195; 12) Department of Genome Sciences, University of Washington, Seattle, Washington, USA 98195; 13) Division of Genetic Medicine, Department of Pediatrics, University of Washington, Seattle, WA, 98195, USA; 14) Division of Genetic Medicine, Seattle Children's Hospital, Seattle, WA, USA 98105.

Spondylocarpotarsal synostosis (SCT) is an autosomal recessive disorder characterized by progressive fusions of the thoracic and lumbar vertebrae, as well as carpal and tarsal bones, and results from nonsense mutations in the gene Filamin B (*FLNB*). Utilizing a *Flnb* global knockout mouse model, we showed that the vertebral fusions are caused by the collapse and ossification of the intervertebral disc (IVD). This collapse is influenced by upregulation of the TGFβ/BMP signaling pathways in the IVD and resembles disc degeneration disorder (DDD), a common condition. The mechanism in part results from FLNB's direct interactions with inhibitory Smads; loss of FLNB leads to lack of normal TGFβ/BMP pathway modulation and increased activity. We have confirmed that increasing TGFβ/BMP signaling in WT spinal cultures can phenocopy loss of FLNB. To further understand the genetic mechanisms of progressive vertebral fusions, SCT-FLNB negative patients underwent exome analysis. In three patients, heterozygosity for mutations in the gene Myosin Heavy Chain 3 (*MYH3*) was identified. Two patients presented with previously unreported *de novo* changes in MYH3, predicting the changes p.Phe646Cys and p.Leu900fs. Both patients presented with vertebral and carpal/tarsal fusions characteristic of SCT. The proband from a three generation family harbored the predicted mutation, p.Ser243del. This change has also been seen in a one family with AD multiple ptyergium syndrome (DA8). While this family did not have large joint contractures, they had mild camptodactyly. Transfected cells with mutated *MYH3* plasmids had an inhibitory effect on both canonical and noncanonical TGFβ signaling and altered BMP2 induced signaling revealing a heretofore unknown regulatory role for MYH3. Furthermore, MYH3 is not only an embryonic myosin; persistent postnatal MYH3 expression was specifically localized to the muscle tissues joining the neural arches of the spine. In the absence of FLNB, matrix and signaling changes within the IVD lead to cell fate changes, transforming IVD cells to bone. However, the newly identified *MYH3* mutations suggest that abnormal extraneous mechanical forces stemming from altered TGFβ signaling activity in muscle can also lead to fusions of the vertebral bodies. Genetic studies of a rare mendelian disorder have led to new insights into the mechanisms of progressive vertebral fusions and disc generation, expanding our understanding of the cellular functions of both FLNB and MYH3.

37

Identification of new genes responsible for syndromic developmental abnormalities using whole exome sequencing. *M. Lefebvre[1,2], Y. Duffour[2], E. Tisserant[2], L. Olivier-Faivre[1,2], D. Lehalle[1,2], N. Jean Marçais[1,2], N. Laurent[3], M-C. Antal[4], S. El Chehadeh[5], E. Schaefer[5], L. Lambert[6], B. Leheup[6], B. Foliguet[7], J-P. Masutti[7], F. Arbez-Gindre[8], C. Quelin[9,10], S. Odent[9], M. Fradin[9], P. Loget[10], N. Bigi[11,12], D. Genevieve[11], M. Willems[11], S. Blesson[13,14], A. Toutain[13], F. Lafargue[15], C. Francannet[15], A-M. Beaufrere[16], P. Dechelotte[16], J. Thevenon[1,2], C. Thauvin[1,2].* 1) Génétique médicale, CHU de Dijon, Dijon, France; 2) Equipe GAD, EA4271, FHU-Translad, Université de Bourgogne, France; 3) Service d'anatomie pathologique, CHU de Dijon, Dijon, France; 4) Service d'anatomie pathologique, CHU de Strasbourg, Strasbourg, France; 5) Génétique Médicale, CHU de Strasbourg, Strasbourg, France; 6) Génétique Médicale, CHU de Nancy, Nancy, France; 7) Service de Foetopathologie, CHU de Nancy, Nancy, France; 8) Service d'anatomie pathologique, CHU de Besançon, Besançon, France; 9) Génétique Médicale, CHU de Rennes, Rennes, France; 10) Service d'anatomie pathologique, CHU de Rennes, Rennes, France; 11) Génétique Médicale, CHU de Montpellier, Montpellier, France; 12) Service d'anatomie pathologique, CHU de Montpellier, Montpellier, France; 13) Génétique Médicale, CHU de Tours, Tours, France; 14) Service d'anatomie pathologique, CHU de Tours, Tours, France; 15) Génétique Médicale, CHU de Clermont-Ferrand, Clermont-Ferrand, France; 16) Service d'anatomie pathologique, CHU de Clermont-Ferrand, Clermont-Ferrand, France.

Multiple Congenital Anomalies (MCA) are defined by the association of at least 2 congenital malformations. The etiologic diagnosis of these conditions is mandatory to allow genetic counseling and prenatal or preimplantation diagnosis. Regarding current diagnostic tests, the diagnostic rate in MCA fetuses is about 30%. Whole exome sequencing (WES) is a powerful tool to identify genetic variants and increase the diagnostic rate to 25-50%. However, it has not been fully evaluated in fetopathology. We aimed to assess the contribution of WES to diagnose fetal MCA and to identify new genes responsible for fetal MCA. We recruited 100 polymalformed fetuses from 10 prenatal diagnosis centers in France and performed WES with the solo strategy. We first performed a targeted analysis of known OMIM disease-causing genes and then extended the analysis to other genes in the negative cases in order to identify candidate genes. To date, 64/100 fetuses have been included and 38 fetal DNA have been sequenced. Causal mutations in 13 known genes in 14 cases have been identified (37%) and 4 candidate genes are suspected in 5 fetuses (13%) and remain to be replicated; leading to a final 50% rate. Among the identified variants of known disease-causing genes, 5 were sporadic *(NIPBL, MYH3, PIGW, FGFR2, ARID1A, COL1A*), 1 was from dominant inheritance (*TFAP2A*), 5 were from recessive inheritance (*ACE, TREX1, B3GLCT, ERCC2* and *NPC1*) and 1 was from X-linked inheritance (*ZIC3*). In numerous cases, phenotypes were atypical such as diaphragmatic hernia in the *NIPBL*-mutated fetus, micromelia in the *PIGW*-mutated fetus, absence of fracture in the *COL1A1* mutated fetus or absence of heterotaxy and genito-urinary anomalies in the *ZIC3*-mutated fetus. Moreover, some phenotypes appeared extreme in few fetuses: while the *NPC1*-mutated fetus presented with hepatic cirrhosis, its mutation had been reported with variable phenotype and had never been described in fetus but histological studies confirmed the diagnosis of Niemman-Pick C disease. This case highlights also the importance of different tissues conservation (fibroblast culture, frozen tissue, paraffine embedded tissue, tissue in toluene for electronic microscopy), for reverse phenotyping. In conclusion, WES is a powerful tool to diagnose fetal MCA and to extend our knowledge of the phenotype spectrum of known disease-causing gene, notably in underexplored populations like fetuses, and to discover new disease-causing genes with data-sharing.

38

Retinoic acid catabolizing enzyme CYP26C1 is a genetic modifier in SHOX deficiency. *A. Montalbano[1], J. Juergensen[2], R. Roeth[1], B. Weiss[1], M. Fukami[3], S. Fricke-Otto[4], G. Binder[5], T. Ogata[6], E. Decker[7], G. Nuernberg[8,9], D. Hassel[2], G.A. Rappold[1,10].* 1) University Hospital Heidelberg, Heidelberg, Germany; 2) Department of Internal Medicine III - Cardiology, Heidelberg University Hospital, Heidelberg, Germany; 3) Department of Molecular Endocrinology, National Research Institute for Child Health and Development, Tokyo, Japan; 4) Children's Hospital Krefeld, Krefeld, Germany; 5) Children's Hospital, University of Tübingen, Tübingen, Germany; 6) Department of Pediatrics, Hamamatsu University School of Medicine, Hamamatsu, , Japan; 7) Bioscientia Center for Human genetics, Ingelheim, Germany; 8) Center for Molecular Medicine, Cologne, Germany; 9) Cologne Center for Genomics, Cologne, Germany; 10) Interdisciplinary Centre for Neurosciences (IZN), University of Heidelberg, Heidelberg, Germany.

Mutations in the homeobox gene *SHOX* cause SHOX deficiency, the most frequent monogenic cause of short stature. The clinical severity of SHOX deficiency varies widely, ranging from short stature without dysmorphic signs to mesomelic skeletal dysplasia (Léri-Weill dyschondrosteosis, LWD). In rare cases, individuals with SHOX deficiency are asymptomatic. To elucidate the factors that modify disease severity/penetrance, we studied a three-generation family with five affected individuals with LWD using whole genome linkage analysis and whole exome sequencing. The variant p.Phe508Cys of the retinoic acid catabolizing enzyme *CYP26C1* co-segregated with the *SHOX* variant p.Val161Ala in the five affected individuals, while the *SHOX* mutant alone was present in three asymptomatic individuals. Two further independent LWD cases with *SHOX* deficiency and damaging *CYP26C1* variants were identified. The identified damaging variants in CYP26C1 affected its catabolic activity, leading to an increased level of retinoic acid. We also provide evidence that high levels of retinoic acid significantly decrease *SHOX* expression in human primary chondrocytes and zebrafish embryos. Individual morpholino knockdown of either gene shortens the pectoral fins, whereas depletion of both genes leads to a more severe phenotype. Together our findings demonstrate that *SHOX* and *CYP26C1* act in a common molecular pathway controlling limb growth and describe *CYP26C1* as the first genetic modifier for SHOX deficiency.

39

SFRP4 **ablation in Pyle disease reveals the differential regulation of trabecular versus cortical bone and highlights the importance of cortical bone in bone stability.** *A. Superti-Furga[1], H. Saito[2], P.O. Simsek Kiper[3], F. Gori[2], S. Unger[1], E. Hesse[2], K. Yamana[2], R. Kiviranta[2], N. Solban[4], J. Liu[5], R. Brommage[5], K. Boduroglu[3], L. Bonafé[1], B. Campos-Xavier[1], G. Nishimura[7], K.M. Girisha[8], H. Takita[10], K. Harshman[6], B. Stevenson[9], C. Rivolta[11], R. Baron[2,12].* 1) Division of Genetic Medicine, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland; 2) Harvard School of Dental Medicine, Dept of Oral Medicine, Infection and Immunity, Boston, MA, USA; 3) Unit of Pediatric Genetics, Department of Pediatrics, Hacettepe University Medical Faculty, Ankara, Turkey; 4) Acceleron Pharma, Inc. Cambridge, Massachusetts, United States; 5) Lexicon Pharmaceuticals, The Woodlands, Texas, United States; 6) Genomic Technologies Facility, Center for Integrative Genomics, University of Lausanne, Switzerland; 7) Dept of Radiology and Medical Imaging, Tokyo Metropolitan Kiyose Children's Hospital, Kiyose 204-0021, Japan; 8) Department of Medical Genetics, Kasturba Medical College, Manipal University, Manipal, India; 9) Swiss Institute of Bioinformatics, Lausanne, Switzerland; 10) Department of Orthopedics, National Hospital Organization Saga Hospital, Japan; 11) Dept. of Computational Biology, University of Lausanne, Lausanne, Switzerland; 12) Harvard Medical School, Dept of Medicine, Endocrine Unit, Massachusetts General Hospital, Boston, MA, 02114, USA.

Pyle disease (MIM 265900) is a recessive disorder characterized by tall stature, moderate bone fragility, wide metaphyses and thin or absent long bone cortex. As remarked by Pyle (1931) and Cohn (1933), there is exuberant formation of spongiotic bone but lack of cortical bone and absence of periosteal bone remodeling leading to paddle-shaped long bones. We have identified biallelic inactivating mutations in the *soluble frizzled-related protein 4* (*SFRP4*) gene in four individuals with Pyle disease and studied the effect of *sfrp4* ablation in mice. *Sfrp4*-ablated mice show wide bones with expanded trabecular component and reduced cortical component. The study of mouse calvarial (cortical) and bone marrow (trabecular) osteoblasts revealed a differential regulation of BMP and WNT signalling. Mechanistically, in cortical osteoblasts, absence of SFRP4 results in activation of BMP signaling and elevation of sclerostin (SOST) expression, and this dysregulation can be counteracted by BMP antagonists or by antibodies to sclerostin. Developmentally, SFRP4 deficiency allows for vigorous formation of trabecular bone but impairs the formation of cortical bone ; in consequence, there is « spongious hypertrophy », thin to absent bone cortex, impaired bone remodeling, and increased bone fragility. The pathogenesis of Pyle disease highlights the differential regulation of trabecular versus cortical bone formation, underlines the importance of cortical bone for mechanical bone stability, and indicates BMP and SOST modulation as a possible way to stimulate cortical bone formation in this and in other conditions.

40

Absence of the ER cation channel *TMEM38B*/TRIC-B disrupts intracellular calcium homeostasis and dysregulates collagen synthesis in recessive osteogenesis imperfecta. *W.A. Cabral[1], M. Ishikawa[2,3], M. Garten[4], E.N. Makareeva[5], B.M. Sargent[1], M. Weis[6], A.M. Barnes[1], E.A. Webb[7,8], N.J. Shaw[8], L. Ala-Kokko[9], F.L. Lacbawan[10,11], W. Högler[7,8], S. Leikin[5], P.S. Blank[4], J. Zimmerberg[4], D.R. Eyre[6], Y. Yamada[2], J.C. Marini[1].* 1) Section on Heritable Disorders of Bone and Extracellular Matrix, NICHD, NIH, Bethesda, MD, USA; 2) Molecular Biology Section, NIDCR, NIH, Bethesda, MD, USA; 3) Department of Restorative Dentistry, Division of Operative Dentistry, Tohoku University, Graduate School of Dentistry, Sendai, Japan; 4) Section on Integrative Biophysics, NICHD, NIH, Bethesda, MD, USA; 5) Section on Physical Biochemistry, NICHD, NIH, Bethesda, MD, USA; 6) Department of Orthopaedics and Sports Medicine, University of Washington, Seattle, WA, USA; 7) School of Clinical and Experimental Medicine, Institute of Biomedical Research, University of Birmingham, Birmingham, United Kingdom; 8) Department of Endocrinology and Diabetes, Birmingham Children's Hospital, Birmingham, United Kingdom; 9) Connective Tissue Gene Tests, Allentown, PA, USA; 10) Department of Medical Genetics, Children's National Medical Center, Washington D.C., USA; 11) Molecular Genetics Pathology Section, Department of Molecular Pathology, Robert Tomsich Pathology and Laboratory Medicine Institute, Cleveland Clinic, Cleveland, OH, USA.

Type XIV osteogenesis imperfecta (OI), a moderately severe recessive form caused by null mutations in *TMEM38B*, was first identified among Bedouins but has now been found in individuals in Europe, North America, Pakistan, and China. *TMEM38B* encodes the ER membrane monovalent cation channel, TRIC-B, proposed to counterbalance IP_3R-mediated Ca^{2+} release from intracellular stores. The molecular mechanisms by which *TMEM38B* mutations cause OI are unknown. We identified 3 probands with recessive defects in *TMEM38B*. Despite the presence of reduced *TMEM38B* transcripts, TRIC-B protein is undetectable in proband fibroblasts and osteoblasts. We used live cell imaging with the cytoplasmic Ca^{2+} indicator Fura-2 and the ER-localized probe D1ER to investigate the effect of absence of the anion channel on calcium flux from ER to cytoplasm and ER stores, respectively. Release of ER luminal Ca^{2+} to the cytoplasm is impaired in proband fibroblasts and osteoblasts with TRIC-B deficiency, although SERCA and IP_3R, the channels for ER calcium uptake and release, have normal stability. Notably, steady state ER Ca^{2+} is unchanged in TRIC-B deficiency. These data support a role for TRIC-B in the kinetics of ER calcium depletion and recovery. The disturbed Ca^{2+} flux activates the PERK pathway of the UPR, increases BiP, and dysregulates synthesis of type I collagen in proband cells. Collagen helical lysine hydroxylation is reduced, despite increased lysyl hydroxylase 1 protein (LH1). In contrast, collagen telopeptide hydroxylation is increased, although proband cells display the expected decrease of the Ca^{2+}-dependent PPIase for LH2, FKBP65. Procollagen chain assembly is delayed, suggested that a significant portion of the ER-resident protein disulfide isomerase (PDI) involved in collagen C-propeptide folding may be sequestered by calreticulin binding, resulting from abnormal Ca^{2+} flux. The resulting misfolded collagen is substantially retained in TRIC-B null cells, consistent with a 50-70% reduction in secreted collagen. Lower-stability forms of collagen that elude proteosomal degradation are not incorporated into extracellular matrix, which contains only normal stability collagen, resulting in matrix insufficiency. These data support a role for TRIC-B in intracellular Ca^{2+} homeostasis, and demonstrate that absence of *TMEM38B* causes OI by dysregulation of calcium flux kinetics in the ER, impacting multiple collagen-specific chaperones and modifying enzymes.

41

Evidence for body mass index gene x environment interaction using 120,000 individuals from the UK Biobank study.

J. Tyrrell[1], *A.R. Wood*[1], *R.M. Ames*[1], *H. Yaghootkar*[1], *R. Beaumont*[1], *S.E. Jones*[1], *M.A. Tuke*[1], *K. Ruth*[1], *R.M. Freathy*[1], *A. Murray*[1], *Z. Kutalik*[2], *M.N. Weedon*[1], *T.M. Frayling*[1]. 1) University of Exeter Medical School, University of Exeter, Exeter, United Kingdom; 2) Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital (CHUV), Lausanne, Switzerland.

Statement of Purpose Susceptibility to obesity in today's environment has a strong genetic component. However, little is known about how genetic susceptibility interacts with modern environments and behaviours to predispose some individuals to obesity whilst others remain slim. Social deprivation is associated with a higher risk of obesity but it is not known if it accentuates genetic susceptibility to obesity. Previous gene-obesogenic environment studies have been limited by the need to perform meta-analyses of many heterogeneous studies and studies have not necessarily corrected for statistical artefacts such as different variances between groups (heteroscedasticity). We aimed to use 120,000 individuals from the UK Biobank study to test the hypothesis that objective measures of relative deprivation in the UK accentuate genetic susceptibility to obesity. *Methods* We used the Townsend deprivation index (TDI) as a measure of deprivation and a 69-variant genetic risk score (GRS) as a measure of genetic susceptibility to obesity. We tested the association of the genetic risk score with BMI in high and low socioeconomic groups and tested for interactions (using the continuous TDI as an exposure measure). To test the specificity of any apparent interactions we repeated analyses using a simulated environment (that was correlated with BMI in the same way as TDI) as an interaction term and using randomly selected groups of individuals of different BMIs. *Results* We found evidence of gene-environment interactions with TDI (Pinteraction=$3\text{x}10^{-10}$). Within the 50% of most deprived individuals, carrying 10 additional BMI-raising alleles was associated with approximately 3.8 kg extra weight in someone 1.73m tall. In contrast, within the 50% of least deprived individuals carrying 10 additional BMI-raising alleles was associated with approximately 2.9 kg extra weight. When we used a simulated environment or randomly selected groups of individuals to be of different BMIs, we observed only nominal evidence of apparent interaction, (simulated environment Pinteraction = 0.04; randomly selected groups: Pinteraction=$9\text{x}10^{-4}$) suggesting the interaction was specific to TDI. *Conclusions* Our findings provide evidence that social deprivation accentuates the genetic predisposition to obesity.

42

A thrifty variant in *CREBRF* strongly influences body mass index in Samoans.

R.L. Minster[1], *N.L. Hawley*[2], *C.-T. Su*[1], *G. Sun*[3], *E.E. Kershaw*[4], *H. Cheng*[3], *O.D. Buhule*[5], *J. Lin*[1], *M. Sefuiva Reupena*[6], *S. Viali*[7], *J. Tuitele*[8], *T. Naseri*[9], *Z. Urban*[1], *R. Deka*[3], *D.E. Weeks*[1,5], *S.T. McGarvey*[10,11]. 1) Dept of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA; 2) Dept of Epidemiology (Chronic Disease), School of Public Health, Yale University, New Haven, CT; 3) Dept of Environmental Health, College of Medicine, University of Cincinnati, Cincinnati, OH; 4) Dept of Medicine, Division of Endocrinology, University of Pittsburgh, Pittsburgh, PA; 5) Dept of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA; 6) Bureau of Statistics, Government of Samoa, Apia, Samoa; 7) Medical Specialist, Apia, Samoa; 8) Department of Health, American Samoa Government, Pago Pago, American Samoa; 9) Ministry of Health, Government of Samoa, Apia, Samoa; 10) International Health Institute, Dept of Epidemiology, School of Public Health, Brown University, Providence, RI; 11) Dept of Anthropology, Brown University, Providence, RI.

Overweight and obesity rates in Samoa are among the highest in the world. Given the high rates of obesity and the relatively isolated nature of this Polynesian population, we conducted a genome-wide association study of body mass index (BMI) in 3,072 adult Samoans and observed significant association between BMI and variants on 5q35.1. Subsequent targeted sequencing and fine mapping highlighted a nonsynonymous coding variant (rs373863828, c.1370G>A, p.Arg457Gln) in *CREBRF* (discovery sample $p = 7.0 \times 10^{-13}$; validation sample of 2,102 adult Samoans $p = 3.5 \times 10^{-9}$). The variant accounts for 1.93% of BMI variance in the discovery sample and 1.08% in the validation sample, and each copy of the minor allele is associated with an increase in BMI of 1.36 kg/m^2 in the discovery sample and 1.45 kg/m^2 in the validation sample. These are the largest effect sizes on BMI observed for a common variant in humans, exceeding the effects of *FTO.* This variant is also positively associated with total and regional adiposity and obesity, but unexpectedly is inversely associated with fasting glucose and odds of diabetes. rs373863828 has a minor allele frequency (MAF) of 0.259 in Samoans, but a MAF of 0.0000412 among individuals in the Exome Aggregation Consortium. This disparity in the allele frequency of the missense variant in Samoans versus other populations, the population history of Samoans, and their high prevalence of obesity leads us to consider the possibility that this is a "thrifty" allele that has risen to this frequency by natural selection. The possibility of a thrifty genotype has been proposed before, but little evidence of one has been observed to date. Analysis of the variation around rs373863828 in Samoans reveals elevated extended haplotype homozygosity in the presumed derived minor allele compared to the ancestral major allele, evidence of positive selection on this allele at this locus. Additional evidence of positive selection is indicated by an integrated haplotype score (iHS) of 2.94 ($p \approx 0.003$) and an nS_L score of 2.63 ($p \approx 0.008$). Additional studies are necessary to fully explore the role of rs373863828 and *CREBRF* in Samoans and its role in obesity in other populations. However, these initial findings provide remarkable evidence of a heretofore unknown contributor to the genetic architecture of body mass.

43

Genetic study of body mass index in 173,430 Japanese identifies 76 new loci and highlights shared heritability with broad spectrum of complex diseases.

M. Akiyama[1], M. Kanai[1], Y. Okada[1,2], A. Takahashi[1,3], Y. Momozawa[4], M. Ikeda[5], N. Iwata[5], S. Ikegawa[6], M. Hirata[7], K. Matsuda[7], T. Yamaji[8], M. Iwasaki[8], K. Sobue[9], M. Yamamoto[10,11], M. Kubo[12], Y. Kamatani[1]. 1) Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan; 2) Department of Statistical Genetics, Osaka University Graduate School of Medicine, Osaka 565-0871, Japan; 3) Laboratory for Omics Informatics, Omics Research Center, National Cerebral and Cardiovascular Center, Osaka 565-8565, Japan; 4) Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan; 5) Department of Psychiatry, Fujita Health University School of Medicine, Aichi 470-1192, Japan; 6) Laboratory for Bone and Joint Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan; 7) Graduate school of Frontier Sciences, The University of Tokyo, Tokyo 108-8639, Japan; 8) Division of Epidemiology Center for Public Health Sciences National Cancer Center, Tokyo 104-0045, Japan; 9) Iwate Medical University, Iwate 028-3694, Japan; 10) Graduate School of Medicine, Tohoku University, Sendai 980-8575, Japan; 11) Tohoku Medical Megabank Organization, Tohoku University, Sendai 980-8573, Japan; 12) RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan.

Obesity, a risk factor for multiple complex diseases, is a highly heritable trait. Previous genome-wide association studies (GWASs) have identified > 100 loci associated with body-mass index (BMI). However, majority of these loci were identified by European population. To identify genetic loci associated with obesity and to expand knowledge on body weight regulation, we conducted a large-scale GWAS for BMI using participants of BioBank Japan project (N = 158,284). After GWAS and meta-analysis with results from two Japanese cohorts (Japan Public Health Center-based Prospective Study and Tohoku Medical Megabank Organization; N = 15,146), we identified 85 loci associated with BMI ($P < 5.0 \times 10^{-8}$), of which 51 were newly identified. Lead variants of these loci accounted for 2.8 % of phenotypic differences. Conditional analysis revealed four significant second signals. Comparison with the publically available GWAS results revealed that BMI susceptibility loci were generally shared between European and Japanese population (Pearson's correlation coefficient [r] = 0.76). Furthermore, replication and meta-analysis for highly associated variants reported in Europeans further identified 25 additional new loci. In total, 76 novel loci were identified through this study, which brought number of known susceptibility loci to 182. Explained variance estimated by common variants using GREML method implemented in GCTA software were 29.8 ± 3.4 % (mean ± standard error) and 2.3 ± 0.8 % for autosomes and X chromosome, respectively. Comprehensive evaluation of positional candidate genes using functional annotations, cis-eQTL, protein-protein interaction, pathway analysis and mouse knockout phenotype prioritized 37 functional candidate genes with multiple biological evidences. Assessments of cell type specific histone modifications suggested involvement of CD19 primary cells, pancreatic islets and central nervous systems in body weight regulation. We also evaluated genetic correlation between BMI and 34 complex diseases using cross-trait LD score regression and found that BMI was genetically correlated with nine diseases, including metabolic, cardiovascular, psychiatric, allergic, bone and joint disorders (false discovery rate < 5%).

44

Exome-wide association meta-analysis for body mass index identifies protein-altering variants revealing new insights in the biology underpinning obesity.

R.J.F. Loos[1], V. Turcot[2], H.M. Highland[3], Y. Lu[1], C. Schurmann[1], M. Graff[3], A.E. Justice[3], R.S. Fine[4], K.L. Young[3], M. Feitosa[5], E. Marouli[6], A. Wood[7], L.A. Cuppens[8], P. Deloukas[6], I.B. Borecki[5], J.A. Pospisilik[9], deCODE Genetics[10], T.M. Frayling[7], G. Lettre[2], C.M. Lindgren[11], K.E. North[3], J.N. Hirschhorn[4] For the BBMRI-NL, the GOT2D, the CHARGE, and the GIANT Consortia. 1) The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 2) Montreal Heart Institute, Montreal, Quebec, Canada; 3) Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; 4) Broad Institute of Harvard and MIT, Cambridge, MA, USA; 5) Department of Genetics Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO, USA; 6) William Harvey Research Institute, Barts and The London School of Medicine and Dentistry Queen Mary University of London, London, UK; 7) University of Exeter Medical School, Exeter, UK; 8) Boston University School of Public Health, Boston, MA, USA; 9) Max Planck Institute of Immunology and Epigenetics, Freiburg, Germany; 10) deCODE Genetics, Reykjavik, Iceland; 11) The Big Data Institute (BDI), The Li Ka Shing Centre for Health Information and Discovery, Roosevelt Drive, Headington, Oxford, UK.

Efforts to elucidate the genetic contribution, and thus underlying biology, to obesity have focused on genome-wide association studies (GWAS), which have identified >200 loci robustly associated with body mass index (BMI) and obesity. GWAS-identified loci have typically small effects, are common, non-coding, intronic or intergenic, and may not directly affect protein function. They are often hard to interpret and inference of the causal gene(s) and/or variant(s) has proven to be challenging. Here, we aim to discover coding variants that are rare or low frequency (R/LF) (minor allele frequency <5%) and have larger effects on BMI. With this approach, we expect to expedite the identification of the causal gene(s)/variant(s). Our samples included up to 526,508 individuals from 123 studies with exome array genotype data. We meta-analyzed summary association statistics for 216,883 R/LF coding variants with inverse normally transformed BMI. We took forward 20 variants with P<2E-6 for validation in an independent sample of up to 192,226 individuals (deCODE, UKBiobank) and subsequently combined all samples (N_{max}=718,734) in a final meta-analysis. We identified 16 rare/LF coding variants that achieved significance (P<2E-7), including 9 variants in 8 novel loci, such as *RAPGEF3* (EAF=1%, P=1.6E-15, 0.30kg.m^{-2}/effect allele) and *ACHE* (EAF=3.9%, P=2.8E-10, 0.13kg.m^{-2}/effect allele). We also found associations for variants in known monogenic obesity genes, including those in *MC4R* and *KSR2*. For example, a stop-codon (rs13447324, Y35X) in *MC4R* (P=2.3E-10, EAF=0.01%) has an effect size equivalent to 2.4kg/m^2 per effect allele (~7 kg in a 1.7m-tall person). For 7 of the 16 R/LF variants, the minor allele was associated with *lower* BMI, including variants in *GIPR*. Meta-analyses of gene-based results, testing aggregate effects of R/LF variants in 16,222 genes, did not identify additional genes, but implicated a second rare variant in *GIPR.* Pathway analyses (DEPICT) based on the newly identified coding variants/genes highlight gene-sets primarily involved in synaptic mechanisms and neurotransmitter secretion and transport, independently confirming the previous analyses from GWAS for BMI that highlighted many of the same gene sets. In summary, exome-wide analyses for BMI reveal protein-altering R/LF variants in *specific* genes that influence previously implicated CNS processes, which may lead to the identification of new therapeutic targets for treatment and prevention of obesity.

45

Identification and characterisation of a novel rare variant for body fat distribution and diastolic blood pressure.

K.E. Schraut[1,2], I. Williamson[3], P.K. Joshi[2], R.H. Stimson[1], P. Gautier[3], V. Emilsson[4], W.A. Bickmore[3], P. Navarro[3], N.M. Morton[1], J.F. Wilson[2,3]. 1) Centre for Cardiovascular Sciences, Queen's Medical Research Institute, University of Edinburgh, Royal Infirmary of Edinburgh, Little France Crescent, Edinburgh EH16 4TJ, Scotland; 2) Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, Scotland; 3) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, Scotland; 4) Icelandic Heart Association, IS-201 Kopavogur, Iceland.

Genome-wide association meta-analyses for total adiposity and fat distribution have identified over 95 loci associated with human obesity. Total adiposity associates predominantly with brain gene networks whereas fat distribution variants often map to metabolic genes expressed in the adipose or other peripheral tissues. To identify novel genes affecting fat mass and distribution we focused on the isolated population of Orkney using phenotypes derived from DXA scans and genetic data imputed to the 1000Genomes Project. 1,200 individuals from the ORCADES cohort were included in the discovery sample, followed by replication in the Icelandic AGES cohort with 3,219 participants. We identified a new locus on chromosome 4 associated with the ratio of android and gynoid fat (p= 4.5x10^{-10}) which replicated in abdominal fat by CT (p=0.003). Compared to other British and Scandinavian populations, the variant showed a 3-fold increase in frequency in Orkney. Per allele, variant carriers show a reduction in android fat by 3% and visceral fat of 140g as well as lowered diastolic blood pressure of 10mmHg. The lead SNPs map to an ENCODE-predicted DNase1 hypersensitivity site within the second intron of the *ENPP6* gene*,* suggesting a role in genome regulation. Marking the areas with sequence-specific probes by 3D fluorescent in situ hybridisation confirmed that the association interval co-localised more frequently with the *ENPP6* promoter than with other gene promoters within the same chromosomal region in SH-SY5Y neurons (p=0.01) but not human SGBS adipocytes. Consistent with this *ENPP6* mRNA levels were extremely low in human subcutaneous and visceral adipose tissue. ENPP6 expression is highest in the brain and kidney, suggesting a novel neuronal – and/or renal – mediated mechanism driving fat distribution and blood pressure. To model the impact of *Enpp6* on adiposity in vivo, *Enpp6*$^{-/-}$ mice were generated and their metabolic profile investigated. *Enpp6*$^{-/-}$ mice showed an increase of overall fat percentage by 4% in comparison to wild-type animals (n= 16, p=0.002). Using the advantage of genetic drift in a population isolate (the jackpot effect) we have identified a novel, rare variant of large effect associated with fat distribution and blood pressure. Our data supports ENPP6 as a new, non-adipose, mechanism regulating fat distribution and blood pressure that could have potential therapeutic significance.

46

Analysis of BMI using whole exome sequence from 49,178 individuals from the Geisinger Health System DiscovEHR study.

A.H. Li[1], C. Gonzaga-Jauregui[1], G.C. Wood[2], U.L. Mirshahi[2], F. Dewey[1], T. Mirshahi[2], I. Borecki[1]. 1) Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY, USA; 2) Geisinger Health System, Danville, PA, USA.

Obesity is associated with increased risk of a variety of diseases and morbidities and represents an increasing public health burden. More than one hundred loci have been associated with obesity or body mass index (BMI), largely localized in non-coding regions. We focus on the role of protein-altering variations influencing BMI, by whole exome sequencing 49,178 participants of European descent from the DiscovEHR study. The mean BMI was 31.2 (48.7% obese); the mean BMI of 2,784 participants from a bariatric cohort was 49.2. This is the largest exome sequenced cohort to date for the investigation of adiposity genetics. We found significant associations with BMI at 13 variant sites (p<1x10^{-7}) within 11 unique genes, and suggestive evidence at an additional 19 sites (p<1x10^{-6}) in 17 unique genes. Eighteen of these variants were associated with increased BMI and 14 with lower BMI - the biology of variants associated with lower BMI in an obesogenic environment is of potential therapeutic interest. Rare LoF variation (MAF<1%) was further investigated by gene-based variant aggregation analysis, with suggestive evidence of association with BMI (p<1x10^{-5}) observed in 5 genes. Coding variants in 31 of 1,330 unique genes annotated in the NGRHI-EBI GWAS catalog for BMI and related traits were associated with BMI in our data; 26 novel genes were identified. The genes associated with BMI in our cohort represent diverse biological functions. Associations with *BDNF* (rs6265), *MC4R*, and other genes with a known role in hypothalamic signaling were found. In addition, we observe novel associations with other metabolic genes, including potential roles in adipocyte metabolism (*KLF3, GPD1L*) and activity in the GI tract (*DHRS11*). These results support the utility of exome sequencing to identify novel genes influencing adiposity.

47

Heritable epimutations associated with breast cancer risk. *M.C. Southey[1,4], J.E. Joo[1], E.M. Wong[1], J.L. Hopper[2], D.R. English[2,4], D.E. Goldgar[3,1], G.G. Giles[4,2], R.L. Milne[4,2], J.G. Dowty[2], kConFab and The Australian Breast Ccancer Family Study.* 1) Department of Pathology, The University of Melbourne, Melbourne, Australia; 2) Centre for Epidemiology and Biostatistics, The University of Melbourne, Melbourne, Australia; 3) Huntsman Cancer Institute, University of Utah Health Sciences Center, Utah, USA; 4) Cancer Epidemiology Centre, Cancer Council Victoria, Australia.

OBJECTIVES While most epigenetic marks are reprogrammed during early embryogenesis, some studies have reported Mendelian-like inheritance of germline DNA methylation in particular cancer-susceptibility genes. In this study, we attempted to identify such heritable epimutations for breast cancer using epigenome-wide methylation data and multiple-case families. **METHODS** We studied 25 families that were recruited through Australian family cancer clinics and that each had multiple cases of breast cancer but no mutations in any known breast cancer-associated gene. Methylation was measured at approximately 480,000 genetic loci in 210 of the 2141 family members using the Infinium HumanMethylation450 BeadChip array. We hypothesized that some heritable epimutations are caused by rare genetic variants which predispose carriers to aberrant patterns of methylation at particular loci. We developed a novel statistical method to identify methylation sites whose measured values are most consistent with a Mendelian inheritance pattern, based on segregation analysis and the expectation-maximization algorithm. Carrier probabilities for the hypothesized rare, autosomal, dominant DNA variants inducing these inheritance patterns were calculated for the 1000 most-Mendelian methylation sites, based on family structure but not affected status. Cox proportional hazards survival analysis was then used to assess associations between these carrier probabilities and breast cancer risk. Probes located on the X-chromosome or within 10 base pairs of known SNPs were excluded. **RESULTS** After correcting for multiple testing, we identified 11 methylation sites whose corresponding carrier probabilities are associated with breast cancer. Three of these sites are clustered within 200 base pairs of a noncoding RNA which is known to have a tumour suppressor role and is suspected to be regulated by DNA methylation. **CONCLUSIONS** We screened almost half a million methylation sites for those with Mendelian inheritance patterns, using a novel statistical method which incorporates family structure but not affected status. We then identified 11 methylation sites which might contain heritable epimutations associated with breast cancer risk.

48

Functional annotation of breast cancer risk-associated loci identified using the OncoArray. *J. Beesley[1], S. Kar[2], K.I. McCue[1], K. Michailidou[2], K.B. Kuchenbaecker[2], L. Fachal[2], D.M. Glubb[1], A. Lemaçon[3], A. Droit[3], P. Soucy[3], A.M. Dunning[2], J.D. French[1], P. Kraft[4,5], M.K. Schmidt[6,7], A.C. Antoniou[2], R.L Milne[8], J. Simard[3], D.F. Easton[2], S.L. Edwards[1], G. Chenevix-Trench[1], Consortium of Investigators of Modifiers of BRCA1/2 and Breast Cancer Association Consortium.* 1) Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane Australia; 2) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 3) Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Lav, Quebec, Quebec, Canada; 4) Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 5) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 6) Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands; 7) Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands; 8) Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia.

Genome-wide association studies (GWASs) of breast cancer susceptibility have identified 107 risk-associated loci. Using the OncoArray (a custom Illumina chip comprising ~570K SNPs with genome-wide coverage), BCAC and CIMBA have recently identified 81 additional breast cancer risk-associated loci, including 65 associated with overall risk, six with estrogen receptor (ER) positive risk, and 10 with risk of ER negative disease. Most of these regions contain large numbers of highly significant variants in strong linkage disequilibrium that cannot be excluded as causal variants using statistical methods. We developed an annotation pipeline to interrogate publicly available data from normal and tumour breast cells in order to highlight potentially functional variants, predict target genes and prioritise experimental validation. We used annotations from the ENCODE Consortium, Roadmap Epigenomics Projects and other published studies which fall into categories relating to putative effects on transcription factor binding, regulatory element activities, and chromatin interactions. In addition, we carried out expression quantitative trait loci (eQTL) analysis using data from The Cancer Genome Atlas (TCGA), METABRIC and the Gene-Tissue Expression (GTEx) projects. We report genomic features potentially impacted by each variant in tabular form with a hyperlink to a customised UCSC Genome Browser session to enable further exploration. At 6q23.1, the only predicted target gene, based on ChIA-PET data, is *L3MBTL3*, and the sentinel SNP, rs6569648, is associated with expression of *L3MBTL3* in METABRIC (P = 4.3x10^{-6}). At 1p34, one of the predicted target genes, based on PreSTIGE predictions and ChIA-PET data, is *NSUN4*, and the sentinel SNP, rs60123014, is associated with expression of *NSUN4* in TCGA (P = 1.3 x 10^{-21}), METABRIC (P = 5.3 x 10^{-3}) and GTEx (P = 1.7 x 10^{-10}). Experimental analysis validated predictions of chromatin interactions between risk-associated SNPs and the promoters of *CRNDE*, *CUX1* and *PRKRIP1*. For the great majority of loci no significant eQTLs were identified but other bioinformatically predicted target genes include *ATM*, *BCL2L11*, *FBXO32*, *GATA3*, *HELQ*, *KDELC2*, *NCOA1*, *NPAT*, *PDCD6*, *SMAD3*, *SOX4* and *ZNF217*. These results highlight the utility of in silico annotation of GWAS-identified variants, in particular for target gene prediction, in formulating hypotheses for experimental follow-up studies.

49

Functional characterization of prostate cancer risk associated SNPs in the *TET2* locus. *S.A. Brodie[1], J. Boland[1], M. Yeager[1], M. Dean[2], M. Nickerson[2].* 1) National Cancer Institute, Division of Cancer Epidemiology and Genetics, Cancer Genomics Research Laboratory, Leidos Biomedical Research Inc., Gaithersburg, MD; 2) National Cancer Institute, Division of Cancer Epidemiology and Genetics, Laboratory of Translational Genetics, Gaithersburg, MD.

Genetic alterations associated with metastatic prostate cancer (PCa) can be discovered by sequencing tumor genomes, identifying molecular markers for this lethal stage of disease. Previously, we characterized somatic alterations in metastatic PCa tumors in the methylcytosine dioxygenase *ten-eleven translocation 2* (*TET2*) gene, which is altered in 5%-15% of myeloid, kidney, colon and prostate cancers. Genome-wide association studies previously identified non-coding risk variants near or within the *TET2* locus that are associated with PCa and melanoma. We performed fine-mapping of PCa risk across *TET2* using genotypes from the PEGASUS case-control cohort and identified six new risk variants in introns 1 and 2. To explore the functional consequences of the newly discovered SNPs, we performed electrophoretic mobility shift assays (EMSAs). Oligonucleotides containing two risk variants were bound by the transcription factor octamer-binding protein 1 (Oct1 – also known as POU class 2 homeobox 1; *POU2F1*). Furthermore, *TET2* and *POU2F1* expression were positively correlated in prostate cell lines suggesting a functional relationship between Oct1 binding and *TET2* expression. We examined *TET2* expression in PCa and found reduced expression in metastatic PCa compared to primary PCa as well as in high versus low Gleason scored tumors. 5-hydroxymethyl cytosine (5hmC) was also depleted in tumor vs normal tissues. Kaplan-Meier analysis reveals that low *TET2* expression correlates with shorter disease-free-survival. *In vitro* studies utilizing siRNA directed against *TET2* increases metastatic potential of PCa cell lines as measured by proliferation, migration and invasion assays. Taken together, noncoding SNPs in the *TET2* locus can confer genetic risk of poor outcome PCa possibly by decreasing *TET2* expression, and thereby driving cancer to a pro-metastatic phenotype.

50

Developmental transcription factor NFIB is a target of oncofetal miRNAs and is linked to tumour aggressiveness in lung adenocarcinoma. *D. Becker-Santos[1], B. C. Minatel[1], K. Thu[1], J. English[2], V. Martinez[1], C. MacAulay[1], W. Lockwood[1], S. Lam[1], W. Robinson[3], I. Jurisica[4], W. Lam[1].* 1) Integrative Oncology Dept, British Columbia Cancer Research Centre, Vancouver, BC, Canada; 2) Pathology Dept, Vancouver General Hospital, Vancouver, BC, Canada; 3) Medical Genetics Dept, University of British Columbia, Vancouver, BC, Canada; 4) Princess Margaret Cancer Centre and Techna Institute, University Health Network, and Departments of Medical Biophysics and Computer Science, University of Toronto, Toronto, ON, Canada.

Rationale: Tumourigenesis recapitulates many aspects of normal development, such as high rates of cell proliferation, vascular restructuring, increased cell motility and invasiveness. Accordingly, genes involved in fetal lung development are thought to play crucial roles in the malignant transformation of adult lung cells. Consequently, the study of lung tumour biology in the context of lung development has the potential to reveal key regulatory pathways reactivated or suppressed in lung cancer. **Methods:** 131 pairs of non-small cell lung cancer (NSCLC) tumour and non-malignant lung tissues and 5 human fetal lung tissue samples were profiled by miRNA-sequencing (all fetal tissues used in this study were obtained with the informed consent of the parent). To investigate protein-coding genes controlled by the *oncofetal* miRNAs identified, miRNA Data Integration Portal (mirDIP) was applied followed by luciferase-reporter assays. Associations between patient survival and mRNA expression of selected *oncofetal* miRNA-gene targets were evaluated in ~1,400 NSCLC cases. Immunohistochemical analysis of *oncofetal* miRNA targets was performed on a lung adenocarcinoma tissue microarray. **Results:** We describe for the first time a comprehensive characterization of miRNA expression in human fetal lung tissue, and identified numerous miRNAs that recapitulate their fetal expression patterns in NSCLC. Nuclear Factor I/B (NFIB), a transcription factor essential for lung development, was identified as being frequently targeted by these *oncofetal* miRNAs. Concordantly, analysis of NFIB expression in multiple NSCLC cohorts revealed its frequent underexpression in tumours (>60%). Remarkably, low expression of NFIB was significantly associated with higher grade, biologically more aggressive kinds of lung adenocarcinoma, and ultimately, poorer survival in patients with this subtype of lung cancer. **Conclusions:** This work has revealed a prominent mechanism for the downregulation of NFIB, a developmental transcription factor essential for lung differentiation, which we found to be associated with aggressive phenotypes of lung adenocarcinoma and consequently, poor patient survival. Restoration of NFIB expression in lung adenocarcinoma could induce lung cell differentiation and has the potential to reduce tumour aggressiveness.

51

The connection between germline risk variants and somatic mutation patterns in sarcoma. *D.L. Goode[1,2], E. Galligan[1], S.M. Rowley[1], M.L. Ballinger[3], D.M. Thomas[3].* 1) Peter MacCallum Cancer Centre, 305 Grattan Street, Melbourne, Victoria, Australia; 2) Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Victoria, Australia; 3) Garvan Institute of Medical Research The Kinghorn Cancer Centre, 370 Victoria Street, Darlinghurst, NSW.

Sarcomas encompass a broad range of tumours of tissues of mesenchymal origin (e.g., bone, muscle, connective and adipose tissue) and are characterized by an earlier age of onset than most cancers. Sequencing 72 tumour suppressors and DNA damage response genes in germline samples from 1153 patients enrolled in the International Sarcoma Kindred Study (ISKS) found excess burden of rare inherited protein-damaging genetic variants in several novel putative sarcoma risk genes relative to ancestry-matched control populations (Ballinger et al, *The Lancet Oncology*, 2016). To investigate the links between variants in these genes and sarcoma formation, we sequenced the same panel of 72 genes in tumours from 50 ISKS patients who carried predicted deleterious germline variants in one or more genes in the panel. We looked for second inactivating somatic mutations in these genes to determine which act as tumour suppressors through the classic ‘two-hit’ mechanism as well as assessed the effects of these genes on overall abundance and patterns of somatic mutations. About 25% of samples displayed clear second hits or loss of heterozygosity in established cancer risk genes such as *BRCA2* and *APC* or non-canonical risk genes *NBN* and *PTCH1*, suggesting the latter are genuinely involved in sarcoma development. Tumours from patients with multiple damaging germline variants usually did not acquire second hits in the affected genes, but often had inactivating somatic mutations in additional tumour suppressors and DNA repair genes that were homozygous wild-type in the germline. Carriers of multiple DNA damage response and repair variants did not display a significantly higher cumulative point mutation burden (p=0.29), though *ATM* and *APC* carriers did tend to have more somatic SNVs than the rest of the cohort (p=0.02). These findings offer strong supporting evidence for *BRCA2*, *NBN* and *PTCH1* as true sarcoma risk genes while reducing confidence in other candidates. Our results also provide intriguing insights into how deleterious germline variants in DNA repair genes may act as drivers of sarcoma formation through their indirect effects on background somatic mutation rates.

52

Functional annotation of ovarian cancer risk loci identifies regulatory mechanisms disrupted by variants and tissue of origin for disease histotypes. *M.R. Jones[1], K. Lawrenson[2], C. Phelan[3], A. Antoniou[4], P. Pharoah[5], S. Gayther[1], D.J. Hazelett[1], Ovarian Cancer Association Consortium, Consortium of Investigators of Modifiers of BRCA1 and BRCA2.* 1) Bioinformatics and Functional Genomics Center, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA; 2) Women's Cancer Program at The Samuel Oschin Comprehensive Cancer Institute, Los Angeles, CA, USA; 3) Department of Cancer Epidemiology, Division of Population Sciences, Moffitt Cancer Center, Tampa, FL, USA; 4) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 5) Department of Public Health and Primary Care, University of Cambridge, Cambridge, Cambridgeshire, UK.

Invasive epithelial ovarian cancer (EOC) has the highest mortality among cancers specific to women. Genetic association studies comprising ~24,000 ovarian cancer cases and 29,000 controls have so far identified 32 EOC risk loci for different histoypes of disease (including high and low grade serous, clear cell, endometrioid, mucinous and low malignant potential EOC) at genome wide levels of statistical significance ($P=5x10^{-8}$). While many cancer loci are enriched in non-coding regulatory elements in cell types reflecting the primary source tissue, a surprisingly large fraction of loci exhibit little to no enrichment at all. To address this for ovarian cancer, we have mapped and annotated all EOC risk loci relative to non-coding features of publically available cell lines (ENCODE) and tissue samples (Roadmap Epigenomics Mapping Consortium), and for regulatory datasets (H3K27ac; H3K4me3; CTCF; PAX8; noncoding RNAs) generated in-house for a range of ovarian cancer and normal precursor tissues. In doing so we have identified the regulatory mechanisms most commonly intersected with EOC risk variants, and for different subtypes of ovarian cancer the likely cells of origin. Our data indicate that ovarian surface and fallopian tube epithelial cells (OSE and FTE) are the most likely precursors of serous ovarian cancer ($P_{enrichment} = 2.4x10^{-23}$ for OSE and $3.8x10^{-30}$ for FTE). Our analyses implicate other tissues and organ systems in the etiology for other ovarian cancer subtypes; in particular mucinous ovarian cancer, which does not appear to derive from tissues of the ovary. For mucinous risk loci we observe enrichment of SNPs co-localizing with enhancers for gastro-intestinal tissues, blood cell types (implicating the immune system), mesenchymal tissues, and even in some stem cell populations. We have also identified a subset of loci for which the top causal SNPs affect active enhancers (marked with H3K27Ac) across the majority of cell types, constituting candidate pleiotropic loci. These results are consistent with a view of ovarian cancer etiology involving disease- and cell-type specific regulatory mechanisms, as well as non-cell autonomous mechanisms that may be shared with other cancers.

53

Identification of thousands of context-dependent eQTLs unravelling cell type-specific signalling networks. *P.A.C. Hoen[1], D.V. Zhernakova[2], P. Deelen[2], M. Vermaat[1], M. van Iterson[3], M. van Galen[1], W. Arindrarto[4], P. van 't Hof[4], H. Mei[4], F. van Dijk[2], H.J. Westra[5,6,7], M.J. Bonder[2], J. van Rooij[8], M. Verkerk[8], P.M. Jhamai[8], M. Moed[3], S.M. Kielbasa[3], J. Bot[9], M.A. Swertz[2], A. Isaacs[10,11], J.B.J. van Meurs[8], R. Jansen[12], B.T. Heijmans[3], L. Franke[2], Biobank-based Integrative -Omics Consortium (BIOS).* 1) Department of Human Genetics, Leiden University Medical Center, Leiden, Netherlands; 2) Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, Netherlands; 3) Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, Netherlands; 4) Sequence Analysis Support Core, Leiden University Medical Center, Leiden, Netherlands; 5) Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, USA; 6) Partners Center for Personalized Genetic Medicine, Boston, USA; 7) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, USA; 8) Department of Internal Medicine, ErasmusMC, Rotterdam, Netherlands; 9) SURFsara, Amsterdam, the Netherlands; 10) Genetic Epidemiology Unit, Department of Epidemiology, ErasmusMC, Rotterdam, Netherlands; 11) CARIM School for Cardiovascular Diseases and Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, Netherlands; 12) Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, Netherlands.

Expression quantitative trait loci (eQTLs) help to unravel the mechanisms regulating gene expression. More mechanistic insights can be derived from knowledge of the context influencing the nature and strength of eQTLs, as it appears that many of the eQTLs are only apparent in specific cell types or tissues, or after specific stimuli. We developed a novel strategy for the systematic identification of such context-dependent eQTLs, which does not require the direct measuring of their modifiers. In this approach, we determine the influence of the expression level of each gene on each eQTL in a statistical interaction model. The genes that modify the eQTL effect can be proxies for a cell type, when expressed in a cell type-specific manner, or proxies for exposure to external stimuli in case they respond to these environmental cues. This method was applied to the peripheral blood RNA-seq data from 2,116 unrelated, healthy Dutch individuals. Out of the 23,060 significant *cis*-regulated genes (false discovery rate ≤ 0.05), 2,743 genes (12%) show context-dependent eQTL effects. The majority of those were influenced by cell type composition, revealing eQTLs that are particularly strong in cell types such as CD4+ T-cells, erythrocytes, and even lowly abundant eosinophils. A set of 145 *cis*-eQTLs were influenced by the activity of the type I interferon signaling pathway. We identified several *cis*-eQTLs that are modulated by specific transcription factors that bind to the eQTL SNPs, such as SREBP2 binding to the *FADS2* promoter and EBF1 to the *MYBL2* promoter, observations supported by ChIP-seq data. This demonstrates that large-scale eQTL studies in unchallenged individuals can complement perturbation experiments to gain better insight in regulatory networks and their stimuli. The BIOS project is funded by BBMRI-NL, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007).

54

Local regulatory networks across two tissues and applications to analyze rare non-coding variants. *O. Delaneau[1], K. Popadin[2], M. Zazhytska[2], S. 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[2], E. Dermitzakis[1].* 1) Department of Genetic Medicine and Development,University of Geneva, Geneva, 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 (eQTL), a great resource to determine the function of non-coding variants. In order to describe the effects of eQTL on regulatory elements such as enhancers and promoters, we quantified gene expression (mRNA) and three key histone modifications (H3K4me1, H3K4me3 and H3K27ac) across two cell types (Fibroblast and Lymphoblastoid Cell Lines) in up to 300 densely genotyped European samples and then performed an integrative analysis of this multi-layered dataset. First, we find that nearby regulatory elements exhibit a strong coordination that exponentially decays with distance; an observation confirming our previous work (Kilpinen et al, Science 2013, Waszak et al, Cell 2015). This forms local chromatin modules that span up to 1Mb, often comprise multiple sub-compartments, overlap remarkably well topologically associating domains (TADs), bring multiple distal regulatory elements in close proximity, vary across cell types and drive co-expression at multiple genes. Next, we show that this regulation layer is under strong genetic control by mapping QTLs: we notably discovered ~40k chromatin QTLs (cQTLs) affecting 40% of the histone marks. In contrast, we also quantified chromatin modules by using principal component analysis and discovered QTLs for up to 70% of them (modQTLs). These large collections of cQTLs and modQTLs represent a new resource of functional variants with downstream effects on higher order phenotypes: they often are eQTLs, cell type specific and enriched for disease associated variants. Finally, we show how chromatin modules can be used to empower association studies of rare variants when whole genome sequencing is available. Specifically, rare alleles are counted within regulatory elements of each chromatin module, resulting in module-based mutational loads that are tested for association with gene expression or disease status, i.e. a burden test on gene expression. We notably applied this approach on the Geuvadis transcriptomic data and discover that expression of a substantial fraction of the genes (~8%) is associated with rare non-coding variants in modules. Overall, this large-scale study integrating gene expression, chromatin activity and genetic variation across two cell types and hundreds of samples provides key insights into the biology underlying eQTLs.

55

Condition specific transcription factor binding with ATAC-seq for GxE interactions. *R. Pique-Regi, D. Watza, M. Estill, S. Chaudhry, F. Luca.* Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI.

Genome-wide Association studies (GWAS) consistently show that a very large portion of loci important in determining human traits and disease conditions are located in non-coding regions of the genome. These regions likely contain specific regulatory sequences that control gene transcription and can also interact with changes in the cellular environment (e.g. drug treatment). Recent technical advances and large scale projects in functional genomics have facilitated the profiling of regulatory sequences across many cell-types and tissues, yet we are still very far from mapping the sequences that control the transcriptional response to many external stimuli. Importantly, much of the missing genetic heritability in GWAS may be polygenic but some of the small effects may also be dampened by latent environmental interactions that may be different at each loci. Supporting this hypothesis, we recently demonstrated that genes with gene-environment (GxE) interactions are highly enriched in GWAS. More precisely, 50% of genes with condition-specific allele specific expression in a study screening 250 cellular environments were also found in the GWAS catalog. Here, we profiled five of the most interesting environmental conditions for transcription factor binding activity. This was accomplished at a genome-wide scale with the recently developed technique ATAC-seq, which utilizes the Tn5 transposase to fragment and tag accessible DNA. When coupled with a computational method such as CENTIPEDE, footprint models for TFs with known motifs can be generated across the genome to detect binding. From our sequencing results we were able to resolve 383 actively bound motifs across all conditions. We were also able to characterize 5236 regions that have significantly changed chromatin accessibility (FDR < 10%) in response to both copper and selenium. We have extended the CENTIPEDE model hierarchical prior to detect motifs that have differences in footprint activity in treatment vs. control experiments. For both metal ions we have detected a significant increase of binding for ETS and CRE motifs. Our results demonstrate that ATAC-seq together with an improved footprint model are excellent tools for rapid profiling of transcription binding factor activity to study cellular regulatory response to the environment. We also detect examples of allele specific Tn5 hypersensitivity (ASH) and conditional ASH (cASH) revealing the mechanism underlying GxE interactions.

56

First depleted, then enriched: The evolution of transposable element co-option into gene regulatory function. *J. Capra[1], C. Simonti[1], M. Pavlicev[2].* 1) Vanderbilt University, Nashville, TN; 2) Cincinnati Children's Hospital Medical Center, Cincinnati, OH.

Transposable elements (TEs) make up more than half of the human genome. Since TEs contain regulatory sequences that promote their transcription and amplification, they provide a fertile landscape of potential gene regulatory elements on which evolution can act. Indeed, many classes of TE contribute to the evolution of gene regulation in different lineages through co-option into alternative gene promoters, enhancers, and even insulators. However, the current understanding of TE-based rewiring of gene regulatory programs comes from a small number of examples, largely in the immune and reproductive systems.   We comprehensively evaluated the prevalence and evolutionary dynamics of the co-option of TEs into gene regulatory enhancer and promoter elements across 112 cell lines and primary tissues from the FANTOM consortium. Overall, TEs are significantly depleted of regulatory enhancer activity compared to the genomic background ($P < 0.0001$). The degree of depletion varied across contexts (1.5–3x), but it was significant in every cellular context considered. Promoters were even more significantly depleted of TEs than enhancers (2.9x vs. 2.3x overall). Thus, in spite of their regulatory potential, TEs are significantly less active than non-TE regions genome-wide. This suggests that cells actively repress the activity of TE sequences, perhaps to protect themselves from the mutagenic properties of active TEs. Nonetheless, we find that enhancers with tissue-specific activity are significantly more likely to be derived from TEs. Furthermore, the likelihood that a TE has enhancer activity increases with its age. Ancient TEs (originating before the divergence of amniotes) are 9.2 times more likely to have enhancer activity than TEs that integrated on the great ape lineage, and young TE-derived enhancers are significantly more likely to be tissue-specific in activity. Nonetheless, we identified a small number of TE families, most notably the endogenous retroviruses (ERVs), with different dynamics that highlight unique functional trajectories and evolutionary innovations.   Our data suggest striking similarity in the evolutionary and functional dynamics of different TE families. TEs appear to be actively repressed upon integration into the genome, leading to the degradation of their sequences over time. However, when a specific element gains regulatory function in a tissue, it becomes protected from degradation, and may gain activity in other tissues.

57

Towards a mammalian atlas of *in vivo* epigenetic state at single cell resolution. *D.A. Cusanovich[1], R. Daza[1], J.B. Berletch[2], G.N. Filippova[2], L. Christiansen[3], F.J. Steemers[3], C.M. Disteche[2], C. Trapnell[1], J. Shendure[1].* 1) Genome Sciences, University of Washington, Seattle, WA; 2) UW Medicine Pathology, University of Washington, Seattle, WA; 3) Advanced Research Group, Illumina, Inc., San Diego, CA.

Epigenomic measures of chromatin accessibility, such as DNase I hypersensitivity site sequencing and the 'assay for transposase-accessible chromatin' (ATAC-seq), have allowed for many insights into the regulatory state of various cell types and tissues. However, these assays generate average measures of accessibility across a population of cells and therefore require either isolating pure cell populations or accepting that measures of tissue-level accessibility will be a composite readout of the relative proportions of cell types present in that sample. These technical considerations limit the types of samples that can be assayed and the conclusions that can be drawn, particularly for complex tissues made up of many different cell types and tissues with dynamic cellular populations. Recently, we developed a single cell resolution assay for chromatin accessibility based on the ATAC-seq protocol that allows us to generate data on thousands of single cells in parallel. With this assay, we have embarked upon creating comprehensive *in vivo* chromatin accessibility maps of mouse tissues at single cell resolution. In addition to adult samples such as bone marrow, spleen, lung and brain, we are mapping chromatin accessibility in whole mouse embryos. These data allow us to not only measure differences in accessibility across tissues but also to observe developmental changes within cell lineages. Thus far we have collected data from more than 10,000 cells – including 3,684 bone marrow cells, 2,329 spleen cells, and 4,042 E14.5 whole embryo cells – and we are currently collecting data from a broader range of tissues and time points. As an example of the data collected to date, cells from the bone marrow and spleen samples form 4 major cell lineages when clustered together on the basis of their accessible site usage. These lineages show cell type-specific accessibility patterns consistent with the major types of white blood cells – including myeloid cells, B cells, T cells, and erythroid cells. Furthermore, the estimated proportions of these cell types are consistent with expectations for the tissues of origin. With these maps we have generated *in vivo* chromatin accessibility maps in the mouse at finer resolution than has previously been possible. They represent a major step towards cataloging the *in vivo* chromatin landscape of all cell types in mammalian systems.

58

Molecular signatures associated with Zika virus exposure in human cortical neural progenitors. *F. Zhang[1], Y. Cheng[2], C. Hammack[2], E.M. Lee[2], S.C. Ogden[2], Z. Wen[3,4], Y. Li[1], B. Yao[1], T. Xu[5], L. Chen[5], H. Feng[5], Z. Wang[1], C. Shan[6], L. Huang[1], Z. Qin[5], K.M. Christian[3,4], P. Shi[6], M. Xia[7], W. Zheng[7], H. Wu[5], H. Song[3,4,8,*], H. Tang[2,*], G. Ming[3,4,8,9,*], P. Jin[1,*].* 1) Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA; 2) Department of Biological Science, Florida State University, Tallahassee, FL; 3) Institute for Cell Engineering, Johns Hopkins University School of Medicine, Baltimore, MD; 4) Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD; 5) Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA; 6) Departments of Biochemistry & Molecular Biology, Pharmacology & Toxicology, and Sealy Center for Structural Biology & Molecular Biophysics, University of Texas Medical Branch, Galveston, TX; 7) National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD; 8) The Solomon Snyder Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD; 9) Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD.

Zika virus (ZIKV) infection causes microcephaly and has been linked to other nervous system abnormalities. Currently, a large-scale ZIKV outbreak is occurring in the Americas and the virus has so far spread to over 60 different countries and territories. To establish a direct link between ZIKV and microcephaly, we and others have shown that ZIKV efficiently infects human neural progenitor cells (hNPCs) in monolayer and three-dimensional organoids derived from induced pluripotent stem cells. Here, we systematically profiled messenger RNAs and small RNAs in hNPCs exposed to Asian ZIKV^{C}, African ZIKV^{M}, and dengue virus (DENV) using massively parallel sequencing. In contrast to the robust global transcriptome changes induced by DENV infection, ZIKV^{M} has a more selective and larger impact on the expression of genes involved in DNA replication and repair. Both ZIKV^{C} and ZIKV^{M} infected hNPCs, stunted cell proliferation, and led to cell death. Many genes involved in cell cycle, DNA replication, and microcephaly were significantly downregulated upon the infection of both strains. While overall expression profiles were similar, ZIKV^{C}, but not ZIKV^{M}, induced upregulation of viral response genes and TP53. Intriguingly, p53 inhibitors were able to block ZIKV-induced apoptosis in hNPCs, and offered cells bigger neuronal protecting effects upon ZIKV^{C} infection. Deep sequencing of small RNAs revealed that 180 and 310 miRNAs were differentially expressed in ZIKV- and DENV-infected hNPCs, respectively. Interestingly, 8 miRNAs that have been shown to control cell cycle regulators were dysregulated in ZIKV-infected hNPCs. Lastly, small RNA sequences were aligned to viral genomes. About 0.5% and 1.7% of filtered reads from ZIKV- and DENV-infected hNPCs, respectively, could be mapped to the corresponding viral genomes, and exhibited distinct patterns. Among Zika viral small RNAs (vsRNAs), the most abundant sequences were found in predicted stem-loop structures in ZIKV genomic RNA. Intriguingly, several abundant vsRNAs themselves could directly target and induce the altered expression of the mRNAs involved in cell cycle, DNA replication, and microcephaly. Thus, our results reveal virus- and strain-specific molecular signatures associated with ZIKV infection. These findings help to understand ZIKV-host interactions and neurovirulence determinants of ZIKV, opening avenues for possible utilization of p53 inhibitors and/or small RNAs to protect human cells upon ZIKV infection.

59

Direct identification of non-coding variants that modulate expression using high-throughput experimental assays. *R. Tewhey*[1,2], *D. Kotliar*[1,2], *D.S. Park*[2], *E.A. Brown*[1,2], *S.K. Reilly*[1,2], *T.S. Mikkelsen*[2], *S.F. Schaffner*[1,2], *P.C. Sabeti*[1,2]. 1) Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA; 2) Broad Institute, Cambridge, MA.

Genome-wide association studies (GWAS) have implicated genetic variation at thousands of loci in various human diseases and traits. Nevertheless, improved understanding of these diseases is significantly hindered by the difficulty of pinpointing the causal alleles in each disease-associated region of the genome. This difficulty is largely due to our inadequate ability to directly test the effects of non-coding variation, which includes the majority of disease-associated variants. To address this challenge, we adapted the massively parallel reporter assay (MPRA) to identify variants that directly modulate gene expression. We applied it to 32,373 variants from 3,642 cis-expression quantitative trait loci (eQTLs) and control regions that were identified in lymphoblastoid cell lines (LCLs) by the Geuvadis Consortium. Variants identified by MPRA show a strong correlation with existing measures of regulatory function, demonstrating MPRA's capability to pinpoint causal alleles. In total, we identified 842 variants showing differential expression between alleles, including 53 variants within annotated regulatory elements that are associated with diseases and traits. One such example that we investigated in detail, rs9283753 a risk allele for ankylosing spondylitis, is a non-coding variant that alters expression of the prostaglandin EP4 receptor (PTGER4). Using CRISPR/cas9 based allelic replacement in LCLs we provide direct evidence that the variant is an eQTL causal allele that interacts with PTGER4 via a long-range (~200kb) interaction. Many of the variants identified, including PTGER4, have cryptic effects on transcription factor binding. By using an MPRA based saturation mutagenesis screen we can identify binding footprints at single nucleotide resolution within individual enhancers to elucidate candidate transcription factors. We have applied this approach to both rs9283753 and rs202125301, a single nucleotide indel in the first intron of BLK that is strongly associated with Systemic lupus erythematosus. This work illustrates the promise of using high-throughput experimental systems for comprehensively interrogating the impact of non-coding polymorphisms on transcriptional regulation and human biology.

60

Asprosin, a fasting-induced glucogenic and orexigenic protein hormone. *A.R. Chopra*[1,2]. 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX.

The ability to survive periods of fasting is the cornerstone of evolution and life on earth. Mammals respond to fasting by activating an enormous cascade of interconnected processes that are precisely coordinated by an array of hormones. Two such processes are appetite stimulation and hepatic glucose release into the circulation, which together, ensure the drive to obtain food, and keep the brain nourished and alert while that is accomplished. Through study of a rare genetic condition in humans – Neonatal Progeroid syndrome, we have discovered a new fasting-induced protein hormone that is highly expressed in adipose tissue, and upon secretion, coordinately stimulates appetite and hepatic glucose release, and we name it Asprosin. Asprosin is a ~30 kDa C-terminal cleavage product of a larger pro-protein (profibrillin) and circulates in plasma and cerebrospinal fluid (CSF) at nanomolar levels. Its plasma and CSF levels rise with fasting, and in a cell-autonomous manner at the liver and arcuate nucleus, it activates the G-protein-cAMP-PKA pathway resulting in rapid glucose release into the circulation and stimulation of appetite. The cumulative effect of increased circulating asprosin concentration is hyperglycemia, hyperinsulinemia and an increase in appetite, adiposity and body weight. Asprosin loss-of-function results in the opposite – hypoglycemia, hypoinsulinemia, and a reduction in appetite leading to reduced adiposity and body weight. Humans and mice with obesity and insulin resistance show pathologically elevated plasma asprosin levels, and its immunologic or genetic depletion results in protection from obesity-associated hyperinsulinemia and hyperphagia. Thus, asprosin represents the first example of a circulating glucogenic and orexigenic protein hormone that is cleaved from a pro-protein that also generates an extracellular matrix component (fibrillin), and therapeutically targeting it may be beneficial in hyperinsulinemic and hyperphagic conditions like metabolic syndrome.

61

Aggregate allelic burden for cancer risk genes associates with age at diagnosis across 8,206 exomes. *J.J. Pitt[1,2], M. Bolt[1,3], D.J. Fitzgerald[1], L.L. Pesce[4], P. Van Loo[5,6], K.P. White[1,2,3].* 1) Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL, USA; 2) Committee on Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, IL, USA; 3) Department of Human Genetics, University of Chicago, Chicago, IL, USA; 4) Computation Institute, University of Chicago, Chicago, IL, USA; 5) The Francis Crick Institute, London, UK; 6) Department of Human Genetics, University of Leuven, Leuven, Belgium.

Cancer is a complex disease with many known environmental and genetic risk factors. 5-10% of cancers cases can be attributed to highly-penetrant, inherited alleles, which often lead to earlier age at diagnosis. However, the polygenic nature of cancer risk loci and its relationship to age at diagnosis is less understood. We generated harmonized variant calls for 8,206 blood germline exomes from The Cancer Genome Atlas, which represented 31 distinct cancer types (Vcfs are available through the Bionimbus Protected Data Cloud). Across all malignancies, we show that increased pathogenic and deleterious allele burden within ClinVar cancer risk genes (n = 60) is associated with earlier age at diagnosis. On average, each known pathogenic or predicted deleterious allele reduced age at diagnosis by 0.89 (95% confidence interval (CI) = 0.35 – 1.44; P = 0.00060) and 0.57 (95% CI = 0.16 – 0.98; P = 0.0031) years, respectively. This pattern was not seen with random gene sets, genes somatically mutated in cancer, and deleterious alleles exome-wide. These effects also remained when adjusting for race (pathogenic P = 0.0028; deleterious P = 0.0073) and cancer type (pathogenic P = 0.012; deleterious P = 0.0026). Strikingly, we show that high allele burden in breast cancer is an independent predictor of age at diagnosis (P = 0.014) and its effect was as strong as mutations in *BRCA1/2*. Based on ExAC data, individuals with cancer were significantly enriched for deleterious alleles in ClinVar cancer genes compared to controls (P = 0.0052). Finally, using age at diagnosis, we identified novel associations between susceptibility genes and rare tumor types, putatively implicating *ATM* with mesothelioma (P = 0.0047; Q = 0.058) and thymoma (P =0.0050; Q = 0.058). Overall, we propose that greater levels of baseline genetic vulnerability likely renders individuals more sensitive to somatic mutation insults, which subsequently manifests in earlier oncogenesis. Evaluating individuals' harmful alleles in aggregate may assist in clinical cancer risk assessment and warrants further large-scale testing to directly assess its value in screening.

62

Ultraconservation of DNA sequence provides a new lens for focusing on chromosomal structural rearrangements in neurodevelopmental disorder genomes. *R.B. McCole[1], C.Y. Fonseka[1], J. Erceg[1], H. Brand[2], R. Collins[2], V. Pillalamarri[2], S. Erdin[2], C. Redin[2], M.E. Talkowski[2,3,4], T. Wu[1].* 1) Department of Genetics, Harvard Medical School, Boston, MA; 2) Molecular Neurogenetics Unit and Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; 3) Department of Neurology, Harvard Medical School, Boston, MA; 4) Program in Medical Genetics and Genomics Platform, Broad Institute of Harvard and MIT, Cambridge, MA.

Structural genomic rearrangements are an increasingly appreciated feature of genomes from patients with neurodevelopmental disorders (NDD), including autism spectrum disorders (ASD). Recent prolific genotyping and sequencing efforts have uncovered a diverse array of DNA rearrangements within the genomes of NDD patients, including copy number variants, especially those forming *de novo* in patient genomes, genomic inversions, translocations, and complex chromosome breakage events often called chromothripsis. My work has uncovered an important common feature shared by these genomic rearrangements − across the human genome they are very significantly associated with ultraconserved elements (UCEs), which are regions of the genome demonstrating extremely high and unexplained DNA sequence conservation. More specifically, I have found using a permutation-based approach in a meta-analysis of ASD patients that UCEs are significantly enriched within 218 genomic regions affected by *de novo* copy number variation in these patient's genomes ($P=4.81\times10^{-4}$, observed/expected = 1.329). Extending this connection to chromosomal rearrangements that are balanced and do not affect copy number, I also discovered enrichment of UCEs within 220 genomic regions that contain balanced chromosome rearrangement breakpoints from NDD patients ($P=1.49\times10^{-7}$, observed/expected = 3.087). Importantly, I have previously demonstrated[1] that in healthy controls, UCEs show an opposite relationship to structural variation, being depleted from *de novo* CNVs, and further, that UCEs are enriched within cancer-associated CNVs[1]. With these studies in hand, I am exploring the potential links between genome fragility and structural rearrangements in both cancer and NDD, including the speculative model that UCEs may be involved in detecting structural genomic aberrations in healthy cells, and that both neurodevelopmental disorders and cancers are possible consequences that arise when this UCE-based mechanism is compromised.
1. McCole, R. B., Fonseka, C. Y., Koren, A. & Wu, C.-T. Abnormal dosage of ultraconserved elements is highly disfavored in healthy cells but not cancer cells. *PLoS Genetics* 10, e1004646 (2014). This work was supported by grants from the NIH/NIGMS (RO1GM085169, 5DP1GM106412) and Quad Seed funding from Harvard Medical School, a William Randolph Hearst Fund award to R.B.M., and an EMBO Long-Term Fellowship to J.E.

63

Pediatric acute myeloid leukemia survival differences revealed by comprehensive miRNA sequence analysis.

E.L. Lim[1], D.L. Trinh[1], R. Ries[2], Y. Ma[1], J. Topham[1], M. Hughes[2], E. Pleasance[1], A. Mungall[1], R. Moore[1], Y.J. Zhao[1], D.S. Gerhard[4], E.A. Kolb[5], A. Gamis[5], M. Smith[6], T.A. Alonzo[7], R.J. Arceci[3], S. Meshinchi[2], M.A. Marra[1]. 1) Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver BC; 2) Fred Hutchinson Cancer Research Center, Seattle WA; 3) Ron Matricaria Institute of Molecular Medicine, Phoenix Children's Hospital, Phoenix AZ; 4) Office of Cancer Genomics, National Cancer Institute, Bethesda MD; 5) Children's Oncology Group, Arcadia, CA; 6) Cancer Therapy Evaluation Program, National Cancer Institute, Bethesda, MD; 7) University of Southern California, Los Angeles, CA.

Background Induction chemotherapy results in complete remission in 80% of children with acute myeloid leukemia (AML). However, many patients either fail to achieve a remission, or relapse after an initial response and subsequently die of their disease. Although large numbers of somatic karyotypic and molecular alterations have been identified, the majority of them do not accurately stratify patients into risk groups or distinct pathway that can be readily exploited for therapeutic intervention. **Materials/Methods** As part of a genome-scale approach to identify prognostic markers and therapeutic targets, we provide a comprehensive characterization of the pediatric AML miRNA expression landscape, detailing miRNA & mRNA expression patterns and miRNA:mRNA interactions that are characteristic of the disease. miRNA-seq was performed on 696 samples (637 primary, 22 refractory and 37 relapse) and mRNA-seq was performed on 224 samples (177 primary and 47 relapse). **Results** Our analysis revealed that miR-106a-363 members were abundantly expressed in relapse and refractory samples and in primary samples of refractory patients (Wilcoxon test q-value<0.05). Integrative miRNA:mRNA analyses and luciferase reporter assays further demonstrated that several candidate targets of miR-106a-5p were involved in oxidative phosphorylation, which is a process that is suppressed in treatment-resistant leukemic cells. Using penalized lasso Cox regression, we developed a miRNA-based event free survival (EFS) predictive model, comprised of 36 miRNAs. Our model effectively stratified patients into Low, Intermediate and High risk groups (Log-rank p-value <0.001) in both our training and test cohorts, and was independent of established indicators of outcome (cytogenetic risk group and white blood cell count) (Cox PH p-value <0.001). **Conclusions** Through a detailed analysis of the miRNA expression landscape, we identified miRNAs whose expression levels were significantly associated with relapse and refractory disease. In particular, we showed that abundant expression of miR-106a-363 might contribute to treatment resistance by modulating genes involved in energy metabolism. Our miRNA-based predictor of EFS may be used in risk and response identification at diagnosis. Overall, our transcriptome expression profiles provide clinically meaningful data for risk and response identification and define novel pathways that may be amenable to therapeutic targeting.

64

Integrated landscape of molecular alterations in uveal melanoma.

H. Anbunathan[1], M. Field[2], W. Harbour[2], A. Bowcock[1]. 1) NHLI, Imperial College London, London, London, United Kingdom; 2) Bascom Palmer Eye Institute, Miami, USA.

Uveal melanoma (UM) is the most common primary malignancy of the eye in adults which arise from neural-derived melanocytes of the uveal tract of the eye leading to tumors of the iris, ciliary body and choroid. Approximately 50% of tumors metastasize within 10 years, most commonly to the liver. These tumors can be classified into two distinct groups based on gene expression profile namely the class 1 tumors with low metastatic potential and class 2 tumors with high metastatic potential. Previously we identified mutations in a tumor suppressor gene BRCA1 Associated Protein-1 (*BAP1*) on chromosome 3p21 that strongly correlate with class 2 metastasizing tumors and loss of the other copy of chromosome 3. In this study we performed an integrated analysis of genome wide chromosomal aberrations, somatic mutations and gene expression to further explore the genomic landscape in uveal melanoma. We identified 3499 protein altering somatic mutations including 2785 missense, 36 splicing, 208 nonsense and 298 frameshift mutations. We combined our mutation set with the data from TCGA (N=121; Our cohort = 41, TCGA = 80) and found only 6 genes to be significantly mutated against the background mutation rate and report frequencies of these known driver genes established in UMs which include *GNAQ, GNA11, CYSLTR2, BAP1*, *SF3B1* and *EIFIAX*. Oncogenic mutations *GNAQ* (44%), *GNA11*(48%) and *CYSLTR2* (6%) in a mutually exclusive manner which accounted for 98% of all cases and were independent of their gene expression classification while *BAP1* was seen in 35%, *SF3B1* in 25% and *EIF1AX* seen in 15% of cases. We identified recurrent copy number alterations involving chromosome 3, 6, and 8 that were consistent with previous findings, in addition we identified focal deletion on 6q and 11q chromosomal arms that were exclusively found in the small subclass of metastasizing class 1 tumors. Gene expression analysis of these loci reveal significantly altered genes that could be potentially associated with metastasis.

65

Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *A. Fujimoto*[1,2], *M. Furuta*[2], *Y. Totoki*[3], *T. Tsunoda*[2], *M. Kato*[3], *H. Yamaue*[4], *K. Chayama*[2,5], *S. Miyano*[6], *H. Aburatani*[7], *T. Shibata*[3], *H. Nakagawa*[2]. 1) Kyoto University, Kyoto, Kyoto, Japan; 2) IMS, RIKEN; 3) National Cancer Center Research Institute; 4) Wakayama Medical University; 5) Hiroshima University School of Medicine; 6) Institute of Medical Science, The University of Tokyo; 7) Research Center for Advanced Science and Technology, The University of Tokyo.

Liver cancer, which is most often associated with virus infection, is prevalent worldwide, and its underlying etiology and genomic structure are heterogeneous. Here we provide a whole-genome landscape of somatic alterations in 300 liver cancers from Japanese individuals. Our comprehensive analysis identified point mutations, structural variations (STVs), and virus integrations, in noncoding and coding regions. We discovered recurrently mutated coding and noncoding regions, such as long intergenic noncoding RNA genes (*NEAT1* and *MALAT1*), promoters, and regulatory regions. STV analysis found a significant association with replication timing and identified known (*CDKN2A*, *CCND1*, *APC*, and *TERT*) and new (*ASH1L*, *NCOR1*, and *MACROD2*) cancer-related genes that were recurrently affected by STVs, leading to altered expression. These results emphasize the value of whole-genome sequencing analysis in discovering cancer driver mutations and understanding comprehensive molecular profiles of liver cancer, especially with regard to STVs and noncoding mutations.

66

Comprehensive analysis of telomere length and telomere maintenance mechanisms across 31 human cancer types. *S. Zheng, F.P. Barthel, R. Verhaak.* Genomic Medicine, MD Anderson Cancer Center, TX.

Telomere caps chromosome ends and prevent chromosomal fusions. In the majority of somatic cells telomere shortens with each cell division. Cancer cells, on the other hand, maintain telomere length (TL) through reactivation of telomerase or alternative lengthening of telomere pathway (ALT). Though closely related to cancer hallmarks such as chromosomal instability, telomere length has not been systematically analyzed in cancer. We used DNA sequencing to infer TL in 18,430 samples across 31 cancer types. Tumor TL was shorter compared to normal TL but tended to be longer in testicular germ cell tumors, sarcomas and gliomas. TL in non-neoplastic leukocyte and solid tissue samples was negatively correlated with patient age and varied between lineages, with kidney samples showing the longest TL and leukocytes the shortest. Amongst tumors, 73% expressed telomerase reverse transcriptase *(TERT*), which was associated with mutations of the *TERT* promoter (23%), focal amplifications (6%), gene fusions (1%) and structural variants of the promoter (4%) and gene body (3%). Distal breakpoint positions involved in *TERT* promoter structural variants demonstrated increased levels of H3K27ac and H3K4me1, suggesting displaced enhancer elements. We additionally detected hypermethylation of the *TERT* promoter in 69%, and found an unexpected association with *TERT* expression. Combined, 95% of *TERT* expressing tumors was found positive for at least one potential genomic or epigenetic regulatory event. Six percent of tumors did not express *TERT* and harbored somatic mutations, deletions, gene fusions or structural variants in *ATRX* or *DAXX*, both of which have been shown to be tightly associated with ALT. These tumors demonstrated decreased of *ATRX* expression in combination with significantly longer TL and expression of telomeric repeat containing RNA (*TERRA*). Interestingly, 21% of the cohort did not express *TERT* and was *ATRX* wild-type. In this double wild-type group, unsupervised analysis identified positive correlations between telomere length, *TP53* and *RB1* alterations. Gene expression was found to be TL dependent, with genes nearby telomeres showing a negative correlation (increased expression with shorter TL) whereas genes far away from telomeres showed a positive correlation and (increased expression with longer TL). Our analysis provides insights into the various modalities associated with *TERT* expression and provides a landscape of determinants of telomere length in cancer.

67

Reversion to stress-induced mutation is a hallmark of cancer. *K.J. Bussey[1,2], L. Cisneros[1,2], A. Orr[1], M. Miocevic[1], C.H. Lineweaver[3], P. Davies[1].* 1) Arizona State University, Tempe, AZ; 2) NantOmics, LLC, Tempe, AZ; 3) Australian National University, Canberra, ACT, Australia.

Cancer can be interpreted as a reversion to single cell behavior. Regulatory mechanisms to suppress cell-level traits that are detrimental to multicellularity should be evolutionarily young, e.g. < 500 million years (MY). We predict genes causally mutated in cancer are evolutionary young and associated with these regulatory mechanisms. To investigate this, we looked at the genomic distribution of mutation via private, presumably germline SNVs from 130 individuals derived from trios in 1000 Genomes Project as well SNVs from 674 tumor samples from seven different tissues with whole genome sequencing from the ICGC database, release 19. We established the evolutionary ages of 19,786 human genes by assigning them to gene families using Ensembl Pan-Taxonomic Compara database, release 22, defining the age of the human member of the gene family as the maximum phylogenetic divergence time between humans and all the species represented in the corresponding gene family according to TimeTree. We analyzed the relationships between mutational frequency, gene age, amniote homologous synteny blocks (HSBs), evolutionarily re-used breakpoint regions (EBRs) and gene function. We observed that under a model of uniform random mutation across the genome controlled for gene size, genes <500 MY were more frequently mutated in both normal and cancer samples. Genes defined as causal in cancer using the COSMIC Cancer Gene Census were depleted in this age group. Functional enrichment analysis to explain this paradox revealed COSMIC genes with recessive phenotypes were enriched for DNA repair and cell cycle control. The non-mutated members of these pathways are orthologous to genes involved in stress-induced mutation in bacteria. Stress-induced mutation results in clustering of SNVs. Both normal and cancer samples demonstrated SNV clustering. In normal samples clusters of clusters (aka hot spots) were enriched in EBRs (OR=1.132, $p<2.2x10^{-16}$) and excluded from HSBs (OR=0.3061, $p<2.2x10^{-16}$) and COSMIC genes (OR=0.625, p=0.00855). In cancer, hot spots were excluded from EBRs (OR=0.5443, $p<2.2x10^{-16}$), HSBs (OR=0.8291, $p<2.2x10^{-16}$), and COSMIC genes (OR=0.4517, p= 0.003278). Our results suggest the mutational response to stress was developmentally co-opted to maintain diversity in the germline and immune system. Reversion to a stress-induced mutational response is a hallmark of cancer that allows it to effectively search “protected” genome space where causally implicated genes are located.

68

DNA fragile site breakage as a measure of chemical exposure and predictor of individual susceptibility to form oncogenic rearrangements. *Y. Wang[1], C. Lehman[1], Y. Nikiforov[2].* 1) University of Virginia, Charlottesville, VA; 2) University of Pittsburgh, Pittsburgh, PA.

Chromosomal rearrangements are common in cancer and frequently lead to formation of oncogenic fusions that are driver events in cancer progression. Radiation exposure is known to initiate DNA breaks but accounts for only a low percentage of tumors, thus, other factors are involved in generating rearrangements. We previously demonstrated that treatment of human thyroid epithelial cells with fragile site-inducing chemicals can cause significant DNA breakage at the *RET* gene and generate the *RET/PTC* rearrangement, a common mutation in papillary thyroid carcinomas (PTC). Here, we evaluated and now demonstrate that treatment with non-cytotoxic levels of environmental chemicals (benzene and diethylnitrosamine) or chemotherapeutic agents (etoposide and doxorubicin) generates significant DNA breakage within *RET* at levels similar to those generated by fragile site-inducing laboratory chemicals. This suggests that chronic exposure to these chemicals plays a role in the formation of non-radiation associated *RET/PTC* rearrangements. Further, we investigated whether the sensitivity of fragile sites could predict the likelihood of rearrangement formation using normal thyroid tissues from patients with and without *RET/PTC* rearrangements. We found that normal cells of patients with thyroid cancer driven by *RET/PTC* rearrangements have significantly higher blunt-ended, double-stranded DNA breaks at *RET* than those of patients without *RET/PTC* rearrangements. This sensitivity of a cancer driver gene in normal cells suggests for the first time that a DNA breakage test at the *RET* region could be utilized to evaluate the propensity to form *RET/PTC* rearrangements. Moreover, the significant increase of blunt-ended DNA breaks, but not other types of DNA breaks, in normal cells from patients with *RET/PTC*-driven tumors suggests that blunt-ended DNA breaks are the direct substrate required for rearrangement formation, and implicate involvement of the non-homologous end joining pathway in the formation of *RET/PTC* rearrangements.

69

Inferring migration and population-size surfaces across time periods. *H. Al-Asadi[1], D. Petkova[2], J. Novembre[1], M. Stephens[1].* 1) University of Chicago, Chicago, IL; 2) University of Oxford, Oxford, UK.

We present a method MAPs (“inferring Migration And Population-size Surfaces) to visualize population sizes and migration rates across space and time periods. We build on our previously published work, EEMS, a Bayesian method for inferring and visualizing effective migration rates given geo-referenced samples and a grid of populations. However, instead of the SNP-based genetic distance used in EEMS, MAPs uses long segments of genetic similarity as a summary statistic. These segments are indicative of recent shared ancestry (often termed “identity by descent”, IBD). Doing this introduces key advantages. First, modeling IBD segments allows the user to infer recent (and not time-averaged) migration rates and population sizes. Second, it allows the user to infer migration and population surfaces for different recent time periods, which is accomplished by varying the IBD length threshold. Finally, modeling IBD segments introduces recombination into the model and consequently allows migration rates and population sizes to be inferred separately rather than as a joint effective migration parameter. Using extensive coalescent simulations, we find MAPs capable of inferring very recent migration barriers, which are not detectable with methods using SNP-based genetic distance summaries, and capable of inferring both population sizes and migration surfaces across time periods. We apply MAPs to various test datasets and find that the migration patterns are consistent with expected geographic barriers. In addition, we find that the MAPs inferred population sizes mirror census sizes in European countries very well – for example, the correlation is 0.84 for the POPRES data-set (individuals with ancestry across Europe).

70

Genetic variation reveals migrations into the Indian subcontinent and its influence on the Indian society. *A. Bose[1,2], D.E. Platt[2], L. Parida[2], P. Paschou[3,4], P. Drineas[1].* 1) Department of Computer Science, Purdue University, West Lafayette, IN; 2) IBM T.J. Watson Research Center, Yorktown Heights, NY; 3) Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupoli, Evros, Greece; 4) Department of Biological Sciences, Purdue University, West Lafayette, IN.

Archaeological excavations revealed artefacts used by homo Erectus as long as 500-200ky. The moistening at the end of the last glacial period brought expanded subsistence; drying then spread agriculture from 8-5kya, marking some of the earliest migrations and expansions. Around 5ky, the Indus Valley civilization began with the much matured Harappan civilization, whose de-urbanization led to the initiation of the Vedic period. Following this, displacements followed as foreign rulers established dominance in the Indian subcontinent: from Greeks and Scythians, to the first seeds of Muslim invasions, followed by the Mughal Empire. In this phase, India had diverse rulers (including Afghans, Turks, and Mongols). The migrations led to widespread admixture of the Indian population, influencing language, culture, caste endogamy, metallurgical technologies, and more, resulting in a complex and differentiated structure. We set out to explore modern genetics correlating with migration routes into the subcontinent, and to study genomic variation in 48,570 SNPs genotyped in 1484 individuals, across 104 population groups. We propose, COGG (Correlation Optimization of Genetics and Geodemographics), a novel optimization method to model genetic relationships with social factors such as castes, languages, occupation, and maximize the correlation with geography. We calculated the shared ancestry between different caste groups in the subcontinent with other reference populations from Eurasia, using a novel approach. We tested different migration theories into the subcontinent using a Linear Discriminant Analysis of redescription clusters and study recombination events shaping the gene pool. Our results demonstrate that COGG gives us significantly higher correlations, with p-values lower than 10^{-8}. Identification of significant components among caste, language and genetics simplifies the complex structure. We identify varnas (Brahmins and Kshatriyas) to be closely related to reference Eurasian populations, whereas tribal groups show no shared ancestry with them and conclude that they resided in India before migration from Eurasia happened. We identify probable migration routes from Mongolia through Central Asia, and another via Anatolia into the subcontinent. Tibeto-Burman speaking populations share some ancestry with populations from East Asia; on the other hand, Austro-Asiatic speakers did not share ancestry with other Mon-Khmer language speaking populations.

71

Ancestry-specific estimation of recent effective population size in the Americas. *S.R. Browning[1], B.L. Browning[2].* 1) Department of Biostatistics, University of Washington, Seattle, WA; 2) Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA.

Many current-day American populations are admixtures of African, European and/or Native American ancestry. For individuals in these admixed populations, the continental ancestry of the individual's genetic material can be estimated at each point in the genome. We have developed a method to estimate the ancestry-specific recent effective population sizes for admixed populations from local ancestry calls and inferred identity by descent (IBD) segments. Using this method, we can estimate the past effective sizes of the ancestral African, European, and Native American populations, as well as the founding sizes at the time of colonization/migration and the post-admixture sizes. Using simulated admixed data, we demonstrate that our method gives accurate estimates of ancestry-specific effective population sizes for the past 200 generations.   We have applied our method to six admixed American populations from the 1000 Genomes Project and to African American populations sampled from two US cities. In each of these populations, we estimate that the ancestral Native American, European, and African populations had much larger effective population sizes (typically 1-3 orders of magnitude larger) than the corresponding ancestry-specific sizes in the Americas immediately after the colonization/migration events. In most cases, ancestry-specific population sizes rebounded quickly after colonization/migration. We also estimate that prior to colonization, the growth rates of the Native American ancestral populations were 2-4% per generation, which is similar to the estimated growth rates for the European and African ancestral populations over the same time periods.

72

Large-scale characterization of admixed populations and extensions of admixture mapping within the Population Architecture using Genomics and Epidemiology (PAGE)-II study. *G.L. Wojcik[1], K. Nishimura[2], A. Reiner[2], C. Hodonsky[3], S. Shringarpure[1], G. Belbin[4], M.P. Conomos[5], J. Haessler[2], T.A. Thornton[5], C. Laurie[5], L. Hindorff[6], R. James[12], C. Haiman[7], L. LeMarchand[8], T. Matise[9], S. Buyske[10], B. Thyagarajan[11], C. Carlson[2], R. Loos[4], K.E. North[3], C. Avery[3], C. Kooperberg[2], C.D. Bustamante[1], C.R. Gignoux[1], E.E. Kenny[4], PAGE-II Study.* 1) Department of Genetics, Stanford University School of Medicine, Stanford, CA; 2) Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA; 3) Department of Epidemiology, University of North Carolina, Chapel Hill, NC; 4) The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; 5) Department of Biostatistics, University of Washington, Seattle, WA; 6) Division of Genomic Medicine, NHGRI, NIH, Bethesda, MD; 7) Department of Preventive Medicine, University of Southern California Keck School of Medicine, Los Angeles, CA; 8) Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI; 9) Department of Genetics, Rutgers University, New Brunswick, NJ; 10) Department of Statistics & Biostatistics, Rutgers University, New Brunswick, NJ; 11) Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN; 12) Division of Clinical Research & Data Management, NIMHD, NIH, Bethesda, MD.

Admixture occurs via the process of human diaspora as the result of meiotic crossover between ancestral populations. Studies using two- or three-way continental ancestry inference in admixed populations, predominately African-American and Hispanic/Latinos, have leveraged genomic signatures of admixture for the discovery of loci differentially distributed in ancestral populations underlying multiple phenotypes. Here we present an expansion of our current admixture mapping paradigms to encompass a greater spectrum of admixed populations arising from the past 500 years of intercontinental travel to the Americas from Africa, Europe, Asia, and Oceania. The Population Architecture using Genomics and Epidemiology (PAGE)-II Study includes over 50,000 participants from over 50 admixed populations, including African American, Native Hawaiian Islander, and Hispanic/Latino groups in the United States, as well as populations from the Caribbean, Central and South America. PAGE participants were genotyped on the Illumina Infinium Multi-Ethnic Genotyping Array (MEGA), which was designed with cross-population tag SNP selection strategy to reduce bias for ancestry inference and ancestry informative markers to improve performance in underrepresented populations. We also genotyped an extensive reference panel (61 groups) of relevant ancestral populations; including West African, Polynesian, Northwestern and Southern European, Meso and Andean Native American, East Asian and South Asian individuals on the same platform. We infer a wide variety of proportions continental and subcontinental ancestry in the PAGE populations; such as East Asian and Polynesian ancestry in Native Hawaiian Islanders, South Asian ancestry in Tobagan and Guyanese populations, as well as evidence of appreciable Native American ancestry in a subset of self-reported African Americans. The large sample size and the dense backbone of the array enable us to reliably estimate local ancestry for admixed populations with a range of ancestral reference populations (k=2 to k=5) and apply a novel admixture mapping model that allows for up to 5 ancestral populations to characterize ancestry-specific associations. We will demonstrate with a cadre of inflammation and blood traits. These expanded frameworks interrogating the complex ancestral landscape of admixed populations will become increasingly important moving forward with the formation of large multi-ethnic cosmopolitan biobanks and epidemiological studies.

73

A complex history of archaic admixture in modern humans. *R. Bohlender[1], Y. Yu[1], C. Huff[1], A. Rogers[2].* 1) Epidemiology, MD Anderson Cancer Center, Houston, TX; 2) Anthropology, University of Utah, Salt Lake City, UT.

The sequencing of complete Neanderthal and Denisovan genomes has provided several insights into human history. One important insight stems from the observation that modern non-Africans and archaic populations share more derived alleles than they should if there was no admixture between them. We now know that the ancestors of modern non-Africans met, and introgressed with, Neanderthals and Denisovans. The estimate of the quantity of shared derived alleles, the mixture proportion, rests on an assumption of no archaic admixture in African populations, and so African populations have been used as the "non-admixed" outgroup in prior analyses. We find that the story is likely more complex, that the history within Africa involves admixture with population(s) related to Neanderthal and Denisova, and that the mixture proportion estimates for non-African populations have been biased, particularly in Melanesia. Here, we present results from a composite likelihood estimator of archaic admixture, which allows multiple sources of archaic admixture. We apply the method to archaic introgression, but it can be used to estimate ancient admixture among any four populations where the modeled assumptions are met. This joint estimate of Neanderthal and Denisovan admixture avoids the biases of previous estimators in populations with admixture from both Neanderthal and Denisova. To correct for dependence in our data, we use a moving blocks bootstrap to calculate confidence intervals. With assumptions about population size and more recent population separation dates taken from the literature, we estimate the archaic-modern separation date at ~440,000 ± 300 years ago for all modern human populations. We also estimate the archaic-modern mixture proportion in the 1000 genomes, and the modern genomes sequenced with the high coverage Neanderthal and Denisovan genomes. We report those estimates here, support several prior findings, and provide evidence for a lower level of Denisovan admixture (0.0191 [0.0184, 0.0197]), relative to Neanderthal (0.0256 [0.0247, 0.0265]), in Melanesia. On the basis of an excess of shared derived alleles between San, Neanderthal, and Denisova we suggest that a third archaic population related more closely to Neanderthal and Denisova than to modern humans introgressed into the San genomes studied here.

74

Ultra-fine structural inference and population assignment using IBD network clustering and classifiers accurately assign sub-continental origins represented in a large admixed U.S. cohort. *E. Han, R. Curtis, P. Carbonetto, K. Noto, J. Byrnes, Y. Wang, J. Granka, A. Kermany, K. Rand, E. Elyashiv, H. Guturu, N. Myres, E. Hong, C. Ball, K. Chahine.* Ancestry.com DNA, LLC, San Francisco, CA.

Motivation & Objectives: Identifying the geographic origin of individuals using genetic data has broad application in forensics, human disease and evolution. There have been multiple methods proposed to achieve this goal, such as Principle Component Analysis (PCA), Spatial Ancestry Analysis (SPA) and Geographic Population Structure (GPS). However, most methods suffer from decreased prediction accuracy outside Europe and do not apply to the US population comprised of admixed immigrants. In this study, we describe a new method and demonstrate its accuracy in predicting geographic origins in the US post-European colonization or internationally for single origin and admixed samples. **Methods:** We use a database of over 1.5 million consented genotype samples collected from the US and internationally, along with samples from public databases such as POBI. We build a genetic network by estimating the amount of identity-by-descent (IBD) sharing between all individuals. By iteratively applying the Louvain method for community detection, we find a hierarchy of genetic clusters in the network. Levering user-generated pedigrees going back 6-8 generations, we annotate each cluster with birth locations that are enriched in historical time periods. The birth locations of these clusters are generally specific to locations in the US or internationally, allowing for concise geographical interpretation. Although community detection results assign samples to only one cluster, we use machine learning classification to assign samples to multiple clusters. Given this classification and enriched birth locations, we identify the likely geographic origins of each sample. **Results:** Our results include over 300 stable clusters, each comprised of more than 1000 samples. Some clusters correspond to narrow geographical regions, such as people descended from southern West Virginia in the 19th century, and others to broader groups, such as European Jews from Poland. By using the associated pedigrees, we demonstrate the accuracy of these predictions: over 95% of the assigned individuals have at least one known ancestor born in the enriched region defined by most clusters. **Conclusion:** By utilizing large-scale genetic data with associated pedigrees, we have developed the first method for predicting the geographic origin of individuals within the US or internationally with high accuracy. This approach can be used for ultra fine scale genetic ancestry mapping in any population.

75

Systematic functional dissection of common genetic variation affecting transcriptional regulation and human disease. *J.C. Ulirsch[1,2], S.K. Nandakumar[1,2], L. Wang[2], F.C. Giani[1,2], X. Zhang[2], P. Rogov[2], A. Melnikov[2], P. McDonel[2], R. Do[3], T.S. Mikkelsen[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; 3) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY.

Genome-wide association studies (GWAS) have identified over 10,000 common single nucleotide polymorphisms (SNPs) associated with hundreds of human traits and diseases. Nevertheless, for the vast majority of GWAS loci, a causal variant remains unknown, as a result of the large number of variants in linkage disequilibrium (LD) and the challenges of assessing the function of non-coding variation. Recent advances in fine mapping are beneficial, but only rarely allow for the identification of a single causal variant. Similarly, large-scale genome editing approaches have proven useful for assessing endogenous elements, but do not allow for interrogation of allele-specific variation in a scalable manner. In order to address the need for high-throughput functional screening of GWAS loci, we have developed and applied a massively parallel reporter assay (MPRA) that can identify regulatory elements that alter transcription from a minimal promoter. As a proof-of-principle, we chose to screen 2,756 variants in high LD with 75 sentinel SNPs identified from the GWAS for red blood cell (RBC) traits. Our MPRA identified endogenous regulatory elements based upon erythroid DNase I hypersensitivity (DHS) and the activity of tested elements was predicted by the presence of key erythroid transcription factor (TF) motifs. We identified 32 functional variants, termed MFVs, that exhibited differential activity by allele at 23 loci. We derived a positive predictive value for causality from our screen of between 32-50% based upon genetics and putative regulatory function from predictive algorithms. Where heterozygous samples were available, 88% of MFVs were directionally consistent with allele-specific reads in either DHS or ChIP-seq. Finally, we used genome editing to verify the endogenous effects of 3 MFVs and identified 1-3 target genes for each. In one case, we linked the target gene, *RBM38*, back to the original GWAS phenotype in primary human hematopoietic cells. Importantly, when we analyzed TF binding motifs, DNA shape features, and allelic skew in ChIP-seq data, we determined that common SNPs associated with RBC traits frequently affect a regulatory pathway involving the erythroid master TF GATA1. Overall, our method provides a novel, scalable, and cost-effective approach for GWAS follow-up and our ongoing computational and experimental analysis, including the measurement of effects from differential promoter usage in MPRA, will likely further refine these results.

76

Massively parallel ChIP-reporter assays reveal synergystic clusters of transcription factor binding across the human genome. *T.E. Reddy[1,2], C.M. Vockley[1,3], A.M. D'Ippolito[1,4], I.C. McDowell[1,5], W.H. Majoros[1,5], A. Safi[1,6], L. Song[1,6], G.E. Crawford[1,6].* 1) Center for Genomic & Computational Biology, Duke University School of Medicine, Durham, NC, 27708, USA; 2) Department of Biostatistics & Bioinformatics, Duke University School of Medicine, Durham, NC, 27708, USA; 3) Department of Cell Biology, Duke University School of Medicine, Durham, NC, 27708, USA; 4) University Program in Genetics & Genomics, Duke University School of Medicine, Durham, NC, 27708, USA; 5) Program in Computational Biology & Bioinformatics, Duke University School of Medicine, Durham, NC 27708, USA; 6) Department of Pediatrics, Division of Medical Genetics, Duke University School of Medicine, Durham, NC 27708, USA.

Genetic variation in regulatory elements is a major contributor to human traits and diseases. However, determining the specific regulatory mechanisms involved in any given disease remains a major challenge. One reason that challenge persists is that there are millions of putative regulatory elements across the human genome, but few have been assayed for regulatory activity. As a step towards systematically evaluating the regulatory effects of all human enhancers, we developed a high-throughput reporter assay to simultaneously measure the activity of every regulatory element bound by a human transcription factor (TF). We demonstrate our ChIP-reporter approach by quantifying the activity of all >10,000 regulatory elements bound by the glucocorticoid receptor (GR). Results from our ChIP-reporter assays indicate that only a small fraction (~13%) of GR binding sites have glucocorticoid-responsive regulatory activity. However, integrating the reporter-assay results with matched data from several complementary assays (e.g. DNase-seq, ChIP-seq, RNA-seq, and Hi-C) suggests an alternative model. In that model, GR binds the genome in a locally-coordinated manner, resulting in clusters composed of different types of GR binding sites that together have synergistic effects on gene expression. We also provide evidence that our enhancer-cluster model is general to other TFs. The enhancer-cluster model helps to explain (i) the well-known clustering of TF regulatory elements genome wide, (ii) the discrepancy between the number of TF binding sites observed with ChIP-seq and the number of regulated genes observed with RNA-seq, and (iii) observations that single nucleotide variants can disrupt distal TF binding without altering the DNA sequence bound by that TF. As such, the enhancer-cluster model revises our understanding of the ways that non-coding genetic variation can alter regulatory mechanisms that contribute to human traits and diseases.

77

Inferring the genetic architecture of *cis*-gene regulation and evolutionary constraint on human gene expression. *E. Glassberg[1], Z. Gao[2,3], J. Pritchard[1,2,3].* 1) Department of Biology, Stanford University, Stanford, CA; 2) Department of Genetics, Stanford University, Stanford, CA; 3) Howard Hughes Medical Institute, Stanford University, Stanford, CA.

To characterize the genetic architecture and evolution of human gene expression, it is important to estimate the space of potential expression-altering mutations and to infer selective pressures on such variants. These goals are difficult to achieve by analyzing statistically significant eQTL signals alone due to linkage disequilibrium and to differences in statistical power between SNPs of varying allele frequencies. Here, we develop a likelihood-based method to estimate the underlying proportion and effect size distribution of sites affecting gene expression from detected eQTLs. We call *cis*-eQTLs from 85 Yoruban individuals from the GEUVADIS project; using forward stepwise regression, we test 314,195 SNPs within 5kb of the transcription start sites (TSSs) of 13,967 genes and find 1,924 significant eQTLs affecting 1,787 genes. Correcting for lost power due to allele frequency variation and assuming neutrality of gene expression, we estimate that approximately 2% of base pairs close to gene TSSs are truly causal; a 3-4 fold increase over uncorrected estimates based on significant eQTLs alone. This suggests that the regulatory architecture of gene expression is both more complex than standard single-eQTL-per-gene models and simpler than highly polygenic models. Further, the joint distribution of minor allele frequency and estimated effect size suggests stabilizing selection on gene expression. From the depletion of common, large effect eQTLs, we quantify the strength of purifying selection against expression-altering variants. Finally, we apply this method in additional datasets to compare estimated parameters across cell types and gene sets. This provides insight into variation in human genome regulatory architecture and constraint on gene expression.

78

Integrating genomic, endophenotypic, and exposure data to identifiy biomarkers of multi-drug treatment response. *M-J. Fave[1,2], H.A. Edgington[2], J-C. Grenier[1], V. Bruat[1], P. Awadalla[1,2,3].* 1) CHU Sainte-Justine, Universite de Montreal, Montreal, Quebec, Canada; 2) Ontario Institute for Cancer Research, Toronto, Ontario, Canada; 3) Department of Molecular Genetics, University of Toronto, Ontario, Canada.

Gene-by-environment (GxE) interactions are thought to be pervasive and may be responsible for a large fraction of the unexplained variance in phenotypic traits. Yet, a general understanding of how gene regulatory variation is modulated by environmental exposures is lacking. Multi-drug exposures are common in the adult population and individual genetic variation modulates the response to treatment. Such genotype-by-drug interactions can serve as the basis for the development of biomarker panels for predicting drug response and efficacy. To survey genetic, drug exposure, and interaction effects on whole blood transcriptome and discover biomarkers for drug response, we combined whole transcriptome RNA-Seq profiling with whole genome genotyping on 1,000 deeply endophenotyped individuals selected from over 40,000 participants in the CARTaGENE resource. We generated a multi-drug exposure profile and response status for each individual, including anti-diabetic, lipidemia, hypertension, and osteopenia drugs. We report several instances of significant transcriptional genotype-by-drug interactions (drug-eQTLs) for each of the four drug exposures. Several eQTL genes we identified are known to be associated with drug metabolism and were also differentially expressed between exposed and non-exposed individuals. We also report a number of novel interactions. To further identify and replicate interactions of drug exposure with personal genetic background, we used EAGLE, a Bayesian approach for identifying GxE interactions based on allele-specific expression (ASE). We discovered ASE genotype-by-drug interactions for hypertension medication exposure, but none for the other drug exposures, indicating possible unique consequences of hypertension drug exposition on gene expression. We replicated 7 drug-eQTLs that also exhibit ASE genotype-by-drug interactions with hypertension drug exposure, including TRIM44, a gene known to be sensitive to chemotherapy. Using multivariate approaches, we describe ASE profiles that are associated with each of the drug exposition profiles, revealing ASE associations with single and multiple drug exposure profiles. These biomarkers for genotype-by-drug interactions may enhance the optimization and design of testing panels for personalized treatments. More broadly, our work illustrates how environmental exposures can interact with individual genetic variation and alter gene regulation and endophenotypic variation in clinically relevant traits.

79

Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms.

M. Guo[1,2], *S. Nandakumar*[2,3], *J. Ulirsch*[2,3], *S. Zekavat*[2,4], *P. Natarajan*[2,4], *R. Salem*[1,2], *A. Metspalu*[5], *S. Kathiresan*[2,4], *J. Hirschhorn*[1,2], *T. Esko*[1,2,5], *V. Sankaran*[2,3]. 1) Division of Endocrinology, Boston Children's Hospital, Boston, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Division of Hematology/Oncology, The Manton Center for Orphan Disease Research, Boston Children’s Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA; 4) Center for Human Genetic Research, Cardiovascular Research Center, and Cardiology Division, Department of Medicine, Massachusetts General Hospital, Boston, MA; 5) Estonian Genome Center of the University of Tartu, Tartu, Estonia.

Genetic variants affecting hematopoiesis can influence commonly assayed blood laboratory measurements. In order to gain additional biological insight into human hematopoiesis, we performed genetic analyses in a population-based Estonian biobank to identify genes associated with various blood cell counts and traits. We performed high coverage whole genome sequencing in 2300 individuals, followed by SNP genotyping and imputation in an additional ~6000 individuals. Our analyses identified 17 genome-wide significant associations across 14 blood cell traits. At several loci, statistical fine-mapping analyses enabled through the WGS nominated a single putative causal variant, including at a novel basophil count-associated locus near the master regulator of hematopoiesis, *CEBPA*. The lead SNP at this novel *CEBPA* locus was strongly associated with basophil counts (rs78744187; MAF 10.4%; p-value 6.19x10^{-38}), and remarkably explained 4.3% of variation. Integration of epigenetic data revealed that this basophil-associated variant overlaps a novel myeloid regulatory element bound by the key hematopoietic transcription factors GATA2 and RUNX1. In luciferase reporter assays in myeloid cell lines, this element displayed strong enhancer activity (~40 fold), and the basophil-decreasing allele exhibited a 1.4 fold reduction in enhancer activity relative to the reference sequence. *In situ* perturbation of this enhancer by CRISPR/Cas9 mutagenesis in human hematopoietic stem and progenitor cells demonstrates that this enhancer specifically regulates *CEBPA* expression. Moreover, perturbation of this enhancer led to a decrease in basophil differentiation, along with a concomitant increase in mast cell production. Thus, we have identified a novel enhancer that provides temporal regulation of *CEBPA* to specify development along either the basophil or mast cell lineage. Our analyses also identified a basophil association at *GATA2* that was also associated with eosinophils, suggesting a broader role of this variant in production of myeloid lineages. As a result of our WGS analysis and functional follow-up, we have generated novel insights into a GATA2/CEBPA regulatory network involved in basophil differentiation, a process that has been poorly understood in this rare cell lineage that plays a key role in human inflammatory and allergic conditions.

80

The dynamics of genome topology in response to glucocorticoid treatment.

A. D'Ippolito[1,2], *I. McDowell*[1,3], *C. Vockley*[1,4], *A. Barrera*[1], *L. Hong*[1], *S. Leichter*[1], *L. Bartelt*[1], *T. Reddy*[1,5]. 1) Center for Genomic and Computational Biology, Duke University, Durham, NC; 2) University Program in Genetics and Genomics, Duke University Medical Center, Durham, NC; 3) Computational Biology and Bioinformatics, Duke University, Durham, NC; 4) Developmental and Stem Cell Biology Program, Duke University, Durham, NC; 5) Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC.

Understanding the interactions between distal regulatory elements and target gene expression is critical for understanding the genetic mechanisms of many human diseases. Recent technological advances such as in situ Hi-C have made it possible to observe the three-dimensional structure of the human genome at high resolution. Although some studies have examined changes in architecture in response to stimulus, few have examined topology dynamics at the resolution of DNA loops. Here, we have measured genome wide changes in chromatin conformation at high resolution across a 12-hour time course of steroid hormone treatment. Specifically, we used the glucocorticoid (GC) receptor (GR) as a representative model system. We found that GC treatment both induces and represses thousands of DNA loops, and that the induced and repressed loops have dramatically distinct properties. First, induced loops are smaller than repressed loops. Second, the anchors of induced loops are enriched for GC-induced genes and transcription factor binding, but depleted for CTCF. Third, induced loops are highly enriched for GC-responsive GR binding sites, as measured with massively parallel reporter assays. In contrast, repressed loops are enriched for GC-repressed genes in the loop interior rather than at the loop anchors, and the anchors of repressed loops were not enriched for specific DNA binding motifs. Together, these results reveal distinct topological dynamics of gene activation and repression across the human genome, with direct implications for understanding how genetic variation at gene regulatory sequences may contribute to changes in gene expression and, subsequently, human phenotypes.

81

Enabling petabyte-scale genomics in the cloud: Lessons from the NCI Cancer Genomics Cloud Pilots. *G. Kaushik, Z. Onder, D. Locke, B. Davis-Dusenbery, D. Kural.* Seven Bridges, Cambridge, MA.

The advent of next generation sequencing has transformed our ability to generate genomic data. Petabytes of multi-dimensional information from thousands of patients have been collected. However, access and analysis of this information only becomes more challenging as the amount of data continues to increase. This difficulty is exemplified when we consider data generated by the efforts of The Cancer Genomics Atlas (TCGA) network. Historically, downloading the complete TCGA repository can require several weeks with a highly optimized network connection. Further, integrated analysis of this data remains out of reach for any researcher without access to the largest institutional compute clusters. The Cancer Genomics Cloud Pilots project seeks to directly address these challenges by co-localizing data with the computational resources to analyze it. The project was born out of the recognition that conducting biological research is increasingly computationally-intensive. New approaches are required to support effective data discovery, storage, computation, and collaboration. Funded by the National Cancer Institute, the Cancer Genomics Cloud Pilot project enables researchers to leverage the power of cloud computing to gain actionable insights about cancer biology and human genetics from massive public datasets including TCGA. We will highlight our approach to optimized computation, data mining, and visualization solutions that address the challenges associated with analysis of petabyte-scale datasets and beyond. To date, more than 600 researchers have accessed and analyzed TCGA and received free computation and storage credits at www.cancergenomicscloud.org.

82

Modeling the subclonal evolution of cancer cell populations. *M. Wilson Sayres[1,2], D. Chowell[3,4], J. Napier[5], R. Gupta[3], L. Faiss[5], C. Maley[3,6].* 1) School of Life Sciences, Arizona State University, Tempe, AZ; 2) Center for Evolution and Medicine, The Biodesign Institute, Arizona State University, Tempe, AZ; 3) Center for Personalized Diagnostics, The Biodesign Institute, Arizona State University, Tempe, AZ; 4) Mathematical, Computational and Modeling Science Center, Arizona State University, Tempe, AZ; 5) Research Computing Center, Arizona State University, Tempe, AZ; 6) Center for Evolution and Cancer, University of California San Francisco, San Francisco, CA.

There is increasing evidence that tumor architectures are often the consequence of a complex branched subclonal processes; however, little is known about the expected dynamics and the extent to which these divergent subclonal expansions occur. Because this is fundamental to understanding the pathogenesis of cancers, here we study subclonal heterogeneity and resistant alleles by developing and implementing more than 80,000 simulations of a stochastic evolutionary model simulating the process. Under different combinations of the population genetic parameter values, including those that have been previously estimated for glioblastoma multiforme and colorectal cancer, our results show that, at the time of tumor detection, the distribution of sizes of subclones carrying driver mutations has a heavy right tail, with only 1-4 dominant clones present at ≥10% frequency, composing most of the tumor cell population. In contrast, our model predicts that the vast majority (between 100's and 1000's) of subclones will be present at <10% frequency. We find that these minor and often undetectable subclones can harbor treatment-resistant mutations. In our analysis, the number of minor subclones is strongly correlated with the number of dominant clones at detectable levels in a tumor. Model predictions are consistent with empirical data on the number of dominant detectable clones in different cancer types. Our results explain why tumors with greater numbers of detectable clonal populations tend to be associated with poorer clinical outcome across multiple cancer types.

83

Cancer gene discovery via network analysis of somatic mutation data. *I. Lee[1], A. Cho[1], J. Shim[1], E. Kim[1], F. Supek[2,3,4], B. Lehner[2,3].* 1) Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul, Korea; 2) EMBL-CRG Systems Biology Unit, Centre for Genomic Regulation (CRG), 08003 Barcelona, Spain; 3) Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain; 4) Division of Electronics, Rudjer Boskovic Institute, 10000 Zagreb, Croatia.

A major challenge for distinguishing cancer-causing driver mutations from inconsequential passenger mutations is the long-tail of infrequently mutated genes in cancer genomes. Algorithm development for mutation-based cancer gene prediction has successfully dealt with false positive candidates with high background mutation frequency. However, cancer gene discovery solely based on exome mutation frequency is intrinsically limited to genes with frequent mutations in coding regions. Here we present a cancer gene prioritization method based on a pathway-centric analysis of mutation data, MUFFINN (MUtations For Functional Impact on Network Neighbors) that integrates mutational information for individual genes and their neighbors in co-functional networks. MUFFINN is highly predictive for known cancer genes, particularly for genes with low mutation occurrence among cancer patients, with the identification of drivers amongst these genes having substantially higher sensitivity than conventional methods based on gene-centric analysis of mutation data. MUFFINN works effectively with both pan-cancer and individual cancer type samples. MUFFINN has only marginally reduced predictive performance when using only 10% of TCGA patient samples, suggesting that it will be a valuable method for small-scale cancer genome projects and in the initial-stages of larger projects. Using mutation frequency data for 18 types of cancers from TCGA (as of August 2014), we identified approximately 200 novel candidates for cancer genes that were not successfully prioritized by conventional gene-centric methods such as MutSig2.0, MutSigCV, and MutationAssessor. We were able to find supporting evidence for many of them being *bona fide* drivers. Furthermore, we provide a companion web-based prediction server (http://www.inetbio.org/muffinn), which allows researchers to prioritize candidate cancer genes by submitting mutation occurrence data.

84

Genetic determinants of translation in humans. *C. Cenik, J.A. Reuter, E. Sarinay Cenik, D. Spacek, C.L. Araya, M.P. Snyder.* Department of Genetics, Stanford University, Stanford, CA, 94305.

Genetic variants altering regulatory sequences in the genome can have profound consequences on normal phenotypic diversity as well as disease. For example, transcriptional regulatory mutations that increase telomerase gene expression affect ~70% of all melanoma patients and are frequent in several other cancers. While previous studies have begun to unravel the connections between genetic variation and RNA levels in humans, little is known about the genetic determinants of translation efficiency. We have recently conducted an integrative analysis of RNA expression, translation and protein levels in a diverse group of individuals (Cenik et al. 2015). We showed that combined analysis of RNA expression and ribosome occupancy improved the identification of individual protein level differences. Furthermore, we provided evidence that genetic variants in the human population can specifically regulate translation. In particular, our results revealed that the Kozak sequence has significant impact on translation efficiency globally and genetic variants modifying the Kozak sequence can lead to translation differences between individuals (Cenik et al. 2015). To test whether genetic differences in the Kozak region may also have a role in human disease, we extended our study and analyzed ~4700 cancer exome sequencing datasets from 21 cancer types. We identified recurrent somatic cancer mutations in the Kozak region of one gene, *TBC1D12,* affecting nearly ~20% of all bladder cancer patients. *TBC1D12* Kozak region mutations were most frequent in bladder cancer; yet, they were also observed in patients with multiple myeloma, lung adenocarcinoma, lung squamous cell carcinoma, head & neck squamous cell carcinoma, and breast cancer. As such, the identified Kozak mutations in *TBC1D12* constitute one of the most frequently mutated noncoding regions in any cancer type. While the position of the mutations implicate potential involvement of translation regulation, we are currently testing this hypothesis experimentally. These studies are expected to delineate specific roles for *TBC1D12* mutations in cancer. Finally, we developed a novel methodology to precisely measure the relative translation rate of heterozygous alleles from the same cell population. We applied this new assay to *TBC1D12* and several additional natural genetic variants. Taken together, our results establish a framework to discover the genetic determinants of translation in health and disease.

85

Prioritization of target drug combinations with immunotherapy using genomic data. *L. Machado Colli, M. Machiela, T. Myers, L. Jessop, K. Yu, S. Chanock.* DCEG-NCI, NIH, Rockville, MD.

Immune checkpoint inhibitor treatment represents a promising approach towards treating cancer. Stratification of patients for therapeutic decisions is the hallmark of the Precision Oncology Initiative. The capacity to use genomic profiles to select patients most likely to respond to therapy is based on the analysis of genetic alterations observed in available studies. In the near future, combination of checkpoint inhibitors with target drugs represents a promising development. Recent studies have suggested that the number of non-synonymous mutations (NsM) can be used to select melanoma and non-small cell lung cancer patients most likely to benefit from checkpoint inhibitor treatment. The ROC analysis based on these data suggests that 192 NsM optimizes sensitivity (74%) and specificity (59.3%) to identify a subset more likely to respond to immune checkpoint inhibitors. We conducted an assessment of NsM regarding mutational status of 7 genes that have target drugs (EGFR, KRAS, NRAS, NF1, PIK3CA, BRAF, and AKT1) across 7,757 tumor samples drawn from 26 cancers sequenced in the Cancer Genome Atlas (TCGA) Project to estimate the subset of cancers (both types and fractions thereof) that could benefit from target drug and checkpoint inhibitor combination. Cases with 192 or more NsM were classified as possible checkpoint inhibitor responders, while those with less than 192 NsM were classified as not responders. We observed 286 patients with mutated EGFR, 449 with mutated KRAS, 185 with mutated NRAS, 396 with mutated NF1, 983 with mutated PIK3CA, and 53 with mutated AKT1. NRAS (OR=4.65, p=1.1e-22), NF1 (OR=1.6, p=2.2e-27), BRAF (OR=2.17; p=1.1e-14), EGFR (OR=1.76, p=1e-4), and KRAS (OR=1.42, p=0.0016) mutated patients had higher chances to be a checkpoint responder. Our results indicate that the presence of strong driver mutations, namely, NRAS, NF1 and BRAF for which targeted therapy is available could be primary candidates for combination with checkpoint inhibitor therapy in future clinical trials.

86

High performance discovery of complex genome-wide rearrangements with single molecule-based barcoded sequence reads. *L.C. Xia[1,2,3], C. Wood[2], B. Lau[2], N.R. Zhang[3], H.P. Ji[1,2].* 1) Med/Oncology, Stanford University, Stanford, CA; 2) Stanford Genome Technology Center, Stanford University, Palo Alto, CA; 3) Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA.

Genomic rearrangements are hallmarks of and causally linked to many genetic disorders and other common disease such as cancers. Sequencing approaches to characterize such rearrangements are hindered by the low sensitivity of short read paired-end sequencing designs. Here we present the results of genome-wide rearrangement analysis of a colon cancer cell line (LS411N) using a high-throughput and sensitive barcoded linked-read sequencing design. In the work, we first studied the barcode statistics and proposed a Poisson field model for the background barcode sharing of the linked-read design as performed using 10X Genomics Chromium Systems. We then developed a genome-wide 2D-grid scan to identify genomic junctions based on local excessive sharing signals deviating from background statistics. Finally, we applied the designed linked-read sequencing and statistical analysis pipeline to LS411N and identified 118 rearrangement events, including deletions, inversions, tandem duplications and translocations. The LS411N cell line is severely aneuploidy. When using traditional whole genome sequencing methods, most of the events we identified have no or very low number of read pair and split read signals thus will be invisible to common rearrangement analysis. Therefore, our results represent a new sensitive and robust approach to characterize highly complex genomic rearrangements.

87

Using whole genome sequence data to identify causal variants affecting gene expression and disease. *A. Brown*[1], *O. Delaneau*[1], *T. Spector*[2], *K. Small*[2], *E. Dermitzakis*[1]. 1) Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland; 2) Department of Twin Research, Kings College London, United Kingdom.

The benefits of whole genome sequence (WGS) data over genotyping arrays for GWAS studies include increased power from better genotype calls and the potential to move beyond associated regions to causal variants (fully observed with sequence data). However, it is still prohibitive in cost to collect the numbers of samples currently necessary for a well-powered study. In contrast, gene expression studies with 100s of samples identify 1000s of genetic associations; at this scale collecting WGS is feasible. Analysing RNA-seq data in 4 tissues from our previous work on the Eurobats study (Buil et al., Nat Genet 2015), we compare eQTLs found using WGS from UK10K to eQTLs found with genotype imputed to 1000 Genomes from SNP arrays. We find WGS adds little additional power: 27,659 eQTLs discovered compared to 26,351. However, we see a 1.2-1.4 enrichment of WGS eQTLs in open chromatin regions relative to SNP arrays, indicating WGS eQTLs are more likely to identify the exact causal variant. Realistic simulations based on the eQTLs show that around 44% of the sequence eQTLs identify the causal variant as the lead SNP, this number is stable for differing sample sizes.Building on this, we developed the CaVEMaN method, using resampling methods to estimate the probability an eQTL lead SNP is causal. We find 1,668 high confidence causal variants (probability > 0.8). Between 20.3% and 42.9% of these act in at least one other tissue. We validate our work using open chromatin experiments in relevant tissue types from the Epigenomics Roadmap and observed a linear increasing relationship between the causal probability and probability of falling in open chromatin. We predict the proportion of truly causal variants in open chromatin regions to vary from 0.25 to 0.76, depending on the experiment used to call the region. Integrating eQTLs with GWAS signals using our RTC method (Nica et al., PLoS Genet 2010), we further propose causal variants for GWAS signals we infer to be mediated by eQTLs.In conclusion, we find that despite complicated outbred LD structure and high phenotype variability, it is possible to resolve causal variants in expression studies and beyond. Our results also place an upper bound on how informative CHiP-seq experiments from a single individual are for resolving regulatory variation. We will also present results on using our analysis to assess how informative both regulatory annotations and eQTL found in accessible tissues are for discovering causal variants.

88

A new method for genetic region association testing with massively different sequencing depths of coverage. *A.E. Hendricks*[1,2,3,4], *S. Billups*[1], *E. Zeggini*[4], *I. Barroso*[4,5], *S.A. Santorico*[1,2,3], *J. Dupuis*[6]. 1) Mathematical and Statistical Sciences, University of Colorado-Denver, CO, USA; 2) Human Medical Genetics and Genomics Program, University of Colorado-Denver, CO, USA; 3) Biostatistics and Informatics, Colorado School of Public Health, CO, USA; 4) Wellcome Trust Sanger Institute, Cambridge, UK; 5) University of Cambridge Metabolic Research Laboratories and NIHR Cambridge Biomedical Research Centre, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK; 6) Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA.

Large sample sizes are needed to achieve adequate power in case-control studies of rare variants. To improve power, one might use increasingly available whole genome sequenced population controls. However, differences in sequencing depth of coverage between cases and controls cause severe bias when traditional case-control statistics are used. An alternative is to use case-only analysis, which is not susceptible to case-control bias, where the rate of variants within a gene region of interest is compared to the genome-wide average. Case-only analysis can achieve higher power than case-control analysis but is highly sensitive to regional differences in the frequency of variants that are not related to case-status (e.g. mutation rate, annotation accuracy, sequencing accessibility, etc.). To address this, we have developed a simple, computationally efficient genetic region test statistic that combines case-only and case-control tests enabling analysis in samples with extremely different sequencing depths of coverage (e.g., high sequencing depth of coverage cases >50x and low sequencing depth of coverage controls <10x). We present results from a wide-variety of simulations and application to a real dataset consisting of 926 whole-exome sequenced cases with high depth of coverage (~80x) and 3,621 whole-genome sequenced controls with low depth of coverage (~7x) from the UK10K project. Our method achieves equivalent power to existing case-control and case-only methods while maintaining appropriate type I error in the context of region-level and/or case-control biases. Specifically, when we simulate the cases to have 30% more rare variants detected compared to controls (a similar bias to what we would expect for high vs. low sequencing depth of coverage), we find the type I error increases to > 20% for traditional case-control tests (e.g. logistic regression, Fisher’s Exact Test, Likelihood Ratio Test) while our method maintains the expected type I error of 5%. As this method can combine datasets with drastically different sequencing depths of coverage, there is the potential to greatly increase the sample size and, in turn, the power to detect association to a genetic region of interest.

89

A novel rare-variant nonparametric linkage method for analysis of complex familial diseases using whole genome and exome sequence data.
L.H. Zhao, Z. He, B. Li, G.T. Wang, S.M. Leal. Baylor College of Medicine, Houston, TX.

Traditionally nonparametric linkage (NPL) methods were used to analyze families segregating complex traits with unknown mode of inheritance. One drawback of NPL methods when applied to common variant data is that the causal loci mapped to large regions often spanning >50MB making gene identification impossible. With the advent of whole genome and exome sequencing it is now possible to cost effectively generate sequence data for families, but there are limited family-based methods to analyze rare-variants (RVs). Therefore we developed a RV-NPL method that was motivated by population-based aggregate RV association methods, since these methods are considerably more powerful than analyzing individual SNVs. Using family data to perform complex trait RV-NPL has some clear advantages over population- and family-based RV association methods: 1) increased power due to causal variants that with familial aggregation have higher odds ratios than those in the general population; 2) only necessary to analyze affected individuals which increases power 3) robust to population substructure; 4) inclusion of non-causal variant does not increase type II errors 5) can readily be applied to whole genome sequence (WGS) data since recombination events are used as the boundaries of regions to aggregate; and 6) if a sufficient number of families are analyzed mapped regions are well defined, e.g. gene. In order to apply the RV-NPL, after phasing of pedigree data, regional markers are generated using the collapsed haplotype pattern method. Regional markers are analyzed in nuclear and extended families with or without missing data by applying the RV-NPL that uses several different statistics with p-values obtained empirically. We evaluated NPL significance levels suggested by Landers & Kruglyak (1995) and found they are not sufficiently stringent to control type I errors for either WGS or exome data. We therefore established new significance thresholds. The RV-NPL is robust to allelic heterogeneity and more powerful than analyzing SNVs, e.g., for extended-pedigrees we could detect an association with 91% of genes in the exome (α=2.5 x 10^{-6}) using RV-NPL compared to 49% of genes for the traditional NPL. The RV-NPL was applied to WGS data from 112 Alzheimer disease pedigrees with 480 cases. Results from this analysis as well as extensive simulation studies will be used to demonstrate the power of the RV-NPL to identify genes involved in the etiology of complex familial diseases.

90

SeqSpark: A complete analysis tool for large-scale rare variant association studies using whole genome and exome sequence data.
D. Zhang[1], B. Li[1], Z. He[1], G.T. Wang[2], S.M. Leal[1]. 1) Center of Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Department of Human Genetics and Statistics, University of Chicago, Chicago, IL.

Massively parallel sequencing technologies provide great opportunities for discovering rare susceptibility variants involved in complex disease etiology via large-scale imputation, exome and whole genome sequence (WGS) based association studies. Power analyses demonstrate that large sample sizes of tens or even hundreds of thousands of individuals, are required for adequately powered studies. Current analytical tools such as R, PSEQ and Variant Association Tools are unfortunately obsolete when it comes to handling large datasets. To address these issues we developed SeqSpark, a new analysis tool for quality control (QC) and rare variant association analysis. Powered by Apache Spark, a distributive data processing engine, we built an ultra fast data quality control pipeline for genotype data based on quality matrices and variant and sample level statistics, e.g. allele specific read depth, genotype quality score, variant missing rate, transition transversion ratio, global ancestry inference, batch effects, etc. Before analysis variants are fully annotated including prediction of functionality. To facilitate accessing and processing both common, rare and imputed variants, we designed an adaptive data structure which stores the after-QC genotype data in a dense or sparse vector. We implemented single variant as well as popular rare variant association tests in a regression framework, e.g. Combined Multivariate and Collapsing (CMC), Burden, Variable Threshold (VT), Sequence Kernel Association Test (SKAT) and SKAT-O, which can now be performed on large sample size dataset, due to the distributive system and sparse data structure for rare variants. For permutation based p-values, we designed an adaptive framework that can evenly split the computation load across processors. We also implemented Raremetal and Raremetalworker, a popular summary statistics based meta-analysis framework for rare variant association tests. SeqSpark is ideal to use for the analysis of large scale genetic epidemiological studies, where current tools fail because of obsolete low efficient database or cumbersome data structure. SeqSpark is the first tool that can easily handle tens of thousands of samples which are required for well powered association studies to discover susceptibility genes with modest effect sizes. The speed and capabilities of SeqSpark will be demonstrated using several large scale WGS data sets as well data imputed using the haplotype reference consortium.

91

FastSKAT: Sequence kernel association tests for large sets of markers and applications for analyzing LDL cholesterol in whole-genome sequencing data. *K.M. Rice[1], J.A. Brody[2], G.M. Peloso[3], L.A. Cupples[3], T. Lumley[4], CHARGE Lipids Working Group.* 1) Dept of Biostatistics, University of Washington, Seattle, WA, USA; 2) Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA; 3) Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA; 4) Department of Statistics, University of Auckland, Auckland, NZ.

Introduction: The Sequence Kernel Association Test (SKAT) is widely used to test for associations between a phenotype and a set of variants. Computing p-values for SKAT requires the eigenvalues of the genotype covariance matrix, or a similar matrix of equal size – an n x n matrix, where n is the number of subjects or variants, whichever is lower. Extracting the full set of eigenvalues has computational complexity proportional to n^3, and currently limits the use of SKAT. To overcome this, we propose fastSKAT, a new computationally-efficient but accurate approximation, in which only the k largest eigenvalues for SKAT are extracted and a remainder term is evaluated using a Satterthwaite approach. For sample sizes seen in current sequencing studies, these innovations make SKAT tests feasible with at least an order of magnitude more variants than current approaches. **Methods**: We applied fastSKAT in analyses of LDL cholesterol using 4,767 whole genomes. To illustrate fastSKAT's validity, we compared its output with standard SKAT tests for regions of typical size (transcript +- 50Kb). To show how fastSKAT permits analysis of much larger regions than SKAT, we also aggregated by topologically associated domains (TADs, typically 1Mb wide) across the genome that mark regions of higher order chromatin interaction. Finally, we used fastSKAT across each chromosome, to examine the relative contribution of variants that fall within regulatory marks of six histones annotated in adult liver and within 500Kb of known lipid loci. Random sets of the same number of SNPs drawn from the same region were tested for comparison. **Results**: In the transcript +- 50Kb analysis (average 1500 SNPs per test), fastSKAT gave almost identical p-values (Correlation> .999). Using fastSKAT tests among TADs, the top signal had p=1.8E-4, and came from a TAD region on chr19 between 45.0Mb and 45.8Mb (hg19) containing the APOE lipid locus. Running fastSKAT for the TAD regions (average 20,000 SNPs per test) is approximately 2400 times faster than SKAT. SNPs aggregated across a set of histone marks (average 11,350 SNPs per test) were strongly associated with LDL. The strongest chromosome-wide association was on chr19 for H3K36me3 (p= 2.3E-05) while the random set of SNPs from the same region was associated at 0.04. **Conclusion**: fastSKAT quickly and accurately implements SKAT analyses for large numbers of markers. Used with sequence data, it will help address questions that were previously intractable.

92

Human evolutionary history has increased the role of rare variants in complex phenotypes. *R. Hernandez[1,2,3], K. Hartman[1], L. Uricchio[4], C. Ye[2], N. Zaitlen[2,5].* 1) BTS, UCSF, San Francisco, CA; 2) Institute for Human Genetics, UCSF, San Francisco, CA; 3) Institute for Quantitative Biosciences, UCSF, San Francisco, CA; 4) Department of Genetics, Stanford University, Stanford, CA; 5) Department of Medicine, UCSF, San Francisco, CA.

Understanding the genetic architecture of complex traits is a central challenge in human genetics. There currently exists a large disparity between heritability estimates from family-based studies and large-scale genome-wide association studies (GWAS), which has been sensationalized as the "missing heritability problem". Among the possible explanations for this disparity are rare variants of large effect that are neither tagged by existing genotyping platforms, nor well imputed from existing reference panels. However, recent population genetic models suggest that the conditions under which rare variants are expected to substantially contribute to heritability may be fairly limited. We have extended existing models of complex traits to incorporate a wider range of plausible evolutionary features, and provide further insights into the role that rare variants play in shaping complex traits. We use these models to investigate the genetic architecture of gene expression levels across European and African individuals using RNA and whole genome sequencing data from the GEUVADIS and 1000 Genomes Projects. In particular, we investigate whether rare variants are likely to be a source of missing heritability in expression across genes. We pioneered a technique for partitioning heritability estimates across allele frequencies using Haseman-Elston (HE) regression. We find that rare variants (MAF £ 1%) contribute significantly more heritability than common variants (MAF > 5%) across most genes. This observation suggests that rare variants play a substantial role in the heritability of gene expression patterns, which is inconsistent with neutral evolutionary forces operating on the *cis* regulatory architecture of most genes. We then interrogate multiple large-scale imputed case-control data sets from the to demonstrate that rare variants are also a pervasive factor driving the genetic architecture of several complex diseases. We develop an Approximate Bayesian Computation (ABC) algorithm to infer the evolutionary parameters that can explain these observations, and find a striking relationship between the evolutionary forces that have shaped human genomes and the phenotypic variation we observe.

93

Type 2 diabetes-associated variants disrupt function of SLC16A11, a proton-coupled monocarboxylate transporter.

S. Jacobs, V. Rusu, E. Hoch on behalf of the SIGMA T2D Consortium. Broad Institute, Cambridge, MA.

Type 2 Diabetes (T2D) affects more than 415 million people and is a leading cause of morbidity and mortality worldwide. While T2D is influenced by environmental factors, it is also a highly heritable disorder, with genetic variation contributing to a disparity in T2D prevalence across populations. An example of this disparity is observed within American populations, where the prevalence of diabetes in individuals of Mexican or Latin American descent is approximately twice that of US non-Hispanic whites. Through a genome-wide association study, we recently identified a variant haplotype in *SLC16A11* that explains ~20% of the increased T2D prevalence in Mexico. Here, we delve deeper into the genetic association at *SLC16A11*, using genetic fine-mapping in ~8,000 individuals along with biochemical, molecular, cellular and physiological studies to delineate mechanisms underlying T2D risk at this locus. Through these efforts, we define a reduced set of tightly linked common variants likely to contain the causal allele, and identify a *cis*-eQTL for *SLC16A11* in human liver that is associated with decreased *SLC16A11* expression in risk haplotype carriers. Additionally, we demonstrate that T2D risk-associated coding variants in SLC16A11 attenuate activity by disrupting a key interaction with an ancillary protein, thereby reducing plasma membrane localization. These two independent mechanisms by which T2D-associated coding and non-coding variants impact *SLC16A11* expression levels and subcellular localization implicate perturbation of *SLC16A11* as causal at this locus, and suggest reduced SLC16A11 activity as the T2D-relevant direction-of-effect. To gain insight into how disruption of SLC16A11 function impacts T2D risk, we investigate the activity of this previously uncharacterized transporter and establish that SLC16A11 functions as a H$^+$-coupled monocarboxylate transporter. Further, we show that disruption of *SLC16A11* is accompanied by alterations in metabolic pathways implicated in T2D pathogenesis. Our findings illustrate the path from genetic association to effector transcript at this locus, confirm the molecular function of its gene product, define the mechanism by which genetic variation affects SLC16A11 action, begin to elucidate the metabolic processes impacted by SLC16A11 perturbation, and suggest that increasing SLC16A11 function could be therapeutically beneficial for people with T2D.

94

Increased alpha tryptase copy number at *TPSAB1* is associated with common elevations in basal serum tryptase level and variably expressive syndromic comorbidity.

J.D. Milner[1], X. Yu[1], J.D. Hughes[2], Q.T. Le[3], G.H. Caughey[6], Y. Bai[1], T. Heller[4], M. Zhao[5], Y. Liu[1], M.P. O'Connell[1], N. Trivedi[6], C. Nelson[1], T. DiMaggio[1], H. Matthews[8], K.L. Lewis[9], A.J. Oler[1], R.J. Carlson[1], P.D. Arkwright[10], C. Hong[9], D.D. Metcalfe[1], T.M. Wilson[1], L.B. Schwartz[3], Y. Zhang[11], J.J. McElwee[2], M. Pao[12], S.C. Glover[13], M.E. Rothenberg[7], R.J. Hohman[5], L.G. Biesecker[9], J.J. Lyons[1]. 1) 1Laboratory of Allergic Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD; 2) Merck Research Laboratories, Merck & Co. Inc., Boston, MA; 3) Department of Internal Medicine, Virginia Commonwealth University, Richmond, VA; 4) Liver Diseases Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD; 5) Research Technologies Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD; 6) Cardiovascular Research Institute and Department of Medicine, University of California San Francisco, San Francisco, CA, and Veterans Affairs Medical Center, San Francisco, CA; 7) Division of Allergy and Immunology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 8) Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD; 9) Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 10) University of Manchester, Royal Manchester Children's Hospital, UK; 11) Laboratory of Host Defenses, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD; 12) National Institute of Mental Health, National Institutes of Health, Bethesda, MD; 13) Division of Gastroenterology, Hepatology, and Nutrition, University of Florida, Gainesville, FL.

Background: Elevated basal serum tryptase is present in 4-6% of the general population, and has been associated with functional pain, gastroenterologic (GI), neurologic, cardiovascular (CV) and cutaneous symptoms, as well as hymenoptera allergy. Recently, families with dominantly inherited, symptomatic elevated tryptase were identified. **Methods:** Symptom patterns were measured in individuals and families who did not have mastocytosis but had elevated basal tryptase. Next-generation sequencing (NGS) and linkage analyses were performed. Copy number genotyping was performed using a novel digital droplet PCR assay. Mast cells were cultured from circulating CD34$^+$ cells. Findings in the referral cohort were extended in two independent cohorts from the general population who were in studies unrelated to mast cell pathology. **Results:** Elevated basal serum tryptase was identified as a monogenic trait in 35 families (96 affecteds) referred for mast cell dysfunction or syndromic allergy. Presenting symptoms included flushing and pruritus, dysautonomic complaints, irritable bowel syndrome (IBS), gastroesophageal reflux, chronic pain and difficulty concentrating. Connective tissue abnormalities were also commonly observed, including scoliosis, pectus excavatum, syndactyly, retained primary dentition, and joint hypermobility. Many of the observed symptoms overlapped with Ehlers-Danlos syndrome type III (EDS-III). NGS of 7 families failed to identify segregating rare variants, but linkage analysis identified a single region mapping to the tryptase locus. Duplications and triplications of the α tryptase isoform at TPSAB1 segregated with disease, with triplications leading to higher tryptase levels and worse symptoms. In two unselected cohorts from the general population, alleles with α tryptase duplications segregated completely with increased basal serum tryptase and were associated with symptoms similar to those observed in the initial referral group. **Conclusions**: Common elevations in basal serum tryptase are associated with increased α tryptase copy number. This dominant, dose-dependent trait is characterized by variable expression of a symptom complex including cutaneous, GI, CV, and neurocognitive complaints, as well as congenital bone, tooth, and joint abnormalities. In addition, the substantial functional and connective tissue phenotype suggests this trait may define discrete subsets of patients with IBS and other functional disorders, and EDS-III.

95

Genetic inactivation of ANGPTL4 is associated with improved glycemic control and reduced risk of Type 2 Diabetes. *C. O'Dushlaine[1], V. Gusarova[2], P. Benotti[3], T. Mirshahi[3], O. Gottesman[1], C. Van Hout[1], M. Murray[3], A. Mahajan[4], J. Nielsen[5,6], C. Emdin[7], R. Scott[8], S. Bruse[1], O. Holmen[9], D. Ledbetter[3], J. Reid[1], J. Overton[1], G. Yancopoulos[2], N. Wareham[8,10], S. Kathiresan[11], O. Melander[12], G. Abecasis[13], J. Florez[14,15,16], M. Boehnke[13], M. McCarthy[4,17,18], D. Carey[3], A. Shuldiner[1], I. Borecki[1], A. Baras[1], J. Gromada[2], F. Dewey[1], DiscovEHR Collaboration.* 1) Regeneron Genetics Center, Tarrytown, NY, USA; 2) Regeneron Pharmaceuticals, Tarrytown, NY, USA; 3) Geisinger Health System, Danville, PA, USA; 4) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; 5) Department of Internal Medicine,Division of Cardiovascular Medicine, University of Michigan, University of Michigan, Ann Arbor, Michigan, USA; 6) Department of Human Genetics, University of Michigan, University of Michigan, Ann Arbor, MI, USA; 7) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA; 8) MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Addenbrooke's Hospital, Cambridge, UK; 9) HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway; 10) Centre for Diet and Activity Research (CEDAR), Medical Research Council Epidemiology Unit, University of Cambridge, UK; 11) Center for Human Genetic Research, Cardiovascular Research Center and Cardiology Division, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; 12) Department of Clinical Sciences, Malmö, Lund University, Malmö, Sweden; 13) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA; 14) Diabetes Unit and Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA; 15) Programs in Metabolism and Medical & Population Genetics, Broad Institute, Cambridge, MA, USADepartment of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA; 16) Department of Medicine, Harvard Medical School, Boston, MA, USA; 17) Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK; 18) Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, UK.

Angiopoeitin-like 4 (*ANGPTL4*), an endogenous inhibitor of lipoprotein lipase, modulates lipid levels. Recently, loss of function variants in *ANGPTL4* were demonstrated to reduce coronary artery disease risk. Because the lipoprotein lipase pathway also modulates glucose homeostasis, we hypothesized that loss of *ANGPTL4* function might also improve glycemic control and lower the risk of type 2 diabetes. To investigate this, we studied protein-altering variants in *ANGPTL4* among 49,178 exome-sequenced participants of European ancestry in the Regeneron Genetics Center-Geisinger Health System DiscovEHR human genetics study, with follow-up studies in 62,301 type 2 diabetes cases and 287,865 controls. Carriers of p.E40K, a variant that reduces the ability of *ANGPTL4* to inhibit lipoprotein lipase, had 14% lower odds of type 2 diabetes (combined odds ratio 0.86, 95% CI 0.82-0.91, p=2.0x10^{-9}). *Angptl4*-deficient mice on high-fat diets had 31% lower non-fasted glucose levels, and improved glycemic control and insulin sensitivity as revealed by oral glucose tolerance and insulin tolerance tests. In conclusion, genetic inhibition of *ANGPTL4* was associated with lowered serum glucose in mice and reduced risk of type 2 diabetes in humans, indicating that *ANGPLT4* may be a promising therapeutic target for reduction of metabolic disease risk in humans.

96

Novel long non-coding RNAs, *CUPID1* and *CUPID2*, mediate breast cancer risk at 11q13 by modulating response to DNA damage. *J.D French[1], M. Moradi Marjaneh[1], Y.C. Lim[1], M. Clark[3,2], N. Bartonicek[2], W. Shi[1], T. Mercer[2], K. Khanna[1], M. Dinger[2], F. Al-Ejeh[1], J.A. Betts[1], S.L. Edwards[1].* 1) Genetics and Computational Biology, QIMR Berghofer Medical Resesarch Institute, Brisbane, Queensland (QLD), Australia; 2) Garvan Institute of Medical Research, Sydney, Australia; 3) MRC Functional Genomics Unit, Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford, UK.

One of the strongest breast cancer associations identified to date via GWAS is with SNP rs614367 at the 11q13 locus. We previously fine-mapped this region and showed that the strongest risk-associated SNPs fall within a distal enhancer located deep within an intergenic region on 11q13. Using chromosome conformation capture (3C) and CRISPRi we show that this distal enhancer regulates two novel estrogen regulated lncRNAs we called *CUPID1* and *CUPID2,* identified by RNA CaptureSeq. Allele-specific 3C between the distal enhancer and the *CUPID1/2* promoter showed preferential looping for the protective alleles, suggesting that risk-associated SNPs may abrogate chromatin looping, resulting in reduced promoter activity and subsequent transcription. Consistent with this, we show allelic imbalance of *CUPID1* expression in tumours heterozygous for the strongest risk signal. *CUPID1* and *CUPID2* localized to different cellular compartments indicating they may function through independent mechanisms. *CUPID1* was chromatin bound and ChIRPseq showed it was significantly enriched at enhancer regions. *CUPID1*-regulated genes more frequently had *CUPID1*-bound enhancers nearby, suggesting that at least one mechanism by which *CUPID1* regulates target genes is through modulation of enhancer activity. Finally, we show that *CUPID1 and CUPID2* are predominantly expressed in hormone receptor-positive breast tumors and that reduced levels result in an impaired homologous recombination mediated DNA repair suggesting a novel mechanism for the involvement of this region in breast cancer.

97

Functional dissection of the Zika genome reveals a coding component responsible for microcephaly. *E. Oh, M. Kousi, K. McFadden, C. Howald, E. Dermitzakis, N. Katsanis.* Center for Human Disease Modeling, Duke University, Durham, NC.

Infection by the Zika virus, a re-emerging mosquito-borne flavivirus, has been associated with developmental abnormalities, including microcephaly. To determine whether one or more of the viral proteins encoded by the Zika genome might be the driver of brain pathology, we expressed each structural and non-structural component in zebrafish embryos. Using light microscopy and whole-mount staining, we discovered a non-structural component that was sufficient to give rise to the microcephaly phenotype in 3 days-old embryos. The change in head size could be induced in a dose-dependent manner, and was accompanied by a decrease in phospho-histone H3 immunopositive neural progenitors. RNAseq analyses of brain tissue revealed an alteration in immune-related genes suggesting a new mechanism in which viral components interact with the host. Our studies will elucidate how the Zika coding genome modifies the neural network and will inform novel therapeutic strategies aimed to ameliorate or at least prolong the onset of symptoms.

98

Characterization of a new class of disease-causing variants unresponsive to current CFTR targeted therapies. *S.T. Han, M.J. Pellicore, T.A. Evans, E. Davis, K.S. Raraigh, G.R. Cutting.* Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD.

Cystic fibrosis (CF) provides an example of the primary goal of precision medicine: using the right drug for the right patient. To achieve this goal, we are systematically evaluating *CFTR* mutants for their response to the clinically approved CFTR-targeted drugs Ivacaftor and Lumacaftor. Of particular interest are a cluster of mutants in the 6th transmembrane domain (TM6) that line the Cl^- channel formed by CFTR. These variants are associated with less severe pancreatic exocrine disease which would predict that CFTR bearing these variants would generate mature protein and retain minimal function (~3%-5% of WT CFTR). However, when expressed in Fischer rat thyroid cells, CFTR bearing TM6 variants displayed <2% of WT function and minimal response to Ivacaftor, a ‘potentiator’ drug that activates mature CFTR protein. To determine if CFTR function would differ in a native context, we generated human CF airway epithelial (CF8) cells expressing a single integrated *CFTR* cDNA containing each of 4 TM6 variants: R334W, I336K, T338I, or R347P. CFTR protein quantity, folding, and processing were evaluated by western blot while Cl^- channel function and drug response were evaluated by measuring short circuit currents(I_{sc}). Each TM6 mutant was processed to mature form and each demonstrated residual CFTR function ranging from 2%-8% of WT, now consistent with clinical presentation; however, these mutants still had minimal (<2% of WT) response to Ivacaftor. Unexpectedly, the mutants responded differently to Lumacaftor, a drug that assists processing of CFTR. I336K exposed to Lumacaftor showed significant increases in protein levels, processing, function, and response to Ivacaftor resulting in ~60-80% of WT function, which is well into the predicted therapeutic range. Conversely, R334W, T338I, and R347P showed no response across multiple independently derived cell lines. Inspection of the predicted structural location of each variant revealed that amino acid R groups at positions 334, 338, and 347 directly influence Cl^- flow while 336 does not. These results indicate that residues directly involved in Cl^- movement (~2% of CF alleles) comprise a unique ‘theratype’ that requires a different strategy for molecular correction. The marked difference in response of these TM6 variants to small molecules illustrate that a deep understanding of the mechanisms underlying each loss of function variant will be required to properly match individual patients to the appropriate treatment.

99

Quantifying constraint on regulatory variation across 48 human tissues improves interpretation of functional variation. *T. Lappalainen[1,2], P. Mohammadi[1,2], S.E. Castel[1,2], H.E. Wheeler[3], H.K. Im[4], GTEx Consortium.* 1) New York Genome Center, New York, NY; 2) Department of Systems Biology, Columbia University, New York, NY; 3) Departments of Biology and Computer Science, Loyola University Chicago, Chicago, IL; 4) Department of Medicine, University of Chicago, Chicago, IL.

Several approaches have been presented to assess the constraint of genic intolerance to coding variation. This study provides much needed insight into regulatory constraint of genes, which is important for interpretation of regulatory variation discovered in growing whole genome sequencing data sets in the context of disease predisposition. Allelic expression (AE) analysis captures the effect of *cis*-regulatory variants and it is minimally obscured by other confounding factors, as we have shown previously (Lappalainen et al. 2013 Nature, GTEx Consortium 2015 Science). Here we present a probabilistic model to describe AE data as net outcome of a set of unobserved regulatory variants. We use this model to estimate the variation in gene expression introduced by *cis*-regulatory genetic variation in population data. Our simulations demonstrate that the model is robust and accurate across a spectrum of allele frequencies, regulatory complexity, and gene expression levels. We applied this method to AE data from 499 individuals in the GTEx data. We quantified the effect of *cis*-regulatory variation and contrasted it to total variation of expression in 13,399 coding genes in 48 tissues. We used these measures to estimate gene expression heritability, with estimates that are highly correlated to those from the much larger Depression Genes and Networks data (*rho=0.49*). We show high concordance of intolerance to *cis*-regulatory variation with coding region constraint (RVIS; $p<10^{-170}$) and better discrimination of haploinsufficient genes than previous estimates of regulatory constraint. Importantly, coding and regulatory constraint complement each other, with some loss-of-function depleted genes showing constraint only for regulatory and not for coding variation. Genes expressed in the brain show higher regulatory constraint, and gene groups with different cellular functions show distinct patterns of genetic versus total regulatory variation. Finally, we show that genes associated to autism are depleted of genetic but not total gene expression variation ($q<10^{-3}$). We conclude that our method can accurately quantify tolerated *cis*-regulatory variation from AE data. The obtained scores for tissue-specific intolerance of regulatory variation are valuable complement to coding variation constraint scores for interpretation of genetic variation in the noncoding genome.

100

Systematic prediction of conserved non-exonic bases from large-scale functional genomics data. *J. Ernst, O. Grujic, Y. Lee, A. Sperlea.* University of California, Los Angeles, CA.

A large majority of genome-wide association study hits fall into non-exonic regions of the genome and it can be a challenge to determine likely causal variants among multiple ones in linkage disequilibrium. Large scale consortium projects as well as the collective effort of many individual labs have produced thousands of genome-wide experiments on regions of open chromatin, locations of transcription factor binding and histone modifications, which can be a resource to prioritize important regulatory locations in the genome. However with the large number of data sets available, large portions of the genome showing signal, and in many applications the relevant cell types uncertain, it is often unclear how to prioritize genomic locations based on the data. Here we produce a single track that prioritizes likely regulatory important bases in the human genome based on integrating more than 10,000 features of locations of histone modifications, transcription factor binding, open chromatin, and chromatin state calls from ChromHMM across many different cell and tissue types derived from data from the NIH Roadmap Epigenomics or ENCODE projects, or curated from individual labs. Our method integrates the functional genomics features using a supervised machine learning method that is trained to predict locations of non-exonic evolutionary conserved elements identified based on multi-species sequence alignments. We applied the predictors genome-wide to make a probabilistic prediction for each individual nucleotide as to whether it would fall into a non-exonic conserved region based on the functional genomics data. We were able to obtain relatively effective predictions of the conserved regions (AUC = ~0.82). Analysis of the false negative predictions of our method and the genes they are proximal to identified cell types and classes of genes that are not adequately captured in current functional genomic data sets suggesting potential additional cell and tissue types for experimental mapping. Our false positive predictions suggest potential important recently evolved regulatory locations in the genome. Heritability analysis of complex traits such as BMI show that locations with higher prediction scores strongly tracked with locations that explained an increasing fraction of disease heritability suggesting our predictions have the potential to be an important resource for interpreting and prioritizing disease associated variants.

101

Identifying highly constrained protein-coding regions using population-scale genetic variation.

J. Havrilla, B. Pedersen, R. Layer, A. Quinlan. Department of Human Genetics, University of Utah, Salt Lake City, UT.

There is a longstanding interest in uncovering critical genes by measuring genomic conservation between species as well as constraint within a species. Numerous screens have been performed to discover essential genes, typically using induced mutations to knock out genes and quantify their fitness effects. More recent computational approaches have been developed to infer functional constraint across entire transcripts based on a dearth of missense variation (Petrovski et al, 2013; ExAC, under review). While gene-wide predictions of constraint are valuable, there is clearly variability in the degree of constraint within a gene, depending on the specific function and structural properties in the resulting protein region. Therefore, we have devised a linear model to identify significantly constrained regions across the entire protein coding exome by leveraging the deep resource of genetic variation observed among 60,706 exomes from ExAC. We model the expected exonic distance between two missense variants based on CpG density. Constrained coding regions (CCRs) arise when the observed distance between missense variants in ExAC — a proxy for constraint — is much greater than predicted based on many other regions with similar CpG content. We validated the predictive power of our model by demonstrating that ClinVar pathogenic variants were significantly (p<0.001) enriched in CCRs via Monte Carlo simulations. Similarly, ClinVar benign variants are significantly depleted (p<0.001) in these regions. Moreover, CCRs are found in many genes known to be associated with severe phenotypes (e.g., *LMNA*), important cellular function (e.g., *MTOR*), and Mendelian disorders (e.g., *RBP7*). We will present our investigation of CCRs in genic regions whose functions were previously unknown, yet the observed constraint suggests function. For example, we observe many CCRs that exhibit low inter-species conservation yet high constraint in humans. We will demonstrate our validation of these putatively functional CCRs by assessing whether the underlying genes exhibited a significant functional effect among the many extensive CRISPR-Cas9 mutagenesis screens conducted to date. We will further compare CCR-containing genes to protein-protein interaction databases to infer biological function based upon their interaction with proteins of known function. We will present our set of high confidence CCRs of previously unknown biological function and their potential consequences for human disease.

102

Regional analysis of variation tolerance improves variant deleteriousness prediction.

K.E. Samocha[1,2,3], J.A. Kosmicki[1,2,3], K.J. Karczewski[1,2], A.H. O'Donnell-Luria[1,2], E.V. Minikel[1,2,3], M. Lek[1,2], D.G. MacArthur[1,2], B.M. Neale[1,2], M.J. Daly[1,2,3], Exome Aggregation Consortium. 1) Massachusetts General Hospital, Boston, MA; 2) Broad Institute of Harvard and MIT, Cambridge, MA; 3) Harvard Medical School, Boston, MA.

The availability of large-scale exome sequencing datasets has permitted the identification of genomic regions that are intolerant of nonsynonymous variation (constrained). We have previously described methods to identify genes that are severely depleted of missense and/or loss-of-function variation using the exome sequencing data from 60,706 individuals from the Exome Aggregation Consortium (ExAC; Lek et al 2016) and found that these constrained genes were enriched for causes of known Mendelian diseases. Given that missense variants can have dramatically different effects depending on their location in the gene, we developed a method to locate regions within genes that are specifically intolerant of missense variation. We observed that the ~15% of the protein-coding genome severely depleted of missense variation in ExAC is (1) significantly enriched for established ClinVar pathogenic variants and (2) has a much higher rate of *de novo* missense variation in cases with a neurodevelopmental disorder than the rate seen in controls. Since these missense constrained regions are depleted of variation due to selective pressure, we proposed that including information about the local missense depletion could improve variant deleteriousness metrics. We first created a measure of the increased deleteriousness of amino acid substitutions when they occur in missense constrained genes and regions, which outperformed similar amino acid substitution matrices (BLOSUM and Grantham) at separating pathogenic from benign variants. The best predictor of variant deleteriousness, however, was the combination of regional missense constraint, the amino acid substitution score we developed, and PolyPhen-2. The joint metric (MPC score) shows a dramatic ability to identify the key subset of *de novo* missense variants involved in disease from background in cases with a neurodevelopmental disorder. For MPC ≥ 2, the rate of *de novo* missense variants is 0.11 per case exome and 0.01 per control exome, giving a higher rate ratio than that found for protein-truncating variants (~7).

De novo missense variants per exome by MPC score

	MPC < 1.5	1.5 ≤ MPC < 2	MPC ≥ 2
Neurodevelopmental case	0.63	0.13	0.11
Control	0.65	0.05	0.01

Incorporating information about regional missense constraint has allowed us to take a class of variants with only a modest enrichment (~1.2) and identify a small subset with a far greater enrichment of true disease association.

103

Annotation of the human genome through conservation states aids interpretation of disease associated genetic variation. *A. Sperlea, J. Ernst.* University of California Los Angeles, Los Angeles, CA.

Genome-wide association studies have identified a large number of non-coding genomic loci in the human genome associated with disease, whose biological significance is poorly understood. Additional annotations largely based on either functional genomics or comparative data have been used to gain insights into such locations and potentially prioritize likely causal variants among those in linkage disequilibrium. A widely used representation of the functional genomics data is through chromatin states produced by methods such as ChromHMM, which provides cell type specific annotations based on the combinatorial and spatial patterns in epigenomic data. Comparative genomic data provides complementary information as it is not dependent on having data from the appropriate cell or tissue type and can provide single nucleotide resolution information. Recent analyses have suggested conserved elements are among the genomic annotations most enriched for disease heritability. However the currently widely used representations of conservation information focus on either binary calls or a single univariate score from phylogenetic models, and thus do not capture potentially valuable information contained in the multi-species alignments of an increasing number of available species. Here we take a novel approach to annotating the human genome based on comparative genomic sequence information by applying an extended version of ChromHMM. We segment the human genome at single nucleotide resolution into a large number of different conservation states based on the combinatorial patterns of which species align to and which match the human reference genome within a 100-way multi-species alignment. The various conservation states show distinct enrichment properties for other genomic annotations such as regions of open chromatin, CpG islands, transcription start sites, exons and existing conserved element calls. Using our approach we are able to isolate approximately 7% of bases within a widely used set of conserved element calls that are likely false positives due to alignment artifacts. We partitioned the heritability of traits such as BMI and showed strong and distinct enrichments for different conservation states for trait-associated variants. The segmentation of the human genome we have produced based on comparative genomic data has the potential to be a widely used resource to interpret and prioritize among potential disease-associated variants.

104

Assessment of genetic burden and constraint in hypertrophic cardiomyopathy genes leveraging protein structural data and large sequencing cohorts. *J.R. Homburger[1], E.M. Green[2], C. Caleshu[3], M.S. Sunitha[4], R.E. Taylor[5], K.M. Ruppel[5,6], R. Metpally[7], S.D. Colan[8], M. Michels[9], S.M. Day[10], I. Olivotto[11], C.D. Bustamante[12], F. Dewey[13], C.Y. Ho[14], J.A. Spudich[4,5], E.A. Ashley[1,3].* 1) Genetics, Stanford University, Stanford, CA; 2) MyoKardia, Inc. South San Francisco, CA, USA; 3) Stanford Center for Inherited Cardiovascular Disease, Stanford University, Stanford, CA, USA; 4) Institute for Stem Cell Biology and Regenerative Medicine, Bangalore, India; 5) Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, USA; 6) Department of Pediatrics (Cardiology), Stanford University School of Medicine, Stanford, CA, USA; 7) Geisinger Health System, Danville, PA, USA; 8) Department of Cardiology, Boston Children's Hospital, Boston, MA, USA; 9) Department of Cardiology, Erasmus MC, Rotterdam, the Netherlands; 10) Cardiovascular Division, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA; 11) Referral Center for Cardiomyopathies, Careggi University Hospital, Florence, Italy; 12) Department of Biomedical Data Sciences, Stanford University, Stanford, CA, USA; 13) Regeneron Inc. Tarrytown, NY, USA; 14) Brigham and Women's Hospital, Boston, MA, USA.

Variants in sarcomere genes can lead to hypertrophic cardiomyopathy (HCM), a heritable disease characterized by cardiac hypertrophy, heart failure, and sudden cardiac death. A key debate is whether there exist hotspots of pathogenic variation within sarcomere genes, especially in β-cardiac myosin (*MYH7*). Furthermore, how specific sarcomere variants alter motor function or disease expression remains incompletely understood. To address these questions, we outline a novel statistical approach based upon a spatial scan statistic to identify three dimensional and surface regions of proteins enriched for disease variation. We combine structural models of myosin from multiple stages of its chemomechanical cycle, exome sequencing data from two population cohorts of 60,706 and 42,930 individuals, and genetic and phenotypic data from 2,913 HCM patients to identify regions of disease-variant enrichment within β-cardiac myosin and other sarcomere genes implicated in HCM. We show a significant enrichment of disease-associated variants in the converter (p=0.002), a kinetic domain that transduces force from the catalytic domain to the lever arm during the power stroke. Focusing our analysis on surface-exposed residues, we identified a larger region significantly enriched for disease-associated variants that contains both the converter domain and residues on a single flat surface on the myosin head (p=0.002) that is theorized to be involved in protein interactions in the sarcomere. Regions enriched for disease-associated variation have higher levels of conservation in primates. Patients with variants in these enriched regions have an earlier age at diagnosis and worse clinical phenotypes. We expand our analysis to other genes implicated in HCM, including thin filament genes such as the myosin light chains, *TNNI3*, and *TNNT2*. We demonstrate that combining sequencing with protein structures can help identify functional domains important in disease etiology, especially in Mendelian diseases such as cardiomyopathies that are caused by a set of rare, highly penetrant variants. Our methods provide a model for integrating protein structure, large-scale genetic sequencing and detailed phenotypic data to reveal insight into genetic burden and disease.

105

Recessive and dominant de novo *ITPR1* mutations cause Gillespie Syndrome.

L. Fares Taie[1], S. Gerber[1], K. Alzayady[2], L. Burglen[3], D. Bremond-Gignac[5], V. Marchesin[4], O. Roche[5], M. Rio[6], B. Funalot[7], R. Calmon[8], A. Durr[9], V. Gil-da-Silva-Lopes[10], M. Ribeiro Bittar[10], C. Orssaud[5], B. Heron[11], E. Ayoub[2], P. Berquin[12], N. Bahi-Buisson[13], C. Bole[14], C. Masson[15], A. Munnich[6], M. Simons[4], M. Delous[16], H. Dollfus[17], N. Boddaert[8], S. Lyonnet[6], J. Kaplan[1], P. Calvas[18], D. Yule[2], J.M. Rozet[1]. 1) Laboratory of Genetics in Ophthalmology (LGO). INSERM UMR1163. Imagine – Institute of Genetic Diseases, Paris Descartes University, 75015 Paris, France; 2) Department of Pharmacology and Physiology, University of Rochester, Rochester, NY 14526, USA; 3) Reference Center for cerebellar malformations and congenital diseases, Department of Genetics, Trousseau Hospital, AP-HP, F-75012 Paris – Inserm U1141, DHU PROTECT, Robert Debré Hospital, 75019 Paris, France; 4) Epithelial biology and disease – Liliane Bettencourt Chair of Developmental Biology. INSERM UMR1163. Imagine – Institute of Genetic Diseases, Paris Descartes University, 75015 Paris, France; 5) Department of Ophthalmology. IHU Necker-Enfants Malades, University Paris-Descartes, 75015 Paris, France; 6) Department of Genetics. IHU Necker-Enfants Malades, University Paris-Descartes, 75015 Paris, France; 7) Department of Genetics. GHU Henri Mondor, 94010 Créteil, France; 8) Department of Neuroradiology. IHU Necker-Enfants Malades, University Paris-Descartes, 75015 Paris, France; 9) Maladies neurodégénératives. Institut du Cerveau et de la Moëlle épinière. CHU Paris-GH La Pitié Salpêtrière-Charles Foix - Hôpital Pitié-Salpêtrière, 75013 Paris, France; 10) Departamento de Genetica Medica, Faculdade de Ciencias, UNICAMP, CEP 13083-887 – Campinas, SP, Brasil; 11) Department of Neuropediatrics, Hôpital Trousseau, 75012 Paris, France and Department of Pediatrics, Hôpital Jean Verdier, 93143 Bondy, France; 12) Poˆle de pe´diatrie, Centre d'activite´ de neurologie pe´diatrique, CHU d'Amiens - Hoˆpital Nord, 80054 Amiens, Cedex 1, France; 13) Embryology and genetics of human malformation. INSERM UMR1163. Imagine – Institute of Genetic Diseases, Paris Descartes University, 75015 Paris, France; 14) Genomics Platform. INSERM UMR1163. Imagine – Institute of Genetic Diseases, Paris Descartes University, 75015 Paris, France; 15) Bioinformatics Platform. INSERM UMR1163. Imagine – Institute of Genetic Diseases, Paris Descartes University, 75015 Paris, France; 16) Laboratory of Hereditary Kidney Diseases, INSERM UMR1163, Imagine – Institute of Genetic Diseases, Paris Descartes University, 75015 Paris, France; 17) Laboratoire de Ge´ne´tique Me´dicale, Institut de Ge´ne´tique Me´dicale d'Alsace, INSERM U1112, Fe´de´ration de Me´decine Translationnelle de Strasbourg (FMTS), Universite de Strasbourg, 67085 Strasbourg, France; 18) Service de Génétique Clinique, Hôpital Purpan, 31300 Toulouse, France.

Gillespie syndrome (GS) is a rare variant form of aniridia characterized by non-progressive cerebellar ataxia, intellectual disability, and iris hypoplasia. Unlike the more common dominant and sporadic forms of aniridia, there has been no significant association with PAX6 mutations in individuals with GS and the mode of inheritance of the disease had long been regarded as uncertain. Using a combination of trio based whole-exome sequencing and Sanger sequencing in five simplex GS-affected families, we found homozygous or compound heterozygous truncating mutations (c.4672C>T [p.Gln1558*], c.2182C>T [p.Arg728*], c.6366þ3A>T [p.Gly2102Valfs5*], and c.6664þ5G>T [p.Ala2221Valfs23*]) and de novo heterozygous mutations (c.7687_7689del [p.Lys2563del] and c.7659T>G [p.Phe2553Leu]) in the inositol 1,4,5-trisphosphate receptor type 1 gene (*ITPR1*). *ITPR1* encodes one of the three members of the IP3-receptors family that form Ca2+ release channels localized predominantly in membranes of endoplasmic reticulum Ca2+ stores. The truncation mutants, which encompass the IP3-binding domain and varying lengths of the modulatory domain, did not form functional channels when produced in a heterologous cell system. Furthermore, ITPR1 p.Lys2563del mutant did not form IP3-induced Ca2+ channels but exerted a negative effect when co-produced with wild-type ITPR1 channel activity. In total, these results demonstrate biallelic and monoallelic *ITPR1* mutations as the underlying genetic defects for Gillespie syndrome, further extending the spectrum of *ITPR1*-related diseases.

106

Photoreceptors restore retinal function and regenerate healthy outer segments after reconstitution of the BBSome in a mouse model of Bardet-Biedl Syndrome.

Y. Hsu[1], J. Garrison[1], G. Kim[1], D. Nishimura[1], C. Searby[1], A. Schmitz[1], P. Datta[2], S. Seo[2], V. Sheffield[1]. 1) Department of Pediatrics, University of Iowa, Iowa City, IA; 2) Ophthalmology and Visual Sciences, University of Iowa, Iowa City, IA.

Photoreceptor outer segments are specialized forms of cilia, which are renewed every 10 days in mice. The BBSome, a protein complex consisting of BBS1, 2, 4, 5, 7, 8, 9, and 18, is required for normal retinal function. Knockout mouse models of BBSome components have retinal degeneration. However, the role of the BBSome in the development and maturation of the retina is not known. Using a mouse model that enables the tamoxifen-inducible deletion of BBS8, a core BBSome protein, we showed that the deletion of *Bbs8* (Bardet Biedl Syndrome 8 [MIM 615985], also known as *Ttc8*, tetratricopeptide repeat domain-containing protein 8 [MIM 608132]), at four different time points causes a decline in retinal function as measured by electroretinogram, shortening of the outer segments, and eventually loss of photoreceptor cells as observed by histology and electron microscopy. In addition, congenital *Bbs8* knockout mice have misoriented and malformed outer segments prior to notable photoreceptor degeneration. These data suggest that the formation of photoreceptor outer segments requires the BBSome. In developing treatments for blindness in Bardet Biedl Syndrome, it is important to know whether photoreceptors that develop without BBSome function can regenerate healthy outer segments and regain retinal function upon restoration of the BBSome in the patient. To address this, we developed a mouse model where *Bbs8* gene expression is suppressed by a gene trap, and the gene trap can be excised at any point in time by tamoxifen inducible FLP recombinase. In *Bbs8*-gene trapped mice, outer segments are malformed and retinal function is decreased, similar to that of congenital *Bbs8* knockout mice. We performed time-dependent rescue of *Bbs8* either during the critical window of disc morphogenesis between postnatal day 9 to 15 or in adult mice to determine whether photoreceptors can regenerate discs after *Bbs8* re-expression. We found that after the excision of the gene trap, and, therefore, reconstitution of the BBSome, healthy outer segments can be regenerated regardless of the time window of tamoxifen administration, even after retinal development is complete. We conclude that the loss of the BBSome disrupts the normal generation and potentially renewal of outer segments in photoreceptors and that this defect is reversible by the reconstitution of the BBSome. This finding has important implications for gene therapy design treating blindness.

107

Mutations in spliceosome-associated protein homolog *CWC27* cause autosomal recessive syndromic retinitis pigmentosa.

M. Xu[1], Y.A. Xie[2], H. Abouzeid[3], D. Babino[4], Z. Sun[5], A. Eblimit[1], I.S. Othman[6], A. Lehman[7], R. Pfundt[8], A. Fiorentino[9], R. Riveiro[10], J. von Lintig[4], V. Kheir[3], G. Pinton[3], N. Allaman-Pillet[3], R. Dharmat[1], S.A. Agrawal[1], Y. Li[1], A. Hardcastle[9], M.A. Lopez[10], H. Li[5], M. Cheetham[9], C. Ayuso[10], R. Chen[1], R. Sui[5], R. Allikmets[2], D.F. Schorderet[3]. 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Department of Ophthalmology, Columbia University, New York, NY, USA; 3) Institute for Research in Ophthalmology, Sion, Switzerland; 4) Department of Pharmacology, Case Western Reserve University School of Medicine, Cleveland, OH, USA; 5) Department of Ophthalmology, Peking Union Medical College, Beijing, China; 6) Rod El Farag Institute, Cairo, Egypt; 7) Department of Medical Genetics, The University of British Columbia, Vancouver, BC, Canada; 8) Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands; 9) Institute of Ophthalmology, University College London, London, UK; 10) Department of Genetics, Instituto de Investigacion Sanitaria-University Hospital Fundacion Jimenez Diaz (IIS-FJD), Madrid, Spain.

Retinitis pigmentosa (RP) is a rare inherited retinal disorder affecting 1 in 3,000 of the population. It may occur in isolated or syndromic forms. The genetic etiology of RP is remarkably heterogeneous, with over 100 disease-causing genes identified. However, the genetic cause of 40% RP cases are still not fully understood, suggesting additional disease-causing genes remain to be discovered. Here, we identified recessive protein-truncating mutations (NM_005869.3: Q7*, R143*, L167Gfs*3, E315*, V335Sfs*13) in the spliceosome-associated gene *CWC27* in six unrelated families with a phenotype spectrum from isolated to syndromic RP. The patients present retinal degeneration, with or without short stature, craniofacial defects, intellectual disability as well as other systematic abnormalities. *Cwc27* complete knockout mice show significant pre-weaning lethality, and the viable mice present growth retardation and retinal abnormalities, recapitulating the syndromic RP phenotype. CRISPR-Cas system generated another *Cwc27* mutant mouse line with a frameshift mutation near the C-terminal. These mice show no lethality with only retinal degeneration, mimicking the isolated RP phenotype. Our finding established a novel disease-causing gene *CWC27* for an RP-related Mendelian phenotype spectrum. This may further help us reveal the role of spliceosome factors in retinal function maintenance and global development process.

108

Patients with Blepharo-Cheilo-Dontic syndrome show mutations in genes of the cadherin-catenin complex.

A. Kievit[1], F. Tessadori[2,3], J. Douben[1], I. Jordens[2], M. Maurice[2], A. Hoogeboom[1], R. Hennekam[4], S. Nampoothiri[5], H. Kayserili[6], M. Castori[7], M. Whiteford[8], C. Motter[9], C. Melver[9], M. Cunningham[10], A. Hing[10], N. Mizue Kokitsu-Nakta[11], S. Vendramini-Pittoli[11], A. Richieri-Costa[11], A. Baas[2], M. Massink[2], K. van Gassen[2], J. Bakkers[3], F. Santos[12], P. Lapunzina[12], V. Gil-da Silva Lopes[13], A. Slavotinek[14], V. Martinez-Glez[12], J. de Klein[1], G. van Haaften[2], M-J. van den Boogaard[2]. 1) Department of Clinical Genetics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands; 2) Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands; 3) Hubrecht Institute, University Medical Center Utrecht, Utrecht, the Netherlands; 4) Department of Pediatrics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands; 5) Department of Pediatric Genetics, Amrita Institute of Medical Sciences and Research Center, Kerala, India; 6) Department of Medical Genetics, Koç University School of Medicine, Istanbul, Turkey; 7) Department of Clinical Genetics, San Camillo-Forlanini General Hospital, Rome, Italy; 8) Department of Clinical Genetics, Queen Elizabeth University Hospital, Glasgow, United Kingdom; 9) Division of Medical Genetics, Akron Children's Hospital, Akron, Ohio, U.S.A; 10) Division of Craniofacial Medicine, University of Washington Department of Pediatrics, Jean Renny Chair of Craniofacial Medicine, Seattle Children's Craniofacial Center, Seattle, U.S.A; 11) Serviço de Genética, Hospital de Reabilitação de Anomalias Craniofacialis, Universidade de São Paulo, Bauru, SP, Brazil; 12) INGEMM, Institute of Medical and Molecular Genetics, Hospital Universitario La Paz, Universidad Autónoma de Madrid, IdiPAZ, CIBERER, ISCIII, Madrid, Spain; 13) Department of Medical Genetics, School of Medical Sciences, State University of Campinas, UNICAMP, Campinas, São Paulo, Brazil; 14) Department of Pediatrics, University of California, San Francisco, Benioff Children's Hospital, San Francisco, U.S.A.

Blepharo-Cheilo-Dontic (BCD) syndrome is an infrequently described autosomal dominant condition with an unknown cause, characterized by orofacial clefts, ectodermal defects and ocular anomalies. Using exome sequencing, we identified heterozygous mutations in two genes of the cadherin-catenin complex, *CDH1*, encoding E-cadherin, and *CTNND1*, encoding p120-catenin delta 1. We subsequently detected, in 29 patients from 18 families, *CDH1* mutations in 67% and *CTNND1* mutations in 17% of the probands. In 60% of the probands the mutation has arisen de novo. The clinical features of the presented 29 and the reported 75 BCD patients, demonstrate in over 50% hypertelorism, euryblepharon, lagophthalmos, ectropion of the lower lid, distichiasis, cleft lip/palate (CLP), hypodontia, delayed dentition, abnormal dental crown form and high frontal hairline. Less frequent findings are ankyloblepharon, hypothyroidism, skin dermoid cysts, imperforate anus and neural tube defects. Correlations of the phenotype with the genotype demonstrate that BCD patients with a *CTNND1* mutation have a milder phenotype with less frequent CLP, and less frequent facial signs such as high frontal hairline, broad forehead, malformed ears and everted lower lip. They also have no associated findings, like imperforate anus, hypothyroidism and neural tube defects. We demonstrate that the ocular phenotype may decrease with age and even disappear completely, which may indicate that the eye signs may not be recognized in patients with CLP and suggests that BCD syndrome can be underdiagnosed. Mutations in *CDH1* have been linked to familial hereditary diffuse gastric cancer. We did not observe gastric cancer or other malignancies in the presented and reported BCD patients, but most of the patients we describe are still young. *CDH1* plays an essential role in epithelial architecture and intercellular adhesion, functioning as cell invasion suppressor. *CTNND1* binds together with the other catenins to the juxtamembrane domain of *CDH1* and controls stability of the cadherin-catenin complex. Functional experiments in zebrafish and cell lines show that the *CDH1* mutations in BCD patients impair the cell adhesion function of the cadherin-catenin complex in a dominant negative manner. Our findings expand the role in human development of the cadherin-catenin complex, and should facilitate diagnosis of BCD syndrome and genetic counseling to patients and families with this entity.

109

803 individuals with retinal dystrophies investigated with targeted NGS of 124 genes. *K. Gronskov[1], M. Fang[2], M. Bertelsen[3], C. Jespersgaard[1], X. Dang[2], H. Jensen[3], Y. Shen[2], N. Bech[3], I. Dai[2], T. Rosenberg[3], J. Zhang[2], L. Moller[1], Z. Tümer[1], K. Brondum-Nielsen[1].* 1) Dept. Clinical Genetics, Kennedy Center, Rigshospitalet, University of Copenhagen, Glostrup, Denmark; 2) BGI-Shenzhen, Shenzhen 518083, China; 3) Eyefunction Kennedy, Dept. of Ophthalmology, Rigshospitalet-Glostrup, University of Copenhagen, Glostrup, Denmark.

Retinal dystrophy covers a range of diagnoses and can occur either isolated or part of a syndrome. In this study we aimed to identify the underlying genetic cause in 823 retinal dystrophy patients for whom we had DNA in our biobank. The individuals were diagnosed with one of the following diagnoses: autosomal recessive retinitis pigmentosa (AR-RP); autosomal dominant retinitis pigmentosa (AD-RP); X-linked retinitis pigmentosa (XL-RP); Leber's Congenital Amaurosis (OMIM 204000); Usher syndrome (OMIM276900); Bardet Biedl Syndrome (OMIM209900); cone- or cone-rod dystrophy; macular dystrophy or Stargardt disease (OMIM248200); or congenital stationary night blindness (310500). All individuals had given written informed consent for genetic analysis. Due to bad quality of the DNA, 21 individuals were omitted from the investigation and 802 individuals were screened for sequence variations in 124 genes using a targeted NGS approach. After alignment and variant calling, data were filtered with a quality, technical and genetic filter. Variants were interpreted using an in- house system based on the ACMG 2015 guidelines and classified as class 1 (benign), class 2 (likely benign), class 3 (variants of unknown significance, VUS), class 4 (likely pathogenic) or class 5 (pathogenic). Class 3, 4 and 5 variants were verified by Sanger sequencing and reported. However, if a probable molecular genetic explanation of the phenotype was found, only these variants and additional class 4 and 5 variants were verified and reported. Our results in 456 of the 802 individuals, revealed 232 (51%) individuals with a likely molecular genetic explanation of their eye symptoms; 136 (30%) had one class 3, 4 or 5 variation in a gene with autosomal recessive inheritance; in 88 (19%) individuals we found no molecular genetic explanation of the eye symptoms. We found nine major genes (*ABCA4* (OMIM 601691), *EYS* (OMIM 612424), *USH2A* (OMIM 608400), *BEST1* (OMIM 607854), *RHO* (OMIM 180380), *RP1* (OMIM 603937), *RPGR* (OMIM 312610), *CRB1* (OMIM 604210), *PRPH2* (OMIM 179605)). Variants in one of these can explain the eye symptoms in 50% of the 232 individuals. We found variations in 67 different genes, and in 36 genes we found variants in only one or two individuals. These findings confirm the vast genetic heterogeneity of RP and show that NGS panel analysis is the right choice for molecular genetic diagnosis of individuals with retinal dystrophies and provide data for establishment of a diagnostic strategy.

110

Genome-wide association analyses using the Haplotype Reference Consortium for imputations reveals 4 novel loci involved in glaucoma endophenotypes. *A. Iglesias Gonzalez[1], P. Bonnemaijer[1,2], H. Springelkamp[1,2], S. van der Lee[1], N. Amin[1], C.C.W. Klaver[1,2], C.M. van Duijn[1].* 1) Epidemiology, Erasmus MC, Rotterdam, Rotterdam, Netherlands; 2) Ophthalmology, Erasmus MC, Rotterdam, Rotterdam, Netherlands.

Purpose: Glaucoma is a heterogeneous and complex eye disease characterised by degeneration of the optic nerve. High intraocular pressure (IOP) and an enlarged vertical cup-disc ratio (VCDR) are well-recognized risk factors and have been used successfully in genome-wide association studies (GWAS) as endophenotypes. The creation of new imputation resources based on the Haplotype Reference Consortium (HRC) allows the imputation of new, in particular, low-frequency variants that were not reliably imputed in the 1000 genomes (1000g). In this study, we have used HRC imputations to identify low-frequency variants associated with IOP and VCDR. Methods: This study was conducted in three independent cohorts from the Rotterdam Study (I, II, III) and the Erasmus Rucphen Family study. We conducted a fixed-effect meta-analysis of GWAS from up to 14.500 Europeans imputed to the HRC reference panel version 1. The association of variants to IOP and VCDR was adjusted for age, sex, the first five principal components or family structure. Functional characterization and eQTL effects were investigated using HaploReg.v4 and GTEx. Results: For IOP, this study identified a novel locus mapped on chr2p16.1 within the *CCDC85A* gene (rs567596280; MAF=0.002, not in 1000g) and close to the *EFEMP1* gene, previously associated with cup area. For VCDR, three new loci were identified; one on chr5q11.2 within *PDE4D* (rs138567055, MAF=0.002, present in 1000g); other on chr9q21.33 within the lncRNA *GAS1RR* (rs10448329, MAF=0.12, present in 1000g); and one on chr13q21.33 close to *DACH1* (rs568293376, MAF=0.01, not in 1000g, and genome-wide significant only when adjusting for disc area). Except for *CCDC85A*, all genes are expressed in the eye. *GAS1RR* regulates the expression of *GAS1*, which is part of the same gene family as *GAS7*, a gene previously associated with IOP and glaucoma. The variant in the lncRNA *GAS1RR* showed a significant cis-eQTL effect in GTEx (P=3.4 x10^{-06}). No eQTL effects were observed for the variant close to *DACH1*, however, the DACH1 protein acts as a transcriptional repressor of a *SIX6*, a well-established glaucoma and VCDR gene. Conclusions: Using HRC as a reference panel, we identified one novel locus associated with IOP and three with VCDR. Two of the novel genes (*GAS1* and *DACH1*) are part of pathways identified earlier. The other genes have been associated to diabetes (*CCDC85A)* and cancer *(PDE4D).* Our study shows that HRC can offer new opportunities to identify genetic variants.

111

Transcriptome-wide association study of thirty complex traits reveals novel risk genes. *G. Kichaev[1], N. Mancuso[2], H. Shi[1], P. Goddard[3], A. Gusev[4,5,6], B. Pasaniuc[1,2,7].* 1) Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA; 2) Department of Pathology & Laboratory Medicine, Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA; 3) Department of Molecular, Cell and Developmental Biology,University of California, Los Angeles, Los Angeles, CA; 4) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA; 5) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; 6) Program in Medical and Population Genetics, Broad institute, Cambridge MA; 7) Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA.

Associations between genetic variation and disease provide insight into where risk is localized, but have yet to explain the underlying biological mechanisms responsible. Intermediate phenotypes such as gene expression, provide a plausible path of causality where genetic variation influences expression levels, which in turn influence trait. To accurately test this hypothesis, large panels consisting of genotype and expression measurements are required. Unfortunately, comprehensive portfolios of gene expression are typically limited in sample size due to prohibitive cost. Recently, several approaches have demonstrated the efficiency of imputing gene expression into large cohorts. Here we use expression predicted in more than 40 tissues as a mediator for risk and perform a transcriptome-wide association study (TWAS) on more than 2.3 million phenotype measurements from 30 large-scale genome-wide association studies (GWAS). Overall, we found 5,490 associations in 27 traits spanning 1,196 genes in total. We quantified how much of our signal was due to GWAS hits by probing a 1Mb flanking region surrounding the gene and found that ~91% overlap a genome-wide significant SNP; however, we consider the remaining genes that lack an overlap to be “novel” candidates for follow-up studies. Interestingly, associated genes that overlapped a significant SNP were typically not the closest possible. This suggests the heuristic approach of assigning GWAS SNPs function based on the nearest gene is sub-optimal. To assess the degree of pleiotropy for each gene, we counted the total number of trait associations. Our results indicate pervasive pleiotropy with multiple genes associated with more than 5 traits. For example, the GSDMB gene was associated in Crohn’s disease ($p=2.49 \times 10^{-9}$), IBD ($p=2.66 \times 10^{-14}$), ulcerative colitis ($p=8.08 \times 10^{-10}$), HDL ($p=3.53 \times 10^{-13}$), total cholesterol ($p=7.67 \times 10^{-7}$), and rheumatoid arthritis ($p=1.93 \times 10^{-8}$). Our results reinforce the power of the TWAS approach in associating genes with risk for disease and shed additional light onto the mechanistic aspects of complex traits and disease.

112

Phasing, imputation and analysis of 500,000 UK individuals genotyped for UK Biobank. *J. Marchini[1,2], C. Bycroft[2], D. Petkova[2], S. Murphy[3], C. Freeman[2], P. Donnelly[2,1].* 1) Dept Statistics, Oxford Univ, Oxford, United Kingdom; 2) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; 3) UK Biobank, Units 1-4 Spectrum Way, Adswood, Stockport, Cheshire, UK.

The UK Biobank project is a large prospective cohort study of 500,000 individuals from the UK, aged between 40-69 at recruitment. For each participant, there are hundreds of health-related phenotypes including physical measurements and bio-assays, medical records, information about diet and exercise, and agreement that their health can be followed over the course of their life. Samples of all the participants' DNA were genotyped by Affymetrix Research Service Laboratory on the custom-designed UK Biobank Affymetrix Axiom array, which contains over 800,000 SNPs. We describe the quality control strategies that we applied to this large multi-batch data set. We also describe various characteristics of the samples including population structure and relatedness that will be critical to many researchers analyses. We will describe a new accurate method (SHAPEIT3) for phasing Biobank-sized datasets that scales O(NlogN) in the number of samples (N) and results in haplotypes with estimated switch error rates as low as 0.2%. Imputation was carried out using a new method (IMPUTE4) that allows accurate low memory imputation of the UK Biobank dataset in <10 minutes per sample genome-wide. Imputation was carried out using a combination of the Haplotype Reference Consortium reference panel and the UK10K reference panel to provide the best combination of imputation at SNPs and short indels. We will illustrate the potential of this resource to detect novel associations and describe how researchers can apply for access to the genetic and phenotypic data.

113

Improved score statistics for meta-analysis in single variant and gene-level association studies. *J. Yang, S. Chen, G. Abecasis, International Age-related Macular Degeneration Genomics Consortium.* Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA.

Meta-analysis is now an essential tool for genetic association studies, allowing these to combine information on 100,000s - 1,000,000s of individuals, and greatly accelerating the pace of genetic discovery. In ideal settings, meta-analysis performs as efficiently as the more cumbersome alternative of sharing individual-level data. However, when the ratio of cases and controls varies among studies in an unbalanced setting, meta-analysis using current standard methods results substantial power loss. The limitations of current methods apply where meta-analysis is based on combining sample sizes and p-values (the Stouffer method), using regression coefficients and their standard errors (the Cochran method), or by combining score statistics. Here, we describe a novel method for meta-analysis that produces similar results as a more cumbersome joint analysis of the combined dataset, even in unbalanced settings. Our method can accurately approximate score statistics obtainable in a joint analysis, which is suitable for both linear and logistic regression models, with or without covariates. In the special cases without covariates, our method is exactly equivalent to sharing individual-level data. Further, we extend it to enable gene-based Burden and SKAT tests. By simulation studies, we show that --- compared to joint analyses sharing individual level data --- current and widely used meta-analysis methods reduce power up to 90% when case-control ratios vary among studies. In contrast, our meta-analysis method is almost equivalent in power to a joint analysis. Similar results were obtained in a real study of Age-related Macular Degeneration (AMD), with 26 individual studies and 33,976 samples, by the International AMD Genomics Consortium. For example, the known AMD risk gene *CFI* has SKAT p-value=1.9x10^{-10} by joint analysis, p-value=1.2x10^{-4} by directly summarizing score statistics as in the current meta-analysis method, and p-value=3.1x10^{-9} by our meta-analysis method. In summary, our approach produces improved single-variant score statistics in the meta-analysis that can be used to construct both single variant and gene-level association studies. Our method provides a useful framework for ensuring well-powered, convenient, cross-study analysesand is now implemented in the RAREMETAL software (https://github.com/traxexx/Raremetal).

114

Searching for novel cross-phenotype associations using Bayesian meta-analysis. *H. Trochet[1], L. Jostins[1], G. McVean[1,3], M. Pirinen[2], C. Spencer[1].* 1) Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom; 2) Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland; 3) Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford.

The number of genome-wide association studies (GWAS) has grown rapidly over the past decade. Whilst the statistical evidence for association at a variant is often combined across closely related phenotypes via standard meta-analysis, less effort has been made to identify shared genetic effects across weakly related or even putatively unrelated traits. Existing methods often rely on individual-level data, use only *p*-values—which are influenced by study size and minor allele frequency—or make strong assumptions about the similarity of effects across traits. To overcome these drawbacks, we have developed a Bayesian meta-analysis approach that allows us to combine results of a set of GWAS using only readily available summary statistics. Using an approximate Bayes factor (ABF) approach we aim to infer which loci affect multiple phenotypes, and to find which phenotypes are affected. We lay out some the statistical issues that arise in interrogating this potentially complex model space. Using extensive simulations we show that this approach can have more power than existing methods when a genetic variant has an effect on only a subset of traits (for realistic scenarios where fewer than half the traits are associated we observe a 50% increase in power), and quantify the accuracy of the posterior distribution on models. To demonstrate the practical utility of our method, we re-analyse summary statistics data from 15 GWAS that form part of the Wellcome Trust Case-Control Consortium 2. These include studies with shared controls; quantitative and case-control phenotypes; imputed and directly genotyped data; and individuals from studies in Asia, South America, Africa and Europe. As examples, we highlight two loci that indicate novel connections between auto-immune and infectious disease.

115

Novel schizophrenia risk genes identified through genic associations in CommonMind Consoritum and GTEx transcriptome imputation. *L.M. Huckins[1], A. Dobbyn[1], D.M. Ruderfer[1,2,3], M. Fromer[1,2], N. Cox[4,5,6], H.K. Im[6], S. Sieberts[7], B. Devlin[8,9], P. Roussos[1,2,10,11], S. Purcell[1,2,3,12,13], P. Sklar[1,2,10,12], E.A. Stahl[1,2], CommonMind Consortium.* 1) Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York City, NY; 2) Institute for Genomics and Multiscale Biology, New York, NY, USA; 3) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4) Division of Genetic Medicine, Vanderbilt University, Nashville, Tennessee, USA; 5) Department of Human Genetics, University of Chicago, Chicago, Illinois, USA; 6) Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, Illinois, USA; 7) Systems Biology, Sage Bionetworks, 1100 Fairview Ave N, Seattle, WA, 98109, USA; 8) Psychiatry, University of Pittsburgh School of Medicine, 3811 O'Hara St, Pittsburgh, PA, 15213, USA; 9) Human Genetics, University of Pittsburgh, 3811 O'Hara St, Pittsburgh, PA, 15213, USA; 10) Friedman Brain Institute, Icahn School of Medicine at Mount Sinai,New York, NY, 10029, USA; 11) Psychiatry, JJ Peters VA Medical Center, 130 West Kingsbridge Road, Bronx, NY, 10468, USA; 12) Department of Genetics and Genomic Sciences, New York, NY, USA; 13) Analytic and Translational Genetics Unit, Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA.

Genome-wide association studies (GWAS) have identified numerous robustly associated genomic regions, however the biological implications of these regions remain poorly understood. Large-scale transcriptomic datasets (for example RNA-sequencing data) have enabled identification of variants regulating gene expression (eQTLs) in specific tissues. These data can be used to impute tissue-specific genetic expression levels (GREx) from genotypes in larger samples, which can be tested to identify potentially novel associated genes [PMID: 26258848]. Here, we use the largest existing transcriptomic profiling of the brain along with ten GTEx brain regions to impute GREx and test for association in GWAS datasets of schizophrenia (SCZ), bipolar disorder and others from the Psychiatric Genomics Consortium (PGC). Following systematic comparison of prediction modelling techniques, models were created for 13,452 genes from 668 individuals with imputed genotype and RNA-seq data (CommonMind Consortium). Predictors accurately imputed GREx in an independent DLPFC dataset of ~400 individuals (40% of genes have R^2 0.01-0.8, in line with previous models). Results were compared across ancestries, and individual- vs. summary-level data. Our initial analysis used CMC and ten GTEx brain region models to impute GREx in PGC SCZ cases and controls (~32,000/42,000). We tested imputed GREx for association with SCZ, and identified 66 genes with significant associations (p<5x10^{-6}) in at least two brain regions. These included novel associations (for example *OR2J2, TTC14;* minimum p-values across brain regions p=1.1x10^{-26}, p=7.4x10^{-7}) and well-established SCZ genes (*C4A, SNX19;* p=1.8x10^{-21}, p=3.6x10^{-13}). Roughly half of these associations lie in the MHC region, which may help elucidate the complex local landscape of SCZ risk. We have used GREx imputation to harness large GWAS sample sizes, resulting in the first prediction models for the DLPFC. These predictors may identify novel genic associations, as shown for SCZ. Of the 108 PGC GWAS loci, 74 had at least one associated gene reaching nominal significance in this study, while 27 reached genome-wide significance. We will expand on these analyses to probe the general relationship between GWAS-loci, differential gene expression, and eQTLs. We will further use these predictors to address questions about the role of gene expression in disease risk and heritability.

116

Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *A. Gusev[1], N. Mancuso[2], L. Song[3], E. Oh[3], S. McCarroll[4], B. Neale[4,5], R. Ophoff[2], M. O'Donovan[6], G. Crawford[3], N. Katsanis[3], P.F. Sullivan[3], B. Pasaniuc[2], A.L. Price[1], Schizophrenia Working Group of the Psychiatric Genomics Consortium.* 1) Harvard TH Chan School of Public Health, Boston MA; 2) UCLA, Los Angeles CA; 3) Duke University, Durham NC; 4) Harvard Medical School, Boston MA; 5) Massachusetts General Hospital, Boston MA; 6) Cardiff University, Cardiff UK.

Despite the many disease-associated variants identified by genome-wide association studies and the well-documented functional enrichment of disease heritability at regulatory elements, elucidating the underlying genes and their regulatory mechanisms remains a challenge. We performed a transcriptome-wide association study (TWAS) for schizophrenia, jointly analyzing expression data from 3,693 individuals (across three tissues including brain) and schizophrenia summary statistics from 79,845 independent individuals from the Psychiatric Genomics Consortium using recently published TWAS methods (Gusev et al. 2016 Nat Genet). We tested both total gene expression and alternative splicing for association to disease. We identified 157 transcriptome-wide significant gene-disease associations, of which 45 did not overlap a genome-wide significant variant. Surprisingly, the greatest number of associations stemmed from differentially spliced introns in brain tissue that were independent of total gene expression; this enrichment persisted for associations not reaching transcriptome-wide significance, highlighting the importance of assaying splice variation in an appropriate tissue. We replicated these associations via internal cross-validation using individual-level data, as well as in an independent schizophrenia cohort. To identify the underlying regulatory elements, we also tested gene expression and alternative splicing for transcriptome-wide association to chromatin phenotypes assayed by ChIP-seq in independent samples. We determined that our chromatin TWAS approach is substantially more powerful than traditional analyses that overlap top expression-QTL and chromatin-QTL, identifying 10x more genes with transcriptome-wide significant associations to nearby chromatin peaks. The predicted gene-chromatin associations were confirmed using measured RNA-seq data in the same samples, explaining a striking 20% of the variance in total expression on average. Of the 157 genes associated to schizophrenia, 42 were also associated to chromatin phenotypes (a 4-fold enrichment relative to all tested genes, P=1x10^{-11}), identifying specific chromatin elements that likely mark drivers of expression and phenotype. Analyses of individual GWAS loci strongly supported an underlying model where expression was mediated by chromatin activity, and demonstrated substantially greater resolution to identify causal variants when integrating chromatin phenotypes.

117

Joint Bayesian inference of risk variants using summary statistics and epigenomic annotations across multiple traits. *Y. Li.* MIT, Cambridge, MA.

Fine-mapping causal variants is challenging due to linkage disequilibrium and the lack of interpretation of noncoding mutations, which contribute to ~70-90% of GWAS SNPs. Existing fine-mapping methods do not scale well on inferring multiple causal variants per locus and causal variants across multiple related diseases. Moreover, Many complex traits are genetically related and potentially share causal mechanisms. Thus, we hypothesize that exploiting the correlation between traits at the epigenomic annotation level may prove useful in fine-mapping for shared causal mechanisms that go beyond the level of individual variants. We develop a new Bayesian fine-mapping model named RiVIERA. The key features of RiVIERA include 1) ability to model epigenomic covariance of multiple related traits; 2) efficient posterior inference of causal configuration; 3) efficient full Bayesian inference of causal annotations; 4) simultaneously modeling the underlying heritability parameters (variance explained per SNP) and leveraging it in the causal inference. We conducted a comprehensive simulation studies using 1000 Genome and Roadmap epigenomic data to demonstrate that RiVIERA compares quite favorably with existing methods yet having unique advantages of jointly inferring and leveraging the underlying tissue-specific functional co-enrichment among traits that are not necessarily associated with the same set of risk loci. In particular, the efficient inference of multiple causal variants per locus from our RiVIERA model led to significantly improved estimation of the causal posterior and functional enrichments compared to the state-of-the-art fine-mapping methods especially for the cases where more than 3 causal variants co-harboured within the same loci. Furthermore, joint modelling multiple traits confers further improvement over the single-trait mode of the same model, which is attributable to the more robust estimation of the enrichment parameters especially when the annotation measurements (e.g., ChIP-seq) themselves are noisy. We demonstrated the application of RiVIERA on jointly modelling several well-powered GWAS datasets of genetically related traits including lipid traits, coronary artery disease, type 2 diabetes, and BMI. Our analyses revealed several putative pleiotropic pathways that are disrupted by separate mutations in those traits, which do not share the same risk loci. RiVIERA is freely available from Github.

118

Tissue-specific trans-ancestral analysis of genetically regulated expression identifies 99 known and 41 novel genes across 14 metabolic and cardiovascular traits. *J.E. Below[1], L.E. Petty[1], H.M. Highland[1,2], H. Hu[3], P.S. de Vries[1], D. Aguilar[4], G.I. Bell[5], C.D. Huff[3], N.J. Cox[6], C.L. Hanis[1], J. Ma[1], E.J. Parra[7], M. Cruz[8], A. Valladares-Salgado[8], H.K. Im[9], A.C. Morrison[1], E. Boerwinkle[1].* 1) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX; 2) Dept. of Epidemiology, University of North Carolina, Chapel Hill, NC; 3) Dept of Epidemiology, MD Anderson Cancer Center, Houston, TX; 4) Dept of Cardiology, Baylor College of Medicine Houston, Texas; 5) Depts of Medicine and Human Genetics, The University of Chicago, Chicago, Illinois 60637; 6) Vanderbilt Genetics Institute Vanderbilt University School of Medicine Nashville, TN; 7) Dept of Anthropology, University of Toronto at Mississauga, Mississauga, Ontario, Canada; 8) Unidad de Investigación Médica en Bioquímica, Hospital de Especialidades, Centro Médico Nacional Siglo XXI, IMSS, Mexico City, Mexico; 9) Section of Genetic Medicine, Dept of Medicine, University of Chicago, Illinois.

Interpretation of genetic association results is difficult devoid of a biological context. To better understand the genetic etiology of a spectrum of complex traits, we predicted individual-level gene expression from common variants with PrediXcan and determined genes with differentially predicted expression by trait. PrediXcan aggregates evidence of functional variation by leveraging transcriptome data from GTEx to collapse common eQTLs into an imputed expression measure in 41 tissues. To explore the effects of tissue-specific genetically regulated predicted gene expression on cardiovascular and metabolic traits, we performed linear regression on 14 traits including blood lipid levels, BMI, height, blood pressure, fasting glucose and insulin, heart rate, fibrinogen, factor VII, and white blood cell and platelet count in ARIC (9,000 Caucasian, 2500 African American) and three Hispanic-ancestry cohorts (4000 total). Ascertainment of genome-wide significant findings was restricted to tissues of trait-specific functional relevance (metabolic traits: pancreas, skeletal muscle, subcutaneous adipose; cardiovascular traits: atrial appendage, left ventricle, and aortic, coronary, tibial arteries; both: liver, adrenal gland, whole blood). Documenting the validity of the method and robustness of the results, we replicated 49 and 50 genes in loci implicated in prior GWAS for metabolic and cardiovascular traits, respectively. We identified 22 novel genes for metabolic and 19 for cardiovascular traits below traditional genome-wide significance (p<2.5e-6). These genes are significantly enriched in gene ontology terms related to lipoprotein remodeling and metabolic processes. We report heterogeneity in effects at genes that may be due to population-specific ancestry. Top hits were assessed for replication using MetaXcan in the MAGIC Consortium. We validated predicted expression levels genome-wide in whole blood RNA-seq data from a subset of 200 Caucasian participants and report correlation of predicted vs. measured gene expression for top gene-trait associations. One association of particular interest between expression in whole blood of a platelet-activating factor, *PAFAH1B2,* with triglyceride level (p=5e-8) replicated in follow-up (p=9e-9). A single rare coding variant in this gene was recently found to have a large effect on triglyceride levels. Our findings complement this discovery and are the first to show significant association for this trait from common eQTLs.

119

Pleiotropic effects on BMI, WHR, fasting glucose, and fasting insulin levels. *H.M. Highland[1], C.M. Sitlani[2], A.A. Seyerle[1], R. Gondalia[1], M. Graff[1], C.L. Avery[1], K.E. North[1].* 1) Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) Department of Medicine, University of Washington, Seattle, WA, USA.

Obesity is a risk factor for developing type 2 diabetes. While the relationship between diabetes and obesity is complex, it has been suggested that they have a shared genetic etiology, with variation in or near *FTO, TCF7L2* and *KCNJ11* plausibly affecting both traits. To identify loci that have possible pleiotropic effects on both obesity (measured by body mass index (BMI) and waist to hip ratio (WHR)) and glycemic measures underlying diabetes development (measured by fasting glucose (FG) and fasting insulin (FI) levels) we conducted a multivariate GWAS of obesity and diabetes related traits using the previously published summary statistics from the large scale GIANT (N_{BMI} = 339,224, N_{WHR} = 226,643) and MAGIC (N_{FG} = 58,074, N_{FI} = 51,750) studies of predominantly European ancestry individuals. We used a novel analytical approach, the adaptive sum of powered score (aSPU) test, that conducts tests for both concordant and discordant associations across all or some of the included traits. After excluding loci that were known to be significantly associated with any single trait, six novel loci (in or near *SPRY4, LOC645434, NFE2L1, NPR3, REEP3,* and *GGNBP2)* were significantly associated with the four quantitative measures in aggregate (P<5.0e-8). This includes a variant near *SPRY4,* which is part of the MAPK pathway. This variant is associated with increased BMI and FG levels. SPRY4 has been shown to suppress both insulin and EGFR transduced MAPK pathways. An intronic variant in *NPR3* is associated with increased BMI, WHR and increased FI. *NPR3*, which has previously been associated with both height and blood pressure, is a natriuretic peptide receptor that is expressed in kidney, aorta, and adipose tissues. In conclusion, we have demonstrated an important utility of leveraging apparent pleiotropy for novel locus discovery of obesity and diabetes related traits.

120

LD Hub and MR-Base: Online platforms for performing LD score regression and Mendelian randomization analysis using GWAS summary data. *D. Evans[1,2], J. Zheng[1], G. Hemani[1], B. Elsworth[1], H. Shihab[1], C. Laurin[1], M. Erzurumluoglu[3], L. Howe[1], K. Wade[1], N. Warrington[2], H. Finucane[4], A. Price[4], V. Anttila[4], L. Paternoster[1], R. Martin[1], C. Relton[1], G. Davey Smith[1], B. Neale[4], T. Gaunt[1], P. Haycock[1].* 1) MRC Integrative Epidemiology Unit, Univeristy Bristol, Bristol, United Kingdom; 2) Diamantina Institute, University of Queensland; 3) Genetic Epidemiology Group, University of Leicester, United Kingdom; 4) Broad Institute of MIT and Harvard, Cambridge, USA.

LD score regression is a reliable and efficient method of calculating the SNP heritability of complex traits and diseases, partitioning this heritability into functional categories, and estimating the genetic correlation between different phenotypes. Mendelian randomization is an epidemiological approach of estimating the causal relationship between a modifiable exposure and a medically relevant outcome. Both methods can be performed using summary data from genome-wide association studies, and the two can be combined effectively to screen hundreds of different traits for putative causal relationships. To facilitate this process, we have created two new platforms called LD Hub (ldsc.broadinstitute.org) and MR-Base (www.mrbase.org) that simplify implementation of LD score regression and Mendelian randomization. The platforms are built on a database of harmonized summary genetic data, roughly corresponding to >1,200,000 individuals and approximately 2000 phenotypes. LD Hub calculates estimates of heritability and the genetic correlation between the selected phenotypes. MR-Base returns an estimate of the causal effect of hypothesized exposures on an outcome (typically a disease phenotype). In order to illustrate the potential of the software, we used LD Hub to estimate the genetic correlation between 107 metabolomic traits and coronary artery disease (CAD). We followed up variables that showed evidence of genetic correlation using MR-Base in order to determine whether the relationship was likely to reflect a causal influence of the metabolite on CAD. In an analysis of metabolites and CAD, we found that 74 traits were significantly genetically correlated with CAD. CAD was positively genetically correlated with very low, low and intermediate density lipoprotein particles, but negatively correlated with high density lipoproteins (HDL). In the follow-up analysis of these 74 metabolites using MR-Base, we found that increasing levels of 45 fatty acids, amino acids and low-density lipoproteins were causally related to increased risk of CAD. In contrast, most of the HDL fractions did not appear to be causally related to CAD, except triglycerides in small HDL that showed a positive effect. In response to the growing availability of accessible GWAS summary-level data, our database and web interfaces will catalyse the integration of high-dimensional datasets, analytical approaches for causal inference and stimulate translational activities aimed at disease prevention.

121

Genome-wide association study of bone mineral density in the UK Biobank Study identifies over 376 loci associated with osteoporosis. *J.P. Kemp[1,2], J.A. Morris[3,4], M. Medina-Gómez[5], C.L. Gregson[6], V. Forgetta[3,4], K. Trajanoska[5], N.M. Warrington[1], J. Zheng[2], S. Kaptoge[7], F. Rivadeneira[5], J.H. Tobias[6], C.L. Ackert-Bicknell[8], J.B. Richards[3,4], D.M. Evans[1,2].* 1) University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland, Australia; 2) MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK; 3) Departments of Medicine, Human Genetics, Epidemiology and Biostatistics, Lady Davis Institute, McGill University, Montréal, Canada; 4) Department of Twin Research, King's College London, London, UK; 5) Department of Epidemiology, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands; 6) School of Clinical Sciences, University of Bristol, Bristol, UK; 7) Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 8) Center for Musculoskeletal Research, University of Rochester, Rochester, New York, USA.

Bone mineral density (BMD) is a highly heritable trait that is used as the primary diagnostic and prognostic marker for osteoporosis and fracture risk susceptibility and is conventionally measured at the hip or spine by dual energy X-ray absorptiometry. Quantitative ultrasound (QUS) is a non-invasive technique that measures the speed and attenuation of ultrasound through bone, and estimates bone mineral density (eBMD). Previous GWAS of BMD/QUS measures (n = 35,000) have identified up to 68 loci which explain < 5% of the trait heritability. To identify genetic loci associated with BMD variation, we used the software package BOLT-LMM to perform a GWAS of eBMD of the heel on 140,623 individuals of European decent (75,275 females and 65,349 males) from the UK Biobank Study. Our GWAS identified 403 independent SNPs from 376 loci (defined as $r^2 < 0.05$ and +/-500 MB) attaining genome-wide significance ($P < 5x10^{-8}$) which jointly explained 12% of the variation in heel eBMD. These included the majority of SNPs previously associated with DXA derived BMD, as well as 308 novel loci, including many containing genes that have not been previously implicated in bone physiology. LD score regression and bivariate G-REML analyses revealed moderate to high positive genetic correlations between heel eBMD and DXA-derived BMD measures of the forearm, femoral neck, spine, and total body, and a negative genetic correlation with fracture. We also found significant genetic correlations with BMI, celiac disease, educational attainment, age at menarche and LDL cholesterol, but not cigarettes per day, despite the observational association between smoking and risk of osteoporosis. We implemented in-silico fine-mapping by constructing credible sets with a Bayesian method, followed by coding and non-coding SNP annotation using VEP, deltaSVM, and CATO. Results from fine-mapping strongly implicated novel and known genes and predicted causal SNPs. Preliminary functional studies showed increased expression of a novel gene *Zhx3* in maturing calvarial osteoblasts and that *Zhx3* KO mice had increased whole body BMD compared to wild type mice. In summary, using UK Biobank, we increased the number of BMD loci 5-fold, identified traits and diseases sharing etiologic pathways with osteoporosis, and implicated novel proteins, which will serve as potential drug targets to improve the care of patients suffering from this common costly disease.

122

Joint re-analysis of GWAS summary statistics identifies new variants associated with human traits and diseases. *B.J. Vilhjalmsson[1,2], H. Finucane[2], A. Gusev[2], A.L. Price[2], P. Kraft[2], H. Aschard[2].* 1) Bioinformatics Research Centre, Aarhus University, Aarhus, Aarhus, Denmark; 2) Harvard TH Chan School of Public Health, Boston, MA.

Genome-wide association studies (GWAS) have proven successful in identifying thousands of significant genetic associations for multiple traits and diseases. This success is largely thanks to the many GWAS meta-analysis consortia where researchers have combined cohorts to increase sample size, which has led to increased statistical power. However, GWAS meta-analyses across different diseases and traits have received limited attention, even though multivariate analysis enables the detection of pleiotropic genetic variants. Indeed, various multivariate GWAS approaches that account for the phenotypic correlation structure have been proposed and shown to increase power to detect associations (Korte et al., Nat Genet 2012; Zhou and Stephens, Nat Meth 2014). Most of these require individual level data, and are thus not applicable to GWAS summary statistics. To address this problem we propose PCMA, a computationally efficient multivariate meta-analysis of GWAS summary statistics for correlated diseases and traits. PCMA estimates the phenotypic correlation structure and accounts for it to derive a joint association test statistic across all phenotypes for each genetic variant. Furthermore, using the GWAS summary statistics we derive the principal components of the phenotypes and corresponding an association test statistics, facilitating interpretation of any associations found. Using simulations, we also examine the conditions under which biased phenotypic correlation estimates lead to false positives, demonstrate such problem in practice, and propose a set of guidelines and statistics to avoid such pitfalls. Finally, we jointly re-analyze GWAS summary statistics for more than 40 traits and diseases with sample sizes of up to 350,000 individuals, and identify more than 50 novel genome-wide significant associations. The novel associations include an association in the *FOXP2* gene, also known as the language gene, when re-analyzing GWAS summary statistics from the psychiatric genetics consortium, and an association in the *IL6R* gene, a rheumatoid arthritis drug target, when re-analyzing GWAS summary statistics for immune-related diseases.

123

Multivariate genetic risk scores can increase risk prediction accuracy for a wide range of traits. *R. Maier, M. Robinson, P. Visscher, N. Wray.* The University of Queensland, Brisbane, QLD, Australia.

BACKGROUND: Polygenic risk prediction has emerged as a powerful research tool in human genetics in the past few years. Typically, summary statistics from genome-wide association studies (GWAS) are used to predict phenotypes by creating a weighted sum of the SNP effects. While prediction accuracy, measured as the amount of trait variation explained, has an upper bound that is set by the heritability captured by genetic markers, current polygenic risk scores achieve far from this value. Recent studies show that pleiotropy is widespread, with evidence for genetic correlations between even seemingly unrelated diseases. This provides an opportunity to improve the accuracy of polygenic risk predictors by appropriately weighting information from independent but correlated data sets. METHODS: We have previously shown that the joint analysis of five genetically correlated psychiatric traits can lead to an effective increase in sample size and thus to increased accuracy of a genetic predictor. The method we developed required full individual-level genotype data for all traits, which is likely to be a limiting factor for many analyses if multiple traits are to be utilized in parallel. Here, we present a novel computationally inexpensive method for multivariate risk prediction, which only requires GWAS summary statistics. RESULTS: We show through theory, simulation and application to psychiatric traits that our new method can yield similar increases in prediction accuracy as an analysis where full individual-level data are available. We further apply this method to 45 diseases and morphological traits for which summary statistics are publicly available. Validation of these predictors in data from the UK biobank shows that in several traits the multi-trait predictor outperforms any single-trait predictor. CONCLUSION: Our results indicate that prediction accuracy of a multi-trait predictor can be higher than that of a single-trait predictor, especially if the additional traits in the multi-trait predictor have high heritability, large sample size and a high genetic correlation with the predicted trait.

124

Building the human wiring diagram from genome-phenome associations across 525 studies, 300 phenotypes and 2.5 million individuals. *P. Donnelly, G. McVean, W. Ali, R. Davies, T. Down, L. Curren, J. Faria, J. Floyd, C. Franklin, J. Hall, L. Jostins, Y. Li, J. Maller, M. Pirinen, H. Taylor, C. Vangjeli, H. Wilman, S. Wilder, M.E. Weale, S. Myers, G. Lunter, M. Simpson, C. Spencer.* Genomics plc, Oxford, United Kingdom.

Large-scale analyses of naturally occurring genetic variation in humans have revealed many specific insights into the genetic and functional architecture of complex disease. However, there has been less progress in combining such data to build an integrated causal map linking genetic perturbation to its impact on molecules, cells, physiology and health. In part, this has been hampered by the challenges of assembling, harmonizing and analysing such large and heterogeneous data sources. We have developed and applied scalable approaches to integrate and then jointly analyse genome-wide association studies currently spanning over 300 distinct molecular and disease phenotypes, at an integrated set of over 7 million common variants, across 525 studies, which together include more than 2.5 million participants. Of the variants with a minor allele frequency of at least 5% in European populations, approximately 2% show compelling evidence of association to at least one phenotype and, of these, 10% show evidence of association to at least two phenotypes from different phenotypic areas. The application of new methods for deconvoluting these association signals suggests that they reflect at least 1,500 independent causal mutations. We have also developed novel methods to combine information from common and rare genetic variants. Using Bayesian approaches to analyse the phenome-wide impact of protein truncating variants and also to link genetic variants to the function of nearby genes, we can assess the phenotypic consequences of genetic perturbations to over 5,000 genes. Direct knowledge, in humans, of the consequences of these perturbations is a powerful tool for improving drug development pipelines. We show that the patterns of association across phenotypes can be classified into a relatively small number of distinct structures reflecting fundamental pathways affecting complex disease. Our approach provides a framework for untangling the human wiring diagram, which will become even more powerful with increasingly deep information from imaging, cellular phenotyping and electronic medical records.

125

***De-novo* reconstruction of more than 6,000 pedigrees discovered from 51K de-identified exomes within the DiscovEHR cohort.**

J. Staples[1], E.K. Maxwell[1], C. Gonzaga-Jauregui[1], I.B. Borecki[1], M.F. Murray[2], D. Carey[2], F.E. Dewey[1], O. Gottesman[1], L. Habegger[1], J.G. Reid[1], Geisinger-Regeneron DiscovEHR. 1) Regeneron Genetics Center, Regeneron Pharmaceuticals Inc., Tarrytown, NY, USA; 2) Geisinger Health System, Danville, PA, USA.

Pedigrees and family-based analyses have been moving back to the forefront of human genetics. However, many of the large-scale sequencing initiatives planned and underway are ascertaining and sequencing hundreds of thousands of de-identified individuals without the ability to obtain accurate family history and pedigree records, precluding many powerful family-based analyses. This is the case for the ~51,000 individuals that have been whole exome sequenced as part of the DiscovEHR collaboration. We demonstrate that we can infer tens of thousands of close familial relationships within the cohort and reconstruct the corresponding pedigrees directly from the genetic data, identifying many familial relationships that can be used for downstream genotype-phenotype analyses enabling both population and family-based analytical approaches. Using PLINK to estimate genome-wide IBD proportions between all individuals in the DiscovEHR cohort, we found that >48% of the individuals were involved in one or more of the ~5,000 full-sibling relationship, ~7,000 parent-child relationships, and ~15,000 second-degree relationships. Subsequently, we reconstructed >6,000 pedigrees containing two or more sequenced individuals using PRIMUS. The largest extended family we identified contains >3,000 individuals (~6% of the dataset). We have also identified 825 nuclear families, containing 948 trios, allowing us to perform a rich set of trio-based analyses. These trios have aided in improving CNV calling, phasing compound heterozygous mutations, and validating rare variant calls. We have used this rich resource of reconstructed pedigree data to distinguish between novel/rare population variation and familial variants and have leveraged it to identify highly penetrant disease variants segregating in families that are underappreciated in population-wide association analyses. We have validated this approach by identifying related individuals segregating highly penetrant Mendelian disease causing variants causing, among other examples, familial aortic aneurysms, cardiac conduction defects, thyroid cancer, pigmentary glaucoma, and familial hypercholesterolemia, including a large pedigree containing 29 related individuals carrying a novel familial hypercholesterolemia-causing tandem duplication in *LDLR*.

126

Insights from expanded carrier screening of >54,000 individuals by next-generation sequencing.

A.H. Birch, G. Cai, G. Mendiratta-Vij, L. Shi, X. Cai, O. Birsoy, S. Sperber, J. Liao, B. Webb, L. Elkhoury, J. McCarthy, M. Dillon, S. Van Den Berg, M. Delio, G. Diaz, F. Suer, R. Kornreich, L. Edelmann. Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

Prenatal carrier screening allows couples without a known family history of disease to be screened for variants causative for autosomal recessive and X-linked genetic diseases for the purpose of reproductive planning. It has traditionally focused on testing the most common disease-causing variants, as well as ethnic-specific founder mutations. However, many people are of mixed ethnicity or cannot accurately identify their full ethnic background. Additionally, testing for specific mutations will miss other pathogenic variants in the gene. Our expanded carrier screening panel, launched in May 2015, has tested over 54,000 individuals for up to 281 genes performed mainly by next-generation sequencing (NGS). Depending on the size of the panel ordered, between 4% and 65% of individuals test positive for at least one disease. In comparison to our previous carrier screening assay, which was performed largely by genotyping of variants in 111 genes, we determined that 56.4% of reportable variants would have been missed by genotyping. Even among individuals self-identifying as 100% Ashkenazi Jewish, approximately 7.2% of reportable variants would have been missed (range: 0% - 29.6% per gene). To date, 58 carrier couples have been identified from 5922 couples internal to Mount Sinai, giving a carrier couple frequency of 1/102. When the 35 female carriers of X-linked diseases, including Fragile X premutations, are added, the risk of a couple being carriers for a disease on our panel is 1/64. Carrier testing may also result in the identification of individuals who may be mildly affected with genetic disease but unaware of their status. To date, we have identified 87 confirmed compound heterozygotes and homozygotes, including 45 who are expected to manifest symptoms of disease and 42 who may or may not exhibit symptoms, based on the presence of at least one allele with reduced penetrance. Additionally, 2946 males have been tested for up to 21 X-linked diseases. The rationale for testing males is that additional genetic factors, including mosaicism or sex chromosome abnormalities, may conceal the effect of a pathogenic variant. Currently, only one male has been determined to carry a reportable variant in an X-linked gene. The results of the current study underscore the benefits of NGS in carrier screening and the potential for complications, including uncovering underlying disease status in some patients.

127

Improving biobank consent comprehension: A national randomized survey to assess the effect of a simplified form and review/retest intervention. *L.M. Beskow, L. Lin, C.B. Dombeck, E. Gao, K.P. Weinfurt.* Duke Clinical Research Institute, Duke University, Durham, NC.

Background: Proposed changes to federal research regulations include new requirements that consent for research use be obtained for nearly all biospecimens. However, informed consent is beset with challenges and the evidence base for interventions to improve it is limited by methodologic issues. To help address these gaps, we conducted a national online survey to determine the effect of a simplified consent form on biobanking consent comprehension. We hypothesized that 1) comprehension given the simplified form would be no worse than that given a traditional form; and 2) among respondents with the least education, comprehension would be better given the simplified form. We further examined the effect of a review/retest intervention on comprehension. Methods: Recruited from GfK's KnowledgePanel (n=1916), participants were randomly assigned to read a simplified or traditional consent form for a hypothetical biobank. Participants then completed a 21-item quiz developed based on a Delphi process to identify the information participants must grasp to give valid consent. For items answered incorrectly, participants were shown the corresponding consent form section and then presented with another quiz item on that topic. Results: Consistent with our first hypothesis, comprehension among those who received the simplified form was not inferior to those who received the traditional form. Contrary to our second hypothesis, comprehension among those with the least education did not differ significantly by consent form. Only 20% of our sample achieved a perfect score—adequate understanding—upon first quiz attempt. Among those who underwent review/retesting, quiz scores improved significantly in every combination of consent form and education level. Ultimately, 65% of the entire sample achieved a perfect score. Discussion: Our results support calls to simplify and shorten consent forms. Even so, these steps may not be sufficient for achieving adequate comprehension in all population groups, and research on other interventions is warranted. Our results also support the use of comprehension assessments to improve understanding. However, they raise profound questions about whether to set a threshold for understanding and the consequences for failure to meet it. Conclusion: Our results contribute to the evolution of effective approaches to biobanking consent, and inform efforts to empirically study and improve informed consent more broadly.

128

Large-scale genomic analyses identify one fifth of the heritable component of puberty timing and widespread non-linear associations. *F. Day on behalf of the ReproGen Consortium.* MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom.

Introduction: The timing of puberty (captured by age at menarche in females) is a heritable trait, which is regulated by diverse biological processes and is related to many adult health outcomes. Methods: Genome-wide array data, on variants with minor allele frequency>=0.1% imputed to the 1000G reference panel, were analysed in up to 329,345 women of European ancestry from: the ReproGen consortium (N=179,117), 23andMe (N=76,831) or the UK Biobank study (N=73,397). Results: 37,925 variants reached the genome-wide significance threshold ($P<5x10^{-8}$) for association with age at menarche as a continuous outcome; these resolved to 389 independent signals, including 26 indel polymorphisms and 12 signals on the X-chromosome. In combination, the 377 autosomal loci explained ~7.2% of the trait variance in UK Biobank, accounting for >20% of the trait heritability. Amongst the identified signals, 33 genes were implicated by non-synonymous variant mapping, including the reproductive axis neurotransmitter encoding genes: kisspeptin (*KISS1*) and gonadotropin-releasing hormone (*GNRH1*). Across the genome there was strong enrichment of signals around genes active in multiple neuronal tissues, and systematic eQTL integration identified >100 genes with expression regulated by age at menarche signals. We identified widespread non-linear genetic effects on puberty timing, exemplified by higher heritability estimates for "Early-normal menarche" (28.8%; age 8-11 inclusive, N=14,922) than for "Late-normal menarche" (21.5%; age 15-19, N=12,290; P_{diff}=0.03). Similarly, 217/377 genome-wide significant signals had larger effect estimates on "Early-normal" compared to "Late-normal" menarche (each compared to the median age 13 group; binomial P=0.004). Notably, variants near to the *RGS4* gene, which encodes a regulator of G-protein signalling, were associated only with Early-normal menarche. Conclusions: These findings increase by more than 3-fold the number of genomic loci associated with puberty timing, double the explained heritability, and identify several novel mechanisms. The observed larger heritability of Early-normal versus Late-normal menarche timing could indicate interactions between genes and age-related environmental factors.

129

A large genome-wide study of age-related hearing impairment using electronic health records. *T.J. Hoffmann[1], B.J. Keats[2], N. Yoshikawa[3], C. Schaefer[4], N. Risch[1,4], L.R. Lustig[5].* 1) Department of Epidemiology and Biostatistics, and Institute for Human Genetics, University of California San Francisco, San Francisco, CA; 2) Department of Genetics, Louisiana State University Health Sciences Center, New Orleans, LA; 3) Department of Head and Neck Surgery, Kaiser Permanente Medical Center, Oakland, CA; 4) Kaiser Permanente Northern California Division of Research, Oakland, CA; 5) Department of Otolaryngology - Head and Neck Surgery, Columbia University Medical Center, Columbia, NY.

We conducted a genome-wide association study of age-related hearing impairment in 6,527 non-Hispanic white Genetic Epidemiology Resource on Adult Health and Aging (GERA) cases and 45,882 non-cases, and identified two common novel genome-wide significant SNPs. The first was rs4932196 (GERA non-Hispanic white hazards ratio 1.15, 95% CI=1.10-1.20, p=1.5x10^{-10}), 52 Kb 3' of *ISG20*, that replicated in a meta-analysis of the other GERA race/ethnicity groups (1,025 Latino, East Asian, and African American cases, 12,385 non-cases; p=0.00012) and p=0.034 in the UK Biobank (30,802 self-reported cases, and 78,586 controls). The second was rs58389158 (GERA non-Hispanic white hazards ratio 1.11, p=4.0x10^{-8}), that replicated in the UK Biobank (p=0.00042) and had p=0.43 in the other GERA race/ethnicity groups (effect size in the same direction). The SNP was just outside exon 7 of *TRIOBP*, and was highly correlated with rs5756795 (r^2=0.96, GERA non-Hispanic white p=4.4x10^{-8}), a missense mutation in exon 7 of *TRIOBP*, a gene previously shown to be associated with prelingual nonsyndromic hearing loss and expressed in human cochlear tissue. We further tested these SNPs in phenotypes from audiologist notes available on a subset of the cohort (4,903 GERA individuals), stratified by case/non-case status to construct an independent test, and found that rs58389158 further replicated in speech reception threshold (overall GERA meta-analysis p=3.8x10^{-6}). These results suggest that large cohorts with genome-wide data and electronic health records may be a powerful method to begin to characterize the genetic architecture of age-related hearing impairment.

130

Identifying disease-causing genes that act through complementary modes of regulatory elements and protein altering variants in DNA samples linked to electronic medical records. *X. Zhong, Q. Wei, R. Chen, Q. Wang, N.J. Cox, B. Li.* Vanderbilt Genetic Institute, Nashville, TN.

It is clear that both regulatory and coding variants contribute to risk of human diseases. Vast majority of GWAS-identified variants are noncoding and exert their roles in disease predisposition through disrupting regulation of the target genes; protein coding variants, on the other hand, influence disease risk through a direct change in the protein sequence. It is largely unknown how coding and noncoding variants predispose disease risk mechanistically. We hypothesized that a disease-causing gene increases disease risk through two complementary modes: disrupting either the protein coding or the regulation of the gene. If proven true, it provides a principle to unify both signals through gene-based approaches for a better power. To assess to what extent this hypothesis holds true, here, we performed a systematic, phenome-wide scan of genes for rare nonsynonymous coding as well as common regulatory non-coding variants, to search for gene-disease pairs that are concordant in associations of both coding and regulatory variants. The effects of regulatory variants were aggregated into their target gene's genetically-regulated gene expression (GReX) using predixcan. We carried out phenome scans in BioVU, a Vanderbilt BioBank linked to medical records, for both GReX and rare coding variants. Specifically, for GReX we used ~10,000 patients with GWAS data, and for coding variants we used ~25,500 samples with exome-chip data, all of European descent. In total, we detected 2180 gene-disease pairs showing concordant phenotype associations for coding and non-coding signals, with ~400 genes associated with multiple phenotypes (e.g., *CD300A, KIF19, EP300* etc.). In particular, the identified diseases are predominantly associated with reduced predicted gene expression, consistent with the association of coding variants, which showed increased risk of diseases. For example, we observed that reduced expression of *CD300A* is associated with 16 medical phenotypes, including hypopotassemia (p=2e-23), respiratory failure (p=5e-20), renal failure (p=7e-10), pulmonary congestion (p=9e-10), paroxysmal ventricular tachycardia (pvalue=2e-9) to name a few, all of which are also associated with the same nonsynonymous coding variant, pointing to a potential role of *CD300A* in immune response. Altogether our analyses supported the complementary roles of coding and noncoding variants in disease susceptibility and implicated the genes and the associated diseases involving such a mechanism.

131

Thousands of novel variants influencing human blood cell variation and function: Insights from UK Biobank and INTERVAL. *H. Elding[1], W.J. Astle[2,3,4], T. Jiang[5], D. Allen[6], D.J. Roberts[6,7,8], W.H. Owehand[1,2,9,10], J. Danesh[1,5,9,10,11], A.S. Butterworth[5,9,10], N. Soranzo[1,2,9,10].* 1) Human Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom; 2) Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge, CB2 0PT, UK; 3) National Health Service (NHS) Blood and Transplant, Cambridge Biomedical Campus, Long Road, Cambridge, CB2 0PT, UK; 4) Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Cambridge Biomedical Campus, Forvie Site, Robinson Way, Cambridge CB2 0SR, UK; 5) Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, Cambridge, CB1 8RN, UK; 6) Blood Research Group, NHS Blood and Transplant, John Radcliffe Hospital, Headley Way, Headington, Oxford, OX3 9BQ, UK; 7) Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, UK; 8) Department of Haematology, Churchill Hospital, Headington, Oxford OX3 7LE, UK; 9) The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge, University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, Cambridge, CB1 8RN, UK; 10) British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0QQ, UK; 11) National Institute for Health Research Cambridge Biomedical Research Centre , Cambridge University Hospitals, Cambridge, CB2 0QQ, UK.

Mature blood cells are essential for oxygen transport, hemostasis and immune response. Understanding the genes and pathways that determine blood cell differentiation and survival is relevant to pathological processes in many fields of medicine. We describe here a large-scale genome-wide association study, testing associations of 29.5 million human genetic variants with Minor Allele Frequency (MAF) greater than 0.01% with 36 blood cell traits for red cells, white cells and platelets in 173,480 European-ancestry participants. Overall, ~200,000 variants showed significant association with at least one blood cell trait and using multivariate conditional analysis, we identify over 2,000 conditionally independent associations, increasing the total number of known variants by more than 10-fold. These novel associations include over 300 low frequency (1%<MAF<5%) and rare variants (MAF <1%). Several loci harbor multiple common and rare independent variants, some with discordant direction of effects reflecting an underlying complex allelic architecture and diverse haplotype structure. Using data from the BLUEPRINT epigenome project we show co-localisation of GWAS variants with expression and histone Quantitative Trait Loci (QTLs) as well as enrichment of variants and genes within cell-type specific regulatory elements. Moreover, we also demonstrate a causal effect of blood trait variation with autoimmune and cardiovascular disease via Mendelian Randomisation analysis. Significant enrichments of the corresponding gene sets within relevant biological pathways, rare hematological and immune disorders, mouse phenotypes, and human complex disease confirm their importance for the understanding of hematopoietic processes and pathogenesis.

132

Decoding 51 thousand individuals to analyze the most common genetic disorder: Hereditary hemochromatosis (HH). *N.S. Kip[1], M. Corbali[1,3], R. Metpally[4], H. Williams[5], S. Krishnamurthy[4], H. Harrison[1], D. Carey[4], M. Williams[4], A. Baras[2], J. Overton[2].* 1) Laboratory Medicine and Pathology, Geisinger Health System, Danville, PA; 2) Regeneron Genetics Center, Tarrytown, NY; 3) Cerrahpasa Medical Faculty, Istanbul, Istanbul, Turkey; 4) Weis Center for Research, Geisinger Health System, Danville, PA; 5) Urology, Geisinger Health System, Danville, PA.

Background: Hereditary hemochromatosis (HH) is the most common single gene disorder which if untreated causes excess iron deposition leading to cirrhosis, cancer, diabetes, and heart disease. HH most commonly results from recurrent mutations in the *HFE* gene (C282Y, H63D and S65C). Many patients are asymptomatic prior to late stage disease leading to underdiagnosis. Penetrance of clinical disease is also low even in patients homozygous for the severe C282Y mutation. The ACMG guideline recommends against routine screening for HH, however as NGS becomes more common, a “Genome First” assessment of *HFE* variants could identify at risk individuals prior to end stage disease. This study tests this approach. Materials/Methods: Whole exome sequencing was performed using DNA extracted from blood of 51,289 participants. We determined *HFE* gene alterations to analyze the frequency of the 3 common mutations, identify novel variants, and assess genotype-phenotype correlations by comparing pertinent clinical/laboratory findings extracted from the electronic health record (EHR). Results: Prevalence of *HFE* genotypes was as follows: H63D/WT (23%), C282Y/WT (8.9%), H63D/H63D (2.2%), S65C/WT (2.2%), C282Y/H63D (1.8%), H63D/S65C (0.4%), C282Y/C282Y (0.3%), C282Y/S65C (0.2%), S65C/S65C (0.02%). Significant differences were noted for serum iron (SI), ferritin, transferrin saturation (TS) and iron binding capacity (IBC) in C282Y/C282Y, C282Y/H63D, C282Y/WT, H63D/H63D and H63D/WT genotypes in both sexes, compared to the WT group. Median AST was significantly higher in males with C282Y/S65C, whereas, ALT was higher in females with H63D/S65C compared to WT. Homozygosity for C282Y was associated with significantly higher cirrhosis risk in both sexes. Cardiac arrhythmia was significantly higher in males with C282Y/C282Y and H63D/H63D, and female H63D carriers, whereas, congestive heart failure was significant in only H63D/H63D females. Bronze diabetes was significantly increased in males carrying C282Y/C282Y, C282Y/H63D, C282Y/S65C, H63D/H63D genotypes; and in females in all genotypes except C282Y/S65C. Various novel mutations were identified correlating with abnormal serum iron indices. Conclusion: This large scale EHR-linked genomic study indicating that the burden of disease associated with *HFE* variants may be underappreciated, may help develop screening guidelines based on genotype, for those undergoing sequencing for other indications, to improve recognition and management of HH.

133

Prevalence of mutations in high/moderate risk cancer genes in 3,488 patients tested using a uniform NGS cancer panel. *G.E. Tiller[1], M. Alvarado[2], J. Goff[2], R. Haque[3].* 1) Dept Genetics, Kaiser Permanente, Los Angeles, CA; 2) Dept Genetics, Kaiser Permanente, Pasadena, CA; 3) Dept Research & Evaluation, Kaiser Permanente, Pasadena, CA.

Using a multigene cancer panel, we tested 3,488 patients (3,240 women and 248 men) whose clinical presentations and/or family histories suggested the possibility of a familial cancer syndrome with more than one candidate gene. All patients received pre-test counseling by a genetic counselor or clinical geneticist. Over 90% of patients tested were referred due to a personal and/or family history of breast or ovarian cancer (HBOC). 68% of patients had both a personal and family history of cancer (PH+FH), 22% had a family history only (FH), and 8% a personal history only (PH). The age range was 15-93 years. Ethnicity distribution was as follows: 33% European, 25% Latino/Hispanic, 14% other/mixed, 11% Asian, 7% African-American, 6% Ashkenazi Jewish, and 5% unknown. All patients underwent testing using the same 20-gene NGS cancer panel, which also employed microarray or MLPA to detect deletions and duplications. We identified 453 pathogenic/likely pathogenic variants (PV/LPV) in 434 patients (12.4 %) in the following genes: BRCA1 (99), BRCA2 (80), MUTYH (71), CHEK2 (57), ATM (41), PALB2 (26), PMS2 (17), MSH6 (12), APC (11), TP53 (10), MSH2 (9), MLH1 (8), PTEN (3), CDKN2A (2), VHL (2), and one mutation in each of the following genes: BMPR1A, CDH1, EPCAM, SMAD4, STK11. Rates of PV/LPV by cancer history were 14.5% for PH+FH, 8.9% for PH, and 8.5% for FH. 26/307 (8.5%) patients with previously negative single gene tests (298 BRCA gene tests and 12 other tests) tested positive for at least one PV/LPV using our 20-gene panel. Latino/Hispanic patients accounted for 31% of BRCA1/2 mutations but comprised only 25% of the overall cohort. 22 patients had two or more PV/LPV results, including 3 patients with biallelic MUTYH mutations. No mutations were detected in 1,994 patients (57%); at least one variant of unknown clinical significance (VUS) was detected in 1,052 patients (30%). The frequency of VUS varied by ethnic group. Our cohort represents one of the largest and most ethnically diverse groups of patients tested using the same multigene cancer panel at a single institution. Although the majority of patients were referred for personal and/or family history of HBOC, 46% (200 patients) of those with a deleterious mutation representing 5.7% of our overall cohort had PV/LPV in genes other than BRCA1/2 and MUTYH. This underscores that testing via multigene cancer panels can offer clinically actionable results beyond single gene testing.

134

Parental perspectives on whole exome sequencing in pediatric cancer: A typology of perceived utility. *J. Malek[1], M.J. Slashinski[2], J.O. Robinson[1], A.M. Gutierrez[1], D.W. Parsons[3], S.E. Plon[3], A.L. McGuire[1], L.B. McCullough[1].* 1) Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX; 2) Department of Community Health Education, School of Public Health and Health Sciences, University of Massachusetts at Amherst, Amherst, MA; 3) Department of Pediatrics, Texas Children's Hospital, Houston, TX.

Background The scientific community’s perspective on the utility of genomic information is evolving. Recently, the definition of utility has narrowed to focus on evidence-based changes to a patient’s treatment plan. ASHG has critiqued this trend for excluding broader benefits for the patient and for others, such as use in reproductive decision-making and the value of a clear diagnosis even if findings are not medically actionable. We explored how parents of children with cancer perceive the value of whole exome sequencing (WES) results and assessed the degree to which that perception is in line with current thinking about utility. **Methods** Baylor Advancing Sequencing into Childhood Cancer Care (BASIC3) is a clinical sequencing project designed to evaluate the impact of integrating information from germline and tumor WES into the care of newly diagnosed pediatric cancer patients. Semi-structured interviews were conducted with parents of enrolled patients at three time points: before the disclosure of WES results (n=64), one to eight months after disclosure (n=33), and around one year after disclosure (n=25). The interviews were transcribed and analyzed using an inductive qualitative method to identify themes. **Results** Parents identified a broad range of ways in which they found or expected to find their child’s WES results valuable. They expressed optimism about the clinical utility of this information, even beyond what was scientifically reasonable in some cases. Parents did not limit their assessment of utility to its value for their child with cancer; they routinely cited benefits for themselves, the patient’s siblings, and other relatives. Finally, many parents reported experiencing or expecting to experience psychological utility including peace of mind, relief of guilt, and the satisfaction of curiosity as well as pragmatic utility such as the ability to plan for the future and to make better reproductive decisions. **Conclusion** These results demonstrate that parents’ perception of the value of WES results goes well beyond the narrow definition of clinical utility that has been gaining support in recent years. This finding raises ethical questions about: 1) which types of utility are sufficient to justify a recommendation for genomic sequencing and 2) the way in which parents of children with serious diseases make decisions about the use of sequencing technology.

135

Combining linked read technology with standard target-enrichment NGS can accurately resolve short reads and distinguish variants in the Lynch/CMMRD syndrome gene, *PMS2*, from its pseudogene, *PMS2CL*. *C. Kao[1], R. Pellegrino[1], F. Mafra[1], J. Garifallou[1], C. Kaminski[1], F. Wang[1], L. Tian[1], S. Garcia[2], R. Mao[3], W. Samowitz[3], C. Vance[3], C. Vaughn[3], S. Wenzel[4], K. Wimmer[4], H. Hakonarson[1].* 1) Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA; 2) 10X Genomics, Inc., Pleasanton, CA; 3) ARUP Laboratories, Salt Lake City, UT; 4) Division of Human Genetics, Dept. for Medical Genetics, Molecular and Clinical Pharmacology Medical University, Innsbruck, Innsbruck, Austria.

Lynch syndrome is an autosomal dominant multicancer susceptibility syndrome arising from heterozygous germline mutations inactivating one of the following mismatch repair (MMR) genes: *MLH1, MSH2, MHS6,* or *PMS2.* Biallelic germline inactivation of one of these genes results in constitutional mismatch repair deficiency (CMMRD). Genetic testing of MMR genes in individuals with a Lynch/CMMRD-associated cancer and/or a family history of such cancers is important for risk assessment, surveillance, and cancer management. Screening of *PMS2* is particularly challenging due to the presence of 15 different pseudogenes, the most problematic of which is *PMS2CL*, homologous to exons 9, 11-15 and located ~0.7Mb centromeric to *PMS2* in an inverted orientation. Sequence exchange between *PMS2* and *PMS2CL* results in hybrid *PMS2/PMS2CL* alleles*,* which creates difficulties for next-generation sequencing (NGS) workflows that reconstruct an individual genome using short reads (~200-400 bases) aligned back to a reference map. Currently, reliable genetic screening of *PMS2* involves combining several customized non-NGS assays that require specialized laboratory know-how with consequences to cost and accessibility. We demonstrate here a cost-effective, bead-in-an-emulsion partitioning method compatible with standard target enrichment and short-read NGS that can accurately resolve variants in *PMS2* versus *PMS2CL*. Input DNA (1.2ng) was partitioned using a Chromium instrument (10X Genomics, Inc.), where fragments (~50-100kb) from ~150 diploid genome equivalents are segregated into ~1M different gel beads in emulsion (GEMs); each GEM contains ~5-10 different input fragments and creates from them a library of smaller complementary molecules tagged with a 16-mer "barcode" unique to each GEM. Barcoded molecules across all GEMs were pooled and underwent whole-exome enrichment using SureSelect baits (Agilent Technologies, Inc.), then sequenced on a HiSeq2500 (Illumina, Inc.). The barcodes provide positional information linking short reads to the original input DNA fragments captured by a specific GEM. We show that these "linked reads" (or "synthetic long reads") can effectively detect *PMS2* variants (point mutations as well as structural rearrangements) within the *PMS2CL* paralogous region in a number of clinically validated "truth" samples, illustrating a strategy whereby standard NGS-based workflows can be adapted for use in Lynch/CMMRD syndrome screening across all known candidate genes.

136

Mismatch repair activity in non-neoplastic biallelic mismatch repair deficient cells: An explanation for the predominance of *PMS2* mutations and rapid diagnosis. *A.Y. Shuen[1,2], S. Lanni[1], Y. Lisa[1], G.B. Panigrahi[1], S. Deshmukh[1,3], B. Campbell[1,3,5], N. Zhukova[4,5], N. Thakkar[1,2], D. Malkin[1,3,4], U. Tabori[1,3,4,5], C.E. Pearson[1,2].* 1) Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada; 2) Department of Molecular Genetics, University of Toronto. Toronto, ON, Canada; 3) Institute of Medical Science, University of Toronto, Toronto, ON, Canada; 4) Division of Hematology/Oncology, The Hospital for Sick Children, Toronto, ON, Canada; 5) Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, ON, Canada.

Biallelic mismatch repair deficiency syndrome (bMMRD) is a childhood cancer syndrome causing hematological, brain and gastrointestinal tumors due to biallelic mutations in DNA mismatch repair (MMR) genes. In contrast, monoallelic mutations in MMR genes causes Lynch syndrome (Hereditary Non-polyposis Colorectal Cancer), predisposing to adult cancers due to somatic inactivation of the non-mutant allele. Improper diagnosis of bMMRD leads to inappropriate treatment paths with deleterious consequences. Diagnosis of bMMRD is hampered by clinical overlap with other more common cancer syndromes such as Li-Fraumeni and NF1. In addition, genetic testing for bMMRD is limited by variants of unknown significance and the strong bias towards *PMS2* mutations - a gene that is difficult to sequence due to its 16 kb size and the presence of ≥15 *PMS2* pseudogenes. A rapid functional assay that could diagnose bMMRD would be extremely beneficial for cancer management and prevention. We previously reported that non-neoplastic lymphoblastoid cells from a bMMRD patient lacked MMR activity using an *in vitro* functional assay (2015, *Nat Genet,* **47**: 257–262). We hypothesized that direct assessment of MMR activity would be diagnostic for bMMRD. We also hypothesized that there could be differential repair activities between deficiencies of the various MMR proteins. This might explain the skewing in mutation prevalence between bMMRD (with biallelic mutations in *PMS2* being most frequent, followed by *MSH6*, *MLH1* and rarely *MSH2*) and Lynch syndrome (skewed in reverse mostly *MSH2*, followed by *MLH1*, *MSH6* and lastly *PMS2*). To test our hypotheses, we assessed the MMR activity of non-neoplastic lymphoblastoid cells of 22 patients with bMMRD defective for *PMS2*, *MSH6*, *MLH1*, or *MSH2* and 10 of their 1st degree relatives. As might be expected, bMMRD cells with *MSH6-/-* or *MSH2-/-* mutations were completely repair-deficient. Strikingly, residual DNA repair activity (< 30% of wild type) was evident in bMMRD *PMS2-/-* or *MLH1-/-* cells. The residual DNA repair activity in *PMS2-/-* bMMRD individuals logically explains the higher genetic prevalence of *PMS2* mutations in bMMRD; allowing for sufficient genome integrity for survival, yet not free of cancer susceptibility. The assay we have developed could be a rapid and useful diagnostic indicator, permitting early cancer management decisions.

137

***PMS2CL*-hybrid alleles containing *PMS2* sequence and other *PMS2CL*-derived large rearrangements: The importance of correct interpretation of dosage alteration analysis in *PMS2*.** *N. Singh, D. Mancini-DiNardo, B. Leclair, K. Brown, E. Goossen, K. Bowles, B. Roa, M. Jones.* Myriad Genetics Laboratories, Inc., Salt Lake City, Ut, UT.

Background: Approximately 15% of Hereditary Non-Polyposis Colorectal Cancer (HNPCC or Lynch Syndrome) is caused by heterozygous mutations in *PMS2*, with 27% of these mutations categorized as large rearrangements (LRs). The interpretation of putative LRs detected in *PMS2* is confounded by frequent and sometimes extensive sequence exchange between *PMS2* and the pseudogene *PMS2CL* within exon 9 and exons 11 through 15. We have previously observed that ~40% of LRs in *PMS2* occur in exons where this gene conversion is common. Although multiplex ligation-dependent probe amplification (MLPA) is commonly used for LR analysis of *PMS2*, additional confirmatory analyses are necessary to determine whether LRs are specific to the functional gene. Here we demonstrate the necessity of confirmatory analyses to differentiate clinically significant alterations in *PMS2* from alterations in *PMS2CL*, which are not. **Methods:** Three representative examples of apparent LRs specific to exon 9, exon 11 and exons 13-14 in *PMS2* detected by MLPA are described. These LRs were identified during hereditary cancer testing using a 25-gene panel. As part of the panel test, all apparent LRs were further investigated with Sanger sequencing using gene-specific and pseudogene-specific primers and/or long-range PCR analysis specific to *PMS2* and *PMS2CL*. **Results:** In the first two cases, MLPA indicated a duplication in *PMS2* (one in exon 9 and one in exon 11). Confirmatory Sanger sequencing determined that a gene conversion event occurred, which transferred a portion of the sequence from the *PMS2* gene onto *PMS2CL*. This allowed binding of *PMS2*-specific MLPA probes onto the *PMS2CL*-hybrid allele, which resulted in the artifactual appearance of a duplication in *PMS2* for both cases. Long-range PCR specific to *PMS2* determined that the active gene was not duplicated in these cases. For the third case, MLPA indicated a deletion of exons 13-14 in *PMS2*. Long-range PCR specific to *PMS2CL* and *PMS2* determined that exons 13-14 were only deleted in the pseudogene. The LRs discussed here have been identified in a total of 342 patients, showing the importance of confirmatory analyses for apparent LRs detected in *PMS2*. **Conclusions:** The implementation of a multi-step confirmation strategy allowed for the correct assessment of *PMS2* dosage in the cases described above. Accurate interpretation and reporting of *PMS2* is imperative for appropriate medical management and in determining whether testing is recommended for family members.

138

ENIGMA quantitative and qualitative classification criteria for evaluating the clinical significance of *BRCA1* and *BRCA2* sequence variants. *A.B. Spurdle[1], M.A. Parsons[1], F. Couch[2], M. de la Hoya[3], S. Domchek[4], D. Eccles[5], E. Gomez-Garcia[6], C. Houdayer[7], A. Mensenkamp[8], A. Monteiro[9], P. Radice[10], M. Southey[11], S. Tavtigian[12], A. Toland[13], M. Vreeswijk[14], B. Wappenschmidt[15], D.E. Goldgar[12], ENIGMA collaborators.* 1) QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia; 2) Mayo Clinic, Minnesota, USA; 3) dISSC-biomedical research institute, Academic Hospital San Carlos, Madrid, Spain; 4) University of Pennsylvania, Philadelphia, USA; 5) Faculty of Medicine, University of Southampton, Southampton, UK; 6) Maastricht University Medical Center, Maastricht, The Netherlands; 7) Institut Curie and University Paris Descartes, Paris, France; 8) Radboud University Medical Center, Nijmegen, The Netherlands; 9) Moffitt Cancer Center, Tampa, USA; 10) Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy; 11) The University of Melbourne, Melbourne, Australia; 12) Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, USA; 13) The Ohio State University, Columbus, USA; 14) Leiden University Medical Center, Leiden, the Netherlands; 15) Center for Hereditary Breast and Ovarian Cancer, University Hospital Cologne, Cologne, Germany.

Genetic testing for germline variants in susceptibility genes for breast and other cancers frequently identifies gene variants of uncertain clinical significance, including missense, small in-frame insertions/deletions, splice and regulatory region variants. Unclassified variants are a major clinical challenge as they complicate test reporting and genetic counselling, and prevent guided clinical management of patients and their relatives. The ENIGMA (Evidence-based Network for the Interpretation of Germline Mutant Alleles) international consortium is undertaking research to improve methods for classification of variants in breast cancer predisposition genes. To promote standardised classification, ENIGMA has established detailed criteria to classify germline variants in *BRCA1* and *BRCA2* using a 5 tier system that reflects probability of pathogenicity. ENIGMA is a ClinGen-designated expert panel for *BRCA1/2* variant classifications, and ENIGMA classification criteria are currently being applied to variants identified by research and clinical testing sites internationally. ENIGMA expert panel and research activities to date have: • provided, for display in ClinVar, the evidence supporting existing robust classification of “not-obviously-truncating” variants as pathogenic or not-pathogenic • accrued clinical information from collaborators for quantitative multifactorial analysis of additional non-truncating sequence variants, with altered classification anticipated for >390 unique variants • shown that the qualitative criterion of 1% allele frequency used to classify a variant as “not pathogenic” is conservative and reliable • provided evidence that spliceogenicity and pathogenicity are not equivalent, and demonstrated that expert review of (likely) spliceogenic variants, including those at acceptor/donor “consensus” dinucleotides, aids variant interpretation. • provided statistical estimates from large-scale studies to utilize breast tumour pathology data for predicting variant pathogenicity • shown that use of multiple splicing bioinformatic tools does not significantly improve prediction of spliceogenicity of exonic variants, and developed schema to identify exonic variants that may disrupt mRNA splicing These findings demonstrate the value of large-scale international collaborations with gene and disease specific expertise to improve variant classification methods and processes, and deliver clinically meaningful standardised disease gene variant classification.

139

Return of incidental results for *BRCA1/BRCA2* to a 50,726 person cohort within a single healthcare provider organization. *M.F. Murray*[1], *K. Manickam*[1], *A.H Buchanan*[1], *D.M. Lindbuchler*[1], *M.L. Barr*[1], *A.L. Lazzeri*[1], *L.M. Gorgol*[1], *C.Z. McCormick*[1], *C.N. Flansburg*[1], *M. Hallquist*[1], *A.K. Rahm*[1], *A. Fan*[1], *W.A. Faucett*[1], *M.A. Giovanni*[1], *D.N. Hartzel*[1], *J.B. Leader*[1], *H.L. Kirchner*[1], *N.S. Abul-Husn*[2], *F.E. Dewey*[2], *R.P.R. Metpally*[1], *D.J. Carey*[1], *T.N. Person*[1], *M.D. Ritchie*[1], *D.H. Ledbetter*[1]. 1) Geisinger Health System, Danville, PA, USA; 2) Regeneron Genetics Center, Tarrytown, NY, USA.

The combined prevalence estimate for pathogenic *BRCA1* and *BRCA2* (*BRCA1/2*) variants in the general population is 1:400. There are significant limitations to our current clinical approaches for identifying individuals with the increased cancer risk associated with these two genes.   The MyCode Community Health Initiative recruits patient volunteers from across the Geisinger healthcare system; 50,726 adult volunteers underwent whole exome sequencing (WES) between September 2014 and September 2015. We established a process to: identify pathogenic/likely pathogenic (P/LP) *BRCA1/2* variants in individual participant WES, confirm variants by orthogonal methods in a CLIA-certified laboratory, communicate results to patients and their providers, insert results into the electronic health record (EHR), assist in establishing a management plan, and support cascade evaluation and testing of at-risk relatives.   Our bio-informatic analysis process identified 250 individuals (1:203) with P/LP variants in *BRCA1/2* with 148 (59%) *BRCA*2 compared to 102 (41%) *BRCA*1. There are 124 unique P/LP variants identified; P/LP was defined as either pLOF or a ClinVar 2 or 3 star assertion. The average age of those with P/LP variants is 57 (age range 22-89 yrs.) and includes 144 (58%) women.   Sixty adults had incidental *BRCA1/2* findings returned to the EHR as of 05/31/2016, and all eligible confirmed results will be delivered prior to 09/01/2016. Amongst the 60 returned result cases there were 32 women (age range 25-86 yrs.) and 28 men (26-83 yrs.); only 6 (10%) reported having BRCA testing prior to this testing. Nineteen (32%) had a personal history of any cancer, including 6 Breast Cancers (16% of women and 4% of men), 2 Ovarian Cancers (6% of women), and 1 Prostate Cancer (4% of men). Thirty-two (53%) had a family history of a relevant cancer in a first or second-degree relative. Two (3%) had died prior to result return neither with a BRCA-related cause. Scalable systems for identifying, communicating, and managing incidental genomic findings are possible. The empiric prevalence of pathogenic *BRCA1/2* in this cohort is double previous population estimates. Ninety percent of *BRCA1/2* variant carriers receiving results via our approach had not otherwise been identified in the course of routine care. More study is required to understand the *BRCA2* predominance in our cohort, as well as the health outcomes for individuals ascertained in this manner.

140

BRCA population screening in unaffected Ashkenazi Jewish women: A randomized controlled trial of different pre-test strategies. *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], *E. Levy-Lahad*[1,2]. 1) Medical Genetics Institute, Shaare Zedek Medical Center, Jerusalem, Israel; 2) Faculty of Medicine, Hebrew University, Jerusalem, Israel; 3) Department of Obstetrics&Gynecology, Shaare Zedek Medical Center, Jerusalem, Israel; 4) Clalit Health Services, Israel; 5) Breast Surgery Unit, Department 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.

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 such screening, given high mutation prevalence (1/40) and >95% sensitivity and specificity of testing for three common mutations. Towards screening implementation, we aim to examine the impact of excluding pre-test face-to-face 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 780 participants (mean age 46 years), we identified 11 carriers (1.4%). Only 3/11 carriers had significant family history. Post-testing, 95% of GC and 94% of WI participants were satisfied/very satisfied with testing (NS). 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 63% and 70% in GC vs. 56% and 65% in WI, at Q1 (p=.0004) and Q2 (p=.01) respectively. The difference in knowledge was 0.5/10 points at both Q1 (p=.0004) and Q2 (p=.04). Carriers had higher PPC and knowledge than non-carriers. At Q2, carriers' stress level was higher (15.4 vs. 5.3, p=.0006), as expected. Within GC, only PPC increased over time (from Q1 to Q2, p=.003), whereas within WI all outcomes improved over time, with greater satisfaction, lower IES and increased knowledge and PPC (p<.001). 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:** Screening using a streamlined process would identify substantially more carriers (regardless of family history) while addressing logistic and cost limitations. These ongoing results suggest that compared to WI, pre-test GC provides a mild, temporary, increase in knowledge, accompanied by a 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.

141

First genome-wide significant locus for pre-eclampsia susceptibility discovered on fetal chromosome 13 near *FLT1*.

R. McGinnis[1], V. Steinthorsdottir[2], N. Williams[1], V. Dolby[3], G. Thorleifsson[2], S. Shooter[1], L. Stefansdottir[2], J. Sigurdsson[2], A. Haugan[4], S. Chappell[5], T. Jääskeläinen[6], G. Silva[7], L.C. Vestrheim Thomsen[8], W.K. Lee[9], E. Staines-Urias[10], J. Kemp[11], F. Dudbridge[10], J.P. Casas[10], T. Hegay[12], N. Simpson[3], J. Walker[3], N. Zakhidova[12], D. Lawlor[11], G. Syvatova[13], D. Najmutdinova[14], P. Magnus[4], A.C. Iversen[7], H. Laivuori[6], L. Morgan[5], InterPregGen Consortium. 1) Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 2) deCODE genetics/Amgen, Reykjavik, Iceland; 3) Leeds Institute of Biomedical and Clinical Sciences, University of Leeds, United Kingdom; 4) Norwegian Institute of Public Health, Oslo, Norway; 5) School of Life Sciences, University of Nottingham, United Kingdom; 6) Haartman Institute, Medical Genetics, University of Helsinki, Finland; 7) Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Bergen, Norway; 8) Department of Clinical Science, University of Bergen, Norway; 9) BHF Glasgow Cardiovascular Research Centre, University of Glasgow, United Kingdom; 10) London School of Hygiene and Tropical Medicine and University College London, United Kingdom; 11) MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, United Kingdom; 12) Institute of Immunology, Uzbek Academy of Science, Tashkent, Uzbekistan; 13) Scientific Centre of Obstetrics, Gynaecology and Perinatology of Ministry of Health, Almaty, Kazakhstan; 14) Republican Specialized Scientific-Practical Medical Centre of Obstetrics and Gynecology, Tashkent, Uzbekistan.

Pre-eclampsia (PE) affects ~5% of pregnancies and causes about 25% of maternal deaths in pregnancies worldwide, making PE one of the two leading causes of death among pregnant women. PE also causes ~500,000 annual fetal deaths and millions of premature deliveries with the resulting burden for neonatal health. Despite strong evidence that PE susceptibility is inherited, genome-wide association studies (GWAS) of maternal PE cases have not identified any chromosomal loci responsible for PE that are genome-wide significant ($p<5\times10^{-8}$) or robustly replicate in independent datasets, perhaps due to inadequate sample sizes (<1000 cases) in published reports. Here we report the first GWAS of children of PE pregnancies (“fetal cases”) and the discovery of the first highly replicable, genome-wide significant locus for PE susceptibility. This chromosome 13 locus was initially discovered in GWAS meta-analysis of fetal cases of European descent from the InterPregGen Consortium (2512 cases, 302137 controls from deCODE and from the UK GOPEC and WTCCC Consortia) augmented by ALSPAC UK GWAS data (146 fetal cases, 6130 controls) yielding a pvalue of $p<3.4\times10^{-8}$ at rs4769613. The position of the locus near the gene Fms-like tyrosine kinase 1 (*FLT1*) provides important biological support since considerable evidence implicates a soluble isoform (sFLT1) as contributing to the pathology of PE by binding and abnormally lowering circulating placental growth factor (PlGF) and vascular endothelial growth factor (VEGF). Followup genotyping strongly confirmed the locus ($p<1.5\times10^{-4}$, rs4769613; $p<7.6\times10^{-5}$, rs7328374) in independent fetal cases from the Norwegian MoBa cohort (1154 cases, 1174 controls) and Finnish FINNPEC cohort (580 cases, 782 controls); and combined meta-analysis of GWAS and replication data provides very robust results ($p<4.7\times10^{-11}$ at rs4769613; $p<6.6\times10^{-11}$ at rs7328374). These SNPs appear to be in a regulatory region of *FLT1*. Furthermore, they also associate with red blood cell count, a downstream target of VEGF signalling. We will discuss these results in relation to key PE characteristics and categories such as late or early onset PE, and birthweight that is normal or small for gestational age; and we will also discuss functional investigation of the *FLT1* locus. Funding: EU FP7 grant 282540 to InterPregGen; Wellcome Trust grants (098051, WT090355/A/09/Z, WT090355/B/09/Z, WT088806, WT094529MA, WT087997MA); MRC (MC_UU_1203/5), Wellcome Trust/MRC(102215/2/13/2).

142

Genome-wide meta-analysis of polycystic ovary syndrome in women of European ancestry identifies novel loci.

T. Karaderi[1], C. Meun[2], T. Laisk-Podar[3,4], F.T.J. Lin[5], W. Wu[6], A. Mahajan[1], B.H. Mullin[7,8], M.R. Jones[9,10], F.R. Day[11] on behalf of the PCOS Consortium. 1) The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, OX3 7BN, Oxford, United Kingdom; 2) Division of Reproductive Medicine, Department of OBGYN, Erasmus MC, ‘s Gravendijkwal 230, 3000CA, Rotterdam; 3) Women’s Clinic, University of Tartu, Tartu 51014, Estonia; 4) Competence Centre on Health Technologies, Tartu 50410, Estonia; 5) Division of Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA, 60611; 6) Department of Epidemiology, Indiana University Richard M. Fairbanks School of Public Health; 7) Department of Twin Research & Genetic Epidemiology, King’s College London, London, UK; 8) School of Medicine and Pharmacology, University of Western Australia, Nedlands, Western Australia; 9) Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA, 90048; 10) Bioinformatics and Computational Biology Research Center, Cedars-Sinai Medical Center, Los Angeles, CA, USA, 90048; 11) MRC Epidemiology Unit, University of Cambridge, Cambridge, CB2 0QQ, United Kingdom.

Polycystic ovary syndrome (PCOS) is a common complex disorder causing reduced fertility affecting 5-15% of reproductive-aged women worldwide. Characterized by metabolic disturbances, hyperandrogenism and chronic anovulation, its etiology is largely unknown but with a clear genetic component. To date, genome-wide association studies (GWAS) have delivered 16 PCOS loci. Here, we perform a large GWAS meta-analysis of PCOS in up to 10,074 cases and 103,164 controls of European ancestry (NIH criteria, 2,540 cases/15,020 controls; Rotterdam criteria, 2,669 cases/17,035 controls; self-reported, 5,184 cases/82,759 controls from 23andMe). This genomic control corrected fixed-effects meta-analysis included 10,637,747 variants (imputed to 1000 Genomes March 2012 all ancestries panel) excluding markers with minor allele frequency <1% and imputation quality <0.3. We identified 15 independent loci ($P<5x10^{-8}$), four of which were novel. All but one association are consistent across the case definitions [near *GATA4/NEIL2*, $OR_{\text{self-reported}}=1.08$ (1.03-1.13); $OR_{\text{Rotterdam}}=1.21$ (1.14-1.28); $OR_{\text{NIH}}=1.33$ (1.26-1.41)]. We find significant genetic correlations ($P<8.9x10^{-4}$) with obesity, fasting insulin, type 2 diabetes, high-density lipoprotein cholesterol, menarche timing, triglycerides and cardiovascular risk factors. In Mendelian randomization analyses, both obesity ($P=1.6\times10^{-23}$) and fasting insulin ($P=1.7\times10^{-5}$) seem to play a causal role in PCOS, independent of each other. On an individual marker level, we note overlapping ($r^2>0.8$) and significant associations ($P<10^{-8}$) with sex hormone levels, vitiligo, inflammatory skin disease and type 1 diabetes suggesting pleiotropy. We see potential neuroendocrine (*FSHB, ZBTB16, GATA4/NEIL2, DENND1A, RAB5B, TOX3*), metabolic (*THADA*) and developmental (*YAP1, ERBB4, ERBB3, MAPRE1*) components to PCOS. Further characterization of the observed PCOS associations in relevant subphenotypes (sex hormone levels, hyperandrogenism, amenorrhea, polycystic ovarian morphology and ovarian volume) will provide more details about the potential functions of these variants. This large-scale study implicates a role of neuroendocrine and metabolic mechanisms in PCOS and is advancing our knowledge about its genetic architecture and etiology.

143

83 rare and low-frequency coding variants implicate specific genes affecting human height variation.

E. Marouli[1], M. Graff[2], C. Medina-Gomez[3,4], K. Sin lo[5], A. Wood[6], T.R. Kjaer[7], R.S. Fine[8], C. Schurmann[9], C. Oxvig[7], Z. Kutalik[10], F. Rivadeneira[3,4], R.J.F. Loos[9], T.M. Frayling[6], J.N. Hirschhorn[8], G. Lettre[5], P. Deloukas[1] For deCODE Genetics, the BBMRI-NL, the GOT2D, the CHARGE, and the GIANT Consortia. 1) William Harvey Research Institute, Barts and The London School of Medicine and Dentistry Queen Mary University of London, London, UK; 2) Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; 3) Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands; 4) Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands; 5) Montreal Heart Institute, Université de Montréal, Montréal, Québec, Canada; 6) University of Exeter Medical School, Exeter, UK; 7) Department of Molecular Biology and Genetics, Aarhus University, Denmark; 8) Broad Institute of Harvard and MIT, Cambridge, MA, USA; 9) The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 10) Institute of Social and Preventive Medicine, Lausanne University Hospital, Lausanne, Switzerland.

Human stature is a highly heritable, classic polygenic trait and primarily the end-result of many complex growth processes in childhood. Genome-wide association studies (GWAS) have identified ~700 variants associated with adult height variation. We screened >700,000 individuals with the Exome-chip array, which harbors over 200,000 coding variants with MAF <5%, and identified 32 rare (MAF<1%) and 51 low-frequency (1%<MAF<5%) coding variants associated with height at array-wide significance (P<2x10^{-7}). We also identified 3 additional genes (*CSAD*, *NOX4*, *UGGT2*) through gene-based tests. Several of the rare variants (e.g. in *IHH*, *STC2*, *AR* and *CRISPLD2*) were associated with differences in height of up to 1.9 cm, more than 10 times the average effect of previously identified common variants. *In vitro* experimental data showed that the effect of height-increasing variants in the *STC2* gene (+1-2 cm/rare allele) is likely due to partial loss of function of this gene leading to increased PAPP-A mediated cleavage of IGFBP4, possibly resulting in higher bioavailability of insulin-like growth factors. These 83 height variants with MAF <5% map to genes implicated in bone biology (*IL11RA*, *NOX4*) and growth defects (*ADAMTS3*, *PTH1R*). Additionally, pathway analysis identified gene sets in both growth-specific pathways (chondrocyte biology and skeletal development) and more global biological processes (transcription factor binding and embryonic size/lethality). In the UK-Biobank, these 83 rare and low-frequency coding variants explain ~1.7% of the heritable variation in height with an average fraction of heritability explained slightly lower than that of known common variants (0.021% versus 0.029%). In total, we found 120 new height loci (MAF≤50%); through conditional analysis in the known and new height loci, we now have identified a total of 968 independent variants that together account for 27.4% of the heritable contribution to height variation. Many height variants have pleiotropic effects but most of these are common (out of 96 variants associated with another trait at P<2x10^{-7},1 was rare and 6 low-frequency). The most pleiotropic height variant was the missense rs13107325 (MAF=6.2%) in the divalent cation transporter gene *SLC39A8* that is associated with BMI, all four lipid traits, systolic and diastolic blood pressure. In conclusion, our results confirm that rare and low-frequency coding variation contributes to the regulation of height and highlights relevant biology.

144

Large-scale genome wide association of human body proportion (sitting height ratio) identifies 161 associated loci.

K. Tsuo[1,2], R.M. Salem[2,3], R. Fine[2,3,4], M. Guo[2,3,4], Y. Chan[2,3,4], S. Vedantam[2,3], J.N. Hirschhorn[2,3,4]. 1) Harvard College, Cambridge, MA; 2) Medical and Population Genetics, Broad Institute, Cambridge, MA; 3) Endocrinology, Boston Children's Hospital, Boston, MA; 4) Department of Genetics, Harvard Medical School, Boston, MA.

Hundreds of common genetic variants have been discovered to be associated with human height, but the genetics underlying body proportion have not been as extensively studied. Genetic analysis of sitting height ratio (SHR), a measure of body proportion, has yielded insights into the biology of skeletal growth, but only a few robust SHR-associated variants have been identified. SHR is calculated by dividing the sitting height, the length from a person's head to the surface on which they are seated, by the total (standing) height. To expand our understanding of skeletal growth, we conducted the largest (to date) genome-wide association study of SHR, performed on approximately 120,000 British individuals of European ancestry from the UK Biobank study. Analyses were stratified by sex and meta-analysis was performed to combine the sex-specific results. We identified 161 independent genome-wide significant loci (p < 5E-8) associated with SHR, of which 41 were previously associated with total height. Of the 161 loci, 143 reached genome-wide significance in the combined meta-analysis of sex-stratified results. Of the remaining 18 loci, 9 loci reached genome-wide significance in men only and 9 loci reached genome-wide significance in women only. These 18 loci were not genome wide significant in the meta-analysis, suggesting they are sex-specific. The loci associated with SHR in males or females may yield novel insights into the sexual dimorphism of skeletal growth; to explore this, we are performing pathway analysis (using DEPICT) on the full set of significant loci, as well as the male- and female-only signals, to identify enriched pathways and tissues for each set of loci. In summary, we have identified many novel loci associated with SHR that together can offer new insights into the genetic and biological basis of body proportion and skeletal growth in women and men.

145

Genome-wide association study with replication implicates *WFS1* in cisplatin-associated hearing loss and reveals an enrichment of SNPs in Mendelian genes for deafness. *H.E. Wheeler[1], E.R. Gamazon[2], R.D. Frisina[3], C. Perez-Cervantes[1], O. El Charif[4], S.D. Fossa[5], D.R. Feldman[6], R. Hamilton[7], D.J. Vaughn[8], C.J. Beard[9], C. Fung[10], L.H. Einhorn[11], C. Kollmannsberger[12], J. Kim[13], T. Mushiroda[14], M. Kubo[14], S. Ardeshir-Rouhani-Fard[11], N.J. Cox[2], M.E. Dolan[4], L.B. Travis[11], The Platinum Study Group.* 1) Departments of Biology and Computer Science, Loyola University Chicago, Chicago, IL; 2) Vanderbilt University, Nashville, TN; 3) University of South Florida, Tampa, FL; 4) University of Chicago, Chicago, IL; 5) Oslo University Hospital, Oslo, Norway; 6) Memorial Sloan-Kettering Cancer Center, New York, NY; 7) Princess Margaret Cancer Centre, Toronto, ON; 8) University of Pennsylvania, Philadelphia, PA; 9) Dana-Farber Cancer Institute, Boston, MA; 10) University of Rochester Medical Center, Rochester, NY; 11) Indiana University, Indianapolis, IN; 12) University of British Columbia, Vancouver, BC; 13) The University of Texas MD Anderson Cancer Center, Houston, TX; 14) RIKEN Center for Integrative Medical Science, Yokohama, Japan.

Cisplatin is widely used and highly ototoxic. Some degree of hearing loss is found in 80% of testicular cancer survivors (TCS) on comprehensive audiometric examination, while 18% have severe to profound hearing loss. We performed a genome-wide association study (GWAS) for cisplatin-associated hearing loss (CAHL) modeled as a quantitative phenotype using air conduction thresholds (4 – 12 kHz) in 511 TCS of European genetic ancestry. One SNP (rs62283056) in the first intron of *WFS1*, which encodes wolframin ER transmembrane glycoprotein, met genome-wide significance for association with CAHL ($P = 1.4 \times 10^{-8}$). Mutations in this gene can cause autosomal dominant deafness and Wolfram syndrome, also known as DIDMOAD (Diabetes Insipidus, Diabetes Mellitus, Optic Atrophy, and Deafness). The minor allele of rs62283056 associates with increased hearing loss and decreased expression of *WFS1* (cis-eQTL) in several human tissues from the GTEx Project. The association between decreased expression of *WFS1* and increased hearing loss was replicated in an independent cohort from the BioVU repository (n = 18,620) using the gene-based method PrediXcan. Decreased predicted expression of *WFS1* associated with increased hearing loss in hypothalamus ($P = 6.6 \times 10^{-4}$), basal ganglia (P = 0.044) and artery (P = 0.036). In a set of 30 central nervous system cancer cell lines, we found lower levels of *WFS1* confer greater sensitivity to cisplatin (Spearman r = 0.33, P = 0.036), but not to other non-ototoxic chemotherapeutics. Beyond this top signal, we show CAHL is a polygenic trait ($h^2 = 0.92 \pm 0.62$, P = 0.039) with a genetic architecture related to Mendelian forms of deafness: SNPs within 50kb of 84 genes known to cause Mendelian nonsyndromic deafness are significantly enriched for low P-values in the GWAS, as indicated by the departure from the null in the quantile-quantile plot and a permutation resampling analysis (empirical P = 0.048). This indicates that some of the same underlying biological mechanisms driving congenital hearing loss (i.e. deafness) appear to contribute to the hearing loss induced by cisplatin. Our approach represents a new model for studies of drug-induced and other provocative phenotypes. Mendelian genes for the severe versions of drug-induced phenotypes may contribute significantly to the genetic architecture of the pharmacological trait. This work is supported by NIH R01CA157823, U19GM061390, R01MH101820, R01MH090937, and P50MH094267.

146

***DNMT3B* SNP contributes to nicotine dependence across 16 GWAS samples of European and African ancestry and influences expression of *DNMT3B* in cerebellum.** *D.B. Hancock[1], G.W. Reginsson[2], S.M. Lutz[3], R. Sherva[4], A. Loukola[5], C. Minica[6], X. Chen[7], C.A. Markunas[1], K.A. Young[3], F. Gu[8], D.W. McNeil[9], B. Qaiser[5], M.T. Landi[6], P. Madden[10], L.A. Farrer[4], J. Vink[6], N.L. Saccone[10], M.C. Neale[11], H.R. Kranzler[12], M.L. Marazita[13], D. Boomsma[6], T.B. Baker[14], J. Gelernter[15], J. Kaprio[5], N.E. Caporaso[8], T.E. Thorgeirsson[2], J.E. Hokanson[3], L.J. Bierut[10], K. Stefansson[2], E.O. Johnson[1].* 1) RTI International, Research Triangle Park, NC, USA; 2) deCODE Genetics / Amgen, Reykjavik, Iceland; 3) University of Colorado Anschutz Medical Campus , Aurora, CO, USA; 4) Boston University School of Medicine, Boston, MA, USA; 5) University of Helsinki, Helsinki, Finland; 6) Vrije Universiteit, Amsterdam, The Netherlands; 7) University of Nevada, Las Vegas, NV, USA; 8) National Cancer Institute, Bethesda, MD, USA; 9) West Virginia University, Morgantown, WV, USA; 10) Washington University, St. Louis, MO, USA; 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) University of Wisconsin, Madison, WI, USA; 15) Yale University School of Medicine, New Haven, CT, USA.

Cigarette smoking is the leading cause of preventable mortality worldwide. Prior genome-wide association study (GWAS) analyses of smoking phenotypes have established a number of genetic loci, most notably loci harboring nicotinic acetylcholine receptor and nicotine metabolism genes. To identify novel loci, we conducted the largest ever GWAS meta-analysis of nicotine dependence (ND), using 16 samples with total N=44,566 (33,841 of European ancestry and 10,725 African Americans [AAs]). We defined ND using cigarettes per day and other components of the Fagerström Test of ND. We tested 19 million 1000 Genomes imputed SNPs and indels for association with mild/moderate/severe ND, separated by sample and ancestry and adjusted for age, sex, eigenvectors for population stratification, and cohort-specific covariates. Results were combined using inverse variance-weighted meta-analysis. In addition to identifying three known nicotinic acetylcholine receptor loci (*CHRNA5-A3-B4* [smallest $P=4.1\times10^{-38}$], *CHRNB3-A6* [$P=2.6\times10^{-8}$] and *CHRNA4* [$P=4.6\times10^{-8}$]), we observed novel genome-wide significant associations in the dopamine β-hydroxylase (*DBH*, $P=3.4\times10^{-8}$ for the upstream SNP rs111280114) and DNA methyltransferase *DNMT3B* ($P=4.2\times10^{-8}$ for the intronic SNP rs910083) genes. *DBH* was implicated by prior GWAS meta-analysis of smoking cessation. No *DNMT3B* variants have been reported before for any addiction phenotype, and we found that rs910083-C was associated with increased ND risk across both ancestries: among AAs, frequency=77% and odds ratio (95% confidence interval) = 1.11 (1.05–1.16) for severe vs. mild ND, and among European ancestry individuals, frequency=44% and odds ratio (95% confidence interval) = 1.05 (1.03–1.08). Moreover, we identified rs910083 as a significant *cis*-expression quantitative trait locus (*cis*-eQTL). In the brain, the highest *DNMT3B* RNA expression levels were found in cerebellum, where structural deficits have been indicated in ND. Rs910083-C was associated with higher *DNMT3B* cerebellar RNA expression in the Genotype-Tissue Expression project (N=89, $P=2.1\times10^{-7}$) and independently in the Brain eQTL Almanac (N=130, $P=8.8\times10^{-5}$). Higher *DNMT3B* expression resulting from exposure to addictive substances has been reported *in vitro* with cigarette smoke and *in vivo* animal models with cocaine. These findings offer new directions to gain a better understanding of the brain-specific genetic regulatory factors that contribute to ND and possibly other drug dependencies.

147

Genetic variants associated with lung function predict chronic obstructive pulmonary disease (COPD) susceptibility. *N.R.G. Shrine[1], L.V. Wain[1], M. Soler Artigas[1], I.P. Hall[2], M.D. Tobin[1], BiLEVE,SpiroMeta,UKHLS,COPDGene,ECLIPSE,NETT/NAS,GenKOLS,LHS,lung eQTL study,DiscovEHR,-deCODE,BioMe.* 1) Department of Health Sciences, University of Leicester, Leicester, UK; 2) Division of Respiratory Medicine, Queen's Medical Centre, University of Nottingham, Nottingham, UK.

COPD is the third leading cause of death globally. Understanding the genetic factors associated with reduced lung function and with COPD risk could inform drug target identification. In order to boost power for discovery of robust genetic associations with COPD, genome-wide association studies (GWAS) for lung function can be carried out in large general population cohorts and associated variants followed up in COPD case-control studies. We undertook a GWAS of quantitative lung function traits in 48,493 samples chosen from the extremes and middle of the lung function distribution in UK Biobank (N=502,682). 81 putatively novel variants ($P < 5\times10^{-7}$) were meta-analysed with an independent subset of UK Biobank participants (N=49,727), a study of lung function in the general population (SpiroMeta consortium, N=38,199) and the UK Household Longitudinal Study (N=7449) giving 43 novel signals of genome-wide significant association with lung function ($P < 5\times10^{-8}$). These 43 novel signals were combined with 52 previously reported lung function signals to create a 95-variant allelic risk score. We tested association of the risk score with moderate-severe COPD in UK Biobank (spirometrically defined, 10,547 cases, 53,948 controls) and with COPD susceptibility in 8 additional studies (8829 cases, 94,025 controls of European ancestry): COPDGene, ECLIPSE, National Emphysema Treatment Trial/Normative Aging Study, GenKOLS, lung eQTL study, Geisinger-Regeneron DiscovEHR Study, deCODE genetics and Mount Sinai BioMe Biobank. In interim analyses the odds ratio for COPD per standard deviation in risk score (~6 risk alleles) was 1.35 ($P=1.3\times10^{-229}$) across all studies. In the subset of studies comprising deeply-characterised COPD cases and controls (6110 cases; 4404 controls), the odds ratio was 1.37 ($P=1.5\times10^{-38}$) compared to 1.42 ($P=3.9\times10^{-205}$) in UK Biobank. Furthermore, in UK Biobank we show a gradation in COPD risk across deciles of allelic risk score, and a COPD risk more than three times higher for individuals in the top decile compared to the bottom decile. The estimated proportion of COPD cases attributable to allelic risk scores above the first decile (population attributable risk fraction) was 45.4% (95% CI 40.6 to 49.8). We demonstrate the benefit of studying lung function as a quantitative trait to identify genetic risk factors for COPD. The 43 novel signals highlight novel genes and pathways which could lead to therapeutic interventions.

148

29 novel associations for male pattern baldness provide new insights into aetiology and genetic correlations. *N. Pirastu[1], P.K. Joshi[1], T. Esko[2,3,4], J.F. Wilson[1,5].* 1) Usher Institute PHSI, The University of Edinburgh, Edinburgh, Midlothian, United Kingdom; 2) Estonian Genome Center, University of Tartu, Riia 23b, 51010, Tartu, Estonia;; 3) Program in Medical and Population Genetics, Broad Institute, Cambridge Center 7, Cambridge, 02242, MA, USA;; 4) Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Cambridge, 02141, MA, USA;; 5) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, Scotland.

Male-pattern baldness or androgenetic alopaecia is the most common cause of hair loss in men, with a prevalence ranging between 50% and 80%, depending on age. Its importance is not limited to physical appearance and related distress, baldness has also been associated with increased risk of cardiovascular disease and type 2 diabetes. Despite its high heritability (~80% estimated in twins), up to now a limited number of associated loci have been identified. We have conducted the largest case-control genome-wide association study (GWAS) on male-pattern baldness to date using UK Biobank (UKB) data. The phenotype was collected as part of the UKB self-assessment questionnaire using 4 different images to categorise the degree of hair loss, ranging between 1 (no sign of baldness), 2 intial signs of alopecia, 3 presence of the vertex and 4 near complete loss of scalp hair. In order to have a better phenotype we used as controls only people in category 1 and as cases people in categories 3 and 4. No age cut-off was used on the samples as male-pattern baldness usually first presents before 30 years of age and people in UKB are selected to be older. For the discovery step we used only people who self-identified as British but were also genetically so defined by UKB. This led to the definition of 25,662 cases and 17,928 controls. Genome-wide association was conducted using age, 15 PCs and the array batch as covariates. GWAS revealed 41 genome-wide significant independent loci replicated on the remaining European UKB samples (3,436 cases vs 2,435 controls): 29 of these were novel, and we confirmed all of the 12 known loci. Further bioinformatics analysis revealed that many of the observed associations overlapped with loci known to influence other common traits paticularly related to hormone and cancer. These results suggest that male-pattern baldness shares a wide-spread pleiotropic genetic background with other traits, shedding a new light not only on its aetiology but also on the links with other diseases and risk factors.

149

Multiplexed super-resolution imaging of chromosome structure *in situ* with DNA-PAINT. *B.J. Beliveau, H.M. Sasaki, F. Schüder, S.K. Saka, M. Dai, P. Yin.* Systems Biology, Wyss Institute / Harvard University, Boston, MA.

Super-resolution microscopy has the potential to transform our understanding of the organization of the genome in individual cells, as it can resolve structures on the order of tens of nanometers size and can provide rich, quantitative information about structures being imaged. DNA-PAINT, similar to methods such as STORM and PALM, is a localization-based technique that generates a large series of single-molecule fluorescence events that are used to construct super-resolution images. In the case of DNA-PAINT, these fluorescence events are the result of the transient hybridization of fluorescently labeled oligonucleotides present in the imaging buffer to complementary strands present on the target to be imaged (Jungmann *et al*. 2010, doi: 10.1021/nl103427w). The programmable nature of nucleic acid hybridization affords DNA-PAINT several unique advantages, including the ability to multiplex to ten or more targets in the same sample without sacrificing imaging performance (Jungmann *et al.* 2014, doi: 10.1038/nmeth.2835), to perform absolute quantification of target molecules (Jungmann *et al.* 2016, doi: 10.1038/nmeth.3804), and to combined with spinning disc confocal microscopy to allow acquisition of single-molecule super-resolution images at depths up to ~10 µm in cell and tissue specimens.   We have harnessed DNA-PAINT to image human chromosomes labeled by the Oligopaints FISH technique (Beliveau *et al.* 2012, doi: 10.1073/pnas.1213818110; Beliveau *et al*. 2015, doi 10.1038/ncomms8147). With this combination, we have achieved six color super-resolution images at 20-30 nm resolution that reveal striking contacts and contours between topologically associating domains (TADs) and their intervening regions. We are now deploying our technology to sample the structures and folding conformations of chromosomes at the scale of TADs as well as the smaller sub-domains that have been recently reported (Rao *et al*., doi: 10.1016/j.cell.2014.11.021), with the goal of understanding how these properties vary within and between cells. We also are combining our multiplexed DNA FISH assay with super-resolution RNA FISH and super-resolution immunostaining in order to examine chromosome structure, gene expression state, and the presence of chromatin-associated proteins simultaneously in the same sample. We anticipate that such a picture will provide critical insights into how chromosome conformation and associated protein complement directly impacts its function and expression state.

150

Unbalanced constitutional chromothripsis are recombinant chromosomes of cryptic parental balanced chromothripsis. *N. Kurtas[1], A. Provenzano[2], V. Orlandini[2], L. Xumerle[3], S. Bargiacchi[2], L. Leonardelli[3], U. Giussani[4], A. Pansa[4], R. Artuso[5], A. Vetro[1], E. Errichiello[1], M. Delledonne[3], S. Giglio[2,5], O. Zuffardi[1].* 1) Department of Molecular Medicine, University of Pavia, Pavia, Pavia, Italy; 2) Medical Genetics Unit, Department of Biomedical Experimental and Clinical Sciences "Mario Serio", University of FLorence, Firenze, Italy; 3) Department of Biotechnologies, University of Verona, Verona, Italy; 4) Laboratorio di Genetica, Ospedali Riuniti di Bergamo, Bergamo, Italy; 5) Medical Genetics Unit, Meyer Children's University Hospital, Firenze, Italy.

Chromothripsis explains complex chromosomal rearrangements (CCRs) confined to a single or a few chromosomes. It is usually characterized by extensive genomic rearrangements consisting in multiple deletions and disordered orientation of the portions of the original chromosome, However, “shattering and stitching” of chromothripsis does not fully explain the occurrence of multiple duplications or concurrent duplications and deletions as reported in some CCRs. We demonstrate that at least some of them reflect recombinant chromosomes derived from the chromothripsis present in one parental chromosome, rather than the primary chromothripsis event.   We studied by whole genome sequencing, WGS, (mate pair sequencing) five cases, three of which with complex chromosome rearrangements, as detected by conventional cytogenetics and array-CGH. The fifth case (case 5) was that of the healthy mother with an apparent normal karyotype. Her affected child carried non-contiguous deletion and duplication along 3q22.1-q24 and 3q26.2-q26.31, revealed by array-CGH.   In four cases, all with intellectual disability and complex phenotype, we detected either a number of deletions or deletions and duplications involving from 1 to three chromosomes, with no less than 15 breakpoints each and a novel reassembly of the chromosomes involved. In the fifth case, FISH with probes from the chromosome regions that were unbalanced in the child, revealed a simple paracentric inversion. WGS showed that the long arm of one chromosome 3 was random reassembly, including breakage of multiple protein-coding genes, without noticeable phenotypic effects, clearly indicating the occurrence of a catastrophic event.   Our findings strongly indicate that apparently *de novo* complex rearrangements can in fact be recombinant chromosomes derived by a cryptic “balanced chromothripsis” present in one healthy parent. According to this hypothesis, we can expect that these parents have more than one unbalanced offspring, which so far it has not been reported. However, considering the complexity of the rearrangement, the probability that it forms a vital recombinant appears very limited and it seems quite likely the occurrence of early abortions.

151

Quantification, sub-family classification and genomic origin of transcribed *Alu* in age-related macular degeneration.

M.E. Kleinman[1,2], J.T. Lowery[3], C. Liu[3], B.J. Fowler[1], D. Lou[1], K. Mohan[1], S.C. Prajapati[1], Y. Hirano[1], A.K. Berner[1], J. Roney[1], J.L. Abney[1], B.D. Gelfand[1,4], M. Keddache[5], A.G. Hernandez[6], J. Liu[3], J. Ambati[1,7]. 1) University of Kentucky, Department of Ophthalmology and Visual Sciences, Lexington, KY, USA; 2) University of Kentucky, Department of Pharmacology and Nutritional Sciences, Lexington, KY, USA; 3) Department of Computer Science, University of Kentucky, Lexington, KY, USA; 4) 4Department of Biomedical Engineering, University of Kentucky, College of Medicine, Lexington, KY, USA; 5) Department of Human Genetics, Cincinnati Childrens' Medical Center, Cincinnati, OH, USA; 6) Carver Biotechnology Center, University of Illinois Urbana-Champaign, Urbana, IL, USA; 7) Department of Physiology, University of Kentucky, College of Medicine, Lexington, KY, USA.

Age-related macular degeneration (AMD) is the most common cause of irreversible blindness in the developed world. We have previously demonstrated increased levels of non-coding RNA derived from *Alu* sequences in human eyes with advanced dry AMD (geographic atrophy, GA) due to loss of the RNA processing enzyme DICER1 using multiple techniques including adaptor-ligation PCR and in-situ immuno-localization. This accumulation of *Alu* RNA is toxic to the retinal pigment epithelium (RPE), a monolayer that is critical for the maintenance of overlying photoreceptors and optimal vision. In this study, we developed a next-generation *Alu* RNA sequencing (*Alu*-Seq) pipeline to quantify RNA copy numbers of all known *Alu* sub-family sequences in complex biological samples and map their origins in the human genome. The *Alu*-Seq pipeline we designed was capable of highly sensitive detection of target sequence in a complex biological sample. Synthetic databases containing known *Alu* reads were analyzed with expression value measurements approaching ground truth (R^2=0.99946). Down-regulation of *Dicer1* expression in primary human RPE isolates led to over a 5-fold increase in *Alu* RNA sequences compared to control. These techniques allowed us to identify a specific pattern of *Alu* subfamily gene expression that with significant up-regulation of *Alu Y* RNA compared to others. Size fractionation of input RNA resulted in improved detection of elevated *Alu* RNA expression but was not required for subfamily assignment or quantification. Macular RPE/Choroid from human eyes with GA harbored 30% higher levels of *Alu* RNA compared to age-matched controls with a similar distribution of subfamily *Alu* sequence expression to the *in vitro* model. Utilizing advanced bioinformatics methods, *Alu* expression was partitioned into transcriptomic and inter-genic regions. Specific loci of *Alu* inserts in the human genome were identified as hotspots of aberrant expression that may be important in the etiology and risk of AMD progression. These data reveal a novel signature of pathogenic *Alu* subfamily expression and provide critical insight into the development bioinformatics pipelines for *Alu* sequencing data.

152

Structural variation landscape across 26 human populations reveals population specific variation patterns in complex genomic regions.

P. Kwok[1], C. Chu[1], A. Hastie[2], E. Lam[2], A. Leung[3], L. Li[3], C. Lin[1], J. McCaffrey[4], Y. Mostovoy[1], A. Naguib[2], S. Pastor[4], A. Poon[1], R. Rajagopalan[4], M. Sakin[1], J. Sibert[4], W. Wang[2], E. Young[4], H. Cao[2], T. Chan[3], K. Yip[3], M. Xiao[4]. 1) Univ California, San Francisco, San Francisco, CA; 2) BioNano Genomics, Inc., La Jolla, CA; 3) Chinese University of Hong Kong, Shatin, Hong Kong, SAR; 4) Drexel University, Philadelphia, PA.

While structural variation (SV) maps based on short-read sequences and statistical phasing have been constructed for samples comprising the 1000 Genomes Project[1], the sensitivity of detection and localization of some classes of SVs (such as long insertions, inversions, copy number variations, and duplications spanning several kb or more) are suboptimal. The challenge of detecting and localizing SVs other than deletions arises from the inherent limitations of short-read sequencing and the imperfections of the human reference genome sequence assembly. We have constructed genome maps[2] for 156 unrelated individuals from 26 human populations with long DNA molecules (>150 kb) fluorescently labeled at specific sequence motifs (nickase recognition sites). These samples consist of 6 individuals (3 males and 3 females) from each of 26 human populations of the 1000 Genomes Collection. As the data are generated from native DNA without amplification and assembled without the use of the human reference genome, the genome maps are *de novo* assemblies of the 156 genomes. All SVs >3 kb are easily visualized and uniquely placed on the map in one experiment. Wehen these genome maps were compared against the *in silico* maps derived from the human reference genome and against each other, we found that there were clear population SV patterns. These population SV patterns are most pronounced in complex regions of the genome where large (>50 kb) inversions and tandem duplications are mixed together in the same loci. These regions include the loci for microdeletion syndromes (such as 7q11.23, 15q13.3, 16p11.2 and 22q11.2) and subtelomeric regions where near identical, long repeats render them hotspots for SV formation and impossible for short-read sequences to assemble into unique contigs. 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 these regions that are associated with important human diseases. The population specific SV patterns may also shed light on the origins of the complex regions and the patterns more closely associated with human disease. References: 1. Sudmant PH et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015; 526:75-81. 2. Mak AC et al. Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays. Genetics. 2016; 202:351-62.

153

Visualizing structural variation at the single cell level to explore human genome heterogeneity. *A.D. Sanders[1], M. Hills[1], D. Porubsky[2], V. Guryev[2], E. Falconer[1], P.M. Lansdorp[1-3].* 1) Terry Fox Laboratory, BC Cancer Research Centre, Vancouver, British Columbia, Canada; 2) European Research Institute for the Biology of Ageing, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands; 3) Division of Hematology, Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada.

Studies of genome heterogeneity and plasticity aim to resolve how genomic features underlie phenotypes and disease susceptibilities. Identifying genomic variants that differ between individuals and cells can help uncover the functional elements that drive specific biological outcomes. For this, single cell studies are paramount, as it becomes increasingly clear that the contribution of rare but functional cellular subpopulations is important for disease prognosis, management and progression. Until now, studying these associations has been challenged by our inability to map structural rearrangements accurately and comprehensively. To overcome this, we employed the template strand sequencing method, Strand-seq, to preserve the structure of individual homologues and visualize genomic variants in single cells. We used this method to rapidly discover, map, and genotype human polymorphisms with unprecedented resolution. This allowed us to explore the distribution and frequency of structural rearrangements in a heterogeneous cell population, identify several polymorphic domains in complex regions of the genome, and locate rare alleles in the reference assembly. We then extended this analysis to comprehensively map the complete set of inversions in an individual's genome and define their unique inversion profile. We predict characterizing inversion profiles of patients will have important implications for personalized medicine. Finally, we generated a non-redundant, global reference of structural rearrangements in the human genome and better characterized their architectural features. Taken together, we describe a powerful new framework to study structural variation and genomic heterogeneity in single cell samples, whether from individuals for population studies, or tissue types for biomarker discovery.

154

Data double take: Three examples of atypical pathogenic alterations detected in exome sequencing data. *J.M. Hunter[1], C. Mroske[1], K. Helbig[1], B. Barrows[1], J. Cook[1], W. Mu[1], J. Capasso[2], A.V. Levin[2], M.J. Butte[3], R.S. Finkel[4], H. Lu[1], K.D.F. Hagman[1], S. Tang[1], W. Alcaraz[1].* 1) Clinical Genomics, Ambry Genetics, Aliso Viejo, CA; 2) Pediatric Ophthalmology and Ocular Genetics, Wills Eye Hospital, Philadelphia, PA; 3) Immunology & Allergy, Child Health Research Institute, Stanford University, Stanford, CA; 4) Division of Pediatric Neurology, Nemours Children's Hospital, Orlando, FL.

Clinical exome sequencing has become a routinely ordered test, especially for pediatric patients with phenotypes that are difficult to assign to specific genetic etiologies. Variant calling algorithms typically identify single nucleotide variants (SNVs) and small insertions/deletions (indels) with accuracy and ease. Because exome data can contain information about more complex alterations such as micro translocations and duplications too small to be detected by micro array, additional steps can be taken to ensure that such alterations are not excluded from analysis. We present three cases in which atypical damaging alterations were identified by analyzing exome data beyond the typical SNVs and small indels. The first case was a 17y old male with bilateral pigmentary maculopathy, cone-rod dystrophy, poor vision, and other signs of severe retinal dystrophy. A single paternally inherited nonsense alteration was identified using our exome analysis pipeline. Haploinsufficiency of *CRB1* is not typically pathogenic. To determine if a second pathogenic alteration was present in *CRB1*, exome sequencing data was analyzed by a fusion detection pipeline, and a second maternally inherited micro duplication disrupting exon 2 was detected. In a second unrelated case, a 14y old male presented with frequent infections, high IgE, and CD4 lymphopenia. A paternally inherited 7bp insertion was identified in the proband in *DOCK8*, a gene associated with autosomal recessive hyper-IgE recurrent infection syndrome (Job syndrome). Again, analysis by our fusion pipeline exposed a maternally inherited translocation between chromosome 3 and intron 2 of *DOCK8*, resulting in loss of exons 1-2 of *DOCK8*. Interestingly, there are several literature reports describing deletion of exon 1-2 of *DOCK8*, suggesting that the translocation may represent a recurrent and relatively common cause of Dock8 deficiency. Finally, a 5y old female presented with elevated creatinine kinase, exercise-related muscle fatigue and pain, and muscle biopsy with decreased calpain3 staining. No definitive alterations were identified until coverage statistics were analyzed, which revealed a homozygous deletion encompassing *CAPN3* and part of *GANC*, confirming the diagnosis of LGMD2A. These results demonstrate that in some cases, exome data contains information beyond SNVs and small indels. Thorough analysis of the data can lead to identification of atypical, but nevertheless important disease causing alterations.

155

CNV and homozygosity mapping from HiSeq X whole genome sequencing data: Fit for clinical use. *B.A. Lundie[1,2], M. Buckley[1,2,7,8], M.J. Cowley[2,8], M.E. Dinger[1,2,7], D. Fatkin[5], M. Field[4], V. Gayevskiy[2], C. Horvat[3], A.E. Minoche[2], G. Peters[3], C. Puttick[2], T. Roscioli[1,2,7,8], A. Zankl[2,3,6].* 1) Genome.One at The Garvan Institute of Medical Research, Darlinghurst, NSW, Australia; 2) Garvan Institute of Medical Research, Darlinghurst NSW 2010 Australia; 3) Sydney Genome Diagnostics, Children's Hospital Westmead, Westmead, NSW 2145, Australia; 4) NSW Health, Royal North Shore Hospital, St Leonards NSW 2065, Australia; 5) Victor Chang Cardiac Research Institute, Darlinghurst NSW 2010 Australia; 6) Sydney Medical School, University of Sydney; 7) St Vincent's Clinical School, UNSW Australia; 8) SEALS Genetics, NSW Health Pathology Service.

Microarray has been in routine clinical use for more than a decade and has dramatically increased the diagnostic yield for many patient cohorts. The rapidly decreasing cost combined with the broad and uniform depth of coverage from Illumina HiSeq X whole genome sequencing (WGS) data presents a unique opportunity to improve detection of CNVs and, in combination with SNV calls further increase this diagnostic yield within a single platform. We have developed a pipeline that identifies regions of CNV utilising split read, discordant read and read depth data. Multiple quality attributes and annotations enable us to obtain a comprehensive high confidence CNV call-set. By adding human population allele frequencies for CNVs, we can distil down to rare disease causing variants. Further, we developed a streamlined visualization procedure that allows the inspection of CNV with their underlying evidence in genome browsers. With the addition of allele frequency plots to highlight regions of homozygosity (ROH) this pipeline is also able to quickly identify regions of interest for SNV analysis pertinent to the referral. Our current pipeline for SNV analysis has recently attained clinical accreditation to the international standard 15189. The next iteration of this clinical pipeline will include CNV and ROH detection to the same standard. To this end we have performed extensive validation against clinical microarrays (50 patients) and NA12878 gold standards. Reproducibility of CNVs >10kb outside of segmental duplications is 96%. The reproducibility of smaller CNVs is currently being assessed and will be validated against orthogonal methods. Preliminary data suggests that the reproducibility is likely to be approximately 90%. We conclude that WGS data provides wider, more uniform and higher resolution coverage than current best-practice use of microarrays in pathology and is superior in terms of both analytic sensitivity and specificity. We present the application of the improved pipeline, which includes SNV, CNV and ROH, in a cohort of clinically referred patients.

156

NGS facilitates identification of retrotransposon insertional mutations in hereditary cancer genes. *Y. Qian, D. Mancini-DiNardo, H.C. Cox, T. Judkins, M. Elias, N. Singh, K. Brown, B. Coffee, K. Bowles, B. Roa.* Myriad Genetics Laboratories, Inc., Salt Lake City, UT.

Background: Retroelements (REs), also known as transposons, are widespread and comprise about 45% of the human genome. Previous studies suggest that insertion of REs into critical regions of genes, including control regions and exons, may disrupt normal gene function and cause genetic disease. Fewer than 100 RE insertion mutations have been reported to be associated with human disorders, including cancer. However, recent evidence suggests that the incidence of pathogenic RE insertions is likely underestimated due to technical challenges in their detection. Here, we investigated the utility of Next Generation Sequencing (NGS) to improve the identification of RE insertions in cancer predisposition genes. **Methods:** Blood and saliva-derived DNA samples were tested with a 25-gene hereditary cancer panel using PCR-based NGS. NGS dosage analysis (NGS LR) was used to identify large rearrangement mutations. Exon-based targeted Microarray CGH and/or multiplex ligation-dependent probe amplification (MLPA) were used as confirmatory assays. Suspected RE insertions were further investigated by targeted PCR and sequencing analyses to determine insertion size and location. The number of pathogenic RE insertions identified with NGS LR (2014-2015) was compared to those identified with traditional genetic testing methodologies used from 2004-2014 (multiplex qPCR or Southern Blot analysis). **Results:** NGS-based dosage analysis identified 10 novel RE insertions over 2 years. These insertions were identified in 6 genes (*ATM*, *BARD1*, *BRCA2*, *MLH1*, *MSH6*, *PALB2*) in 32 tested individuals. This accounts for 34% (10/29) of all unique RE insertion mutations identified by our laboratory over 12 years of testing. Among all RE insertion mutations, we identified several potential founder mutations that were enriched in patients of specific ancestries. **Discussion:** Our results show that PCR-based NGS, in conjunction with confirmatory assays, facilitates the identification of pathogenic RE insertions, with more than a third of all RE insertions being identified since the launch of NGS LR testing. This study provides evidence that the incidence of RE insertional mutations in human cancers and other disorders may be higher than previously known. This added knowledge is of great importance for early diagnosis and preventive management for high risk patients and their families, particularly for those who might have been reported as negative using traditional technologies.

157

Identification of 5 new genes and characterization of 3 ciliary modules implicated in oro-facial-digital syndromes.

A. Bruel[1], B. Franco[2], Y. Duffourd[1], J. Thevenon[1], L. Jego[1], E. Lopez[1], J. Deleuze[3], R. Giles[4], C. Johnson[5], M. Huynen[6], L. Burglen[4], M. Morleo[2], G. Pierquin[7], B. Doray[8], I. Panigrahi[9], D. Gaillard[10], B. Aral[11], S. Phadke[12], L. Pasquier[13], S. Saunier[14], A. Mégarbané[15], O. Rosnet[16], M. Leroux[17], J. Wallingford[18], O. Blacque[19], M. Nachury[20], T. Attie-Bittach[21], J. Rivière[1], L. Faivre[1], C. Thauvin-Robinet[1]. 1) Team EA4271 GAD, Université de Bourgogne, Dijon, Bourgogne, France; 2) Department of Translational Medicine, Medical Genetics Ferderico II University of Naples, Italy; 3) Centre National de Génotypage, Evry, France; 4) Department of Nephrology and Hypertension, University Medical Center Utrecht, Utrecht, The Netherlands; 5) Section of Ophthalmology and Neurosciences, Leeds Institute of Molecular Medicine, University of Leeds, Leeds, LS9 7TF, UK; 6) Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud university medical center, Geert Grooteplein 26-28, 6525 GA Nijmegen, Netherlands; 7) Service de Génétique, CHU, Liège, Belgium; 8) Service de Génétique Médicale, Hôpital de Hautepierre, CHU, Strasbourg, France; 9) Genetic-Metabolic Unit, Department of Pediatrics, Advanced Pediatric Centre, Pigmer, Chandigarh, India; 10) Service de Génétique, Hôpital Maison Blanche, CHRU, Reims, France; 11) Laboratoire de Génétique Moléculaire, PTB, CHU, Dijon, France; 12) Department of Medical Genetics, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, Uttar Pradesh, India; 13) Centre de Référence Maladies Rares « Anomalies du Développement et Syndromes malformatifs » de l'Ouest, Unité Fonctionnelle de Génétique Médicale, CHU Rennes, France; 14) INSERM U983, Institut IMAGINE, Hôpital Necker-Enfants Malades, Paris, France; 15) Al Jawhara Center, Arabian Gulf University, Manama, Bahrain; 16) Centre de Recherche en Cancérologie de Marseille, INSERM UMR1068, F-13009 Marseille, France; 17) Department of Molecular Biology and Biochemistry and Centre for Cell Biology, Development and Disease, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada; 18) Department of Molecular Biosciences, Center for Systems and Synthetic Biology, and Institute for Cellular and Molecular Biology, University of Texas at Austin; 19) School of Biomolecular and Biomedical Science, University College Dublin, Belfield, Dublin 4, Ireland; 20) Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA; 21) INSERM UMR1163, Université de Paris-Descartes-Sorbonne Paris Cité, Institut IMAGINE, Hôpital Necker-Enfants Malades, Paris, France.

Oral-facial-digital syndromes (OFDS) are characterized by the association of oral, facial and digital anomalies. The different modes of inheritance and additional features lead to clinically delineate 13 subtypes. For a long time, only the *OFD1* gene, responsible for OFDI subtype and coding for a centrosomal protein, has been known, suggesting the involvement of the primary cilium in OFDS. Mutations have recently been reported in the *TMEM216*, *DDX59*, *WDPCP*, *SCLT1*, *TBC1D32* and *TCTN3* genes in anecdotic cases. We recruited an international cohort of 115 OFDS index cases. Almost half of them had an *OFD1* causal mutation (59/115). To identify new genes involved in OFDS, we performed whole-exome sequencing in 23 patients. In 13/23 cases (56.5%), we identified 5 novel genes (*C2CD3*, *TMEM107*, *INTU*, *KIAA0753, IFT57),* enlarged the clinical spectrum of OFDS of 3 known genes responsible for other ciliopathies (*C5orf42*, *TMEM138*, *TMEM231*) and confirmed the involvement of 3 known genes in OFDS (*OFD1*, *DDX59, WDPCP*). Functional studies (*in vitro*, murin, zebrafish, xenopus and c. elegans models) demonstrated the involvement of the centriolar growth, the transition zone and the intraflagellar transport, through the characterization of 3 major protein complexes: the KIAA0753/OFD1/FOPNL complex controlling the centriole elongation, the MKS module (TMEM107/TMEM231/TMEM216), an essential component of the transition zone, and the CPLANE complex (INTU/FUZ/WDPCP) enabling in the IFT-A assembly. We demonstrated the large clinical and genetic heterogeneity of OFDS, yielding the initial classification in 13 subtypes obsolete, extending the number of 15 causal genes, and confirming OFDS as a new full subgroup of ciliopathies.

158

Rare genetic variations in *MEPE* are associated with otosclerosis and craniofacial bone disorder with facial paresis and mixed hearing loss.

H. Van Bokhoven[1,2], I. Schrauwen[3,4], L. Tomas-Roca[1,2], U. Altunoglu[5], M. Wesdorp[6,7], H. Valgaeren[3], M. Sommen[3], M. Rahmouni[1,2], E. van Beusekom[1,2], M.J. Huentelman[4], E. Offeciers[8], I. dHooghe[9], R. Vincent[10], A. Huber[11], P. Van de Heyning[12], F. Di Berardino[13], E. De Leenheer[6,9], C. Gilissen[1], C.W. Cremers[6], B. Verbist[14,15], A.P.M. de Brouwer[1], G.W. Padberg[16], R. Pennings[2,6], H. Kayserili[17], H. Kremer[1,2,6], G. Van Camp[3]. 1) Human Genetics, 855, Radboud university medical center, Nijmegen, Netherlands; 2) Donders Institute for Brain, Cognition and Behaviour, Radboud university medical center, 6500 HB Nijmegen, The Netherlands; 3) Department of Medical Genetics, University of Antwerp, Prins Boudewijnlaan 43, 2650 Edegem, Antwerp, Belgium; 4) Neurogenomics Division, Translational Genomics Research Institute, 445 N 5th str, 85004 Phoenix, AZ, USA; 5) Medical Genetics Department, Istanbul Medical Faculty, Istanbul University, Millet cad. Fatih, 34093 İstanbul, Turkey; 6) Department of Otorhinolaryngology, Hearing & Genes, Radboud university medical center, 6500 HB Nijmegen, The Netherlands; 7) Radboud Institute of Molecular Life Sciences, Radboud university medical center, 6500 HB Nijmegen, The Netherlands; 8) University Department of Otolaryngology, St-Augustinus Hospital Antwerp, Antwerp, Belgium; 9) Department of Otolaryngology, Ghent University Hospital, Ghent, Belgium; 10) Causse Ear Clinic, Colombiers, France; 11) University Hospital Zurich, Department of Otorhinolaryngology, Head and Neck Surgery, Zurich, Switzerland; 12) Department of ORL, University Hospital of Antwerp, Wilrijkstraat 10, 2650 Edegem, Belgium; 13) Dept. of Clinical Sciences and Community Health, Audiology Unit. University of Milan, I.R.C.C.S. Fondazione "Cà Granda", Osp.le Maggiore Policlinico, Milano Italy; 14) Department of Radiology, Radboud university medical center, Nijmegen, the Netherlands; 15) Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands; 16) Radboud university medical center, Department of Neurology, Donders Institute for Brain, Cognition and Behaviour, PO Box 9101, 6500 HB Nijmegen, The Netherlands; 17) Medical Genetics Department, Koç University School of Medicine (KUSOM), Topkapi, 34010 Istanbul, Turkey.

Craniofacial bone disorders comprise a group of diseases caused by abnormal growth and/or development of the head and facial bones, and can be associated with hearing loss due to abnormalities of the outer, middle or inner ear. In this study, we identified a heterozygous frameshift variant c.1273delC (p.Gln425Lysfs*38) in the *MEPE* gene in a family with non-progressive Hereditary Congenital Facial Paresis (HCFP) and mixed conductive-sensorineural hearing loss associated with diploic thickening and sclerosis of the skull. *MEPE* encodes a matrix extracellular phosphoglycoprotein and plays an inhibitory role in bone mineralization. Next, we hypothesized that this gene might also be important in otosclerosis bone remodeling disorder and screened for mutations in *MEPE* in 91 individuals with familial otosclerosis. We identified an additional heterozygous frameshift variant, c.199_202delGAAA (p.Lys70Ilefs*26), that segregated with the phenotype in two apparently unrelated families with otosclerosis, albeit with some reduced penetrance which is typically observed in otosclerosis families. Furthermore, we screened 1398 unrelated cases with otosclerosis and 1447 ethnically-matched controls. We observed the rare c.209_212del frameshift variant in eight affected individuals with otosclerosis only. None of the controls carried this variant (Fisher's Exact p = 0.003). Two other rare variants (c.184G>T: p.Glu62* and c.229G>A: p.Ala77Thr) were identified in cases and not in controls (cumulative Fisher's Exact p = 0.0008). Our results pinpoint MEPE as a key player in temporal bone and middle ear mineralization, which is involved in the pathogenesis of otosclerosis and other craniofacial bone disorders associated with mixed hearing loss.

159

***KIAA1109* variants are associated with a severe syndromic brain development disorder with arthrogryposis.**

N. Voisin[1], H. Shamseddin[2,3], F. Tran Mau Them[4], E. Preiksaitiene[5], R. Fish[6], L. Gueneau[1], L. Ambrozaityte[5], A. Morkuniene[5], N. Guex[1,7], B. Roechert[8], S. Pradervand[1,7], I. Xenarios[1,7], M. Neerman-Arbez[6], C. Shaw-Smith[9], V. Kucinskas[5], J. Chelly[4], F.S. Alkuraya[2,3], A. Reymond[1], Deciphering Developmental Disorders (DDD) Study. 1) Center for Integrative Genomics, University of Lausanne, Lausanne, Vaud, Switzerland; 2) Department of Genetics, King Faisal Specialist Hospital and Research Center, Riyadh 11211, Saudi Arabia; 3) Saudi Human Genome Program, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia; 4) Laboratoire de Diagnostic Génétique, Hôpitaux Universitaire de Strasbourg, Strasbourg, France; 5) Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University, Santariskiu St. 2, LT-08661, Vilnius, Lithuania; 6) Department of Genetic Medicine and Development, 1211 University of Geneva Medical School, Geneva, Switzerland; 7) Swiss Institute of Bioinformatics (SIB), University of Lausanne, 1015 Lausanne, Switzerland; 8) Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, Geneva, Switzerland; 9) Clinical Genetics Department, Royal Devon and Exeter National Health Service Foundation Trust, Exeter, UK.

We delineate a new genetic multi-system syndrome with neuronal migration defects. Whole exome sequencing of 7 individuals from 6 unrelated families from 4 continents with overlapping clinical manifestations identified biallelic loss-of-function and missense variants in *KIAA1109*. Shared phenotypic features representing the cardinal characteristics of this syndrome combine brain atrophy, hydrocephaly and Dandy-Walker malformation with club foot and arthrogryposis. Severe cases were incompatible with life, whereas carriers of milder missense variants presented with severe global developmental delay, absence of language, syndactyly of 2nd and 3rd toes, hypermetropia and severe muscle hypotony resulting in incapacity to stand without support. The 5005 amino acid KIAA1109 protein is evolutionarily conserved and interacts with proteins previously associated with intellectual disability and regulation of cell division such as CTNNB1, PPP2R4 and BUB3. Consistent with a causative role for *KIAA1109* loss-of-function and hypomorphic variants in this new syndrome, knockdowns of the zebrafish orthologous gene with morpholinos or CRISPR-cas engineering resulted in embryos with hydrocephaly and abnormally curved notochords and overall body shape. Similarly, absence of tweek, the *Drosophila* ortholog of *KIAA1109*, resulted in lethality or severe neurological defects reminiscent of proband features; mutants had seizures and were unable to walk or stand upright for long periods. We suggest naming this new neuronal migration disorder Alkuraya-Kucinskas syndrome as they first described affected individuals at the severe and mild end of the phenotype, respectively. This work has received funding from Lithuanian-Swiss cooperation program to reduce economic and social disparities within the enlarged European Union under project agreement NoCH-3-□MM-0. (AR).

160

Mutations in *CDC45*, encoding an essential component of the pre-initiation complex, cause Meier-Gorlin Syndrome and craniosynostosis.

L.S. Bicknell[1,2], A. Fenwick[3], M. Kliszczak[3,4], F. Cooper[2], J. Murray[2], L. Sanchez-Pulido[2], S.R.F. Twigg[3], A. Goriely[3], S.J. McGowan[5], K. Miller[3], I.B. Taylor[3], C. Logan[2], M. Koopmans[7], C.P. Ponting[2], A.P. Jackson[2], A.O.M. Wilkie[3,6], W. Niedzwiedz[4], Meier-Gorlin Syndrome Clinical Consortium. 1) Department of Pathology, University of Otago, Dunedin, Otago, New Zealand; 2) MRC Human Genetics Unit, IGMM, University of Edinburgh, Edinburgh EH4 2XU, UK; 3) Clinical Genetics Group, MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK; 4) Department of Oncology, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK; 5) Computational Biology Research Group, MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK; 6) Craniofacial Unit, Department of Plastic and Reconstructive Surgery, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford OX3 9DU, UK; 7) Department of Clinical Genetics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands.

DNA replication precisely duplicates the genome to ensure stable inheritance of genetic information. Impaired licensing of origins of replication during the G_1 phase of the cell cycle has been implicated in Meier-Gorlin syndrome (MGS), a disorder defined by the triad of short stature, microtia and a/hypoplastic patellae. Biallelic partial loss-of-function mutations in multiple components of the pre-replication complex (preRC; *ORC1*, *ORC4*, *ORC6*, *CDT1* or *CDC6)* as well as *de novo* stabilizing mutations in the licencing inhibitor, *GMNN,* cause MGS. Here we report the identification of mutations in *CDC45* in 17 affected individuals from 13 families with MGS and/or craniosynostosis. *CDC45* encodes a component of both the pre-initiation (preIC) and CMG helicase complexes, required respectively for initiation of DNA replication origin firing and ongoing DNA synthesis during S-phase itself, hence is functionally distinct from previously identified MGS-associated genes. The phenotypes of affected individuals range from syndromic coronal craniosynostosis to severe growth restriction, fulfilling diagnostic criteria for Meier-Gorlin syndrome. All mutations identified were biallelic and included synonymous mutations altering splicing of physiological *CDC45* transcripts, as well as amino acid substitutions expected to result in partial loss of function. Functionally, mutations reduce levels of full-length transcripts and protein in subject cells, consistent with partial loss of CDC45 function and a predicted limited rate of DNA replication and cell proliferation. Our findings therefore implicate the preIC as an additional protein complex involved in the etiology of MGS, and connect the core cellular machinery of genome replication with growth, chondrogenesis and cranial suture homeostasis.

161

Phenotype and genotype in 52 patients with Rubinstein-Taybi Syndrome caused by EP300 mutations. *J. Van Gils[1], P. Fergelot[1], M. Van Belzen[2], D. Lacombe[1], R.C. Hennekam[3], EP300 Working Group.* 1) Department of Medical Genetics and INSERM U1211, University Hospital of Bordeaux, Bordeaux, France; 2) Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands; 3) Department of Pediatrics, Academic Medical Center, Amsterdam, The Netherlands.

Rubinstein-Taybi syndrome (RSTS) is a developmental disorder characterized by a typical face and distal limbs abnormalities, intellectual disability and a vast number of other features. Two genes are known to cause RSTS, CREBBP in 60% and EP300 in 8-10% of clinically diagnosed cases. Both paralogs act in chromatin remodeling and encode for transcriptional co-activators interacting with >400 proteins. Up to now 25 individuals with an EP300 mutation have been published. Here we describe the phenotype and genotype of 42 unpublished RSTS patients carrying EP300 mutations and intragenic deletions and offer an update on another 10 patients. We compare the data to 308 individuals with CREBBP mutations. We demonstrate that EP300 mutations cause a phenotype that typically resembles the classical RSTS phenotype due to CREBBP mutations to a great extent, although most facial signs are less marked with the exception of a low-hanging columella. The limb anomalies are more similar to those in CREBBP mutated individuals except for angulation of thumbs and halluces which is very uncommon in EP300 mutated individuals. The intellectual disability is variable but typically less marked whereas the microcephaly is more common. All types of mutations occur but truncating mutations and small rearrangements are most common (86%). Missense mutations in the HAT domain are associated with a classical RSTS phenotype but otherwise no genotype-phenotype correlation is detected. Preeclampsia occurs in 12/52 others of EP300 mutated individuals versus in 2/59 mothers of CREBBP mutated individuals, making pregnancy with an EP300 mutated fetus the strongest known predictor for pre-eclampsia.

162

A synergistic effect of laminin and *P4HA2* mutant genes deregulates ECM remodeling causing a novel developmental syndrome. *F. Napolitano[1], S. Sampaolo[2], A. Tirozzi[1], F. Gianfrancesco[1], G. Di Iorio[2], T. Esposito[1].* 1) Institute of Genetics and Biophysics, Italian National Research Council , Naples, Italy; 2) Neurology Clinic II, Department of Medical Sciences, Surgery, Neurology, Metabolic Diseases and Geriatrics, Second University of Naples, Italy.

The extracellular matrix (ECM) is a critical component of the human tissues microenvironment. ECM and ECM components, are important in phenomena as diverse as developmental patterning, stem cell niches, cancer, and genetic diseases. We report a novel dominant developmental syndrome characterized by laxity of the visceral ligaments, impaired wound healing, serum negative arthritis, mild alopecia and teeth defects. Most of the patients exhibit severe myopia associated with retinal detachment and night blindness. We discovered that this disorder is caused by a synergistic effect of two novel mutations in one of the laminin genes and in the Prolyl 4-hydroxylase alpha-2 (P4HA2) gene. The laminin mutation segregates with the majority of the symptoms, while P4HA2 mutation is associated with the ocular defects. Patients carrying both the mutations show a more severe phenotype due to the synergistic effect of mutations in two proteins involved in ECM function. We showed that the P4HA2 mutation causes a significant decrease of expression of both P4HA2 RNA and protein and impairs the collagen deposition. The laminin mutation is located in the LG globular domain. This domain is crucial for protein folding and for Sonic hedgehog (Shh), Wnt and PI3Kinase pathways induction through integrin’s binding. We demonstrated that the laminin mutation alters the amount of peptides derived from protein cleavage and perturbs the activation of the epithelial-mesenchymal signaling producing an unbalanced expression of the SHH-GLI1 pathway, which is up regulated in patients cells, and of ECM proteins (COL1A1, MMP1 and MMP3) which are strongly inhibited. This suggests a condition resembling fibrosis, which is the result of defective repair processes, often seen after chronic injury and/or inflammation. Moreover, the cells of the patients are more responsive to treatment with TGFβ1 and Ilβ1 cytokines that are involved in the matrix remodeling. Studies carried out on human skin biopsies showed alteration of dermal papilla with a reduction of the germinative layer and an early arrest of hair follicle down growth mainly observed in the patients carrying the two mutations. A similar defect is observed in the knock-in mouse model generated in our laboratory, the study of which is ongoing.

163

Identification of a RAI1-associated disease network. *M.N. Loviglio[1], C.R. Beck[2], T. Harel[2], W. Bi[2], M. Leleu[3,4], N. Guex[4], A. Niknejad[4], E.S. Chen[2], S. Gu[2], J. White[2], I. Crespo[4], J. Yan[2], W. Charng[2], C.A. Shaw[2], Z. Coban-Akdemir[2], J. Rougemont[3,4], I. Xenarios[1,4], J.R. Lupski[2,5,6,7], A. Reymond[1].* 1) Center for Integrative Genomics (CIG), University of Lausanne, Lausanne, Vaud, Switzerland; 2) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; 3) School of Life Sciences, EPFL (Ecole Polytechnique Fédérale de Lausanne), CH-1015 Lausanne, Switzerland; 4) Swiss Institute of Bioinformatics (SIB), CH-1015 Lausanne, Switzerland; 5) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; 6) Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA; 7) Texas Children's Hospital, Houston, TX 77030, USA.

Smith-Magenis syndrome (SMS) is a developmental disability/multiple congenital anomaly disorder resulting from haploinsufficiency of *RAI1*. We investigated a cohort of 149 individuals presenting the constellation of SMS features, and focused on 15 individuals showing neither hemizygosity in the SMS critical region nor variants in *RAI1*. Using whole-exome sequencing we identified potentially deleterious variants in nine patients. Eight of these variants affect *KMT2D, ZEB2, MAP2K2, GLDC*, *CASK, MECP2, KDM5C* and *POGZ,* known to be associated with Kabuki 1, Mowat-Wilson, cardiofaciocutaneous, glycine encephalopathy, mental retardation and microcephaly with pontine and cerebellar hypoplasia, X-linked mental retardation 13, X-linked mental retardation Claes-Jensen type and a recently described intellectual disability syndrome, respectively. The ninth individual carries a *de novo* variant in *JAKMIP1*, a regulator of neuronal translation that was recently found deleted in a patient with autism spectrum disorder. Analyses of co-expression networks and biomedical text mining suggest that these pathologies and SMS are part of the same disease network, potentially explaining the overlapping phenotypes. Further support for this hypothesis comes firstly from transcriptome profiling of 10.5 dpc embryos that shows that the expression levels of both *ZEB2* and *MAP2K2* are perturbed in *Rai1*-/- mice. Secondly, chromatin conformation capture revealed contacts between *RAI1* and the *ZEB2* and *GLDC* flanking loci, as well as between *RAI1* and human orthologs of the genes that show perturbed expression in a *Rai*-/- mouse model, in particular genes mapping to MMU11. This finding possibly explains the enrichment of MMU11-mapping genes within genes differentially expressed in *Rai1-/-* mouse embryos. These holistic studies of *RAI1* and its interactions allow insight into SMS and other disorders associated with intellectual disability and behavioral abnormalities, demonstrating the utility of a comprehensive genomic approach even in the diagnosis of distinctive disorders.

164

Combined next generation sequencing techniques untangle the genomic structure of complex nonrecurrent deletions in subjects with Smith-Magenis syndrome and reveal a strong bias to paternally deleted chromosomes. *C. Fonseca[1], C.R. Beck[1], Z.C. Akdemir[1], Z. Chong[2], E.S. Chen[1], P.C. Thorton[1], P. Liu[1], B. Yuan[1], M. Withers[1], S.N. Jhangiani[3], A.C. English[3], D.M. Muzni[3], R.A. Gibbs[3], C.A. Shaw[1], P.J. Hastings[1], J.R. Lupski[1,3,4,5].* 1) Dept Molecular Human Genetics, Baylor Col Medicine, Houston, TX; 2) 2Department of Bioinformatics and Computational Biology, the University of Texas MD Anderson Cancer Center, Houston, TX; 3) Human Genome Sequencing Center, BCM, Houston, TX; 4) Department of Pediatrics, BCM, Houston, TX; 5) Texas Children's Hospital, Houston, TX.

The majority of patients (~70-80%) with Smith-Magenis syndrome carry a *de novo* 3.6 Mb deletion spanning 17p11.2 that includes *RAI1,* a gene that encode a putative transcription factor. Formation of this recurrent deletion occurs through ectopic crossover between flanking large low-copy repeats, leading to a decreased copy number of genes mapping within the common deleted segment. However, ~18% of the deletions in SMS cases are unique to the carrier individual (nonrecurrent events) and occur through a distinct molecular mechanism. In the latter cases, occasional complex genomic rearrangements (CGR) can occur resulting in multiple copy-number alterations accompanied by further large structure rearrangements of the 17p region such as inversions and an increased number of single-nucleotide variants (SNVs) in *cis* with the CGR. Defining the *de novo* local structure of the CGR in patients can be challenging due to the ubiquitous presence of LCRs distributed along 17p. Here we use a combination of next generation and third generation sequencing platforms (PacBio long-reads and Illumina short-reads targeted to 7 Mb of 17p11.2) along with cytogenetic techniques (e.g. custom array comparative genomic hybridization) to uncover the genomic structure of nonrecurrent deletions in nine individuals with SMS. Remarkably, all nine deletions occurred on the paternal chromosome. In eight of the nine cases the rearrangement breakpoint junctions were identified using our combined strategy. Importantly, we were able to resolve three cases in which further complexities (i.e. inversions) accompanied the deletions. In these cases, the deletions involved very large repeats at one of the junctions, which initially hampered our ability to obtain precise breakpoint data. Furthermore, we detected 1 to 3 *de novo* SNVs (9 in total) within the 7 Mb captured region in 6 individuals (1.43-4.29 x 10^{-7}/bp), i.e. a 10X increase compared with the spontaneous *de novo* rate in germline (~1 x 10^{-8}/bp/generation). Point mutations occurred mainly in homonucleotide A runs (6 out of 8) with a prevalent occurrence of transversions as opposed to transitions. In summary, our data provide evidence that challenging genomic rearrangements can be resolved using a combination of genomic technologies and strategies and indicate that a higher local SNV mutational rate accompanies formation of nonrecurrent rearrangements.

165

Comprehensive analysis of RNA-sequencing to find the source of every last read across 544 individuals from 53 tissues. *S. Mangul[1], H. Yang[1], N. Zaitlen[2], S. Shifman[3], E. Eskin[1].* 1) University of California, Los Angeles, Los Angeles, CA; 2) University of California, San Francisco, San Francisco, CA; 3) Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem.

High throughput RNA sequencing technologies have provided invaluable research opportunities across distinct scientific domains by producing quantitative readouts of the transcriptional activity of both entire cellular populations and single cells. The majority of RNA-Seq analyses begin by mapping each experimentally produced sequence (i.e., read) to a set of annotated reference sequences for the organism of interest. For both biological and technical reasons, a significant fraction of reads remains unmapped. We use 5000 RNA-Seq samples corresponding to 53 tissues from GTEx project. We have applied Read Origin Protocol (ROP) for discovering the source of all reads, to profile repeats, circular RNAs, gene fusion, trans-splicing, variable B/T-cell receptor sequences and microbial communities. ROP protocol consists of six step and is able to account for 99% of all reads. We find that the vast majority of unmapped reads are human in origin and originate from diverse sources, including repetitive elements, non-co-linear elements or recombined B and T cell receptors (BCR/TCR). In addition to human RNA, a large number of reads were microbial in origin, often occurring in sufficient numbers to study the taxonomic composition of microbial communities. We assess combinatorial diversity of the antibody repertoire by looking at the recombinations of the of VJ gene segments of BCR and TCR loci detected from unmapped reads. Spleen and small intestine yield increased number of combinations of gene segments from immunoglobulin kappa locus (IGK) and T cell receptor beta (TCRB) locus with 200-250 combinations, on average, per sample for IGK and 75-150 recombinations for TCRB respectively. Some tissues yeld increase recombinations of IGK locus and dicreased recombinations of TCRB (e.g. salivary gland). We observe no gene recombinations of IGK/TCRB locus for 10 tissues (e.g. brain, heart, and muscle). This study is the first that systematically accounts for almost all reads in GTEx RNA-seq data. We demonstrate the value of analyzing unmapped reads present in the RNA-seq data to study the non-co-linear, immunological and microbiome profiles of a tissue of interest. The ‘dumpster diving’ profile of unmapped reads output by ROP is not limited to RNA-Seq technology and may apply to whole-genome sequencing. The ROP pipeline is available at https://sergheimangul.wordpress.com/rop/.

166

The incorporation of whole blood and fibroblast RNAseq with whole exome sequencing implicates *LZTR1* in a novel syndrome with features of rasopathy and mitochondrial dysfunction. *M. Jain[1], L.C. Burrage[1], J.A. Rosenfeld[1], B.C. Dawson[1], X. Yang[1], C.A. Bacino[1], A. Balasubramanyam[2], P.M. Moretti[1,3], S.K. Nicholas[4], J.S. Orange[5], E. Roeder[6], L.T. Emrick[1], B.H. Graham[1], J.W. Belmont[1], N. Hanchard[1], W.J. Craigen[1], B.H. Lee[1], Members of the Undiagnosed Diseases Network.* 1) Molecular and Human Genetics, Baylor college Medicine, Houston, TX; 2) Medicine, Diabetes, Endocrinology and Metabolism, Baylor College of Medicine, Houston, TX; 3) Neurology, Baylor College of Medicine, Houston, TX; 4) Pediatrics-Allergy & Immunology, Baylor College of Medicine, Houston, TX; 5) Pediatrics-Rheumatology, Baylor College of Medicine, Houston, TX; 6) Pediatric Genetics, Baylor College of Medicine, San Antonio, TX.

Exome sequencing has revolutionized the practice of clinical genetics, with studies demonstrating that sequencing provides a molecular diagnosis in known disease genes in up to 30% of cases. However, this leaves a large percentage of cases with an unknown etiology. The Baylor College of Medicine (BCM) Clinical Site of the Undiagnosed Diseases Network (UDN) is performing RNAseq in whole blood and/or fibroblast with the goal of using these data to guide variant interpretation, define new candidates and evaluate for downstream functional impact. Characterization of control data (whole blood, n=28; fibroblast, n=8) demonstrates that expression is detectable for 58.7% of known genes and 64.3% of OMIM disease genes when examining both tissues. In the initial phase of this study, we have completed evaluation of 6 whole blood and 3 fibroblast samples from 6 patients. The analysis pipeline involves a tiered approach where top candidate variants from exome sequencing are reviewed. We next identify potential novel splice sites (1839.2/sample), markedly increased or decreased expression (126.3/sample), and allele specific expression (1928.3/sample). We prioritize candidates that have diminished expression in cases versus controls and that either have potential novel splice sites, evidence of allele specific expression, or annotated loss of function variants by exome to classify potential loss of function alleles (average of 76.3 genes/sample). We tested male siblings evaluated at the UDN site at BCM using this approach; they share a striking phenotype incorporating clinical features suggestive of a rasopathy and mitochondrial dysfunction and have undergone extensive evaluation without a diagnosis. Whole blood and fibroblast RNAseq on both brothers demonstrate rare features of *LZTR1* which include significantly reduced expression (Zscore < -6) and presence of a stopgain that escapes nonsense mediated decay. *LZTR1* has previously been implicated in familial schwannomatosis and Noonan syndrome and is recognized as a CUL3 ubiquitin ligase adaptor. Pathway analysis identified *HRAS* as an upstream regulator and reduced expression of mitochondrial oxidative phosphorylation genes. Functional evaluations are ongoing. We demonstrate a transcriptomic strategy for evaluation of genetic variation in rare disorders and have applied it to identify a strong candidate. We propose that systematic RNAseq analysis can further increase the diagnostic rate in clinical genetic evaluations.

167

Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *B.B. Cummings[1,2,3], M. Lek[2,3], T. Tukiainen[2,3], J. Marshall[2,3], F. Zhao[2,3], B. Weisburd[2,3], S. Donkervoort[4], R. Foley[4], L. Waddell[5], S. Sandaradura[5], G. O'Grady[5], E. Oates[5], J. Dowling[6], N.F. Clarke[5], S.T. Cooper[5], C. Bonnemann[4], D.G. MacArthur[2,3].* 1) Harvard Medical School, Longwood, MA; 2) Broad Institute of Harvard and MIT, Medical and Population Genetics, Boston, MA; 3) Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 4) National Institute of Neurological Disorders and Stroke, NIH, Neuromuscular and Neurogenetic Disorders of Childhood Section, Bethesda, MA; 5) The Children's Hospital at Westmead, Institute for Neuroscience and Muscle Research, Sydney, Australia; 6) 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 Mendelian disease diagnosis. However the current diagnosis rate for exome analysis across a variety of rare diseases is approximately 25-50%. One promising approach to increase the discovery of disease-causing variants is RNA sequencing (RNA-seq) from patient tissue, which provides direct insight into transcriptional perturbations caused by genetic changes. RNA-sequencing can be used for the detection of aberrant splicing, allelic imbalance and somatic variation, all types of events rarely detectable from genotype data alone. Such analyses can empower molecular diagnosis by validating the transcriptional effects of variants proposed by exome data or by identifying novel variants. Here we describe an integrated approach of patient muscle RNA sequencing in over 50 exome-unsolved cases with severe neuromuscular disease, leveraging an analysis framework focused on the detection of transcript level changes that are unique to the patient compared to a collection of over 180 control skeletal muscle RNA-seq samples. 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 overall diagnosis rate of 30% in our patients. In addition, we report the discovery of a highly recurrent de novo intronic mutation in COL6A1 that results in a splice-gain event disrupting a critical glycine repeat region. We have now confirmed this pathogenic variant in a total of 30 undiagnosed collagen myopathy patients, thus explaining a substantial fraction of all undiagnosed patients with this disease. This study represents the first systematic application of RNA-sequencing in a disease-relevant tissue to rare disease diagnosis, and highlights the utility of RNA-sequencing for the detection and interpretation of variants missed by standard exome-based approaches.

168

Implementation and clinical utility of transcriptome sequencing: Experience from neuromuscular disorders. *M.R. Hegde[1], B.R. Nallamilli[1], G. Gibson[2], D. Arafat[2], H.P. Subramanian[2], C. da Silva[1].* 1) Human Genetics, Emory University, Atlanta, GA; 2) Georgia Institute of Technology, Atlanta, GA.

DNA based assays result in detection of a large number of variants of unknown clinical significance (VUS). According to the recent ACMG variant interpretation guidelines, several pieces of data including functional evidence are needed for VUS reclassification. We implemented a clinical transcriptome assay for limb-girdle muscular dystrophies (LGMDs) which constitute an important genetically and clinically heterogeneous subgroup. In order to overcome the diagnostic odyssey encountered by LGMD patients a large scale study for clinical trial enrollment through Jain Foundation and MDA, we sequenced over 2000 individuals clinically suspected with LGMD and confirmed the diagnosis in about 40% of these individuals. Around 20-25% cases had VUS, which include promoter, silent and missense variants as well as some intronic variants. NMD clinical transcriptome analysis was performed on the target muscle tissue of these individuals to functionally assess the consequence of the identified VUSs. We are able to interrogate alternate splice effects, frameshift effects and allele loss due to non-sense mediated decay of the expressed mutant allele. Our NMD transcriptome assay interrogates transcripts of 79 NMD genes, using a commercially designed RNA probe library. With 40,332 baits of 120 bases each, the entire library is about 4.8Mb in size. High quality RNA extracted from muscle biopsies is sequenced in 150bp pair-end reads using the Illumina RNA seq kit and Illumina sequencers. Here, we show the results from our NMD transcriptome assay and discuss the clinical utility of the next next-generation sequencing based diagnostic test- targeted transcriptome assay. We functionally characterized multiple VUSs and have shown evidence for pathogenicity in various genes including DMD (3), DYSF (2), GNE (1) and SGCB (1) and have successfully reclassified these variants as pathogenic. The variants in the DMD and SGCB genes were found to introduce new splice sites thereby resulting in frameshift mutations. The deep-intronic DYSF variants created pesudoexons, whereas the variant in the promoter region of the GNE gene resulted in a complete loss of the corresponding allele, making the affected individual transcriptionally hemizygous for the second causative mutation on the opposite allele. It is expected that complete transcriptome assays will eventually make their way into the clinic for functional assessment of genomic variants.

169

The impact of genome structural variation on human gene expression. *A.J. Scott[1,6], C. Chiang[1,6], J.R. Davis[2,3], E.K. Tsang[2], X. Li[2], Y. Kim[4], F.N. Damani[4], S.B. Montgomery[2,3,5], A. Battle[4], D.F. Conrad[6,7,8], I.M. Hall[1,6,8], GTEx Consortium.* 1) McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO; 2) Department of Pathology, Stanford University School of Medicine, Stanford, CA; 3) Department of Genetics, Stanford University School of Medicine, Stanford, CA; 4) Department of Computer Science, Johns Hopkins University, Baltimore, MD; 5) Department of Computer Science, Stanford University, Stanford, CA; 6) Department of Medicine, Washington University School of Medicine, St. Louis, MO; 7) Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO; 8) Department of Genetics, Washington University School of Medicine, St. Louis, MO.

Structural variants (SVs), including copy number variants (CNVs), balanced rearrangements, and mobile element insertions (MEIs), are an important source of human genetic diversity but their contribution to traits, disease, and gene regulation remains unclear. The Genotype-Tissue Expression (GTEx) project provides an unprecedented opportunity to address this question due to the availability of deep whole genome sequencing (WGS) and multi-tissue RNA-seq data. Here, we used comprehensive SV detection methods – including split-read mapping, paired-end mapping, and read-depth analysis – to map structural variation in 147 GTEx individuals, resulting in 24,157 high confidence SVs. We mapped *cis* expression quantitative trait loci (eQTLs) in 13 tissues via joint analysis of SVs, single nucleotide (SNV) and short insertion/deletion (indel) variants and identified 24,801 eQTLs affecting the expression of 10,101 distinct genes. Based on haplotype structure and heritability partitioning, we estimate that SVs are the causal variant at 3.3-7.0% of eQTLs, which is nearly an order of magnitude higher than previous estimates from low coverage WGS, and represents a 26- to 54-fold enrichment relative to the scarcity of SVs in the genome. Expression-altering SVs also showed a 1.2-fold larger median effect size on gene expression than SNVs and indels. We identified 787 putatively causal SVs predicted to directly alter gene expression, most of which (88.3%) are noncoding variants that show significant enrichment at enhancers and other regulatory elements. By evaluating linkage disequilibrium between SVs, SNVs, and indels, we nominate 70 SVs as plausible causal variants at published genome-wide association study (GWAS) loci. Remarkably, 29.9% of the common SV-eQTLs are not well tagged by flanking SNVs and we observe a notable abundance (relative to SNVs and indels) of rare, high impact SVs associated with aberrant expression of nearby genes. These results suggest that comprehensive WGS-based SV analyses will increase the power of both common and rare variant association studies. We further present highlights of our ongoing work aimed at expanding this analysis to more than 600 individuals in the GTEx dataset.

170

Systematic computational identification and experimental verification of variants that activate exonic and intronic cryptic splice sites. *M. Lee[1], P. Roos[2], N. Sharma[1], T.A. Evans[1], M.J. Pellicore[1], S. Stanley[3], S. Khalil[3], A.N. Lam[1], B. Vecchio-Pagan[1], M. Armanios[3], G.R. Cutting[1].* 1) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD; 2) Miner & Kasch, Severna Park, MD; 3) Department of Oncology, Johns Hopkins University, Baltimore, MD.

Cryptic splice activation is considered a rare and exotic event; however, it has long been suspected that SNVs which directly activate splice sites are more common than currently appreciated. We developed and tested a novel variant annotation method that combines machine learning classification with a context-dependent selection algorithm to identify splice variants from complete gene sequences. A unique strength of our approach is the context-dependent selection algorithm that compares the splice potential of a sequence bearing a variant to the splice potential of the reference sequence as a high quality filter. The classifier model was trained exclusively with nucleotide feature data from >4000 canonical splice sites and accurately identified 93.3% (168 of 180) of canonical splice sites from the complete sequences of 5 disease genes: *BRCA2*, *CFTR*, *DKC1*, *HEXB*, and *LMNA*. Using variants known to cause disease as well as variants with unknown disease liability in *CFTR*, a gene whose loss of function alleles cause cystic fibrosis (CF), we demonstrate the efficacy of our method in identifying known as well as novel cryptic splice variants. The method correctly predicted the effect of 18 of 21 (86%) known *CFTR* splice variants. Of 1423 unannotated *CFTR* disease-associated variants, the method identified 32 novel cryptic splice variants. Two predicted exonic cryptic splice variants distant from canonical splice sites and labeled as missense mutations were experimentally verified as activating "deep exonic" cryptic sites. Full sequencing of the *CFTR* exons and introns in 14 CF patients with incomplete genotypes revealed 1 known (in 3 patients) and 3 novel (in 1 patient each) intronic cryptic splice variants that we experimentally verified. Application of the method to 6 individuals with dyskeratosis congenita and incomplete *DKC1* genotypes identified two intronic splice variants which caused pathologically aberrant splicing that we experimentally validated in patient lymphoblasts. Context-dependent selection generated high confidence candidate splice variants from ClinVar "pathogenic" SNVs or "variants of uncertain significance" and revealed that 28.1% (129 of 458) and 21.6% (75 of 348), respectively, were predicted to activate exonic or intronic cryptic splice sites. Our findings indicate that cryptic splice site activation appears to be a more common disease mechanism than previously expected and should be routinely considered for variants segregating with disease.

171

Major changes in mitochondrial RNA processing in human cancers. *A. Hodgkinson*[1], *Y. Idaghdour*[2]. 1) Department of Medical and Molecular Genetics, King's College London, London, United Kingdom; 2) Biology Department, Division of Science and Mathematics, New York University Abu Dhabi, United Arab Emirates.

The role of mitochondria in cancer has long been controversial. Despite the well-known Warburg effect and the presence of cancer-linked mutations in nuclear genes associated with mitochondrial processes, variation in mitochondrial DNA (mtDNA) itself has never been conclusively linked to tumor initiation or progression. Recently, work has moved beyond mutations in the mitochondrial genome to other important genetic processes, finding altered mitochondrial copy number in tumor tissue when compared to adjacent normal samples, as well as mutations in mtDNA that have altered frequencies in mtRNA - suggestive of altered RNA processing. Despite this, no large-scale analysis of tumor-specific mitochondrial post-transcriptional events has ever been carried out. Here, using integrated genomic analysis we consider changes to mitochondrial RNA processing in human cancers by analysing paired tumor and normal samples from over 600 individuals and 12 cancer types from The Cancer Genome Atlas. In doing so, we find strong and consistent patterns of altered mitochondrial processing in cancers that are also associated with changes in nuclear gene expression. Furthermore, we identify genetic markers that potentially modulate the cell's response to these changes in a tumor specific context and we link levels of altered mitochondrial RNA processing to patient survival outcomes, showing that these events are a hallmark of cancer.

172

An IL-6 targeted therapeutic modulates eQTL in whole blood. *E.E. Davenport*[1,2], *M. Gutierrez-Arcelus*[1,2], *K. Slowikowski*[1,2], *J.S. Beebe*[3], *B. Zhang*[3], *M. Vincent*[3], *S. Raychaudhuri*[1,2,4]. 1) Division of Genetics and Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA; 3) Pfizer Inc., Cambridge, MA; 4) Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK.

In order to understand the relationship between IL-6 exposure and gene regulation, we sought to identify genes whose eQTL effects were modulated by IL-6 blockade. We took advantage of data on 157 systemic lupus erythematosus (SLE) patients recruited to evaluate the efficacy and safety of a neutralizing IL-6 monoclonal antibody in a phase II clinical trial (Wallace et al. Arth Rheum 66(12) 3529-40, 2014). For this study we obtained whole blood high-depth RNA-seq profiling at 0, 12, and 24 weeks in anti-IL-6 exposed and unexposed individuals. All individuals were also genotyped genome-wide (Illumina HumanOmniExpressExome-8v1.2). We assessed whether anti-IL-6 alters the relationship between genomic variation and gene expression by identifying drug-eQTL interactions. We aligned data to the reference genome and quantified gene expression using Subread and featureCounts respectively. We conducted eQTL analysis on 379 RNA-seq samples from 157 patients. We used a linear mixed model with patient as a random effect to identify 4,363 cis eQTL genes ($p<2.3E-8 = 0.05/2,184,435$ tests). We used the most significant SNP for each of these genes to identify drug-eQTL interactions. We observed 128 drug-eQTL interactions at the $p<0.01$ level, whereas we would expect 44 by chance. To confirm that this relative enrichment was not a statistical artifact, we permuted drug exposure status 1,000 times and found no instances of >115 interactions suggesting that the number of observed interactions is highly unlikely to have happened by chance. Intriguingly we observed interactions for *IL10* and *NOD2,* suggesting that these genes are regulated by IL-6 targets. We also explored the effect of other factors such as the interferon status of the patients and cell counts. We found particularly great enrichment of eQTL interactions between high and low interferon status (164 with $p<0.01$, no instances >120 across 1,000 permutations). By using a linear mixed model and taking advantage of multiple time points for the same patient, we have improved our power to detect modulations of eQTL in response to different factors including the treatment with a targeted therapeutic. Drug-eQTL interactions in the context of a specific targeted therapeutic have the potential to lead to biological insight about the targeted moiety, and about the cascade of downstream elements regulated by it. In this specific instance, we have derived potential novel insights into the mechanism of action of IL-6 signaling in SLE.

173

Heterozygous mutations in *MAFB* cause Duane retraction syndrome and inner ear defects. *J.G. Park[1,2,3,4], M.A. Tischfield[1,2], A.A. Nugent[1,2], L. Cheng[1,2], S.A. Di Gioia[1,2], W.-M. Chan[1,2,3], G. Maconachie[5], T.M. Bosley[6], C.G. Summers[7], D.G. Hunter[1,2], C.D. Robson[1,2], I. Gottlob[5], E.C. Engle[1,2,3,8].* 1) Boston Children's Hospital, Boston, MA; 2) Harvard Medical School, Boston, MA; 3) Howard Hughes Medical Institute, Chevy Chase, MD; 4) Duke University School of Medicine, Durham, NC; 5) University of Leicester, Leicester, UK; 6) Johns Hopkins School of Medicine, Baltimore, MD; 7) University of Minnesota, Minneapolis, MN; 8) Broad Institute, Cambridge, MA.

Duane retraction syndrome is a congenital eye movement disorder defined by limited outward gaze of the eye, and retraction of the eye on attempted inward gaze. MRI and EMG studies of patients have found that the abducens nerve is absent or hypoplastic, and the lateral rectus extraocular muscle is aberrantly innervated by a branch of the oculomotor nerve instead of the abducens nerve. Although this pathology has been well characterized, the developmental etiology of Duane syndrome remains unknown. In this study, we use a combination of human genetics and a new mouse model to provide evidence that the primary cause of the disorder is a failure of the abducens nerve to innervate the lateral rectus muscle. By targeted screening of our cohort of 400 patients with Duane syndrome for mutations and copy number variations in *MAFB*, we identify three heterozygous loss-of-function mutations causing Duane syndrome and a dominant-negative mutation causing both Duane syndrome and inner ear defects. Through genotype-phenotype correlations in humans and *Mafb* knockout mice, as well as *in vitro* luciferase studies of the dominant-negative mutation, we propose a threshold model for variable loss of MAFB function causing Duane syndrome and inner ear defects. We use embryonic whole mount preparations and orbital dissections in *Mafb* knockout mice to demonstrate that selectively disrupting abducens nerve development is sufficient to cause secondary innervation of the lateral rectus muscle by aberrant branches of the oculomotor nerve. We observe that these aberrant branches of the oculomotor nerve form at developmental decision regions in close proximity to target extraocular muscles. In conclusion, we present genetic and developmental evidence that the primary etiology of Duane syndrome is failure of the abducens nerve to fully innervate the lateral rectus muscle in early development. This has important clinical implications, since our findings suggest that a wide variety of insults to the developing abducens nerve *in utero* may cause Duane syndrome.

174

Mutations in *SMCHD1* are the predominant cause of arhinia and a form of muscular dystrophy. *H. Brand[1,2,3], N. Shaw[4,5], Z.A. Kupchinsky[6], H. Bengani[7], T.I. Jones[8], L. Plummer[4], S. Erdin[1], K.A. Williamson[7], J. Rainger[7], K.E. Samocha[3,9], R.L. Collins[1], D. Lucente[1], C. Seabra[1,10], Y. An[1,11], A. Lek[12,13], S. Pereira[14], T. Kammin[14], M. Nassan[15], J.K. Rainger[7], E.C. Liao[16,17,18], C.C. Morton[3,14,19], N. Katsanis[6], J.F. Gusella[1,2,3,13], J.A. Marsh[7], D.G. Macarthur[3], W. Crowley[4], P.L. Jones[8], E.E. Davis[6], D.R. FitzPatrick[7], M.E. Talkowski[1,2,3].* 1) Molecular Neurogenetics Unit and Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, 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, Broad Institute of MIT and Harvard, Cambridge, MA; 4) Harvard Reproductive Endocrine Sciences Center and NICHD Center of Excellence in Translational Research in Fertility and Infertility, Reproductive Endocrine Unit of the Department of Medicine, Massachusetts General Hospital, Boston, MA; 5) National Institute of Environmental Health Sciences, Research Triangle Park, NC; 6) Center for Human Disease Modeling, Duke University Medical Center, Durham, NC; 7) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh Western General Hospital, Edinburgh, UK; 8) Department of Cell and Developmental Biology, University of Massachusetts Medical School, Worcester, MA; 9) Analytic and Translational Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; 10) GABBA Program, University of Porto, Porto, Portugal; 11) Institute of Biomedical Sciences, Medical School, Fudan University, Shanghai China; 12) Genetics and Genomics, Manton Center for Orphan Disease Research, Children's Hospital, Boston, MA; 13) Department of Genetics, Harvard Medical School, Boston, MA; 14) Department of Obstetrics, Gynecology, and Reproductive Biology, Brigham and Women's Hospital, Boston, MA 02115, USA; 15) Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, USA; 16) Center for Regenerative Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA; 17) Division of Plastic and Reconstructive Surgery, Massachusetts General Hospital, Boston, MA; 18) Harvard Stem Cell Institute, Cambridge, MA; 19) Department of Pathology, Brigham and Women's Hospital, Boston, MA.

Arhinia, or the complete absence of an external nose, is a rare congenital malformation whose etiology is unknown (see companion abstract by Shaw et al. for analyses of the associated phenotypic spectrum). We established an international consortium of 35 arhinia cases (~20% of all reported cases in the medical literature and 18 new cases) to identify the genetic cause of this rare malformation using whole-genome, whole-exome, and targeted gene sequencing. We compared the rare mutation burden in cases for every gene in the genome to 60,706 subjects from the Exome Aggregation Consortium (ExAC) and identified only one gene, *SMCHD1* that exceeded genome-wide significance (p=2.9x10^{-17}). We identified heterozygous missense mutations in *SMCHD1* that were not observed in ExAC among 77% of arhinia cases, including five recurrent sites. SMCHD1 functions to maintain X-inactivation, silence autosomal repeat sequences, and regulate genomic imprinting. Intriguingly, loss-of-function (LoF) mutations in *SMCHD1* cause type 2 facioscapulohumeral dystrophy (FSHD2) in a complex genetic model in which hypomethylation of the 4q35 D4Z4 repeat array in the presence of a disease-permissive haplotype and reduced D4Z4 array size lead to disease. Using ExAC, we found that SMCHD1's intolerance to missense variation (constraint) showed regional variability; the region of the gene that includes the 5' GHKL-type ATPase domain is constrained, while the 3' region of the gene is not. Notably, all arhinia associated mutations are tightly clustered in proximity to this ATPase domain, whereas FSHD2-specfic variants span the entire gene. We explored the functional mechanisms by which *SMCHD1* mutations cause arhinia through RNAseq in arhinia subjects, CRISPR/Cas9 studies in mouse embryos and zebrafish models, and methylation analyses of the D4Z4 repeat array associated with FSHD2. These analyses suggest that alteration of *SMCHD1* can recapitulate the nasal, ocular, and reproductive defects observed in subjects with arhinia. *SMCHD1* mutations associated with arhinia also displayed FSHD2-like DNA hypomethylation signatures at the D4Z4 repeat array, although there have been no reported cases of comorbidity between these disorders, suggesting pleiotropic effects of *SMCHD1* and the possibility of a modifier locus for arhinia. The results from this study indicate distinct phenotypic outcomes associated with functional mutations in *SMCHD1*, and further studies to understand this mechanism are ongoing.

175

The phenotypic spectrum of arhinia associated with mutations in *SMCHD1*: From isolated arhinia to Bosma arhinia microphthalmia syndrome.

N.D. Shaw[1,2], H. Brand[2], Z.A. Kupchinsky[3], J.M. Graham[4], J.R. Willer[3], A. Verloes[5], A. Rauch[6], K. Steindl[6], L.A. Schimmenti[7], B. Brasseur[8], C. Cesaretti[9], J.E. Garcia-Ortiz[10], T.P. Buitrago[11], O.P. Silva[12], B. Loeys[13], A. Kaindl[14], C.H. Cho[15], J. Law[16], N. Ferraro[17], D. Sato[18], C. Jacobsen[17], J. Tryggestad[19], J.D. Hoffman[20], V. van Heyningen[21], S.B. Seminara[2], W.J. Crowley[2], A. Lin[2], D.R. Fitzpatrick[21], M.E. Talkowski[2,22], E.E. Davis[3]. 1) National Institute of Environmental Health Sciences, Research Triangle Park, NC; 2) Massachusetts General Hospital, Boston, MA; 3) Duke University Medical Center, Durham, NC; 4) Cedars Sinai Medical Center, Los Angeles, CA; 5) Robert Debré Hospital, Paris, France; 6) University of Zurich, Zurich, Switzerland; 7) Mayo Clinic, Rochester, Minnesota; 8) University of Miami Leonard M Miller School of Medicine, Miami, Florida; 9) Fondazione IRCCS Ca` Granda, Ospedale Maggiore Policlinico, Milan, Italy; 10) Centro de Investigación Biomédica de Occidente, Guadalajara, México; 11) Fundación Hospital Infantil Universitario de San José, Bogota, Columbia; 12) Private Plastic Surgical practice, Columbia; 13) University Hospital of Antwerp, Antwerp, Belgium; 14) Charité University Medicine, Berlin, Germany; 15) The University Hospital of Bern, Bern, Switzerland; 16) University of North Carolina at Chapel Hill, Chapel Hill, NC; 17) Boston Children's Hospital, Boston, MA; 18) Hokkaido University Graduate School of Medicine, Sapporo, Japan; 19) University of Oklahoma Health Sciences Center, Oklahoma City, OK; 20) Boston Medical Center, Boston, MA; 21) University of Edinburgh Western General Hospital, Edinburgh, UK; 22) Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA.

Arhinia is a severe craniofacial anomaly of unknown genetic etiology. In a companion abstract (Brand et al.), we report the identification of missense mutations in a constrained domain of *SMCHD1* in 76% of arhinia cases. Here, we review the complex phenotypic spectrum associated with these mutations through an exhaustive literature review and detailed phenotyping of a newly assembled international cohort of 35 cases (17 previously reported, 18 new) as well as associated phenotypes from zebrafish models. Nearly all of the 80 arhinia cases in the literature had extensive craniofacial defects including colobomatous microphthalmia, cataracts, absent nasolacrimal ducts, choanal atresia, and cleft palate. Seventeen subjects had ocular defects and reproductive failure due to hypogonadotropic hypogonadism (HH), a triad called Bosma arhinia microphthalmia syndrome (BAM; OMIM 603457). All cases had abnormal palates, paranasal sinuses, nasolacrimal duct stenosis, and/or maxillary deficiency, yet ocular and reproductive phenotypes were variable. The most common ocular findings were microphthalmia (59%), colobomas (50%), and cataracts (27%), but 3 subjects (9%) had normal eye anatomy and vision. Defects were restricted to the craniofacial region with the exception of inguinal hernias and cryptorchidism. There was great intra- and interfamilial variability in the severity of the phenotype. Among the 26 subjects (18 male; 8 female) for whom the reproductive axis could be assessed, all demonstrated reproductive failure due to HH and absent olfactory structures on brain MRI, consistent with gonadotropin releasing hormone (GnRH) deficiency; 23/26 subjects also had ocular defects, suggesting a high prevalence (88%) of BAM in the presence of arhinia. Overall, 78% of subjects meeting diagnostic criteria for BAM harbored a *SMCHD1* mutation, similar to the overall cohort. Modeling of altered *smchd1* in larval zebrafish recapitulated phenotypes observed in BAM, including maldevelopment of pharyngeal skeleton cartilage, small eyes, and blunted GnRH neuronal projections, each of which could be rescued by injection of wild-type *SMCHD1* mRNA. *SMCHD1* encodes an epigenetic repressor implicated in a rare muscle disease (FSHD2). Intriguingly, we identified a p.G137E mutation previously described in a patient with FSHD2 in one subject with BAM. These data indicate that similar *SMCHD1* mutational mechanisms may result in pleiotropic effects leading to diverse phenotypic outcomes.

176

***De novo* gain-of-function mutations in the epigenetic regulator *SMCHD1* cause Bosma arhinia microphthalmia syndrome.**

C. Gordon[1], S. Xue[2], G. Yigit[3], H. Filali[1], K. Chen[4], N. Rosin[3], K. Yoshiura[5], M. Oufadem[1], T. Beck[4], A. Sefiani[6], H. Kayserili[7], J. Murphy[4], C. Chatdokmaiprai[8], A. Hillmer[9], D. Wattanasirichaigoon[8], S. Lyonnet[1], A. Javed[9], M. Blewitt[4], J. Amiel[1], B. Wollnik[3], B. Reversade[2]. 1) Institut Imagine, INSERM U-1163, Paris, France; 2) Institute of Medical Biology, A*STAR, Singapore; 3) Institute of Human Genetics, Göttingen, Germany; 4) The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia; 5) Nagasaki University, Nagasaki, Japan; 6) Institut National d'Hygiène, Rabat, Morocco; 7) Koç University School of Medicine, Istanbul, Turkey; 8) Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand; 9) Genome Institute of Singapore, A*STAR, Singapore.

Bosma arhinia microphthalmia syndrome (BAMS) is an extremely rare and striking condition characterized by complete absence of the nose (arhinia) with or without ocular defects. Arhinia is presumed to result from a specific defect of the nasal placodes or surrounding neural crest-derived tissues during embryonic development. By exome sequencing we identified missense mutations in the extended ATPase domain of the epigenetic regulator Structural Maintenance of Chromosomes Flexible Hinge Domain Containing 1 (*SMCHD1*) as the cause of BAMS in all 14 cases studied. All mutations were de novo where parental DNA was available. ATPase assays using wild-type or mutant versions of purified SMCHD1 protein indicated that the BAMS mutations increase the catalytic activity of the protein. In overexpression assays in Xenopus embryos we observed that injection of *SMCHD1* RNA harboring a BAMS mutation resulted in more severe frontonasal and eye hypoplasia than injection of wild-type *SMCHD1*. These functional assays suggest that the BAMS mutations behave as gain-of-function alleles. This is in contrast to loss-of-function mutations in *SMCHD1* that have been associated with facioscapulohumeral muscular dystrophy (FSHD) type 2, a disorder with no phenotypic overlap with BAMS. In FSHD type 2, loss of the epigenetic silencing activity of *SMCHD1* results in pathogenic misexpression of the transcription factor *DUX4* in skeletal muscles. Our results establish *SMCHD1* as a key player in nasal development and provide biochemical insight into its enzymatic function that may be exploited for development of therapeutics for FSHD.

177

Identification of non-coding variants at 1p22 that are pathogenic for nonsyndromic orofacial clefting. *R. Cornell[1], H. Liu[1], E. Leslie[2], M. Dunnwald[1], M. Marazita[2], A. Lidral[3].* 1) Dept of Anatomy and Cell Biology, College of Medicine, University of Iowa, Iowa City, IA; 2) Cenetr for Craniofacial and Dental Genetics, Dept. of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania; 3) Dept of Orthodontics, College of Dentistry, University of Iowa, Iowa City, Iowa.

Genome wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) in which particular alleles are strongly associated with risk for orofacial clefting. However, it is challenging to identify the biological mechanisms underlying such risk. First, only a fraction of risk-associated SNPs are pathogenic whereas the rest are “rider SNPs,” and, second, most risk-associated SNPs lie in non-coding DNA. In a resequencing project, we identified many orofacial-clefting-associated SNPs that are in linkage disequilibrium with the GWAS lead SNP at 1p22. To distinguish between pathogenic and rider SNPs at this locus, we first amplified DNA elements containing the 10 most-strongly associated SNPs, engineered the elements to contain risk or non-risk alleles of the resident SNPs, and tested their enhancer activity *in vitro* in oral-epithelium and palate mesenchyme cells. Three SNPs (rs4147828, rs2275035, and rs560426) had allele-dependent effects on the enhancer activity of elements in oral-epithelium cells; in every case, the risk allele reduced activity. Chromatin configuration capture confirmed that the SNP-encompassing elements interact with the promoter of *ARHGAP29*. In one case, this interaction was disrupted when the resident SNP, rs4147828, harbored the risk allele. In transgenic zebrafish, elements containing two of the identified SNPs had enhancer activity in cranial epidermis, oral epithelium and oral mesenchyme, consistent with the pattern of *arhgap29b* expression. Using CRISPR/Cas9-mediated genome editing in oral-epithelium cells, we found that expression of *ARHGAP29* was reduced when elements containing these candidate pathogenic SNPs were deleted, or when such elements were engineered to harbor risk instead of non-risk alleles. Finally, chromatin immuno-precipitation experiments revealed that the risk allele of rs4147828 reduces binding of MAFB whereas that of rs2275035 elevates binding of KLF4, and over-expression analyses indicated that both events reduce the activity of the enhancers that contain these SNPs. Cumulatively these results indicate that risk-associated alleles of rs4147828, rs2275035, rs560426 reduce expression of *ARHGAP29* in oral epithelium by altering the binding of specific transcription factors; haplotype analyses suggest they act in concert. This study has helped to translate genetic risk, a statistical observation, into an understanding of mechanisms that underlie pathogenesis.

178

Biallelic sequence variants in *INTS1* in patients with developmental delays, cataracts and craniofacial anomalies. *A. Slavotinek[1], D. Lessel[2], A.M. Innes[3], M. Krall[1], D. Schneidman[4,5], R. Lamont[3], D. Baillat[6], E. Wagner[6], G. Mancini[7].* 1) Dept Pediatrics, Division Genetics, Univ California, San Francisco, San Francisco, CA; 2) Institute of Human Genetics, Univ Medical Center Hamburg Eppendorf, Hamburg, Germany; 3) Dept. Medical Genetics and Alberta Children's Hospital Research Institute, University of Calgary, Alberta, Canada; 4) Dept Bioengineering and Therapeutic Sciences, Univ California San Francisco, San Francisco, CA; 5) Dept Pharamceutical Chemistry, Univ California San Francisco, San Francisco, CA; 6) Dept Biochemistry and Molecular Biology, Univ Texas Medical Branch at Galveston, Galveston, TX; 7) Dept Clinical Genetics, Erasmus Univ Medical Center, Rotterdam, The Netherlands.

We present 7 children with biallelic sequence variants in the Integrator complex subunit 1 (*INTS1*) gene. Ages ranged from 1 to 19 years and there were 2 sib pairs. All children had significant developmental delays and 5 had minimal speech, 5 had hypotonia and 5 had an abnormal gait. All children had hypertelorism and a broad nasal tip and 6 children shared similar dysmorphic features, with downslanting palpebral fissures, low-set, simple ears and epicanthic folds. Other distinctive findings were short stature, frontal bossing, downturned corners of the mouth, a wide gap between the upper incisors or cleft lip and palate, pectus deformities and broad overlapping toes. Six children developed juvenile cataracts. Microphthalmia and colobomas, unilateral renal agenesis, renal dysplasia and horseshoe kidney, ventricular septal defect and severe pulmonary hypertension were also observed. There were 4 missense, one frameshift and one nonsense variant that was also predicted to disrupt splicing. All variants were predicted to be disease-causing and none were present in 1000 Genomes or as homozygous variants in the ExAC browser. Five of the 6 variants affected the C-terminus of the protein and preliminary modeling showed that the variant p.Pro1874Leu may interfere with helix folding and p.Arg2100Cys with disulfide bond formation. *INTS1* transcripts are present in human fetal brain and the corpus callosum and cerebellum of adult brain. We used *in-situ* hybridization to demonstrate expression of *ints1* in the developing zebrafish eye. The integrator (INTS) complex is conserved from Drosophila to human and comprises 14 subunits that associate with RPB1, the largest subunit of RNA polymerase II, to catalyze endonucleolytic cleavage of nascent snRNAs and to assist RNA polymerase II in promoter-proximal pause-release on protein coding genes. Homozygous knockout mice for *Ints1* have arrested development at the early blastocyst stage, with increased levels of unprocessed, primary snRNA transcripts compared to heterozygotes. A severe neurodevelopmental disorder has been observed in a sibship with biallelic *INTS8* mutations (Oegema et al., in preparation). In these 7 children with *INTS1* sequence variants, a distinctive phenotype was present in 6 individuals from 4 unrelated families. The phenotypes associated with an integrator complex gene point to a new mechanism for intellectual disability, eye defects, craniofacial and skeletal anomalies in humans.

179

Mutations in *MYT1*, encoding the myelin transcription factor, are a rare cause of OAVS, within the RA signaling pathway.

C. Rooryck[1,2], M. Berenguer[1], E. Lopez[1], A. Tingaud-Sequeira[1], S. Marlin[3], A. Toutain[4], F. Denoyelle[5], A. Picard[6], S. Charron[1], G. Mathieu[1], H. de Belvalet[1], B. Arveiler[1,2], P.J. Babin[1], S. Bragagnolo[7], A.B. Perez[7], M.I. Melaragno[7], M. Colovati[7], D. Lacombe[1,2]. 1) Univ. Bordeaux, Maladies Rares: Génétique et Métabolisme (MRGM),U 1211 INSERM, F-33000 Bordeaux; 2) CHU de Bordeaux, Service de Génétique Médicale, Centre de Référence Anomalies du Développement et Syndromes Malformatifs, F-33000, Bordeaux, France; 3) Hôpital Universitaire Necker-Enfants-Malades, Département de Génétique, Centre de Référence des Surdités Génétiques, F-75015, Paris, France; 4) Hôpital Bretonneau, Service de Génétique, Centre Hospitalier Universitaire, F-37044, Tours, France; 5) Hôpital Universitaire Necker-Enfants-Malades, Service d'ORL pédiatrique et de chirurgie cervicofaciale, Centre de Référence des malformations ORL rares, F-75015, Paris, France; 6) Hôpital Universitaire Necker-Enfants Malades, Service de chirurgie maxillo-faciale, F-75015, Paris, France; 7) Genetics Division, Universidade Federal de Sao Paulo, Sao Paulo, Brazil.

Goldenhar syndrome (GS) or Oculo-Auriculo-Vertebral Spectrum (OAVS) is a developmental disorder involving first and second branchial arches, characterized by asymmetric ear anomalies, hemifacial microsomia, ocular defects, and vertebral malformations. Although numerous chromosomal abnormalities have been associated with OAVS, no causative gene has been identified so far. Among other environmental factors, Retinoic Acid (RA) has already been described as a teratogenic agent leading to OAVS features in humans. As sporadic cases are mostly described in GS, we have performed whole exome sequencing on selected affected individuals and their unaffected parents, looking for *de novo* mutations. Consequently, we identified a heterozygous nonsense mutation in one patient in the *MYT1* gene. Further, we detected two other heterozygous missense mutations in two unrelated patients, from a cohort of 240 OAVS patients. One of these missense mutations also arose *de novo*, while the other was inherited from a father with incomplete phenotype. This gene encodes the Myelin Transcription factor 1 which is highly expressed in the developing central nervous system. Functional studies by transient knockdown of *myt1a,* homolog of *MYT1* in zebrafish, led to specific craniofacial cartilages alterations and to the up-regulation of Neural Crest Cells marker *sox10*. Moreover, cells studies confirmed close links between MYT1, RA and the RA receptor beta (RARB). Indeed, All-trans RA (ATRA) treatment led to an upregulation of cellular endogenous *MYT1* expression. Additionally, cellular wild-type *MYT1* overexpression induced a down-regulation of *RARB* leading to a negative feedback of the RA signaling pathway, whereas mutated *MYT1* did not, confirming the pathogenic effect of the mutations. Overall, we report *MYT1* as a candidate gene implicated in OAVS, within the RA signaling pathway.

180

A genotype-first approach identifies gain-of-function mutations of *TFE3* in a novel syndrome with intellectual disability, seizures, facial dysmorphism, short stature and obesity.

D. Lehalle[1,2], M. Avila[2], L. Duplomb-Jego[2], Y. Duffourd[2], P. Kuentz[1], J. St-Onge[2], T. Jouan[2], J. Thevenon[1,2], C. Thauvin-Robinet[1,2], P. Vabres[1], L. Faivre[1,2], J. Betschinger[4], J.B. Rivière[1,3,5]. 1) Fédération Hospitalo-Universitaire Médecine Translationnelle et Anomalies du Développement (TRANSLAD), Centre Hospitalier Universitaire Dijon, 21079 Dijon, France; 2) Centre de Génétique et Centre de Référence Anomalies du Développement et Syndromes Malformatifs de l'Interrégion Est, Centre Hospitalier Universitaire Dijon, 21079 Dijon, France; 3) Child Health and Human Development Program, Research Institute of the McGill University Health Centre, Montreal, QC H4A 3J1, Canada; 4) Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland; 5) Department of Human Genetics, Faculty of Medicine, McGill University, Montreal, QC H3A 1B1, Canada.

Whole exome sequencing (WES) has proven to be a powerful tool for discovering *de novo* germline and postzygotic variants responsible for sporadic developmental disorders. It has also facilitated the identification of novel syndromes using a “genotype-first” approach. In the present study, the index case was a 19-year-old female from unrelated healthy parents. She had severe intellectual disability, seizures, coarse facial features, short stature, obesity, and skin hypopigmentation along Blaschko's lines. This combination of features did not match any known syndrome. We performed WES in the patient and her parents, and identified a *de novo* germline variant in *TFE3* (NM_006521.4:c.356A>C; p.Gln119Pro), which is located on the X chromosome and encodes a key regulator of stem cell pluripotency. We identified a second *de novo* missense change of *TFE3* (NM_006521.4:c.557C>T; p.Pro186Leu) in an unrelated female patient from our cohort of clinical WES. Her clinical presentation was strikingly similar to the index case, except for the skin hypopigmentation. It has been shown that nuclear exclusion of TFE3 is required for embryonic stem (ES) cell differentiation, thus providing a functional assay for testing activity of mutant TFE3 alleles. We generated TFE3 knockout (KO) mouse ES cells using CRISPR/Cas9. Doxycycline-inducible wildtype (WT) and mutant transgenes (with the corresponding murine mutations p.Gln119Pro and p.Pro186Leu) were stably transfected into these cells. Doxycycline-dependent expression of both TFE3 mutants induced transcription of the TFE3 targets APOE and TRPM1, indicating production of functional proteins. We analyzed subcellular localization of the transgenes. While the WT protein, similar to endogenous TFE3, was detected in cytoplasm and nucleus, both mutants were predominantly nuclear. Since ectopic nuclear TFE3 impairs ES cell differentiation, we tested the ability of TFE3 mutants to interfere with differentiation. Exit from the ES cell state is accompanied by loss of self-renewal, but this process was severely impaired upon overexpression of mutant but not WT TFE3. Taken together these findings suggest that patient mutations give rise to nuclear confined TFE3 gain-of-function alleles. In conclusion, our data demonstrate that gain-of-function mutations of *TFE3* in females cause a previously unrecognized syndrome characterized by severe intellectual disability, seizures, coarse facial features, short stature and obesity.

181

Histone methylation-demethylation defects in forms of intellectual disability and refractory epilepsy. *L. Poeta[1], A. Padula[1], A. Ranieri[1], B. Attianese[1], M. Valentino[1], C. Shoubridge[2], K. Helin[3], J. Gecz[2], E. Di Schiavi[4], S. Filosa[4], C. Schwartz[5], L. Altucci[1,6], H. van Bokhoven[7], M.G. Miano[1].* 1) Institute of Genetics and Biophysics National Council of Research, Naples, Naples, Italy; 2) Department of Paediatrics, University of Adelaide, South Australia 5006, Australia; 3) Centre for Epigenetics, University of Copenhagen, Copenhagen DK-2200, Denmark; 4) Institute of Bioscience, CNR, Naples 80131, Italy; 5) Greenwood Genetic Center, Greenwood, South Carolina. 29646, USA; 6) Second University of Naples, Naples 80100, Italy; 7) Radboud University Nijmegen Medical Centre, 6500 HB, Nijmegen, The Netherlands.

Mistakes in histone methylation-demethylation rounds have been directly involved in several forms of Intellectual Disability (ID) with Epilepsy and/or Refractory Epilepsy (RE). Lysine-specific demethylase 5C (KDM5C) is an X-linked gene, which encodes a chromatin JmjC eraser with H3K4me2/3 demethylase activity. KDM5C is frequently mutated in a spectrum of X-linked ID (XLID) and/or RE. It functions as a transcriptional repressor that is critical for transition of neural progenitors to neurons. We identified a disease path, linking functionally KDM5C to another XLID/Epilepsy gene, encoding the homeotic transcription factor ARX, whose mutations impair severely KDM5C transcript regulation. Furthermore, we analysed two additional XLID proteins that also bind KDM5C promoter. They are PHD Finger Protein 8 **(**PHF8), a H3K9 demethylase; and Zinc Finger Protein 711 (ZNF711), a transcriptional factor, which role is almost unknown. We observed that PHF8 and ZNF711, which co-occupy the target promoter, induce cooperatively the KDM5C stimulation. This activity seems to be ARX-independent and we propose that the transcriptional induction by ARX does not synergize with the action of the PHF8/ZNF711 complex. Moreover, in patient-derived cell lines mutated in the KDM5C path, we found a global defect of H3K4me3 signalling, potentially due to a compromised KDM5C activity. Because chromatin modifications are reversible, it is possible that epigenetic drugs could compensate for the KDM5C-H3K4me3 deregulation. We screened a number of compounds targeting chromatin enzymes. We used as cell disease model neuronally-differentiated Arx KO/Kdm5C-depleted ES cells that show GABAergic abnormalities in association with a global increase of H3K4me3 signal. A strong compensation of KDM5C downregulation has been obtained at crucial time-point of neuronal maturation. We tested the epi- treatments in C. elegans mutants of Alr-1 and Rbr-2, the homologous counterparts of human ARX and KDM5C, respectively. Both mutants present specific defects in neuronal structures and functions. Ongoing efforts will allow us to identify druggable hallmarks that could open up towards the exploitation of potential strategies to treat the growing group of ID and RE diseases caused by defects in chromatin and/or transcriptional regulators.

182

Shared therapeutic approaches are justified by common morphological and transcriptome changes in Rett spectrum disorders. *E. Landucci[1], L. Bianciardi[1], S. Daga[1], A.M. Pinto[1,2], E. Frullanti[1], M. Brindisi[3], S. Butini[3], V. Imperatore[1], F. Ariani[1,2], S. Brogi[3], G. Campiani[3], A. Renieri[1,2], I. Meloni[1].* 1) Medical Genetics, University of Siena, Siena, Italy; 2) Genetica Medica, Azienda Ospedaliera Universitaria Senese, Siena, Italy; 3) European Research Centre for Drug Discovery & Development Dept. Biotechnology, Chemistry and Pharmacy University of Siena, Siena, Italy.

Rett spectrum (RTT) disorders are among the most frequent and extensively studied intellectual disabilities but specific therapies are still missing. In 2008, we unraveled that, in addition to the classic *MECP2*-related form, a *FOXG1*-related congenital variant showing a shorter perinatal normal period and more severe microcephaly also spans RTT spectrum. Both *MECP2* and *FOXG1* encode transcriptional regulators playing fundamental roles during brain development. Here we change the perspective of RTT field identifying a common molecular signature and common druggable targets. We took advantage of the breakthrough genetic reprogramming technology to generate iPSCs as patient-specific human disease model. Five different clones of iPSC-derived neurons coming from 4 patients (2 mutated in *MECP2* and 2 in *FOXG1* with different mutations) were analyzed and compared with 2 controls. Such analysis led to the identification of a unique neuronal morphological phenotype showing impairment of neuronal networking capability. RNA-seq transcriptome analysis identified hundreds of deregulated genes in common to both *MECP2* and *FOXG1*-mutated neurons. Overall, we identified a disruption of few driver pathways explaining the morphological alterations and including neuronal projection morphogenesis, extracellular matrix assembly, GABA-ergic signaling and microtubules assembly. For the first time we demonstrate here that the similarity of patient signs/symptoms corresponds to a similarity of phenotype in iPSC-derived neurons and that both are due to underlying common molecular defects. Indeed, in vitro experiments with drugs selective for the tubulin acetylation pathway (a newly developed HDAC6 inhibitor) and the GABA-ergic circuits (a repurposed drug) showed a significant reversal of the unique discovered morphological phenotype. These findings completely change the molecular view of RTT pathomechanisms and open a real possibility of an efficacious common treatment for RTT disorders.

183

Diet rescues lethality in a model of *NGLY1* deficiency, a rare deglycosylation disorder.

C.Y. Chow, K.G. Owings. Department of Human Genetics, University of Utah, Salt Lake City, UT.

Autosomal recessive loss-of-function mutations in *N-Glycanase 1* (*NGLY1*) cause *NGLY1* deficiency, the only known human disease of deglycosylation. *NGLY1* deficiency is a devastating, extremely rare, neglected disease. Patients with *NGLY1* deficiency present with developmental delay, movement disorder, seizures, hypotonia, liver dysfunction, and alacrima. NGLY1 is a conserved component of the endoplasmic reticulum associated degradation (ERAD) pathway. ERAD is responsible for degrading misfolded proteins that accumulate in the lumen of the ER. NGLY1 deglycosylates misfolded proteins in the cytoplasm as they are translocated from the ER lumen for degradation. While little is known about the pathogenesis underlying *NGLY1* deficiency, it is thought that loss of NGLY1 activity results in accumulation of highly N-glycosylated misfolded proteins in the cytoplasm, acting as a 'sink' for free UDP-GlcNAc. In turn, this might deplete the circulating pool of UDP-GlcNAc in the cell, resulting in disease. We hypothesized that restoring the levels of UDP-GlcNAc in the cells might rescue some of the phenotypes associated with *NGLY1* deficiency. We used ubiquitous RNAi knockdown of *Pngl* (*Drosophila* ortholog of *NGLY1*) in *Drosophila* to model complete loss of NGLY1 activity seen in human patients. We show that supplementing the normal *Drosophila* diet with GlcNAc can rescue lethality associated with loss of Pngl activity. When raised on normal food, without GlcNAc supplementation, we observed significant lethality, with eclosion of only 18% of expected *Pngl* knockdown adults. *Pngl* knockdown lethality occurs throughout larval and pupal development. When diet is supplemented with 100 ug/ml of GlcNAc, we observed significant rescue of the developmental lethality, raising the adult *Pngl* knockdown eclosion rate to nearly 70%. We also demonstrate that genetic alterations in the ERAD and cytoplasmic heat shock pathways can influence the lethality of *Pngl* knockdown. Finally, through a natural genetic modifier screen in the *Drosophila* Genetic Reference Panel (DGRP), we provide further evidence that the ERAD and cytoplasmic heat shock pathways are defective in *NGLY1* deficiency. These data suggest a plausible pathophysiology for *NGLY1* deficiency. More importantly, our study points to a potential therapy through a simple, readily available, diet supplement.

184

A platform for genotype-phenotype correlation in iPSC-derived oligodendrocytes identifies patient-specific defects in children with Pelizaeus-Merzbacher Disease.

Z. Nevin[1], R. Karl[1], D. Factor[1], P. Douvaras[3], J. Laukka[4], V. Fossati[3], G. Hobson[2], P. Tesar[1]. 1) Genetics and Genome Sciences, Case Western Reserve University School of Medicine, Cleveland, OH; 2) Alfred I. duPont Hospital for Children, Nemours Biomedical Research, Wilmington, DE; 3) The New York Stem Cell Foundation, New York, NY; 4) Departments of Neuroscience and Neurology, University of Toledo College of Medicine and Life Sciences, Toledo, OH.

Pelizaeus-Merzbacher Disease (PMD [MIM 312080]) is a pediatric leukodystrophy that presents with a wide spectrum of clinical severity. Although PMD is a rare disease, hundreds of different mutations in the X-linked myelin gene *proteolipid protein 1* (*PLP1* [MIM 300401]) have been identified in patients. Attempts to identify a common pathogenic process connecting this multitude of mutations to patients' diverse clinical presentations have been complicated by an incomplete understanding of the function of PLP1 and limited access to primary human oligodendrocytes (OLs). To address these issues, we developed a platform to efficiently model patient-relevant mutations that recapitulate the clinical and genetic heterogeneity seen in PMD. We generated a panel of induced pluripotent stem cells (iPSCs) from 12 patients with point mutations, duplication, triplication, and deletion of *PLP1* and differentiated two iPSC clones per patient into OLs and oligodendrocyte progenitors (OPCs). RNA sequencing in iPSCs identified a *PLP1* splicing defect caused by an intron mutation and revealed that duplication and triplication lines express *PLP1* mRNA 2- and 3-fold higher, than controls ($p<0.05$). Differentiation of the panel to OPCs revealed a 50% reduction in the number of PDGFRa+ OPCs averaged across PMD lines compared to controls ($p<0.01$), with high correspondence between clones of a given patient, but high variability between patients (3% to 46% of culture). Differentiation of OPCs to OLs identified four categories of PMD OL morphology defects: no O4+ OLs (n=2), O4+/PLP1- OLs (n=6), disorganized O4+/PLP1+ OLs (n=1), and O4+ OLs with perinuclear retention of PLP1 (n=3). This last category was suggestive of protein misfolding, so a duplication and severe point mutation were treated with two small molecule modulators of endoplasmic reticulum stress, both of which improved OL morphology in both PMD lines. These compounds were further tested on OLs co-cultured with rat dorsal root ganglion neurons (DRGs), a model of myelination. Under these conditions, each compound only rescued either the duplication or the point mutation, suggesting that although both mutations result in perinuclear retention of PLP1, there are distinct pathogenic processes involved. These studies highlight the utility of an iPSC-based platform for discerning patient-specific developmental, molecular, and cellular defects in myelin diseases and inform the pursuit of new patient-specific therapies.

185

Exon inclusion for the treatment of splice site mutation in merosin-deficient congenital muscular dystrophy. *D.U. Kemaladewi, E. Hyatt, M. Ding, X. Zhu, E.A. Ivakine, R.D. Cohn.* Program in Genetics and Genome Biology, Hospital for Sick Children, Toronto, Ontario, Canada.

Merosin-deficient congenital muscular dystrophy (MDC1A) is caused by mutations in the *LAMA2* gene encoding α2 chain of Laminin, a protein that is important for maintaining skeletal muscle stability. Consequently, MDC1A patients suffer from severe muscle degeneration and accumulation of fibrotic tissue, leading to the loss of skeletal muscle functions. Efforts have been directed toward the development of auxiliary therapies targeting fibrosis, yet there have been no studies aimed at correction of the genetic defect in MDC1A. The discovery of CRISPR/Cas9 genome editing technology has opened up avenues for the development of novel treatment strategies. Here, we sought to correct the splicing defect in the MDC1A mouse model dy^{2J}/dy^{2J} by creating a novel splice donor site using CRISPR/Cas9 system and assess its therapeutic potential. The dy^{2J}/dy^{2J} mice carry a c.417+1 g>a splice site mutation in the *Lama2* gene, causing exclusion of exon 2 and production of a shorter, nonfunctional Laminin α2 protein. We utilized *S. aureus* Cas9 and two guide RNAs to excise an intronic region flanking the mutation and a putative splice donor site in the *Lama2* gene. Following the precise non-homologous end-joining, we successfully created a functional, alternative splice donor site in the *Lama2* intron 2 region in dy^{2J}/dy^{2J} myoblasts, resulting in an inclusion of the previously missing exon 2 in *Lama2* mRNA. Subsequently, using Adeno-associated virus serotype 9 (AAV9) as a delivery vehicle, we observed exon 2 inclusion and restoration of the full-length Lama2 protein in skeletal muscles of dy^{2J}/dy^{2J} mice. Importantly, we observed significant improvement in muscle histopathology and locomotion activity in the treated mice, indicating the therapeutic potential of this strategy. Collectively, our data demonstrate the feasibility and therapeutic benefit of CRISPR/Cas9-mediated correction of the splice site mutation in *Lama2 in vitro* and *in vivo*, which opens up an entirely new treatment strategy for approximately 30% of the MDC1A patient population. Finally, our novel splicing modulation approach can potentially be applied to disease-causing intronic mutations that alter splice site recognition, providing an attractive therapeutic strategy for various inherited diseases.

186

Antisense oligonucleotide therapy for the fatal epilepsy Lafora disease. *T.R. Grossman[1], S. Ahonen[2], J. Turnbull[2], L.A. Hettrick[1], H. Kordasiewicz[1], M. Katz[1], M.L. McCaleb[1], P. Wang[2], X. Zhao[2], B.A. Minassian[2].* 1) Antisense Drug Discoverey, IONIS PHARMACEUTICALS, Carlsbad, CA; 2) Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Canada.

Lafora disease (LD) is an autosomal recessive, progressive myoclonus epilepsy (OMIM #254780), exhibiting cognitive decline and escalating myoclonic, visual, convulsive, and other seizures, with onset typically in teenagers followed by decline and death usually within 10 years. LD is caused by mutations in either EPM2A (laforin) or EPM2B (malin) genes. The disease is characterized by the presence of Lafora bodies (LB) which contain polyglucosan, a poorly branched form of glycogen, in neurons, muscle and other tissues. LD mouse models deficient in laforin or malin, exhibit similar pathologies to LD patients including LB formation, widespread degeneration of neurons, impaired behavioral responses, ataxia, spontaneous myoclonic seizures and EEG epileptiform activity. Previous work has shown that knockout of the glycogen synthase Gys1 could rescue neurological defects in LD mouse model. To evaluate the therapeutic potential of brain Gys1 inhibition, we used antisense oligonucleotides (ASO) to knockdown Gys1 in two LD mouse models. Gys1 ASO administration resulted in a dramatic reduction in brain Gys1 mRNA and protein leading to a dramatic reduction in brain glycogen levels, inhibited LB formation and a robust reduction in neurodegeneration. ASOs with similar chemistries directed to targets in the CNS are currently in clinical trials for other neurological diseases and are presently delivered to patients by lumbar puncture. Our results provide a positive proof of concept that supports development of ASO therapy for Lafora disease patients that could potentially provide much needed relief to one of the severest diseases of adolescence. .

187

Minimal cerebellar genetic modification associated with ASO reduction of ATXN2 expression and delayed SCA2 mouse motor and neurophysiological phenotypes. *S.M. Pulst[1], M. Schneider[1], P. Meera[2], K. Figueroa[1], F. Rigo[3], F. Bennett[3], T. Otis[2], D.R. Scoles[1].* 1) Department of Neurology, University of Utah, Salt Lake City, UT; 2) University of California Los Angeles, Los Angeles, CA; 3) Ionis Pharmaceuticals, Carlabad, CA.

Spinocerebellar ataxia type 2 (SCA2) is caused by CAG repeat expansion in the ATXN2 gene resulting in polyglutamine expanded ATXN2 protein leading to pathogenic gain of toxic functions. We have generated two SCA2 mouse models with well-characterized behavioral, electrophysiologic and trancriptomic changes. **Objective:** To develop an ASO lowering *ATXN2* expression and modifying motor and electrophysiological phenotypes in SCA2 mouse models with minimal off target effects. **Methods:** We performed in vitro screens of 152 ASOs designed to lower ATXN2 expression. We then performed single intracerebroventricular (ICV) ASO (175-210 mcg) injections into one lateral ventricle in SCA2 mice identifying the ASO best modifying motor and electrophysiological phenotypes. Cerebellar molecular phenotypes were determined by qPCR, Western blotting, and RNA-seq. The SCA2 mouse models used were ATXN2-Q127 with ATXN2 expressed in Purkinje cells (PCs), and BAC-ATXN2-Q72 (BAC-Q72) (Hansen et al., 2013, Dansithong et al., 2015). **Results:** ATXN2-Q127 and BAC-Q72 mice treated with the lead ASO7 from age 8 wks to age 21 wks or 18 wks, respectively, had up to 75% reduced cerebellar *ATXN2* expression. *Aif1* expression, a marker for microglial activation, was not increased. In both models, we observed delayed onset of age-dependent rotarod phenotypes, vs. saline-injected control mice (ATXN2-Q127, n=15, P<0.01; BAC-Q72, n=11-14, P<0.01). ASO7 treatments also restored the PC firing frequency (FF) phenotypes for both models to FFs in age matched WT mice (ATXN2-Q127, Tx from age 8-21 wks, ASO treated 42±2 Hz, saline treated 16±1 Hz P<0.001) (BAC-Q72, Tx from age 30-40 wks, ASO treated 49±2 Hz, saline treated 36±1 Hz, P<0.001). Transcriptome analysis in ATXN2-Q127 and BAC-Q72 mice identified a shared age-dependent reduction of Pcp2 and Rgs8 expression. Treatment with ASO7 resulted in increased *Pcp2* and *Rgs8* mRNAs by qPCR and significantly increased protein levels by western blot analysis. **Conclusions:** A single treatment with ASO7 after disease onset lowered *ATXN2* expression and resulted in improvement of motor behavior. In addition, steady state levels of two key PC marker proteins were restored and firing frequency of Purkinje cells was normalized. These preclinical studies provide support for development of ASO-based therapies in human dominant ataxias.

188

Treatment of Niemann-Pick disease, type C1 subjects with intrathecal VTS-270 (2-hydroxypropyl-β-cyclodextrin). *F.D. Porter[1], N.Y. Farhat[1], E.A. Ottinger[2], J.C. McKew[2], L. Weissfeld[3], B. Machielse[4], E.M. Berry-Kravis[5], C.H. Vite[6], S.U. Walkley[7], D.S. Ory[8], TRND Team.* 1) NICHD, National Institutes of Health, DHHS, Bethesda, MD; 2) NCATS, National Institutes of Health, DHHS, Bethesda, MD; 3) Statistics Collaborative, Washington, DC; 4) Vtesse Inc., Gaithersburg, MD; 5) Departments of Pediatrics, Neurological Sciences and Biochemistry, Rush University Medical Center, Chicago, IL; 6) Department of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA; 7) Rose F. Kennedy Intellectual and Developmental Disabilities Research Center, Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY; 8) Diabetic Cardiovascular Disease Center, Washington University, St. Louis, MO.

Niemann-Pick disease, type C1 (NPC1) is a recessive lysosomal storage disorder with storage of unesterified cholesterol and progressive neurodegeneration. Preclinical testing demonstrated significant potential of VTS-270, a specific formulation of 2-hydroxypropyl-β-cyclodextrin, to significantly delay cerebellar Purkinje cell loss, slow progression of neurological signs, and increase lifespan in both murine and feline models of NPC1. Safety and clinical efficacy of intrathecal VTS-270 were evaluated in an open-label, dose escalation Phase 1/2a study. Intrathecal doses ranging from 50-1200 mg were evaluated monthly in 12 subjects treated for 18 months and in 2 subjects treated for 12 months. Three additional subjects were treated every two weeks. 24-hydroxycholesterol was used as a biomarker of target engagement and CSF protein biomarkers were evaluated. NIH Neurological Severity Scores (NSS) were used to evaluate clinical progression. A control group (n=15) consisted of a cohort of subjects being followed in a longitudinal Natural History study of similar age and disease severity. No drug-related serious adverse events were observed. Mid to high frequency hearing loss, an expected adverse event, was observed. Transient fatigue and ataxia was a significant, but variable, issue at doses ≥ 900 mg. Although both were clinically apparent, quality of life was not significantly impacted. Biomarker studies with 24-hydroxycholesterol confirmed target engagement. Levels of both calbindin D and FABP3, markers of neuronal damage, were significantly decreased in the majority of subjects after treatment. The total NSS for the 14 subjects treated monthly increased at a rate of 1.22 points/year (se= 0.34) compared to 2.89 points/year (se=0.36, p=0.001) for the control group. Inclusion of the 3 patients treated every two weeks decreased the annual progression rate to 1.02 points/year (se=0.33, p<0.001). Analysis of individual components of the total score showed significant decrease disease progression for ambulation, cognition and speech. Responder analysis showed disease progression in 100% of the natural history cohort versus only 50% in the VTS-270 cohort. This Phase 1/2a trial of intrathecal VTS-270 for the treatment of NPC1 demonstrates an acceptable safety profile and potential clinical efficacy. These data were used to support FDA Breakthrough designation and initiation of a controlled multicenter, multinational Phase 2b/3 clinical efficacy trial.

189

Identification of novel susceptibility loci and genes for breast cancer risk: A large transcriptome-wide association study in 119,000 cases and 101,000 controls of European descent. *L. Wu[1], J. Long[1], X. Guo[1], P. Kraft[2], R. Milne[3,4], K. Michailidou[5], J. Beesley[6], A. Dunning[7], P. Pharoah[7], J. Simard[8], G. Chenevix-Trench[6], D. Easton[5,7], W. Zheng[1] on behalf of the Breast Cancer Association Consortium.* 1) Vanderbilt University Medical Center, Nashville TN, USA; 2) Harvard School of Public Health, Boston, MA, USA; 3) Cancer Council Victoria, Melbourne, Australia; 4) School of Population and Global health, The University of Melbourne, Melbourne, Australia; 5) Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 6) QIMR Berghofer Medical Research Institute, Brisbane, Australia; 7) Department of Oncology, University of Cambridge, Cambridge, UK; 8) Laval University, Québec City, Canada.

Background Common genetic risk variants in approximately 180 loci have been found to be associated with breast cancer risk in genome-wide association studies (GWAS). These variants, however, explain only a small fraction of breast cancer heritability, and genes responsible for the association observed in these loci remain largely unknown. To identify novel loci and genes for breast cancer risk, we performed a transcriptome-wide association study evaluating associations of genetically predicted gene expressions with breast cancer risk. **Methods** We used genotyping and gene expression data obtained in normal breast tissues from female subjects of European descent included in the Genotype-Tissue Expression Project and established models using the elastic net method to predict gene expression using genetic variants. We evaluated model performance using data from The Cancer Genome Atlas and selected 6,343 genes for association analyses with breast cancer risk, using data obtained from approximately 119,000 cases and 101,000 controls of European ancestry included in the Breast Cancer Association Consortium. **Results** We identified 42 genes showing a significant association with breast cancer risk at $P < 7.88\times10^{-6}$, a Bonferroni-corrected significance level for multiple comparisons, including 17 genes in regions not yet reported for breast cancer risk, 16 previously unreported genes in known risk loci with expression either showing an association with breast cancer risk independent of the known risk variant (n=7) or correlated with the known risk variant SNP (n=9), and nine previously reported genes. Using less stringent *P*-value of 1.05×10^{-3} (adjusting for false discovery rate), we provided evidence for 21 additional breast cancer susceptibility gene candidates in 18 known loci, including 15 genes showing an association independent of the known risk variant(s). **Conclusion** In a transcriptome-wide association study, we identified multiple novel susceptibility loci and genes for breast cancer risk. Our study provided substantial new information towards the understanding of breast cancer genetics and biology.

190

A GWAS including 30,882 estrogen receptor negative or *BRCA1* mutation-related breast cancer cases and 110,088 controls identifies 10 new susceptibility variants. *R.L. Milne[1,10], K.B. Kuchenbaecker[2,11], K. Michailidou[2,12], J. Beesley[6], S. Kar[2], S. Lindström[5,13], S. Hui[3], A. Lemaçon[9], P. Soucy[9], A. Droit[9], G.D. Bader[3], P.D.P. Pharoah[2], F.J. Couch[4], D.F. Easton[2], P. Kraft[5], G. Chenevix-Trench[6], M. Garcia-Closas[7], M.K. Schmidt[8], A.C. Antoniou[2], J. Simard[9], Breast Cancer Association Consortium & Consortium of Investigators of Modifiers of BRCA1/2.* 1) Cancer Council Victoria, Melbourne, Victoria, Australia; 2) University of Cambridge, Cambridge, UK; 3) University of Toronto, Toronto, Canada; 4) Mayo Clinic, Rochester, MN, USA; 5) Harvard T.H. Chan School of Public Health, Boston, MA, USA; 6) QIMR Berghofer Medical Research Institute, Brisbane, Australia; 7) National Cancer Institute, Rockville, MD, USA; 8) The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands; 9) Laval University, Québec City, Canada; 10) University of Melbourne, Melbourne, Victoria, Australia; 11) Wellcome Trust Sanger Institute, Cambridge, UK; 12) Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus; 13) University of Washington School of Public Health, Seattle, WA, USA.

Breast cancer GWAS have identified more than 100 susceptibility SNPs. Most of these studies have included predominantly estrogen receptor (ER)-positive cases. GWAS focused on ER-negative disease, or *BRCA1* mutation carriers, who are more likely to develop ER-negative disease (70-80% of cases), have identified 11 SNPs. We aimed to discover additional ER-negative breast cancer susceptibility SNPs by performing a GWAS of European women using data from the Breast Cancer Association Consortium (BCAC) and the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). New genotyping data were generated for 9,655 ER-negative cases and 45,494 controls from BCAC and 15,566 *BRCA1* mutation carriers (7,784 with breast cancer) from CIMBA using the Illumina OncoArray beadchip, a 570K SNP custom array with genome-wide coverage. Imputation to the 1000 Genomes Project (Phase 3) generated data for ~11.5M SNPs with minor allele frequency (MAF)>0.005 and imputation r^2>0.3. For BCAC data, we applied logistic regression, adjusting for country and principal components. For CIMBA data, we used a retrospective cohort analysis framework, stratifying on country and birth cohort. These analyses were also applied to an independent set of previously generated data from other genome-wide genotyping of an additional 11,813 ER-negative cases and 55,100 controls from BCAC and 3,342 *BRCA1* mutation carriers (1,630 with breast cancer) from CIMBA. Fixed-effects meta-analysis was used to combine results across genotyping initiatives and consortia. We identified independent associations at $P<5\times10^{-8}$ with 10 variants at nine novel ER-negative breast cancer susceptibility loci. At P<0.05, we replicated associations with 10 variants previously reported in ER-negative or *BRCA1* mutation carrier GWAS, and confirmed ER-negative disease associations for 105 susceptibility variants identified by other breast cancer GWASs. These 125 variants explain approximately 16% of the familial risk of this breast cancer subtype. There was high genetic correlation (0.78) between risk of ER-negative breast cancer and breast cancer risk for *BRCA1* carriers. These findings will inform improved risk prediction, both for the general population and for *BRCA1* mutation carriers. Fine-mapping and functional studies should lead to a better understanding of the biological basis of ER-negative breast cancer, and perhaps inform the design of more effective preventive interventions, early detection and treatments for this disease.

191

Prostate cancer meta-analysis from more than 143,000 men identifies 57 novel prostate cancer susceptibility loci.

F. Schumacher[1,2], A. Amin Al Olama[3], S. I. Berndt[4], S. Benlloch[3], M. Ahmed[5], X. Sheng[6], D. F. Easton[3], F. Wiklund[7], P. Kraft[8,9], S. J. Chanock[4], B. E. Henderson[6], D. V. Conti[6], Z. Kote-Jarai[5], C. A. Haiman[6], R. A. Eeles[5,10] On behalf of ELLIPSE, PRACTICAL, CaPS, BPC3, PEGASUS. 1) Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH; 2) Seidman Cancer Center, University Hospitals, Cleveland, OH; 3) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 4) Division of Cancer Epidemiology and Genetics, National Cancer Institute, US National Institute of Health, Bethesda, MD; 5) Institute of Cancer Research, London, UK; 6) Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA; 7) Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden; 8) Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard School of Public Health, Boston, MA; 9) Department of Biostatistics, Harvard School of Public Health, Boston, MA; 10) Royal Marsden National Health Service (NHS) Foundation Trust, London and Sutton, UK.

Currently, genome-wide association studies (GWAS) and fine-mapping efforts have identified over 100 prostate cancer (PrCa) susceptibility loci, capturing 35% of the PrCa familial relative risk (FRR) within European-ancestral populations. In order to discover additional PrCa susceptibility loci we designed a custom high-density genotyping array, the OncoArray, based on the largest PrCa meta-analysis to date. The OncoArray platform consists of 310K SNPS selected from several cancer GWAS meta-analyses and fine-mapping studies as well as a 260K GWAS backbone for additional discovery (http://epi.grants.cancer.gov/oncoarray/).   The OncoArray genotypes (~47,000 PrCa cases and ~28,000 controls) were imputed to the October 2014 release of the 1000 genomes project in conjunction with several previous PrCa GWAS of European ancestry: UK stage 1 (1,906 cases/1,934 controls) and stage 2 (3,888 cases/3,956 controls); CaPS 1 (498 cases/502 controls) and CaPS 2 (1,483 cases/519 controls); BPC3 (2,137 cases/3,101 controls); NCI PEGASUS (4,622 cases/2,954 controls); and iCOGS (21,209 cases/ 20,440 controls). Risk analyses for overall PrCa risk, aggressive PrCa (defined by PrCa clinical characteristics), early age of onset and Gleason score were performed limited to individuals of European ancestry. Logistic and linear regression summary statistics were meta-analyzed using an inverse variance fixed effect approach.   The PrCa meta-analysis identified novel loci significantly associated (*$P<5.0x10^{-8}$)* with overall PrCa (N=57), advanced PrCa (Gleason ≥8, death from PrCa, PSA>100, or disease stage “Distant”; N=3), and early-onset PrCa (≤55 years of age; N=3). Although overall and stratified PrCa analyses identified unique SNPs, several regions overlap and will require further investigation. When combined multiplicatively, the 57 novel PrCa loci captures 5.6% of the FRR among the OncoArray samples. Overall, nearly 41% of the FRR is explained by a combination of novel and previously identified PrCa loci. In comparison to the median group, men in the top 10%-ile of the genetic risk score (GRS) group have a relative risk of 2.91 for developing PrCa. Furthermore, the PrCa risk is nearly 5x greater than the median group for the top 1%-ile of the GRS. While further functional annotation is underway, these novel loci will provide insight into the underlying biology of PrCa susceptibility.

192

Meta-analysis of genome-wide association data for 51,978 women identifies four new susceptibility loci for endometrial cancer.

T.A. O'Mara[1], D.D. Buchanan[2,3], T. Dörk[4], P.A. Fasching[5,6], E.L. Goode[7], P. Hall[8], D. Lambrechts[9,10], R.J. Scott[11,12,13,14], E. Tham[15,16], J. Trovik[17,18], D.F. Easton[19,20], I. Tomlinson[21], A.B. Spurdle[1], D.J. Thompson[19], ECAC, BCAC. 1) Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia; 2) Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Melbourne, Vic, 3010, Australia; 3) Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Vic, 3010, Australia; 4) Hannover Medical School, Gynaecology Research Unit, Hannover, 30625, Germany; 5) University of California at Los Angeles, Department of Medicine, Division of Hematology/Oncology, David Geffen School of Medicine, Los Angeles, CA, 90095, USA; 6) Department of Gynecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, 91054, Germany; 7) Department of Health Sciences Research, Mayo Clinic, Rochester, MN, 55905, USA; 8) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, SE-171 77, Sweden; 9) Vesalius Research Center, VIB, Leuven, 3000, Belgium; 10) Laboratory for Translational Genetics, Department of Oncology, University Hospitals Leuven, Leuven, 3000, Belgium; 11) Hunter Medical Research Institute, John Hunter Hospital, Newcastle, NSW, 2305, Australia; 12) Hunter Area Pathology Service, John Hunter Hospital, Newcastle, NSW, 2305, Australia; 13) Centre for Information Based Medicine, University of Newcastle, NSW, 2308, Australia; 14) School of Biomedical Sciences and Pharmacy, University of Newcastle, Newcastle, NSW, 2308, Australia; 15) Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, SE-171 77, Sweden; 16) Clinical Genetics, Karolinska University Hospital Solna, Stockholm, SE-17176 77, Sweden; 17) Centre for Cancerbiomarkers, Department of Clinical Science, The University of Bergen, 5020, Norway; 18) Department of Obstetrics and Gynecology, Haukeland University Hospital, Bergen, 5021, Norway; 19) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK; 20) Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, CB1 8RN, UK; 21) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK.

Endometrial cancer is the most common cancer of the female reproductive system in developed countries. To investigate genetics of endometrial cancer, we have established the Endometrial Cancer Association Consortium (ECAC), comprising fourteen study groups from Europe, the USA and Australia. Genome-wide association studies (GWAS) by ECAC and others have previously identified eight loci associated with endometrial cancer susceptibility. We have since conducted the largest GWAS meta-analysis for this disease. A total of 9,077 cases and 42,901 controls were genotyped using the Illumina OncoArray platform (570K custom array, nine studies), the Illumina iCOGS platform (200K custom array, five studies), and other commercial GWAS platforms (three studies). We imputed up to 14 million genetic variants from baseline genotypes, using the October 2014 release of the 1000 Genomes Project as a reference. Association testing by logistic regression was performed for each study and combined by inverse variance fixed-effects meta-analysis. Four novel loci were found to be associated with endometrial cancer risk at $P < 5 \times 10^{-8}$. Genes within close proximity of the newly identified regions include *SKAP1*, *NF1* and *SSPN* and the *HLA* gene complex. Additionally, the *SH2B3* locus, originally identified by meta-analysis of endometrial cancer and colorectal cancer GWAS, was confirmed at $P < 10^{-8}$ as an endometrial cancer risk locus. Assuming a log-additive association with risk, SNPs at the twelve identified genome-wide significant risk loci, combined with four additional variants reaching study-wide significance ($P<10^{-5}$), explain 6.7% of the familial risk of endometrial cancer. These results provide insight into the biology of endometrial carcinogenesis and enhance information required for future risk stratification models.

193

Assessing pleiotropy among common cancers in the UK Biobank. *J.D. Hoffman, R.E. Graff, M.N. Passarelli, J.S. Witte.* Department of Epidemiology & Biostatistics, University of California, San Francisco, San Francisco, CA.

Genome-wide association studies (GWAS) have identified a number of genetic variants that are associated with cancer risk. Among these variants, several have been observed to be associated with more than one type of cancer. It remains unclear, however, whether such genetic pleiotropy is more evident for certain subgroups of cancers, among individuals with particular traits, or among individuals diagnosed with multiple cancers. We set out to examine pleiotropy in 14 of the most common cancers diagnosed in the UK Biobank cohort. SNPs associated with any cancer at a significance threshold of $p \leq 5e\text{-}5$ were aggregated from the NHGRI-EBI GWAS catalog. We excluded SNPs identified from studies of survival, treatment response, and/or continuous phenotypes. SNPs with ambiguous strand assignment or no reported effect allele were excluded. From the UK Biobank genotyped cohort currently available, subjects were divided into age and sex matched unrelated case-control cancer sets using a 1:4 case to control ratio. Pleiotropy was assessed using a polygenic risk score (PRS) approach as well as by evaluating independent SNP associations. We generated cancer specific PRSs by weighting the effect allele by the log-odds ratio as reported in the literature for each individual cancer and then summing into a single score. Each cancer specific PRS was applied across all cancer phenotypes in logistic regression models adjusting for the first 10 principal components. Logistic regression was similarly performed in single variant analyses across the 14 cancers. We observed 11 nominally significant PRS cross-cancer associations. The most strongly significant association was between the bladder cancer PRS and kidney cancer risk ($p = 0.002$). The colorectal PRS was nearly as significant when applied to prostate cancer ($p = 0.003$). In single variant analyses, rs17023900 (3p12.1), rs4444235 (14q22.2), rs877529 (22q13.1), and rs9273363 (6p21.32) each had four cancer associations. An additional 25 independent loci were observed to have three cancer associations. Our results highlight the shared genetic basis of many of the most common cancers. Further work exploring the intersection of these variants' pathogenic role in cancer risk is in progress.

194

Risks of breast, ovarian and contralateral breast cancer for *BRCA1* and *BRCA2* mutation carriers from a prospective cohort of 10,015 mutation carriers. *A.C. Antoniou[1], K. Kuchenbaecker[1,2], D. Barnes[1], N. Andrieu[3], C. Nogues[4,5], M.J. Blom[6], C. Engel[7], D. Goldgar[8], K. Kast[9], F. van Leeuwen[6], R.L. Milne[10], T. Mooij[6], K.A. Phillips[12], M.B. Terry[11], J.L. Hopper[13], M. Rookus[6], D.F. Easton[1,14], International BRCA1/2 Carrier Cohort Study, Breast Cancer Family Registry, kConFab.* 1) Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; 2) The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; 3) Institut Curie, PSL Research University, INSERM, U900, Paris, France; 4) Hôpital René Huguenin-Institut Curie, Oncogénétique Clinique, Saint-Cloud, France; 5) Institut Paoli Calmettes, Pole Clinique Consultation d'Oncologie Génétique, Marseille, France; 6) Netherlands Cancer Institute, Amsterdam, The Netherlands; 7) Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany; 8) Department of Dermatology, Huntsman Cancer Institute, University of Utah School of Medicine, Salr Late City USA; 9) Department of Gynecology and Obstetrics, University Hospital Carl Gustav Carus, TU Dresden, Dresden; German Cancer Consortium (DKTK), Dresden and German Cancer Research Center (DKFZ), Heidelberg, Germany; 10) Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia; 11) Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, USA; 12) Peter MacCallum Cancer Centre, Melbourne, Australia; 13) Centre for Epidemiology and Biostatistics, University of Melbourne, Australia; 14) Department of Oncology, University of Cambridge, Cambridge, United Kingdom.

The clinical management of *BRCA1* and *BRCA2* mutation carriers requires precise estimates of age-specific cancer risks. However, a wide range of penetrance estimates has been reported with some mutation carriers having high cancer risk and others a much lower risk. Reasons for the large range of estimates include: 1) the small size of many studies; 2) the retrospective design of most studies, leading to ascertainment and testing bias; 3) inconsistent measurement or adjustment for potential risk modifiers. We used *prospective* data on *BRCA1*/*2* mutation carriers to estimate precise age-specific risks of breast (BC), ovarian (OC) and contralateral breast (CBC) cancer and to evaluate differences in risks by family history (FH). We used an international cohort of 10,015 of mutation carriers (6133 *BRCA1;* 3882 *BRCA2*) with a median follow-up of 5 years. We estimated annual incidences, standardised incidence ratios (SIR) and cumulative risks of BC, OC and CBC. We used Cox regression to assess whether risks differ by FH. The cumulative BC risk by age 80 was estimated to be 72% (95%CI: 65-79%) for *BRCA1* and 69% (95%CI: 61-77%) for *BRCA2* mutation carriers. The estimated SIRs decreased with increasing age in both *BRCA1* and *BRCA2* mutation carriers ($P<10^{-7}$). The age-specific BC incidences increased rapidly in early adulthood, until ages 30-40 for *BRCA1* and 40-50 for *BRCA2* mutation carriers, then remained at a similar, constant rate (20-30 per 1000 person-years) over the remaining years of life. The cumulative OC risk by age 80 was 44% (95%CI: 36-53%) for *BRCA1* and 17% for *BRCA2* carriers (95%CI: 11-25%). The estimated cumulative risk of CBC 20 years after the first BC diagnosis was 40% (95%CI: 35-45%) for *BRCA1* and lower, 26% (95%CI: 20-33%, P-diff=0.001), for *BRCA2* carriers. BC risk estimates for both *BRCA1* and *BRCA2* mutation carriers increased with the number of relatives diagnosed with BC (P-trend: 0.0001 for *BRCA1;* 0.02 for *BRCA2*). These estimates, obtained using the largest prospective cohort of mutation carriers, confirm the patterns of cancer risk indicated by retrospective analyses, and demonstrate the importance of family history in modifying cancer risk in carriers. Given the increase in *BRCA1/2* mutation screening through widely accessible gene panel testing, these risk estimates will be critical to the better management of *BRCA1* and *BRCA2* mutation carriers.

195

When is a coding variant association not a coding variant signal? *A. Mahajan on behalf of the ExT2D Exome Chip Consortium, for PROMIS, CHARGE, T2D-GENES/GoT2D, and DIAGRAM.* WTCHG, University of Oxford, Oxford, United Kingdom.

To evaluate the contribution of coding variation to type 2 diabetes (T2D) risk, we aggregated exome variant data from 73,033 T2D cases and 362,353 controls of diverse ancestries including (i) exome-array data from 51 studies and (ii) genome-wide association (GWA) data from UK BioBank and Genetic Epidemiology Research on Aging (GERA), supplemented with UK10K and/or 1000 Genomes imputation. A total of 69 coding variants mapping to 35 loci attained exome-wide significance (P<4.3x10^{-7}). Given the incomplete coverage of regional variants on exome-array, functional inference requires that we determine whether the coding variant is driving an association (and not simply tagging a nearby non-coding causal variant). To do so, we performed fine-mapping (FM) in a trans-ethnic meta-analysis of GWA data (41,284 T2D cases and 311,715 controls) from UK Biobank, GERA, and 18 additional 1000 Genomes-imputed studies from the DIAGRAM consortium. Genome-wide analyses demonstrate substantial enrichment of GWA signals within coding sequence: these priors can be incorporated into evaluations of the probability that a given variant is causal. Accordingly, we performed FM analyses using two different prior distributions based on variant annotation: (i) uniform, giving equal weight to all variants, irrespective of annotation; and (ii) weighted, based on Sveinbjornsson et al., giving elevated weight to non-synonymous variants. Within each locus, we used these alternate priors to construct credible sets of variants that collectively account for 99% posterior probability (PP) of including the causal variant. Across loci, the weighted prior typically improved FM resolution: smaller credible sets (20% contain <10 variants vs 7% with uniform priors) mapped to smaller genomic intervals (1.2-fold reduction). At four loci (including *GCKR* and *SLC30A8* where the coding variant is previously considered to drive the GWA signal) the informative prior analysis confirmed >90% chance that the coding variant(s) were causal. At five loci (e.g. *ARAP1*, *MRPS35*), FM clearly indicated that the coding variant we had identified was not driving the signal under either prior. At other loci, evidence was inconclusive yielding coding variant PPs between 20% (e.g. *THADA*, *PPARG*) and 90% (e.g. *POC5*, *PAM*). Our analysis indicates the importance of interpreting coding variant association discoveries in the context of the wider regional perspective afforded by GWA data to avoid false attributions of functional impact.

196

An integrative platform to uncover the mechanisms of the association of *TCF7L2* and Type 2 Diabetes. *M. Nobgrea[1], D.R Sobreira[1], S. Strobel[2,3], N. Sinnott-Armstrong[2], M. Claussnitzer[2,4,5].* 1) Human Genetics, University of Chicago, Chicago, IL; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Technical University Munich; 4) Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA; 5) Computational Biology, MIT, Cambridge, MA.

Noncoding variants within *TCF7L2* remain the strongest genetic association with polygenic Type-2 Diabetes (T2D). We present an integrated approach to mechanistically dissect this association, providing a paradigm framework for the functional follow-up of Genome-Wide Association Studies (GWAS). Using epigenomics and comparative genomics we computationally predicted that the tag SNP rs7903146 is the causal variant of the association (C/T, MAF=0.31 EUR, 1000 Genomes). This SNP is predicted by the Epigenome Roadmap data to reside within an active enhancer in pancreatic islets, enteroendocrine cells and adipose tissues, which we confirm using reporter assays in appropriate cell lines that associate the risk haplotype with increased enhancer activity. Using genome-wide ChIP-Seq and conditional gene expression studies we show that Gata3 binds to the risk allele of this SNP and that Gata3 is required for the adipocyte enhancer activity of the risk allele. Using primary human adipocytes from 26 homozygous risk and 19 non-risk allele carriers we show that the risk allele is associated with increased *TCF7L2* expression and a down-regulation of genes involved in adipogenesis. Utilizing genetically modified mice we showed that increased adipocyte Tcf7l2 expression results in impaired adipogenesis, with insulin resistance and hepatic steatosis ensuing upon high fat diet feeding. We show that Tcf7l2 regulates major players in adipogenesis in mouse adipocytes. Using genomic editing technology we demonstrate that rs7903146 is sufficient to regulate *TCF7L2* expression in primary human adipocytes, directly linking this SNP to *TCF7L2* expression levels. Finally, we showed that in humans, *TCF7L2* over-expression in adipocytes is seen only in obese individuals with the risk allele at rs7903146, uncovering a genetic by environment interaction. Our integrative approach posits that the mechanistic basis for the association of rs7903146 and T2D involves pleiotropic functions of an enhancer active in multiple tissues with important roles in glucose metabolism as well as a gene x environment interaction that predicts that risk allele carriers that become obese will have an additional phenotypic insult stemming from rs7903146.

197

Large-scale association study of predicted gene expression implicates novel T2D genes. *J.M. Torres[1], A. Barbeira[1], A. Morris[2], K. Shah[3], H. Wheeler[4], G.I. Bell[5], D. Nicolae[1], N.J. Cox[6], H. Im[1].* 1) Medicine/Genetic Medicine, The University of Chicago, Chicago, IL; 2) Institute of Translational Medicine, University of Liverpool, Liverpool, UK; 3) Tempus Inc., Chicago, Il; 4) Department of Biology, Loyola University Chicago, Chicago, IL; 5) Department of Medicine, University of Chicago, Chicago, IL; 6) Division of Genetic Medicine, Vanderbilt University, Nashville, TN.

Most genes implicated by GWAS are done so by proximity to associated markers that mostly reside in non-coding genomic regions. However, this assumption is undermined by long-range interactions between eQTLs and promoters of target genes that are not necessarily the reported trait genes closest to associated regions. In order to incorporate regulatory genetic information to address the challenge of gene-mapping, we apply MetaXcan - an integrative approach that inputs summary statistics from GWAS to test for association between estimates of the genetic component of gene expression and disease. Leveraging 42 predictive models corresponding to a diverse set of primary human tissues and cell types from the DGN and GTEx projects, we performed a series of gene-based tests with summary data from the DIAGRAM trans-ethnic meta-analysis of T2D, including over 100K individuals. We find evidence that not only supports reported T2D genes at associated GWAS loci, but identified a set of novel T2D gene candidates that comprise the majority of significant MetaXcan associations. Moreover, we found support for 8 genes that map to "unknown" T2D loci that have not been previously reported (e.g. *SH3D21*, *ZNRD1*, *FAH*). We also replicated associations in an independent cohort from the GERA study and found that of the 22 genes meeting genome-wide significance, 13 represent novel candidate disease genes. Lastly, novel genes discovered in our analysis overlap with putative genes for traits related to T2D. This study shows that more insight about the genetic architecture of T2D can be gleaned by incorporating regulatory genetic information in gene mapping studies and represents an important step forward in the post-GWAS era.

198

Multiple phenotypes applied to whole-exome sequence data to identify functional variants and implicate genes in metabolic disease. *M.S. Udler[1,2], A.K. Manning[1,2], A. Mahajan[3], J.C. Florez[1,2], J. Flannick[1,2] on behalf of T2D-GENES, LUCAMP, SIGMA, and ESP.* 1) Medicine, Massachusetts General Hospital, Boston, MA; 2) Broad Institute, Cambridge, MA; 3) Oxford University, Oxford, UK.

Exome sequence data can identify high-impact variants that inform on the pathophysiology of human diseases. However, few such variants have emerged to date from large-scale exome-wide studies of many traits, including type 2 diabetes (T2D). Here we present an approach to identify variants putatively causal for hypothesized disease subtypes, based on analysis of associations across multiple T2D-related phenotypes. We applied our approach to whole-exome sequences in 25,982 multi-ethnic T2D cases/controls, performing (a) aggregate analysis of nonsynonymous variants within candidate genes and (b) exome-wide searches for novel genes. We defined two phenotypic patterns: 1) a "lipodystrophy" pattern with *increased* serum total cholesterol (TC), low-density lipoprotein cholesterol (LDL) , and triglycerides (TG) but *decreased* high-density lipoprotein cholesterol (HDL) and BMI, and 2) a "liver efflux" pattern based on reported associations at the fatty liver-associated locus *TM6SF2* (decreased TC, HDL, LDL, and TG). Within *PPARG*, a gene implicated in T2D and lipodystrophy, 4 rare (MAF <0.1%) variants clustered together under the "lipodystrophy" pattern. All 4, but no others in our data, have previously been functionally validated as affecting adipocyte differentiation (hypergeometric P=5x10^{-6}). These variants together with 4 additional in Autosomal Dominant Partial Lipodystrophy genes (*PLIN1*, *LMNA*, and *AKT2*) all sharing the "lipodystrophy" pattern were enriched for T2D risk (Collapse OR=4.09, P=0.01). Exome-wide, 3 liver-expressed genes matched the "liver efflux" pattern: *TM6SF2, PNPLA3,* and *HNF1A*. *PNPLA3* is an established fatty liver gene, and its association with T2D was confirmed in a larger-scale study. *HNF1A* variants cause Maturity Onset Diabetes of the Young type 3 (MODY3) and T2D; while *HNF1A* is not known to cause fatty liver, case reports describe liver steatosis in patients with MODY3. Within these genes, variants harboring the "liver efflux" pattern were enriched for T2D risk (OR=1.10, P=8x10^{-4}), while variants clustering with the opposite pattern (increased TC, HDL, LDL, TG) were enriched for T2D protection (OR=0.75, P=0.01); by contrast, traditional filters of rare nonsynonymous variants exhibited no association with T2D (OR=1.02, P=0.64). Although functional work is needed to validate these associations, our results show that hypothesis-driven joint analysis of multiple phenotypes can identify novel variants of relevance to specific disease pathways.

199

Identification of a functional p.Thr280Met variant in *RBPJL* which associates with T2D in Pima Indians and potentially affects the established T2D locus *CTRB1/2*. *A. Nair, J. Sutherland, P. Kumar, P. Piaggi, Y. Muller, M. Traurig, S. Kobes, R. Hanson, C. Bogardus, L. Baier.* National Institute of Diabetes and Digestive and Kidney Diseases, NIH, Phoenix, AZ.

Recombination Signal Binding Protein for Immunoglobulin Kappa J-region like (RBPJL) forms a complex with PTF1A to regulate gene transcription during pancreatic development. In adult pancreas, the RBPJL/PTF1A complex regulates expression of pancreatic exocrine enzymes. For e.g., RBPJL regulates *CTRB1* which encodes the pancreatic enzyme chymotrypsinogen. The objective of the current study was to assess a missense variant detected in *RBPJL*. Recent analysis of whole genome sequence data in Pima Indians identified a Thr280Met (rs200998587) SNP in exon 8 of *RBPJL*. This SNP has a frequency of 3% in Pima Indians. In 1000G data, the Met allele has only been reported in populations with Amerindian heritage (not present in Europeans, Asians or Africans). The Thr allele is highly conserved and our analysis of a human cDNA tissue panel suggests that *RBPJL* expression is pancreas specific. Genotyping of this variant in 7227 Pima Indians identified a modest association with type 2 diabetes (T2D; OR=1.60[1.21-2.13] per Met allele, *P*=0.001 with genomic control, covariates- age, sex, birth year and principal components 1-5 derived from GWAS). To determine if the T2D-associated variant affects RBPJL function, both wild type and Met-containing RBPJL were over-expressed in HEK293 cells. Western blot showed lower protein levels of the Met-containing protein, while real-time PCR detected same level of RNA expression, suggesting that the Met-containing protein is less stable. Impairment of the Met-containing RBPJL was further seen in luciferase assays which utilized two different RBPJL responsive promoters cloned into the pGL3 basic vector. Luciferase activity was significantly lower (P<0.0001, effect=36% and 15%) with co-transections of the Met-expressing vs. wild type RBPJL in HEK293 cells. Since RBPJL is known to regulate expression of *CTRB1*/*2*, and the G allele at a SNP in *CTRB1*/*2* (rs7202877) has been associated with lower T2D risk, increased GLP-1 stimulated insulin secretion and higher expression of CTRB1/2 in other ethnic groups (in Pima Indians there is modest replication for T2D, OR= 0.79[0.64-0.98] per G allele, *P*=0.03), we performed luciferase assays in HEK293 cells co-transfected with *RBPJL* and *CTRB1* promoter in pGL3 basic vector. Our study confirmed that *CTRB1* promoter is responsive to RBPJL. In summary, we have identified a functional Thr280Met SNP in RBPJL that associates with T2D in Pima Indians and may regulate *CTRB1*/*2*, a known T2D locus in other ethnic groups.

200

Genetic variation modulates multiple dimensions of molecular phenotypes in T2D patients. *A. Viñuela[1,2,3], J. Fernandez[4], A. Kurbasic[5], H. Krogh Pedersen[6], M.G. Hong[7], A.A. Brown[1,2,3], M. Abdalla[4], C. Howald[1,2,3], C. Groves[4], A. Mahajan[4], P.K. Davidsen[6], R. Gupta[6], C. Brorsson[6], K. Banasik[6], C. Prehn[8], A. Artati[8], M. Haid[8], M. Roßbauer[10,11], H. Grallert[10,11], J. Adamski[8], J.M. Schwenk[7], E. Pearson[9], S. Brunak[4], P.W. Franks[5], M.I. McCarthy[4], E.T Dermitzakis[1,2,3] for the IMI DIRECT consortium.* 1) Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Geneva, Switzerland; 2) Swiss Institute of Bioinformatics, Geneva, Switzerland; 3) Institute of Genetics and Genomics in Geneva, University of Geneva, Geneva, Switzerland; 4) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, Oxford; 5) Department of Clinical Sciences, Genetic and Molecular Epidemiology, Malmö, Sweden; 6) Center for Biological Sequence Analysis Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, DENMARK; 7) SciLifeLab Stockholm and School of Biotechnology, KTH Karolinska Institutet Science Park, Solna, Sweden; 8) Helmholtz Zentrum Munich, Genome Analysis Center, Neuherberg, Germany; 9) Institute of Epidemiology II, Research Unit of Molecular Epidemiology, Helmholtz Zentrum München Research Center for Environmental Health, Neuherberg, Germany; 10) German Center for Diabetes Research (DZD), Neuherberg, Germany; 11) Molecular and Clinical Medicine, Ninewells Hospital and Medical School, University of Dundee, Dundee, Scotland, United Kingdom, DD1 9SY.

While GWAS studies have produced many associations between genetic variants and type 2 diabetes (T2D), these have proved of limited use for patient risk stratification and understanding disease aetiology. Instead, a study design that combines multiple dimensions of data can help explain the genetic basis of complex traits. Here we present the DIRECT project, where pre-diabetics and newly diagnosed subjects with T2D underwent extensive genotyping and phenotyping. Blood and plasma samples from ~3,100 individuals were used to produce baseline transcriptomic (RNA-seq), proteomic (multiplexed immunoassays) and metabolomic (both targeted and non-targeted) data to be combined in an integrative framework. Hundreds of clinical and lifestyle quantitative phenotypes were also measured at time of recruitment and in follow up visits at 18 months. We first mapped QTL in each molecular data-type. Using RNA-seq from 2,878 of the individuals we identified 17,714 eQTL (FDR 5%, π_1=0.89%), including a significant eQTL for all protein coding genes and lincRNAs (FDR 3.3%, 15,044 genes). Out of 263 proteins, we found 32 (12%) to be associated with a local genetic variant (pQTL, FDR 5%, π_1=0.07). Using targeted metabolite data in the same individuals we measured 154 metabolites, of which 136 (89%) were significantly associated to at least one SNP (metQTL, FDR 5%, π_1=0.83). When combining these results, we estimated that 23.8% of metQTL and 15.4% of the pQTL were also associated with the expression of a gene. One strong example of a genetic variant acting in more than one trait is rs174530, which affects expression of *FADS1* and the levels of circulating diacyl-phosphatidylcholines C36:1. This metabolite has previously been associated to decreased T2D risk. In contrast, we identified rs10445391 as a pQTL for CCL16 but not an eQTL, as the relevant gene is expressed at very low levels in whole blood. The GTEx data shows the gene is expressed and its protein secreted only in liver, where rs10445391 is a significant eQTL. Finally, we also looked at how relationships between genotype and molecular phenotypes can be perturbed by disease status. A scan for eQTL with different expression effects in subjects with T2D and pre-diabetics identified 250 significant interactions (FDR 1%). By understanding the complex ways genetic effects can manifest and how these can be perturbed by disease, we hope to better understand and stratify those at highest risk of developing T2D and other complications.

201

Rewiring of enhancer-gene interactions drives *PLAU* overexpression in the pathogenesis of Quebec Platelet Disorder. *M.D. Wilson[1,5,9], M. Liang[1,5,8], A. Soomro[7,8], J.S. Waye[3,7], A.D. Paterson[1,2], G.E. Rivard[4], C.P.M. Hayward[3,6,7,9].* 1) The Hospital for Sick Children, Toronto, ON, Canada M5G 0A4; 2) The Dalla Lana School of Public Health and Institute of Medical Sciences, University of Toronto, ON, Canada M5T 3M7; 3) Hamilton Regional Laboratory Medicine Program, Hamilton, ON, Canada L8N 3Z5; 4) Hematology/ Oncology, Centre Hospitalier Universitaire Sainte-Justine, Montreal, QC, Canada H3T 1C5; 5) Molecular Genetics, University of Toronto, Toronto, ON, Canada M5S IA8; 6) Department of Medicine, McMaster University, Hamilton, ON, Canada L8N 3Z5; 7) Department of Pathology and Molecular Medicine, McMaster University, Hamilton, ON, Canada L8N 3Z5; 8) These authors contributed equally; 9) Co-corresponding authors.

Quebec platelet disorder (QPD [OMIM 601709]) is a rare bleeding disorder caused by a unique gain-of-function defect in fibrinolysis. The hallmark of QPD is a marked increase in the expression of urokinase-type plasminogen activator (*PLAU*) in platelets and megakaryocytes, leading to accelerated clot degradation and delayed onset bleeding following hemostatic challenges. The genetic cause of QPD is a heterozygous tandem duplication of a 77 kb region of 10q22 that includes *PLAU* and its putative regulatory elements. However, the markedly increased (>100 fold) and cell-type specific expression of *PLAU* in QPD cannot be explained simply by a copy-number gain. To establish the molecular mechanism of *PLAU* overexpression in QPD, we performed transcriptomic (mRNA-seq) and epigenetic (ChIP-seq for histone modifications) profiling of primary blood cells (granulocytes and cultured megakaryocytes) from QPD patients and controls. Analysis of RNA-seq data showed that QPD *PLAU* transcripts were consistent with reference gene models, with a significantly greater fraction of reads originating from the disease chromosome in megakaryocytes than granulocytes. We identified a putative enhancer ~40 kb downstream of *PLAU* that is highly enriched for the histone modification H3K27ac in megakaryocytes compared to granulocytes. Chromosome confirmation capture experiments support that this enhancer normally interacts with the promoter of the downstream vinculin (*VCL*) gene. *VCL* is upregulated during megakaryocyte differentiation and the QPD duplication places one copy of *PLAU* downstream of this enhancer. We propose that the re-positioning of *VCL*-related regulatory elements contributes to the cell-type specific, dramatic increase in *PLAU* expression in QPD megakaryocytes.

202

Capture Hi-C identifies compelling candidate causal genes and enhancers for multiple sclerosis in the 6q23 region. *P. Martin[1], A. McGovern[1], K. Duffus[1], A. Yarwood[1], S. Schoenfelder[2], A. Barton[1,3], P. Fraser[2], J. Worthington[1,3], S. Eyre[1], G. Orozco[1].* 1) Arthritis Research UK Centre for Genetics and Genomics. Centre for Musculoskeletal Research. Institute of Inflammation and Repair. Faculty of Medical and Human Sciences. Manchester Academic Health Science Centre. The University of Manchester. Stopford Bui; 2) Nuclear Dynamics Programme, The Babraham Institute, Cambridge CB22 3AT, UK; 3) NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester Foundation Trust, Manchester Academic Health Science Centre, Oxford Road, Manchester M13 9WL.

Genome wide association studies (GWAS) have been tremendously successful in identifying variants associated with complex disease, including multiple sclerosis (MS). Many of these variants lie in enhancer regions, although it is often unclear which gene they affect. Capture Hi-C (CHi-C) can be used to study long-range interactions at high resolution between enhancers and target genes. The 6q23 region is a pan-autoimmune locus associated with multiple autoimmune diseases. Our aim was to identify MS causal genes in the 6q23 region by studying chromatin interactions involving MS associated variants and further refining the causal variants using bioinformatics.Chromatin interactions with MS associated regions in the 6q23 locus, defined as SNPs in $r^2 \geq 0.8$ with the lead association, were analysed as part of a larger CHi-C study in T- and B-cell lines that included all known risk loci for rheumatoid arthritis (RA), juvenile idiopathic arthritis (JIA), psoriatic arthritis (PsA) and type 1 diabetes (T1D). Complex long-range interactions were observed between MS associated regions and eight gene promoters, as well as between each region. These interactions cluster in two regions, the first implicating MS candidate genes including *AHI1*, *SGK1* and *BCLAF1* and the second implicating autoimmune related genes such as *IL20RA*, *IL22RA2*, *IFNGR1* and *TNFAIP3*. Bioinformatics analysis of SNPs involved in these interactions further refined the SNPs to enhancer regions and identified many of the target genes as actively transcribed regions.This investigation has strengthened the case for the *AHI1* gene candidate, identified other potential MS gene targets, such as *SGK1*, *BCLAF1*, *IL20RA*, *IL22RA2*, *IFNGR1* and *TNFAIP3* and shown a possible co-regulation of MS GWAS associations in the 6q23 region, which could help elucidate the pathogenesis of MS as well as other autoimmune diseases. Whilst these targets will require further functional investigation, which has been informed by the bioinformatics analysis, this work has the potential to provide novel therapeutic targets or drug repositioning to improve patient outcome.

203

Fine-mapping of obesogenic *cis*-regulatory eQTL variants using high resolution capture Hi-C. *D.Z. Pan[1,2], K. Garske[1], M. Alvarez[1], C.K. Raulerson[3], K.L. Mohlke[3], M. Laakso[4], P. Pajukanta[1,2,5].* 1) Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, USA; 2) Bioinformatics Interdepartmental Program, UCLA, Los Angeles, USA; 3) Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA; 4) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland; 5) Molecular Biology Institute at UCLA, Los Angeles, USA.

Obese individuals' adipose tissue is known to have a different composition from that of lean individuals. In obese adipose tissue, adipocytes continue to enlarge and eventually burst, causing a cellular inflammatory response, which increases tissue heterogeneity. To identify regulatory variants in a tissue-specific manner across the genome, we performed a *cis* expression quantitative trait locus (eQTL) study using RNA-sequence data on 795 subcutaneous adipose tissue biopsies from the Finnish METabolic Syndrome In Men (METSIM) cohort. However, linkage disequilibrium and the sheer number of eQTLs make it difficult to pinpoint specific variants for functional study. We hypothesize that many regulatory *cis*-eQTL variants fall inside distal enhancers, looping to physically interact with the promoter of the eQTL and enhancer target gene. We sought to fine-map these eQTL variants by searching for the subset of blood-derived *cis*-eQTL variants located in regulatory elements based on published capture Hi-C (CHi-C) data from $CD34^+$ hematopoietic progenitor cells. We used the HiCUP and GOTHiC Hi-C pipelines to identify genomic elements significantly interacting with promoters at *Hin*dIII fragment-level resolution with an average fragment size of 4 kb. We compared the location of these genomic elements with METSIM adipose *cis*-eQTLs to highlight regulatory variants at enhancer sites that may be functionally associated with blood-derived cells in adipose tissue. Our preliminary results suggest that a non-trivial proportion of gene promoters (4,101 of 15,429 (26.6%)) interact with at least one putative enhancer that contains an adipose *cis*-eQTL variant, identifying thus a set of genes for which the *cis*-eQTL and enhancer target gene are the same (permutation test, *p*-value=0.0125). Furthermore, analysis of the same set of genes for CTCF Binding Sites (CBSs) using public ChIP-seq data on CD34+ cells (GEO ID GSM651541) showed 4,099 out of 4,101 genes had CBSs in both the enhancer and promoter of looping pairs. Finally, this subset of genes showed a significant enrichment in a pathway related to immune response and lysosomal activity (adjusted p<0.0007) known to be increased in macrophages in obese individuals, suggesting the blood-derived origin of these eQTLs. These results can help uncover functional variants in critical regulatory blood-derived enhancer-promoter looping interactions relevant for transcriptional regulation in heterogeneous human obesogenic adipose tissue.

204

Regulatory activity of non-coding variants in multiple tissue and cell types at the *VEGFA* metabolic trait GWAS locus. *J.P. Davis[1], S. Vadlamudi[1], C. Trevino[2], T.S. Roman[1], Y. Wu[1], M. Engle[1], J. Kuusisto[3], M. Civelek[4,5], A.J. Lusis[4], M. Laakso[3], K.L. Mohlke[1].* 1) Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA; 2) Department of Genetics and Molecular Biology, Emory University, Atlanta, Georgia, USA; 3) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, 70210 Kuopio, Finland; 4) Departments of Medicine, Cardiology, Human Genetics, Microbiology, Immunology & Molecular Genetics, University of California, Los Angeles, Los Angeles, California, USA; 5) Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, USA.

Understanding the molecular and biological role of non-coding variant signals discovered in genome-wide association studies (GWAS) remains a major challenge. We investigated a signal located in an intergenic region 3.7 kb from the 3' end of *VEGFA* previously shown to be associated (P<5x10^{-8}) with triglycerides (TG), HDL cholesterol (HDL), body mass index, adiponectin, and coronary heart disease. At this signal, there are five candidate variants based on linkage disequilibrium r^2>0.7 with the GWAS lead variant rs998584. To determine if this signal may affect expression of nearby genes, we evaluated expression quantitative trait locus (eQTL) results in adipose tissue from 770 participants in the METabolic Syndrome In Men (METSIM) study. The GWAS risk allele rs998584-A was associated (P=2.6x10^{-9}) with decreased expression of *VEGFA* but no other gene within 1 Mb (P>.05). Reciprocal conditional analyses between GWAS and eQTL variants confirmed the signals are coincident. Decreased *VEGFA* expression level was significantly (P<5x10^{-8}) associated with increased TG and insulin, and decreased adiponectin, HDL, and insulin sensitivity. To determine variant overlap with epigenomic marks of transcription regulation, we analyzed available datasets of open chromatin and transcription factor and histone modification ChIP-seq. All five of the variants overlap regulatory marks in blood, adipose, and/or liver. rs998584 is located in a region of open chromatin (DNase-seq) for 39 cell types, while the other four variants overlap such regions in ≤2 cell types. We tested DNA elements containing 1-2 variants for enhancer activity using reporter assays in SGBS adipocyte, 3T3-L1 mouse adipocyte, THP-1 monocyte, HUVEC endothelial, and/or HepG2 liver cell lines. In THP-1 cells, elements containing rs998584, rs11967262, and rs4711750-rs998584 all had activity >4-fold higher than an empty vector (EV) control. In one orientation, risk allele rs998584-A showed significantly (P<0.05) lower expression than the non-risk C allele (27- vs. 36-fold more than EV), a direction that is consistent with the eQTL. In electrophoretic mobility shift assays using THP-1 nuclear lysate, the risk allele rs998584-A showed stronger specific protein binding than the C allele. Together, these results suggest that rs998584 decreases the transcription of *VEGFA* in blood and adipose cells by altered binding of a transcriptional repressor that may lead to increased TG, insulin, and decreased insulin sensitivity.

205

Epigenetic fine-mapping of cardiovascular disease loci in the liver. *C. Brown[1], M. Caliskan[1], Y. Park[1], M. Trizzino[1], M. Beltrame[1], C. Radens[1], K. Wiles[1], S. Elwin[1], K. Olthoff[2], A. Shaked[2], D. Rader[1], B. Engelhardt[3].* 1) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 2) Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 3) Department of Computer Science, Princeton University, Princeton, NJ.

Genome-wide association studies (GWAS) have identified more than 200 loci that contribute to cardiovascular disease risk (CVD). As with other complex phenotypes, the majority of the heritability of CVD risk lies within the non-coding regions of the genome. This has led to the hypothesis that the causal variants at GWAS associated loci lead to changes in local gene expression. As a result of linkage disequilibrium and the fact that cis-regulatory elements (CREs) may target genes over large distances, it is often unclear which variant or gene affects disease risk, however their identification will improve understanding of disease etiology and identify targets for novel therapeutic development. Recent work has demonstrated that histone modification state data can be used to identify CREs harboring disease-associated variants that are more likely to be causal than linked variants that do not overlap functional elements. Existing studies have focused on easily ascertained cell types, while the liver, which plays a critical role in regulating cholesterol and lipid metabolism, and where many CVD associated variants likely affect gene expression, has remained understudied. To identify the specific variants and genes that affect CVD risk, we have deeply phenotyped 283 liver biopsies, collecting RNA-seq along with histone modification and transcription factor ChIP-seq data. We have used these data to identify thousands of genetic variants associated with allele-specific transcription factor binding, histone modification, gene expression, and splicing. Comparison to data from the GTEx and Roadmap Epigenomics projects demonstrate that many of these associations are specific to the liver. We demonstrate that multi-phenotype molecular trait mapping improves statistical power to detect associations and results in improved resolution at identified loci. We have integrated these data with CVD GWAS data using a novel multi-phenotype causal inference framework based on Mendelian randomization to predict the precise variants, CREs, and genes that underlie CVD risk. We have validated the allele-specific regulatory activity of many of these predicted causal variants with a novel parallelized reporter assay. These analyses identify likely disease genes at dozens of previously uncharacterized GWAS loci. We also demonstrate that, at many GWAS loci, candidate genes have been falsely implicated based on proximity to the lead SNP.

206

Genetic determinants of chromatin accessibility predict variation in T cell activation and autoimmunity across human individuals. *R. Gate[1,3], C. Cheng[4], D. Lituiev[3], M. Subramaniam[1,3], A. Siba[4], E.L. Aiden[7,8,9,10,4], I. Machol[7,8,9], M. Shamim[7,8,9], M. Beer[11,12], M. Tabaka[4], K. Hougen[12], C.J. Ye[2,3], A. Regev[4,5,6].* 1) Biological Medical Informatics Graduate Program, University of California, San Francisco, San Francisco, CA. 94158, USA; 2) Department of Epidemiology and Biostatistics, Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94143, USA; 3) Institute of Human Genetics, University of California, San Francisco, CA, 94143, USA; 4) Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard, Cambridge, MA 02142, USA; 5) Department of Biology, MIT, Cambridge, MA 02139, USA; 6) Howard Hughes Medical Institute; 7) The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA; 8) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; 9) Department of Computer Science, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA; 10) Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA; 11) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, USA; 12) Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA.

Over 90% of disease-associated loci identified by genome-wide association studies are located in non-coding regions of the genome likely to play a role in gene regulation. Previous efforts to annotate non-coding variants have focused on overlapping disease associations with genetic variants that modulate gene expression. However, because gene regulation is a complex process involving multiple *cis* and *trans* regulatory elements, pinpointing causal variants for expression and disease remain challenging. Genetic analysis of variability in chromatin states provides a powerful approach to identify genetic variants that directly affect *cis* regulation. We used ATAC-seq to profile the chromatin states of activated $CD4^+$ T cells in 105 healthy individuals of European descent. We found 3-fold more regions of open chromatin in activated versus unstimulated cells. Further, activation specific regions are enriched for known enhancers important for T cell differentiation and activation. At single base pair resolution, transcription factor (TF) footprinting of BATF, ISRE, and BATF-IRF4 shows increased accessibility when activated. Mapping genetic variants associated with variability in chromatin accessibility, we found 1,842 ATAC-QTLs, of which 18% are located within the associated peak, 25% intersect GWAS SNPs, and 62% are contained within either a BATF, ETS1, or CTCF motif. Interestingly, TF footprints between genotypes of associated peaks containing BATF and ETS1 motifs show striking differences in accessibility. Additionally, we called 1,510 ATAC-seq multi-peaks, 695 of which are ATAC-QTLs. Leveraging in situ HiC data generated on pooled cells, we found multi-peak ATAC-QTLs are more likely to be super enhancers and reside in the same contact domain. Finally, we compared ATAC-QTLs with expression QTLs mapped by analyzing RNA-seq data collected from 96 matched individuals. We found 816 eQTLs, 104 of which are also ATAC-QTLs with effect sizes correlated at R=0.49. Thus, variability in chromatin accessibility as measured by ATAC-seq in primary immune cells is determined by genetic variation in a manner affected by the 3D organization of the genome and contributes to gene expression variation. Our results provide insights into how genetic variants modulate chromatin state and gene expression in affecting human disease.

207

Inference of dominance and selection coefficients from large-scale population data identifies recessive genes. *D.J. Balick[1], D.M. Jordan[2], S. Sunyaev[1], R. Do[2].* 1) Department of Medicine, Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 2) Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

Despite revived interest in the human genetics community, the quantification of diploid selection coefficients in specific human variants and genes remains largely elusive. Unlike model organisms, dominance (**h**) and selection (**s**) coefficients in humans must be inferred from natural population data. We present a method to estimate coarse average selection and dominance coefficients per gene by comparing Exome Aggregation Consortium population genetic data in ~35,000 Europeans to simulations of a range of selection and dominance coefficients in a realistic demography. We match putatively deleterious variants (nonsense and damaging missense) via informative summary statistics of the per-gene frequency spectrum. We classify genes as strong selection recessives ($h<0.1$), strongly selected “non-recessives” ($h>=0.1$), under weak selection, nearly neutral, or sub-drift. To validate our recessive and non-recessive gene sets, we demonstrate significant enrichment in genes under recessive selection (and/or depletion of non-recessives) for autosomal recessive disease genes, hearing loss genes, and in genes identified in consanguineous individuals with depleted homozygous LOF variants. We reproduce classical predictions with significant enrichment in large metabolic pathways (e.g. TCA), consistent with Wright’s theory of the kinetic/physiological origin of dominance, GO annotated extracellular secreted enzymes, and enrichment in the non-recessive set (including co-dominant genes) for GO transcription factors. We find significant enrichment for infertility, GO meiosis, and GO spermatogenesis genes in the recessive strong selection category, suggesting a large autosomal recessive component to male-specific infertility consistent with mammalian studies in cattle. To our knowledge this is the first large set of recessive genes in humans (~1500) inferred from panmictic population data. This is qualitatively consistent with recessivity observed in deadly variants in flies and yeast. Notably, a large recessive component in many human genes is inconsistent with the assumption of additivity in previous estimates of selection against non-synonymous variants, as recessive genes under strong selection are inferred under weak selection due to prevalent neutral heterozygotes. Thus, a dominance-aware marginal distribution of fitness effects may substantially increase the average selection strength of deleterious human variants.

208

Global shared natural selection for increased stature in recent human history. *Y. Field[1,2], E.A. Boyle[1], N. Telis[3], Z. Gao[1,2], A. Bhaskar[1,2], J.K. Pritchard[1,2].* 1) Genetics, Stanford University, Stanford, CA; 2) Howard Hughes Medical Institute, Stanford University, Stanford, CA; 3) Program in Biomedical Informatics, Stanford University, Stanford, CA.

There is a growing understanding that the genetic basis for many human traits is polygenic, whereby complex phenotypes depend on the aggregated contribution of many small-effect variants across the genome. The study of the evolution of complex traits thus became a major goal in human evolutionary biology. Work in recent years has shown that variability in height across Europe is likely due in part to recent polygenic adaptation. This work was based on comparison of allele frequencies between populations. However, this approach is poorly suited to detect either very recent or very ancient selection, as well as any adaptation shared across populations.   We have recently introduced (bioRxiv; unpublished) a non-comparative approach to detect signals of polygenic adaptation in very recent history, which is based on the Singleton Density Score (SDS). SDS is a novel measure of recent frequency change of a common allele, inferred from the distribution of linked rare variants in large whole-genome-sequencing cohorts. Applied to data of ~3,000 genomes from the UK10K project, we identified signals of polygenic adaptation during the past ~2,000 years in ancestors of modern British that affected several traits. Above all traits, we found a strikingly strong signal for increased stature. Here we extend this work in two dimensions. First, we use SDS to analyze whole-genome-sequencing data for additional populations. In particular, we examine all 15 non-admixed populations from the 1000-genomes project (including populations from Africa, East Asia and Europe). Compared to the large UK10K cohort, SDS applied to the smaller samples of 1000-genomes (~100 individuals) detects adaptation on a larger timescale, of approximately 7,000 years. Second, we extend our method to infer the temporal dynamics of polygenic adaptation. We focus our analysis on the signal for polygenic adaptation of height.   In agreement with earlier studies, we find that selection for increased height has recently elevated in Northern Europe. However, surprisingly, we find that all examined populations, independently across continents, had experienced detectible selection for taller stature. This apparent global shared selective pressure could not have been detected with earlier comparative approaches. Our work thus sheds new light on recent evolution of human complex traits.

209

Population structure of UK Biobank and ancient Eurasians reveals adaptation at genes influencing blood pressure. *K. Galinsky*[1,2]*, P. Loh*[2,3]*, S. Mallick*[2,4]*, N.J. Patterson*[2]*, A.L. Price*[1,2,3]*.* 1) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; 3) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; 4) Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.

Analyzing genetic differences between closely related populations can be a powerful way to detect recent adaptation. The very large sample size of the UK Biobank is ideal for detecting selection using population differentiation, and enables an analysis of UK population structure at fine resolution. In our analyses of 113,851 UK Biobank samples with strictly defined UK ancestry, population structure in the UK is dominated by 5 principal components (PCs) spanning 6 clusters: Northern Ireland, Scotland, northern England, southern England, and two Welsh clusters. The eigenvalues of these PCs corresponded to discrete population F_{ST} values of 1.76×10^{-4} down to 3.63×10^{-5}, indicating exceedingly subtle structure that is ideal for detecting selection in large samples. Analyses incorporating data from ancient Eurasians show that populations in the northern UK have higher levels of Steppe ancestry, and that UK population structure cannot be explained as a simple mixture of Celts and Saxons.   We performed a scan for unusual population differentiation along top PCs by computing χ^2 (1 dof) statistics equal to the squared correlation between genotypes for each SNP and each PC, suitably scaled to account for genetic drift. We identified a genome-wide significant signal of selection at the coding variant rs601338 in *FUT2* (p=9.16×10^{-9}). In addition, we combined evidence of unusual differentiation within the UK with evidence from an independent selection scan of ancient Eurasians (Mathieson et al. 2015 Nature), by summing the χ^2 statistics from each scan. We identified new genome-wide significant (p≤5×10^{-8}) signals of recent selection at two additional loci: *CYP1A2*/*CSK* and *F12*. We detected strong associations to diastolic blood pressure in the UK Biobank for the variants with new selection signals at *CYP1A2*/*CSK* (p=1.10×10^{-19}) and for variants with ancient Eurasian selection signals in the *ATXN2*/*SH2B3* locus (p=8.00×10^{-33}), implicating recent adaptation related to blood pressure.

210

Quantifying selection and demographic effects on quantitative genetic variation: An application to anthropomorphic traits. *Y.B. Simons, G. Sella.* Biological Sciences, Columbia University, New York, NY.

The genetic architecture of a quantitative phenotype (i.e., the number, frequency and effect size of alleles underlying variation in its value) arises from genetic and population genetic processes. Mutations affecting the trait appear at a rate that reflects the target size, and their trajectory through the population is determined by demographic processes and by the selection acting on them. Many phenotypes, including human height and body mass index (BMI), appear to be under stabilizing selection, either because of selection on the trait itself or through the effects of genetic variation on other traits (i.e., via pleiotropy). With these considerations in mind, we introduce and solve a generative model for the genetic architecture of a continuous trait under direct and pleiotropic stabilizing selection. We derive simple and robust predictions for the distribution of additive genetic variation among loci. We then relate these predictions to observations from GWAS, accounting for how the power to detect a locus depends on its contribution to additive genetic variation. This new theory allows us to make inferences about the population genetic processes that underlie genetic variation for height and BMI in Europeans. We find an extremely good fit to GWAS findings (Wood et al. *Nature Genetics* 2014, Locke et al. *Nature* 2015): by fitting a single parameter for each trait, we are able to explain the distribution of additive genetic variation over genome-wide significant associations. Accounting for the demographic history of European populations suggests that the current GWAS is well powered to identify only loci under moderate selection. This relatively weak selection explains why the majority of loci that have been associated with height and BMI in Europeans are also segregating in African populations. We estimate the target size and distribution of selection coefficients of mutations affecting height and BMI within the range in which the current GWAS is well powered. We then employ these estimates to predict the expected increase in explained heritability with GWAS size due to variants in this range. Our results also suggest how increasing study size will enable the discovery of loci experiencing a wider range of selection effects. The framework presented here can be applied much more broadly, to investigate the genetic and selection parameters governing variation in other quantitative traits.

211

Estimating the selective effect of heterozygous protein truncating variants.

C. Cassa[1], D. Weghorn[1], D. Balick[1], D. Jordan[2], D. Nusinow[1], K. Samocha[3], A. O'Donnell Luria[3], D. MacArthur[3], M. Daly[3], D. Beier[4], S. Sunyaev[1]. 1) Division of Genetics, Harvard Medical School/Brigham and Women's Hospital, Boston, MA; 2) Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 3) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, MA; 4) Center for Developmental Biology and Regenerative Medicine, Seattle Children's Research Institute, Seattle, WA.

The dispensability of individual genes for viability has interested generations of geneticists. For some genes two functional copies are required, while others may tolerate the loss of one or both copies. 60,706 exomes now provide sufficient rare protein truncating variants (PTVs) to make genome-wide estimates of selection. In each gene, the cumulative frequency of rare PTVs is primarily determined by the balance between incoming mutations and selection rather than by reassortment of alleles by stochastic drift. This enables the estimation of the distribution of selection coefficients for heterozygous PTVs (s_{het}) and corresponding estimates for individual genes. We parametrize the distribution of selective effects using an inverse gamma and fit it using the sum over squared deviation of PTV counts. The resulting distribution allows the estimation of selection coefficients for each gene using the posterior probability generated with specific observables for individual genes*.* The mode s_{het} value corresponds to a fitness loss of 0.5%, concordant with estimates from flies. 3,558 genes have $s_{het} > 0.05$, also concordant with estimates of loss of function intolerance derived from population data. Genes under strong selection are enriched in embryonic lethal mouse knockouts [Sanger 703 genes $p=3.48x10^{-11}$; MGI 5,072, $p=3.27x10^{-77}$]. There is also a dramatic enrichment in putatively cell-essential genes from human tumor cell lines [$p=9.51x10^{-65}$] and yeast gene trap assay [$p=1.16x10^{-49}$] in genes with high s_{het} values. In Mendelian disorders, s_{het} significantly differentiates between mode of inheritance in disease genes (N=2,708, $p=3.25x10^{-65}$). In 504 clinical exome cases (Baylor), over 90% of novel dominant variants are associated with $s_{het} > 0.03$, demonstrating potential in new gene discovery. In haploinsufficient disease genes (ClinGen), genes under the strongest selection are enriched in disorders that are severe, highly penetrant, associated with congenital onset, and the result of de novo variation. Notably, there are many genes with high estimated fitness costs that have not been well-studied. We have created a prioritized list of genes using a heuristic score developed from functional evidence to indicate the most promising candidates for future functional screening. We find that key developmental pathways are dramatically enriched in genes with high s_{het} values, indicating that there may be many new genes of importance in this set of unannotated genes.

212

Negative selection in modern human populations involves synergistic epistasis.

M. Sohail[1,2], O. Vakhrusheva[3], J.H. Sul[4], P. de Bakker[5], A. Kondrashov[6], S. Sunyaev[2]. 1) Systems Biology PhD Program, Harvard University, Boston, MA; 2) Division of Genetics, Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA; 3) Department of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia; 4) Department of Psychiatry and Biobehavioral Sciences, UCLA, Los Angeles, CA, USA; 5) Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands; 6) Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA.

Little is known about the role of genetic interactions or epistasis in fitness. Synergistic epistasis, whereby multiple deleterious genetic variants have a larger cost on fitness than expected from their multiplicative effects, is suggested to play an important role in evolutionary dynamics, e.g. by maintaining sexual reproduction and reducing the deleterious mutation load. Its prevalence in natural populations, however, especially in humans still remains unknown. We developed a statistical method to quantify the magnitude and direction of epistasis in an organism’s fitness using DNA sequencing data. Our method uses the distribution of the number of deleterious variants carried by each individual, or the deleterious mutation burden. We compute an interaction statistic *I*, the difference between the empirical variance of the deleterious mutation burden and the additive variance expected under no epistasis. Our test relies on the depletion of the empirical variance due to observed “repulsion” between synergistically interacting deleterious variants. We performed forward simulations and used analytical results to show that *I* is zero under no epistasis, negative under synergistic epistasis and positive under antagonistic epistasis. We applied our method to six next-generation sequencing datasets – three European, Genome of the Netherlands, Alzheimer’s Disease Neuroimaging Initiative, controls from a amyotrophic lateral sclerosis study, and three non-European populations from the 1000 genomes Phase II Project. We observe a significant under-dispersion signal in rare Loss-of-Function (nonsense and splice disrupting) variants across human datasets, signifying that negative selection in modern human populations involves synergistic epistasis. The under-dispersion signal remains significant after correcting for confounders of linkage and population structure. We also replicated our findings in two *D. melanogaster* populations - Zambian flies from the *Drosophila* Population Genomics Project and American flies from the *D. melanogaster* Genetic Reference Panel. To our knowledge, our results are the first to show that empirical data points towards the presence of synergistic epistasis between deleterious variants involved in human fitness. The computed value of *I* for deleterious variants is within theoretical expectation, and helps explain the evolutionary maintenance of sexual reproduction, and how human populations survive their deleterious mutation load.

213

A molecular mechanism underlying the development of ventricular septal defects in individuals with 1p36 deletions.

B. Kim[1]*, H.P. Zaveri*[1]*, V.K. Jordan*[2]*, A. Hernández-García*[1]*, B. Fregeau*[3]*, E.H. Sherr*[3]*, D.A. Scott*[1]. 1) Molecular & Human Genetics, Baylor College Med, Houston, TX; 2) Molecular Physiology & Biophysics, Baylor College Med, Houston, TX; 3) Dept. of Neurology, University of California, San Francisco, San Francisco, CA.

Deletions of chromosome 1p36 are the most common terminal deletions in humans and affect 1 in 5000 newborns. Approximately 70% of infants with 1p36 deletions have congenital heart defects (CHD) with septal defects being the most common. We have recently demonstrated that haploinsufficiency of the *RERE* is sufficient to cause the majority of symptoms associated with proximal 1p36 deletions including septal defects. RERE is nuclear receptor coregulator that positively regulates retinoic acid signaling and is highly expressed in the developing heart. *Rere*-null embryos die around E9.5 with failure of cardiac looping. To overcome this early lethality, we generated RERE-deficient mice that carry an *Rere* null allele, *om,* and an *Rere* hypomorphic allele, *eyes3* (*Rere*$^{om/eyes3}$). On a pure C57BL6 background, 100% of *Rere*$^{om/eyes3}$ embryos develop septal, valve and outflow tract anomalies that are similar to those seen in children with 1p36 deletions. Further studies revealed that the number of atrioventricular canal (AVC) mesenchymal cells—which ultimately will form parts of the atrioventricular septum—were significantly reduced in E10.5 *Rere*$^{om/eyes3}$ embryos. AVC explants from *Rere*$^{om/eyes3}$ embryos also produced fewer migrating mesenchymal cells than explants from control littermates suggesting that RERE-deficiency causes a defect in epithelial mesenchymal transition (EMT). To confirm the cell autonomous role of RERE in the endocardium, we generated endothelial-specific *Rere* conditional knockout embryos using a Tie2-Cre. We found that 33% (2/6) *Rere*$^{flox/flox}$;Tie2-Cre embryos harvested at E15.5 had ventricular septal defects. To determine the molecular mechanism by which RERE functions to regulate EMT in the endocardium, we looked for evidence of dysregulation among a number of genes known to impact EMT. Quantitative RT-PCR and immunohistochemical analyses revealed decreased expression of GATA4—a retinoic acid responsive transcription factor required for EMT—in *Rere*$^{om/eyes3}$ embryo hearts at E10.5. The transcriptional expression of *Erbb3*—a known target of GATA4 that encodes an EGF-family receptor that is essential for EMT—was also reduced. Taken together, these results suggested that RERE plays a critical role in cardiac development by modulating the expression of genes—like *Gata4* and *Erbb2*—that are required for EMT and that dysregulation of these genes contributes to the development of septal defects in individuals with 1p36 deletions that include *RERE*.

214

Phenotyping pipeline for bicuspid aortic valve with/without ascending aortic aneurysm highlights pathological relevance of *ROBO4* to cardiovascular function.

C.E. Woods[1]*, R.A Gould*[1,2]*, C.R. Moats*[3]*, R.J. Rose*[1]*, H.C. Dietz*[1,2,4]*, A.S. McCallion*[1,3]*, MIBAVA Leducq Consortium.* 1) 1 McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 2) 2 Howard Hughes Medical Institute, Baltimore, MD, USA; 3) 3 Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 4) 4 Department of Pediatrics, Division of Pediatric Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

Bicuspid aortic valve (BAV) is a congenital heart defect, affecting 1-2% of the general population. Approximately 30% of individuals with BAV develop ascending aortic aneurysm (AscAA). Although the etiology of BAV/AscAA is unknown, previous family-based studies suggest that BAV is highly heritable. Using whole-exome sequencing, we are systematically screening ≥180 patients to identify causative variants. *ROBO4* is one gene in which we have detected multiple independent predicted deleterious alleles showing appropriate familial segregation. We have established a robust pipeline using zebrafish to functionally evaluate the biological relevance of genes/variants implicated in BAV/AscAA. Collectively, it provides an unprecedented capacity to interrogate the cardiovascular structural and functional correlates of BAV/AscAA in genetically defined developing and adult zebrafish. First, we establish mutant fish lines for selected genes using CRISPR/Cas9 or identify functionally equivalent mutants in public resources. For *robo4,* we generated a 7bp deletion in exon 6 of 19. As predicted, this mutation results in nonsense mediated mRNA decay, significantly reducing expression as quantified by qRT-PCR. Second, we have generated and applied transgenic reporter lines that mark critical structures including the aortic valve and vasculature, making developmental phenotyping possible. We have also optimized echocardiographic protocols that allow sequential survival assessment of aortic valve function and dimension of the *bulbus arteriosus*, the fish equivalent of the ascending aorta. We have completed a longitudinal study of wild type fish, establishing normative values for 3-11 months of age – a critical resource to document genetic predisposition and progression of disease. These data enabled development of a linear mixed model, accounting for the influence of specific biometrics on cardiovascular measurements in zebrafish. Echocardiography of *robo4* mutants at four months reveals extreme outflow turbulence and regurgitation (7/16), not seen in wild type fish (0/32). Importantly, our echocardiograms from mice homozygous for a *Robo4* loss-of-function mutation display a concordant phenotype. These data establish *ROBO4* as a gene critical for the normal development and homeostasis of the left ventricular outflow tract and establish a robust system and protocols for functional validation of other candidate genes of BAV/AscAA.

215

A rare pediatric Mendelian presentation of abdominal aortic aneurysm informs the predisposition for a common but complex cardiovascular disease. *R.A. Gould[1,2], D.T. Au[3], M. Migliorini[3,4], G. MacCarrick[1], N.L. Sobreira[5], J.C. Lopez-Gutierrez[6], S.C. Muratoglu[3,7], D.K. Strickland[3,4,7], H.C. Dietz[1,2,8].* 1) Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, USA; 2) Howard Hughes Medical Institute, Baltimore, Maryland, USA; 3) Center for Vascular and Inflammatory Diseases, University of Maryland School of Medicine, Baltimore, Maryland, USA; 4) Department of Surgery, University of Maryland School of Medicine, Baltimore, Maryland, USA; 5) Center for Inherited Disease Research, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 6) Vascular Anomalies Center, La Paz Children´s Hospital and Associate Professor of Pediatrics, Universidad Autonoma, Madrid, Spain; 7) Department of Physiology, University of Maryland School of Medicine, Baltimore, Maryland, USA; 8) Department of Pediatrics, Division of Pediatric Cardiology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.

Abdominal aortic aneurysm (AAA) affects 1-12% of the population with age-dependent penetrance, high heritability, and a strong influence of common environmental risk factors including smoking and obesity. While great progress has been made in our understanding of the genetic basis of thoracic aortic aneurysm, (identification of over 20 disease genes) our knowledge regarding the genetic determinants of AAA predisposition remain rudimentary. GWA studies have implicated dozens of loci, with few that have replicated and less that are mechanistically understood. Multiple studies have suggested enrichment for a specific allele of *LRP1*, the gene encoding low-density lipoprotein receptor-related protein 1, among patients with AAA. Animal studies have suggested that *LRP1* contributes to vessel wall integrity through a variety of mechanisms including inhibition of *PDGFR* signaling or endocytosis of matrix-degrading proteases, and that *LRP1* deficiency associates with enhanced predisposition for angiotensin II-induced dilatation of the ascending (but not abdominal) aorta. We performed whole exome sequencing for a patient requiring surgery for a massive (2.7 X 4.7cm) infrarenal abdominal aortic aneurysm in early childhood. Smaller aneurysms were seen in branch vessels to the upper and lower extremities without involvement of the thoracic aorta or cerebral vasculature; there was no family history of aneurysm and no manifestation of a systemic connective tissue disorder. Trio analysis revealed compound heterozygosity for a de novo mutation (p.A3487T) and a maternally inherited variant (p.V1291I) in *LRP1* that had not been previously described in databases of normal variation. Both mutations substitute evolutionarily conserved residues and are predicted to be highly deleterious. Patient-derived aortic vascular smooth muscle cells (VSMCs) showed impaired binding, internalization, and degradation of alpha-2 macroglobulin, a known ligand for *LRP1*. We also demonstrate that conditional silencing of *LRP1* expression in mouse VSMCs is sufficient to cause fully penetrant, spontaneous, and discrete AAA by 6 months of age (p<0.001 when compared to controls). Complementation of GWAS data with an aggressive early-onset Mendelian presentation of disease adds to conviction that *LRP1* function is essential for homeostasis of the abdominal aorta and that elucidation of distal pathogenic events will identify promising therapeutic targets for a complex and common human disease.

216

Loss-of-function mutations in the X-linked gene *BGN* cause a severe syndromic form of thoracic aortic aneurysms and dissections. *B. Loeys[1], J.A.N. Meester[1], G. Vandeweyer[1], I. Pintelon[2], K. Waitzman[3], L. Young[3], L.W. Markham[4], J. Vogt[5], J. Richer[6], L. Beauchesne[7], S. Unger[8], A. Superti-Furga[9], E. Reyniers[1], A. Verstraeten[1], L. Van Laer[1].* 1) Antwerp University/University Hospital of Antwerp, Antwerp, Belgium; 2) Department of Cell Biology and Histology, University of Antwerp, Antwerp, Belgium; 3) Department of Pediatric Cardiology, Lurie Children's Hospital of Chicago, Chicago, IL, United States; 4) Divisions of Pediatric and Adult Cardiology, Vanderbilt University, Nashville, TN, United States; 5) Birmingham Women's NHS Foundation Trust, Birmingham, United Kingdom; 6) Department of Medical Genetics, University of Ottawa Heart Institute, Ottowa, ON, United States; 7) Division of Cardiology, University of Ottawa Heart Institute, Ottowa, ON, United States; 8) Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland; 9) Service of Medical Genetics, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland.

Thoracic aortic aneurysm and dissection (TAAD) is typically inherited in an autosomal dominant manner but rare autosomal recessive and X-linked families have been described. So far, *FLNA* (filamin A) is the only X-linked gene associated with a syndromic form of TAAD, namely the periventricular nodular heterotopia type of Ehlers-Danlos syndrome. However, *FLNA* only explains a small number of the X-linked TAAD families. We performed targeted resequencing of 368 extracellular matrix and TGFβ transforming growth factor beta (TGFβ) related genes in a cohort of 11 Marfan-like probands without known causal *FBN1* mutation. We identified two male patients with respectively, a nonsense and a missense with potential effect on splicing mutation in BGN, an X-linked gene encoding the extracellular matrix small leucine-rich proteoglycan biglycan. Subsequent Sanger sequencing of *BGN* in 360 male and 155 female TAAD-patients, negative for known TAAD genes, identified an additional splice site mutation in a male index patient and suggested a deletion in two other male probands. The latter were confirmed by micro-array analysis. Overall, the nature of the identified *BGN* mutations indicates loss-of-function as the underlying pathogenetic mechanism. The clinical phenotype is characterized by early onset thoracic aortic aneurysm (as young as age 1 year) and dissection (as young as age 15 years). Clinical features overlap with Marfan and Loeys-Dietz syndrome and include hypertelorism, bifid uvula, malar hypoplasia, frontal bossing, pectus deformity, joint hypermobility, contractures and mild skeletal dysplasia. The cardiovascular status of *BGN* mutation carrying women ranges from normal, over mild aortic root dilatation to aortic dissection. Histological stainings of the patients' aortic wall revealed a low-normal collagen content, while elastin fibers appeared normal. Immunohistochemistry demonstrated an increase in TGFβ signaling activity, evidenced by a marked increase in nuclear pSMAD2 in aortic wall. Biglycan deficiency in male BALB/cA mice leads to sudden death from aortic rupture, indicating that biglycan is both structurally and functionally essential for the integrity of the aortic wall. These results show that *BGN* gene defects in human cause an X-linked syndromic form of severe TAAD.

217

Mutations in a novel cardiac specific exon of *FLNA* cause X-linked congenital heart disease. *C. Preuss[1], S. Yang[2], F. Wünnemann[1], P. Chetaille[3], M. Samuels[1], H. Björck[4], P. Eriksson[4], S. Mohamed[5], G. Andelfinger[1].* 1) Cardiovascular Genetics, Department of Pediatrics, Centre Hospitalier Universitaire Sainte-Justine Research Centre, Université de Montréal, Montreal, Canada; 2) Department of Cardiology, Nanjing Children's Hospital, Nanjing Medical University, Nanjing, China; 3) Department of Pediatrics, Centre Mère Enfants Soleil, Centre Hospitalier de l'Université (CHU) de Québec, Quebec City, Canada; 4) Atherosclerosis Research Unit, Center for Molecular Medicine, Department of Medicine, Karolinska Institutet, Stockholm, Sweden; 5) Department of Cardio and Thoracic Vascular Surgery, University Clinic of Schleswig-Holstein, Luebeck, Germany.

Congenital heart disease (CHD) is the most common birth defect with a population prevalence of around 1%. Despite a strong male predominance for the highly heritable trait, the genetic basis for these cardiac malformations remains poorly understood. Here, we report the mapping of a novel X-linked locus in three French Canadian pedigrees with multiple affected male individuals with septal defects, aortic valve lesions and arrhythmia. Reconstruction of ascending genealogies supports the notion of a founder effect for this novel X-linked syndrome, dating a common founder couple back to 1788. Linkage analysis and genetic fine mapping in a large pedigree (> 100 individuals) using the Illumina Omni 5.0 genotyping platform revealed a significant linkage interval on chromosome Xq28 (LOD score = 3.29) harboring a 260kb haplotype co-segregating with disease in 17 affected patients. Whole-exome and targeted re-sequencing identified a single mutation in the disease associated haplotype in a conserved intronic region of *FLNA* in the absence of other rare deleterious coding mutations co-segregating with disease. This highly conserved mutation (GERP score > 4) is absent among public data sets (dbSNP144, 1000 Genomes) and in 960 French Canadian controls. Transcriptomic analysis of fetal heart RNA-seq data revealed that the disease associated mutation resides in a novel cardiac specific exon coding for a previously unknown FLNA isoform. Proteomic analysis using a specifically designed antibody targeting the novel FLNA isoform showed a distinct band at the expected size in fetal heart tissues. High-throughput yeast two-hybrid screening for the novel FLNA isoform revealed a specific protein-protein interaction with the cardiac muscle protein CASQ2 (OMIM:114251) that causes an overlapping phenotype. This suggest a potential important functional role of the new tissue specifc isoform in anchoring membrane-bound proteins to the cell membrane cytoskeleton in the course of fetal heart development. Taken together, our study highlights the importance of novel exons and tissue-specific isoforms of candidate genes for the identification of pathogenic coding mutations in Mendelian disease traits.

218

The role of loss-of-function mutations on death and development of rejection after heart transplantation. *J. van Setten[1], B.S. Cole[2], Y.R. Li[2], N. de Jonge[1], M.V. Holmes[2], C.C. Baan[3], O.C. Manintveld[3], A.M.A. Peeters[3], F. Dominguez[4], K.K. Khush[5], P. Garcia-Pavia[4], J.W. Rossano[2], R.A. de Weger[1], J.H. Moore[2], B. Keating[2], F.W. Asselbergs[1].* 1) University Medical Center Utrecht, Utrecht, Netherlands; 2) University of Pennsylvania, Philadelphia, PA; 3) Erasmus Medical Center, Rotterdam, Netherlands; 4) Puerta de Hierro University Hospital, Madrid, Spain; 5) Stanford University, Stanford, CA.

Introduction: Currently, donor/recipient (D-R) matching is suboptimal. Besides HLA, also other genetic factors play a role in graft rejection. Possible sources of genetic variation underpinning rejection are homozygous deletion CNVs spanning whole gene or exon regions and LoF variants ablating two copies of a given gene, resulting in incompatibility across the proteomes of donor and recipient. We have developed a pipeline to identify human knockouts and aim to associate the detected knockout genes with acute rejection and death after heart transplantation. **Methods:** iGeneTRAiN is a large-scale international consortium, which consist of over 11,500 solid organ D-R pairs, including 758 heart transplant D-R pairs. All samples were genotyped and untyped variants were imputed using a combined reference panel of the Genome of the Netherlands and the 1000 Genomes Project. Effects of genetic variants were annotated with ENSEMBL's Variant Effect Predictor using the LOFTEE plugin. CNVs were annotated as LoF if they remove a gene's normal start codon, more than 50% of its coding sequence, or create an internal disruption predicted to result in a frameshift. Important features of our pipeline include the use of phased haplotypes to detect compound heterozygous LoFs and the ability to detect both common and rare variants. For each D-R pair, genes were identified that are inactive in both copies in a transplant recipient but present in at least one functional copy in the corresponding donor. These genes were tested for association with death and rejection using cox proportional hazard models with donor and recipient age and sex, study center, year of transplantation and 10 PCs for donor and recipient as covariates. **Results:** In total, 827 genes were inactive in both copies in the recipient but active in the donor in at least one D-R pair. We identified one gene associated with rejection and one gene associated with death after heart transplantation ($P<3x10^{-5}$ to correct for the number of tests). **Conclusions:** We identified two genes associated with acute rejection and death after heart transplantation. We aim to increase our sample of heart transplant D-R pairs and to conduct cross-organ meta-analyses including lung, liver, and kidney transplants, maximizing statistical power to identify novel genes. We ultimately aim to translate genetic data into clinical applications such as more optimal genomic compatibility matching of D-R pairs and immune suppression therapy dosing.

219

A complex structural variant at the glycophorin gene cluster is associated with strong protection against severe malaria in East Africa. *G. Band*[1,2], *E. Leffler*[1,2], *K.A. Rockett*[1,2], *Q.S. Le*[1], *C.C.A. Spencer*[1], *D.P. Kwiatkowski*[1,2], *MalariaGEN.* 1) Wellcome Trust Centre for Human Genetics, Oxford, Oxfordshire, United Kingdom, OX3 7BN; 2) Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA.

Glycophorins, encoded by a cluster of three paralogous genes on human chromosome 4, form some of the most abundant proteins on the erythrocyte surface, and act as receptors for the malaria parasite *P.falciparum* during red blood cell invasion. A complex system of SNPs, gene conversions and structural variation in these genes determines the diverse MNS blood group system. Motivated by our previous observation of association with severe malaria near this gene cluster, we here use whole-genome sequence data from 765 individuals from sub-Saharan Africa together with publicly available data from the 1000 Genomes Project to survey structural variation across the region. This analysis reveals multiple large deletions and duplications, corresponding both to known blood groups and novel variation. We use imputation and direct typing to infer genotypes for these variants in a genome-wide association study of 8,500 cases of severe malaria and 8,000 population controls from 11 populations across sub-Saharan Africa, Asia and Oceania. We show that a single structural variant, a complex duplication affecting all three paralogs, is associated with strong protection against severe malaria in East Africa, and explains the primary signal of association in the region. We characterise the haplotypic structure of this duplication, which is only observed at high frequency in parts of East Africa, raising questions as to its evolutionary history. The observation of association of this variant with malaria susceptibility adds to the growing body of evidence implicating structural variation in the genetic basis of common diseases, and suggests specific functional hypotheses that will form the basis of future work.

220

Comparative genomics of innate immunity in human and non-human primates. *J.C. Teixeira*[1,2], *L. Quintana-Murci*[1,2]. 1) Institut Pasteur, 25 Rue du Dr. Roux, 75015 Paris, France; 2) CNRS URA3012, 75015 Paris, France.

Innate immunity constitutes the front line of host defence and provides a valuable model for the study of the selective pressures imposed by microorganisms on host genomes. Population genetic studies in humans have shown that the impact of selection on some families of innate immune receptors and downstream signaling molecules varies considerably. Despite humans and closest relatives share most of their genome, increasing evidence suggests that humans and other apes exhibit important differences in susceptibility to, and severity of, infectious diseases (e.g. HIV, malaria). In fact, it is possible that many of such differences emerged as a result of the action of natural selection leading to species-specific adaptations in response to environmental changes. Nevertheless, the effects of natural selection in shaping innate immunity in our closest relatives remain largely unknown. In this study, we aim at uncovering shared and unique signatures of natural selection acting on innate immunity genes in great apes, and identify species-specific adaptations that might be essential for individual and population survival. We first implemented coalescent simulations on great ape demographic history and demonstrate that the different statistics used present high power to uncover targets of purifying, positive, and balancing selection in the genomes of great apes. We then analyzed whole-genome sequence data for different populations of great apes, covering a total of 11 great ape subspecies. To do so, we focused on comparing selection signatures on a set of more than 1,500 loci involved in innate immunity functions with the remainder of the genome. Our analyses show that, taken as a whole, innate immunity genes are privileged targets of natural selection in the genomes of great apes. Moreover, we provide evidence that some loci, and their related biological functions, exhibit extensive differences in the form and intensity of selection across species. These findings suggest important differences across primates in the mechanisms involved in host adaptation to pathogen pressures, informing about the pressures imposed by their respective ecological habitat. Together, our results provide a thorough understanding of selective forces shaping the evolution of innate immunity in great apes. To our knowledge, this study represents the first attempt of bringing to light the evolutionary mechanisms that operated for millions of years as a response to pathogen infection.

221

Colorectal cancer mutational and transcriptome profiles shape the microbiome of the tumor microenvironment. *R. Blekhman[1], M. Burns[1], E. Montassier[2], D. Knights[2].* 1) Genetics, Cell Biology, and Development, University of Minnesota, St. Paul, MN; 2) Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA.

Understanding the interaction between colorectal cancer (CRC) and the microbiota is critical for the development of microbiome-based prognostics and therapies. To that end, we have investigated the association between a tumor's genomic landscape and its associated microbial communities by performing whole-exome sequencing, RNA-seq, and microbiome profiling in tumors and normal colorectal tissue samples from the same patient. We find a strong association between loss-of-function mutations in relevant tumor genes or pathways and shifts in the abundance of specific sets of bacterial taxa. Similarly, we find correlations between abundance of individual bacterial taxa and tumor expression profiles, highlighting pathways of interaction between the tumor and its microbiota. By constructing a risk index classifier, we show that we can use microbiome data to accurately predict the expression and mutational profiles of tumors. For example, we can accurately predict the existence of loss-of-function mutations in cancer-related genes and pathways, including MAPK and Wnt signaling, solely based on the composition of the microbiota. These results can serve as a starting point for development of colon cancer prognostics and individualized microbiota-targeted therapies.

222

GWAS of cellular and clinical traits reveals *VAC14* regulates *Salmonella* invasion and typhoid fever susceptibility. *D. Ko[1,2], M. Alvarez[1], S. Dunstan[3], P. Luo[1], L. Glover[1], S. Oehlers[1], E. Walton[1], L. Wang[1], D. Tobin[1].* 1) Dept. of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC; 2) Dept. of Medicine, Duke University Medical Center, Durham, NC; 3) University of Melbourne, Melbourne, Australia.

Background: Human genetic variation plays a major role in susceptibility to infection. Our lab has developed a cell-based screen of human genetic variation (Hi-HOST: High-throughput human in vitro susceptibility testing) for identifying genetic differences that impact host-pathogen interactions. We hypothesized genetic differences that regulate cellular phenotypes of *Salmonella* infection may reveal cell biological mechanisms and genetic determinants of infectious disease. Specifically, we examined genetic variation in *S.* Typhi invasion of cells and the impact this has on typhoid fever risk.**Methods and Results:** We carried out a Hi-HOST cellular GWAS of genotyped cell lines for invasion by *S.* Typhi. Deviation from neutrality was observed in a stratified QQ plot of eQTLs with a SNP in the gene encoding the phosphoinositide scaffolding protein VAC14 showing association with invasion (p=0.0006). While this association suggested that VAC14 regulates invasion, we experimentally verified this with RNAi, CRISPR knockout cell lines, and plasmid complementation. CRISPR knockout of *VAC14* abolished protein expression and doubled invasion (p=0.005). In systematically testing each of the steps in early *S.* Typhi infection, only docking was increased in *VAC14*$^{-/-}$ cells (p=0.0002). Furthermore, *VAC14*$^{-/-}$ cells exhibited a 35% increase in cellular cholesterol (p=0.009), a host molecule bound by the *Salmonella* type-III secretion system for docking. The results demonstrate phosphoinositides regulate cholesterol to impact *S.* Typhi docking and invasion. To extend these studies *in vivo*, *VAC14*$^{-/-}$ zebrafish are being evaluated by tracking *S.* Typhi dissemination and outcome. Finally, the *VAC14* SNP was tested for association with typhoid fever in a case-control study in Vietnam (n=500 cases, 496 controls). Remarkably, the allele that decreases invasion is associated with protection against typhoid fever (p=0.0098, OR=1.4).**Conclusion:** In this multi-disciplinary project, we combined cellular GWAS, mechanistic studies, zebrafish modeling, and human association testing to discover genetic diversity in *VAC14* regulates cholesterol levels, *Salmonella* invasion, and typhoid fever risk. The results demonstrate the value of functional characterization of genetic variants revealed through GWAS of cellular traits for understanding basic biological mechanisms as well as genetic susceptibility to infectious disease.

223

Genetic control of RNA splicing contributes to inter-individual variation in transcriptional responses to bacterial infection. *A.A. Pai[1], O. Tastet[2,3], J.C. Grenier[3], Y. Nedelec[2,3], V. Yotova[3], C.B. Burge[1,4], L.B. Barreiro[3,5].* 1) Department of Biology, MIT, Cambridge, MA; 2) bDepartment of Biochemistry, Faculty of Medicine, University of Montreal, Montreal, QC, Canada; 3) Sainte-Justine Hospital Research Centre, Montreal, QC, Canada; 4) Department of Biological Engineering, MIT, Cambridge, MA; 5) Department ofPediatrics, Faculty of Medicine, University ofMontreal, Montreal, QC, Canada.

Changes in gene regulation have long been known to play important roles in both innate and adaptive immune responses. These changes likely contribute to variation in susceptibility to immune-related diseases across individuals. Recent work has identified genetic loci that are associated with variation in both susceptibility to infectious diseases and focused on the transcriptional response to immune activation. However, despite emerging examples of post-transcriptional mechanisms as regulators of immune defenses, changes in alternative splicing and mRNA processing are less well characterized in the context of immune responses. Previously, we found overall increases in isoform diversity following bacterial infection, coupled with changes of alternative splicing and remodeling of the 3' UTR landscape. Here, we leverage natural sequence variation to understand how RNA processing responses coordinate with transcriptional responses to regulate inter-individual susceptibility to bacterial infection. Using RNA-seq data from macrophages before and after infection with live bacterial pathogens across from 161 genotyped individuals, we identify SNPs that are associated with variation in RNA processing responses to infection (infection response splicing quantitative trait loci – ir-sQTLs). Overall, we find hundreds of ir-sQTLs, many of which are associated with the RNA processing of genes involved in critical immune system response pathways. Our QTL mapping strategy at the level of individual exons rather than overall isoform usage allows us to interrogate the likely causal mechanisms for sets of ir-sQTLs. Notably, we find that the greatest proportion of ir-sQTLs regulate variation in cassette exon usage or alternative 3' UTR usage and enrichments of ir-sQTLs in splice sites and polyadenylation signals, respectively. These ir-sQTLs are more likely to also be associated with variation in overall gene expression levels and we estimate that a significant proportion of variation in transcriptional responses to infection is mediated by a QTL regulating changes in RNA processing after infection.

224

An atlas of genomic variants connecting cellular host-pathogen traits to human infectious disease. *L. Wang[1], D. Ko[1,2].* 1) Department of Molecular Genetics & Microbiology, Duke University, Durham, NC; 2) Department of Medicine, Duke University, Durham, NC.

GWAS have successfully identified hundreds of genetic variants associated with human diseases and traits. However, elucidating the mechanisms whereby genetic differences impact genes and pathways to affect risk and severity of disease remains a formidable challenge. Cellular phenotypes of infection simplify the complexity of GWAS of infectious disease by providing uniform pathogen exposure and environment, while also providing a system for experimental validation and mechanistic studies. We developed a novel discovery platform of GWAS of cellular traits, named Hi-HOST (high throughput human *in vitro* susceptibility testing) that utilizes nearly a thousand genotyped lymphoblastoid cell lines derived from different human populations. With this platform, we measured and quantified cellular responses of infectivity and replication, cytokine levels, and host cell viability using 9 different pathogens, and consequently carried out family-based GWAS analysis on 148 cellular traits. Nearly all traits (95%) had repeatability of measurement between 50-95% and most (85%) had significant heritability more than 0.1. A total of 69 common genetic variants exceeding the commonly used threshold for genome-wide significance () were found in traits for *Yersinia pestis*, *Chlamydia trachomatis, Toxoplasma gondii* and *Salmonella* species. Follow-up experiments validated the functional significance of some of these SNPs and the mechanisms of these genetic variants in infection. We connected the cellular GWAS to GWAS of infectious disease risk (for typhoid fever, *Salmonella* bacteremia, trachoma, and *Staph.* infection) through enrichment analyses, revealing cellular responses and pathways that play a significant role in risk of specific infectious disease. Furthermore, heritability estimates and partitioning of infectious disease risk by cellular phenotypes reveals the common genetic architecture among cellular and clinical traits of infection. In summary, we 1) generated an atlas of genetic variants associated with 9 pathogens using cellular GWAS; 2) succeeded in identifying and validating known and new genes in susceptibility to infection; and 3) confirmed that cellular pathways play an essential role in infectious disease traits. Our development and validation of this novel discovery platform serves as a first step in interpreting human disease through the lens of cell biology to advance understanding of pathogenesis and further diagnostics and therapy.

225

Novel polygenic risk prediction using individual-level data and summary statistics for secondary traits. *W. Chung*[1,2], *J. Chen*[3], *C. Chen*[1,2], *S. Lindstrom*[1,2], *P. Kraft*[1,2,4], *L. Liang*[1,2,4]. 1) Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, MA; 2) Department of Epidemiology, Harvard School of Public Health, Boston, MA; 3) Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN; 4) Department of Biostatistics, Harvard School of Public Health, Boston, MA.

Recent studies have shown that human complex traits may share similar genetic architecture with other clinical or disease-related phenotypes. Combining information across such secondary traits may increase the prediction accuracy for the primary trait of interest. In this paper, we develop a new statistical framework for genotype-based prediction for complex traits using multiple related phenotypes. Based on penalized least squares methods, we propose a novel cross trait penalty (CTP) function to incorporate the shared genetic effects across multiple traits. Our approach has several advantages: (1) it extracts information from the secondary traits that is beneficial for predicting the primary trait but tunes down information that is not; (2) the primary trait and secondary traits can come from same set or different sets of samples; (3) it can incorporate secondary traits based on individual-level genotypes or summary statistics from large scale GWAS studies; (4) our novel implementation of a parallel computing algorithm makes it feasible to apply penalized least squares methods on millions of SNPs from tens of thousands of samples. Our methods are evaluated on GWAS datasets using the Lasso and the minimax concave penalty (MCP) to induce a sparse solution. We show that our multi-trait methods significantly increase the prediction accuracy using extensive simulations and real GWAS datasets. For example, we take human height as the primary trait and age at menarche (AAM) or body mass index (BMI) as the secondary traits. We use 11,473 individuals from the Nurses' Health Study as a training set and 74,757 individuals from the UK Biobank as an independent validation set. With ~2 million SNPs, the relative gain in prediction R^2 of our multi-trait approach is 69.5% for Lasso (112.4% for MCP) using AAM and 70.4% (121.6%) using BMI. We further show that the use of summary statistics for AAM or BMI from large scale GWAS can substantially improve the prediction accuracy. The relative gain in prediction R^2 is 48.2% (89.2%) using AAM GWAS summary statistics from the ReproGen Consortium and 56.2% (100.5%) using BMI GWAS summary statistics from the Giant Consortium. We finally compare our prediction methods with the recently proposed multi-trait genomic BLUP (MTGBLUP) method. The prediction accuracy for height improves from R^2=4.12% (MTGBLUP) to 5.26% (Lasso+CTP) or 5.13% (MCP+CTP) using AAM and from R^2=4.15% (MTGBLUP) to 5.29% (Lasso+CTP) or 5.35% (MCP+CTP) using BMI.

226

Why real biological interactions are usually not detectable in genetic association analyses. *N. Kodaman, S.M. Williams.* Department of Epidemiology and Biostatistics , Case Western Reserve University, Cleveland, OH.

Association studies assessing interactions between genetic variants and background factors (such as environment, sex, or other physiological traits) have failed to explain much additional genetic variance, despite strong evidence indicating that genetic effects are context-dependent at the organismal level. A possible explanation for this paradox is that the "main effect" term in statistical regression models is more sensitive to biological interactions than the interaction term. We developed population genetic models of context-dependent genetic effects to explore this possibility. First, we show that the conventional interaction term used in linear regression equations inadequately models context-dependent genetic effects when the interacting variable does not induce a change in the *direction* of the genetic variant's effect at the biological level (e.g. from deleterious to benign). When that constraint is imposed on the model, we show that the ratio of the proportion of variance explained by the main effect vs. the interaction effect in a regression analysis can be expected to have a lower bound of 2/π. We also modeled a continuous outcome (*Y*) as a linear function of a normally distributed "context" variable (*Z*), a normally distributed random effect, and genotype (*X*), such that the slope of the genetic effect (B_x) increases at incremental quantiles of *Z*. According to this model, as the number of quantiles increases from 2 to ∞, the ratio of expected proportion of variance explained by the main effect vs. the interaction effect ranges from 9π/2 to π, and does not vary with other model parameters (sample size, minor allele frequency, genetic effect size, or the correlation between *Y* and *Z*). Because the ratio converges rapidly, the model can be well approximated using a single equation, which differs from the conventional linear interaction model only with respect to the cross-product term ($X \cdot \Phi(Z)$ vs. $X \cdot Z$, where Φ is the cumulative distribution function). Our models indicate that highly heterogeneous genetic architecture can account for the large number of small additive effects observed in genetic association studies, but that relying solely on interaction p-values will fail to reveal this complexity. Statistical methods that jointly consider main effects and interaction effects should be most powerful. To test such methods, we provide a flexible way to simulate realistically structured and biologically plausible interaction data.

227

Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls. *Y.J. Hu[1], P. Liao[1], H.R. Johnston[1], A.S. Allen[2], G.A. Satten[3].* 1) Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, United States of America; 2) Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, United States of America; 3) Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America.

Background: Next-generation sequencing of DNA provides an unprecedented opportunity to discover rare genetic variants associated with complex diseases and traits. However, the common practice of first calling underlying genotypes and then treating the called values as known is prone to false positive findings, especially when genotyping errors are systematically different between cases and controls. This happens whenever cases and controls are sequenced at different depths, on different platforms, or in different batches. **Methods:** We provide a likelihood-based approach to testing rare variant associations that directly models sequencing reads without calling genotypes. We consider the (weighted) burden test statistic, which is the (weighted) sum of the score statistic for assessing effects of individual variants on the trait of interest. Because variant locations are unknown, we develop a simple, computationally efficient screening algorithm to estimate the loci that are variants. Because our burden statistic may not have mean zero after screening, we develop a novel bootstrap procedure for assessing the significance of the burden statistic. **Results:** We demonstrate through extensive simulation studies that the proposed tests are robust to a wide range of differential sequencing qualities between cases and controls, and are at least as powerful as the standard genotype calling approach when the latter controls type I error. An application to the UK10K data reveals novel rare variants in gene *BTBD18* associated with childhood onset obesity. The relevant software is freely available.

228

Novel methodology to detect SNP association and heterogeneity in allelic effects between diverse populations via trans-ethnic meta-regression. *A.P. Morris[1,2,3], R. Mägi[2], M. Horikoshi[3], T2D-GENES Consortium.* 1) Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom; 2) Estonian Genome Center, University of Tartu, Tartu, Estonia; 3) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK.

Trans-ethnic meta-analysis of genome-wide association studies (GWAS) across diverse populations can increase power to detect complex trait loci when the underlying causal variants are shared between ancestry groups. However, heterogeneity in allelic effects between GWAS at SNPs in these loci can occur due to ancestry for reasons including: (i) differences in patterns of linkage disequilibrium with the causal variant across diverse populations; and (ii) differential exposure to an interacting environmental risk factor across ethnic groups. Whilst random-effects meta-analysis allows for heterogeneity in allelic effects between GWAS, it is not possible to assess the contribution of ancestry to this variability. Here, a novel approach is presented to detect SNP association and quantify the extent of heterogeneity in allelic effects that is due to ancestry via trans-ethnic meta-regression. Multi-dimensional scaling on the distance matrix of mean pairwise effect allele frequency differences between GWAS across thinned SNPs is first used to derive principal components (PCs). Meta-regression then proceeds by modelling allelic effect sizes as a function of PCs, weighted by their standard errors, providing a test of SNP association and an assessment of the contribution of ancestry to heterogeneity in allelic effects between GWAS. The methodology has been implemented in the MR-MEGA software. Simulations were undertaken for a range of scenarios of allelic effect heterogeneity between 26 populations in the 1000 Genomes Project. Increased power to detect SNP association with a binary trait was observed for MR-MEGA over fixed- and random-effects meta-analysis when heterogeneity in allelic effects was due to ancestry. Improved fine-mapping resolution was also observed across seven distinct type 2 diabetes association signals (at *CDKAL1*, *CDKN2A-B*, *IGF2BP2* and *KCNQ1*) in 22,086 cases and 42,539 controls from 19 GWAS of African American, East Asian, European, Hispanic and South Asian ancestry, imputed to the 1000 Genomes Project reference panel. The strongest evidence for heterogeneity due to ancestry was observed at *CDKAL1* (p=2.0x10^{-7}), where the index SNP demonstrated weaker allelic effects in GWAS of South Asian ancestry than in other ethnic groups, and the 99% credible set of variants driving the association signal was reduced from 4 (mapping to 11.2kb) and 16 (69.0kb), respectively, through fixed- and random-effects meta-analysis, to just 2 (2.3kb) with MR-MEGA.

229

Efficiency and accuracy of fine-mapping using GWAS summary data. *C. Benner[1,2], A. Havulinna[3], V. Salomaa[3], S. Ripatti[1,2,4], M. Pirinen[1].* 1) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; 2) Department of Public Health, University of Helsinki, Helsinki, Finland; 3) National Institute for Health and Welfare (THL), Helsinki, Finland; 4) Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

During the last two years, genetics research has seen a surge of computational approaches that work directly on summary data from Genome-Wide Association Studies (GWAS) to avoid privacy concerns and logistics of sharing individual-level genotype data and to cope with ever increasing sample sizes. Recently, also statistical fine-mapping for identifying causal variants has been extended to use GWAS summary data (CAVIAR, CAVIARBF, PAINTOR). However, computationally expensive exhaustive search restricts the existing approaches to only a few hundred variants. Furthermore, although these approaches require an estimate of Linkage Disequilibrium (LD) between the variants, the impact of LD estimates has not been comprehensively studied and current practices may actually perform poorly. We introduce a software package FINEMAP that replaces the exhaustive search by an ultrafast stochastic search. We demonstrate that (1) FINEMAP opens up completely new opportunities by fine-mapping the HDL-C association of the *LIPC* locus with 20,000 variants in less than 90 seconds while exhaustive search would require thousands of years. By jointly modeling the whole locus, (2) FINEMAP can identify more plausible variant combinations than standard conditional analysis. At the *LIPC* locus we identify a 3-SNP configuration with 190-fold higher likelihood than the top configuration from conditional analysis suggesting that a missense variant and a promoter polymorphism are likely to be causal whereas the lead variant in single-SNP testing has less evidence than a regulatory variant correlated with it. With extensive simulations we further show that (3) FINEMAP is as accurate as exhaustive search when the latter can be completed and (4) achieves even higher accuracy when the latter must be restricted due to computational reasons. We also report important practical results showing that (1) LD estimation from 1000 Genomes Project (1000G) data should be avoided due to its small size, (2) LD shrinkage improves the performance of 1000G panels and (3) 1,000 individuals from the target population is typically large enough for reliable fine-mapping of variants with minor allele frequency above 0.01. Our results are based on comprehensive simulations using up to 15,000 Finns over 100 GWAS regions from coronary artery disease, Crohn's disease, lipids, schizophrenia and type 2 diabetes and on Finnish data of the *APOE* region that we fine-map in detail discovering a novel variant associated with LDL-C.

230

Leveraging the diploid genome to increase power in *QTL studies. *M. Subramaniam, N.A. Zaitlen, C.J. Ye.* UCSF, San Francisco, CA.

Next generation sequencing coupled with molecular assays has enabled unprecedented opportunities to quantitatively measure genome function. When combined with dense genetic data, quantitative trait locus (QTL) mapping of functional genomic traits is a fundamental tool for understanding the genetic basis of processes such as transcription regulation. In standard *QTL analysis, genotypes are tested for association with estimated abundance of each genomic feature. While powerful, this approach ignores the diploid nature of our genomes, testing combined abundances across both alleles of every feature. In this work, we develop a new phase aware test for *QTL analysis (PhAT-QTL) leveraging allele specific estimates of genomic features. Briefly, we phase the genotypes of all individuals in a study and then test for association between the haploid count of each SNP and the allele-specific expression (ASE) of the gene on the same haplotype. We use a linear mixed model to account for the correlated environments of haplotypes from the same individual. Through analytical derivations and simulations, we show that power increases relative to standard genotype based tests as a function of the number of heterozygotes, the noise correlation between haplotypes, and the number of samples with detectable ASE. Simulations show that phasing error and ASE quantification error result in decreased power as opposed to increased false QTLs. Read simulations on 1000G phased genomes demonstrate that PhAT-QTL is able to detect QTLs with varying haplotype structures at a low false positive rate (AUC =0.89). Unlike previous ASE aware *QTL methods, our approach easily scales to thousands of individuals. In order to compare to the leading QTL detection method RASQUAL, while accommodating the increased computational burden, we apply PhAT-QTL to a reduced GEUVADIS dataset of 50 individuals. We assess performance with respect to previously discovered QTLs, observing a true positive rate (TPR) of 44% at a false positive rate of 10% in comparison RASQUAL (TPR=41%) and linear regression (TPR=38%). PhAT-QTL is also able to detect isoform-specific effects and autoregulatory effects in which ASE is detected but the total expression level across both alleles is constrained. With denser genetic maps and technological advances to obtain longer reads, we expect PhAT-QTL to be broadly applicable to a number of other *QTL analyses and will greatly impact the discovery of previously undetectable signals.

231

De novo loss-of-function mutations in _SON_ disrupt RNA-splicing of genes essential for brain development and metabolism, causing an intellectual disability syndrome. *L.E.L.M. Vissers[1], J-H. Kim[2], D.N. Shinde[3], M.R.F. Reijnders[1], N.S. Hauser[4], R.L. Belmonte[5], G.R. Wilson[5], D.G.M. Bosch[1], P.A. Bubulya[6], V. Shashi[7], S. Petrovski[8,9], J.K. Stone[2], E.Y. Park[2], J.A. Veltman[1,10], M. Sinnema[10], C.T.R.M. Stumpel[10], J.M. Draaisma[11], J. Nicolai[12], H.G. Yntema[1], K.L. Lindstrom[13], B.B.A. de Vries[1], T. Jewett[14], S.L. Santoro[15,16], J. Vogt[17], K.K. Bachman[18], A.H. Seeley[18], A. Krokosky[19], C. Turner[19], L. Rohena[20,21], S. Tang[3], D. El-Khechen[3], M.T. Cho[22], K. McWalter[22], G. Douglas[22], B. Baskin[22], A. Begtrup[22], T. Funari[22], K. Schoch[7], A.P.A. Stegmann[10], S.J.C. Stevens[10], D-E. Zhang[23,24,25], D. Traver[25], X. Yao[23], D.G. MacArthur[26,27,28], H.G. Brunner[1,10], G.M. Mancini[29], R.M. Myers[30], T.M. Strom[31], D. Wieczorek[32], M. Hempel[33], F. Kortuem[33], F. Laccone[34], L.B. Owen[2], S-T. Lim[35], D.L. Stachura[5], E-Y.E. Ahn[2,35], University of Washington Center for Mendelian Genomics; The Deciphering Developmental Disorde.* 1) Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, 6500 HB, Nijmegen, The Netherlands; 2) Mitchell Cancer Institute, University of South Alabama, Mobile, AL 36604, USA; 3) Ambry Genetics, Aliso Viejo, CA 92656, USA; 4) Medical Genetics and Metabolism, Valley Children's Hospital, Madera, CA 93636, USA; 5) Department of Biological Sciences, California State University Chico, Chico, CA 95929, USA; 6) Department of Biological Sciences, Wright State University, Dayton, OH 45435, USA; 7) Department of Pediatrics, Division of Medical Genetics, Duke University School of Medicine, Durham, NC 27710, USA; 8) Department of Medicine, The University of Melbourne, Austin Hospital and Royal Melbourne Hospital, Victoria, 3010, Australia; 9) Institute for Genomic Medicine, Columbia University, New York, NY 10027, USA; 10) Department of Clinical Genetics and School for Oncology & Developmental Biology (GROW), Maastricht University Medical Center, 6202 AZ, Maastricht, The Netherlands; 11) Department of Pediatrics, Radboudumc Amalia Children's Hospital, 6500 HB, Nijmegen, The Netherlands; 12) Department of Neurology, Maastricht University Medical Center, 6299 HX, Maastricht, The Netherlands; 13) Division of Genetics and Metabolism, Phoenix Children's Hospital, Phoenix, AZ 85016, USA; 14) Department of Pediatrics, Section on Medical Genetics, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA; 15) Nationwide Children's Hospital, Columbus, OH 43205, USA; 16) Ohio State University College of Medicine, Columbus, OH 43210, USA; 17) West Midlands Regional Genetics Service, Birmingham Women's NHS Foundation Trust, Birmingham, United Kingdom; 18) Geisinger Medical Center, Danville, PA 17822, USA; 19) Department of Pediatrics, Division of Genetics, Walter Reed National Military Medical Center, Bethesda, MD 20889, USA; 20) Department of Pediatrics, Division of Genetics, San Antonio Military Medical Center, Fort Sam Houston, TX 78234, USA; 21) Department of Pediatrics, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA; 22) GeneDx, Inc, 205 Perry Parkway, Gaithersburg, MD 20877, USA; 23) Moores Cancer Center, University of California San Diego, La Jolla, CA 92093, USA; 24) Department of Pathology, University of California San Diego, La Jolla, CA 92093, USA; 25) Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093,USA; 26) Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard, Cambridge, MA 02142, USA; 27) Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; 28) Department of Medicine, Harvard Medical School, Boston, MA 02115, USA; 29) Department of Clinical Genetics, Erasmus University Medical Center, 3015 CN, Rotterdam, The Netherlands; 30) HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA; 31) Institute of Human Genetics, Helmholtz Zentrum Muenchen, GmbH, Neuherberg, Germany; 32) Universitätsklinikum Düsseldorf, Düsseldorf, Germany; 33) Institute of Human Genetics, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; 34) Institut für Medizinische Genetik, Medizinische Universität Wien, Vienna, Austria; 35) Department of Biochemistry and Molecular Biology, College of Medicine, University of South Alabama, Mobile, AL 36688, USA.

The overall understanding of the molecular etiologies of intellectual disability (ID) and developmental delay (DD) is increasing with next-generation sequencing technologies identifying genetic variants in individuals with such disorders. However, detailed analyses conclusively confirming these variants, as well as the underlying molecular mechanisms explaining diseases, are often lacking. Here we report on an ID syndrome caused by *de novo* heterozygous loss-of-function (LoF) mutations in *SON*. Through international collaboration we collected 19 patients with LoF mutations in *SON*. Deep-phenotyping of the patients revealed common features of ID/DD, malformations of the cerebral cortex, epilepsy, vision problems, musculoskeletal abnormalities and congenital malformations. Population genetic signatures for *SON* indicate that *SON* belongs to the 2% most intolerant human protein-coding genes and that *SON* is depleted from LoF mutations in large databases such as ExAC, suggestive for such mutations being under strong purifying selection. To further examine the effect of *SON* haploinsufficiency on embryonic development, we injected zebrafish embryos with *son* morpholinos, which resulted in a host of developmental defects. Embryos surviving 72 hours post injection progressed to more severe phenotypes, with extreme spinal malformations (22%), head and eye malformations with edema of the brain (37%), and profound developmental abnormalities (10%), mimicking features observed in affected individuals. SON is a nuclear speckle protein with dual abilities to bind to DNA and RNA, and its cellular functions includes the regulation of RNA splicing and gene transcription. Hallmark features of *SON* knockdown in HeLa cells and human embryonic stem cells are intron retention and exon skipping. Importantly, our analyses of RNA from patients with *SON* LoF mutations revealed that genes critical for neuronal migration/cortex organization (*TUBG1, FLNA, PNKP, WDR62, PSMD3,* and *HDAC6*) and metabolism (*PCK2, PFKL, IDH2, ACY1* and *ADA*) are significantly downregulated due to accumulation of mis-spliced transcripts resulting from erroneous SON-mediated RNA splicing. In summary, we identified *de novo* LoF mutations in *SON* as a cause of a novel complex neurodevelopmental disorder characterized by ID/DD and severe brain malformations. Moreover, our data highlight SON as a master regulator governing neurodevelopment, and demonstrate the importance of SON-mediated RNA splicing in human development.

232

Using predictive models to expand the locus heterogeneity of tRNA synthetase-related inherited disease. *R. Meyer[1], S. Oprescu[1], A. Beg[2], A. Antonellis[1].* 1) Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109; 2) Department of Pharmacology, University of Michigan, Ann Arbor, MI 48109.

Mutations in genes encoding aminoacyl-tRNA synthetases (ARSs) have been implicated in autosomal dominant, axonal peripheral neuropathy [*i.e.*, Charcot-Marie-Tooth (CMT) disease], which is a heterogeneous class of neurodegenerative diseases characterized by impaired motor and sensory function in the distal extremities. Thus far, mutations in glycyl-(*GARS),* tyrosyl-(*YARS),* alanyl-*(AARS)*, and histidyl-tRNA synthetase (*HARS)* have been implicated in dominant CMT disease. The identified mutations show impaired enzyme function in biochemical and yeast complementation assays and dominant toxicity to axons when over-expressed in *C. elegans*. The above models have considerable predictive power when applied to ARS mutations; data from these functional analyses have been successfully employed to confirm the pathogenic properties of newly identified mutations in patient populations. We now use these models to determine if mutations in any ARS-encoding gene can cause peripheral neuropathy, which would support a role for impaired tRNA charging in disease onset and provide a common framework for studying ARS-related disease. To test this approach we studied threonyl-tRNA synthetase (*TARS)*, which has yet to be implicated in any inherited disease phenotype. Like all four ARS enzymes implicated in CMT disease, TARS acts as a homodimer and charges tRNA in the cytoplasm, making it a strong candidate for a CMT-causing ARS. We performed a forward screen in the yeast ortholog of *TARS* (*THS1*) by introducing 10 missense mutations at highly conserved residues. Three of these mutations (N419Y, R440H, and G548R) show reduced yeast growth, indicating impaired protein function. We then modeled these mutations in the *C. elegans* ortholog of *TARS* (*tars-1*) and showed that G548R *tars-1* is dominantly toxic to axons. These data show that G548R *TARS* is similar to R329H *AARS*, which is a well-characterized, recurrent mutation identified in multiple large pedigrees with CMT disease. Here, we present: (1) our unpublished data indicating that *TARS* mutations should be considered excellent candidates for axonal CMT disease; (2) our unpublished work to develop a humanized yeast assay to more broadly study human *TARS* variants; and (3) our working model for how loss-of-function missense mutations in ubiquitously expressed, essential genes can cause dominant axonal peripheral neuropathy.

233

Establishing a phenotypic model of *MBD5-***Associated Neurodevelopmental Disorder (MAND) by utilization of patient-derived neural stem cells and RNA-seq.** *S.V. Mullegama[1], J.T. Alaimo[2], S.R. Williams[3], L. Chen[4], J.W. Innis[5], F.J. Probst[2], C. Haldeman-Englert[6], T. Ezashi[7], S.H. Elsea[2].* 1) UCLA Clinical Laboratories , UCLA , Los Angeles, CA; 2) Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX, 77030, USA; 3) Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22902, USA; 4) Department of Cellular and Genetic Medicine, School of Ba sic Medical Sciences, Fudan University, Shanghai 200032, China; 5) Departments of Human Genetics and Pediatrics and Communica ble Diseases, University of Michigan, Ann Arbor, MI 48109, USA; 6) Department of Pediatrics, Section on Medical Genetics, Wake Forest University Health Sciences, Winston-Salem, NC, 27157, USA; 7) Division of Animal Sciences, University of Missouri, Columbia, Missouri 65211, USA.

MBD5-Associated Neurodevelopmental Disorder (MAND) is an umbrella term that describes 2q23.1 deletion syndrome, 2q23.1 duplication syndrome, and MBD5 variants, in which altered dosage of *Methyl-CpG-binding domain protein 5* gene (*MBD5*) is responsible for the myriad of clinical features that includes intellectual disability, autism, seizures, speech impairment, ataxia, sleep and behavioral problems and microcephaly. The pathological mechanisms and pathways underlying the MAND phenotype are not well delineated. To understand the MAND phenotype through identifying altered genes and pathways, we utilized stem cell technology and RNA-sequencing (RNA-seq). We present three patients with similar *MBD5* specific deletions that were identified by chromosomal microarray (CMA) and have identical phenotypes. Fibroblasts were obtained from the patients and controls to derive induced pluripotent stem cells (iPSCs), which were then differentiated to neural stem cells (NSCs). Analysis of the iPSCs and NSCs by CMA and *MBD5* mRNA expression testing revealed successful modeling of MAND. To gain insight into the pathways and genes altered that underlie the MAND pathology, we explored the transcriptional networks that MBD5 regulates in the NSCs by integrating RNA-seq. Overall, 300 transcripts were differentially expressed as a consequence of reduced *MBD5* dosage ($q<0.05$). We hypothesized that these transcripts were responsible for the phenotypic manifestations of our patients. Many of these dysregulated genes are essential in cellular processes including transcriptional regulation (*NFIA*), chromatin modification (*CHD8, MEOX1*), and neuronal function (*NTNG1*, *LAMC3*). We also identified altered genes that are involved in phenotypically similar syndromes such as 5q14.3 deletion syndrome (*MEF2C*) and congenital variant of Rett syndrome (*FOXG1*). Through a pipeline of pathway programs, we were able to identify strong statistical enrichment for pathways such as PTEN signaling, Wnt/Notch signaling, and circadian rhythm that allowed us to explain the phenotype of MAND (ID and sleep). Finally, we queried the genes that were not linked to OMIM disorders through the utilization of DECIPHER, ExAC, and clinical CMA databases to identify novel dosage sensitive genes that have an important role in the neuronal development phenotype. Overall, our data clearly demonstrates the utilization of NSCs and RNA-seq in advancing the understanding of the phenotype of a complex neurodevelopmental disorder.

234

The contribution of *de novo* mutations in enhancers and ultra-conserved non-coding elements to severe developmental disorders in 8,000 children.

P.J. Short, S. Gerety, J. McRae, E. Coomber, J.C. Barrett, M.E. Hurles on behalf of the Deciphering Developmental Disorders study. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

The Deciphering Developmental Disorders (DDD) study has collected detailed clinical phenotypes and exome sequence data from from 13,500 children with undiagnosed developmental disorders and their parents to understand the genetic architecture of these disorders. Over half of the 4,294 children analysed thus far have no likely disease-causing variants within protein-coding genes.We predict that some de novo mutations and rare variants in non-coding elements may contribute to developmental disorders through their effect on the timing and duration of gene expression in critical developmental pathways. To test this hypothesis, we sequenced the most highly conserved 4,500 non-coding elements and 800 experimentally-validated enhancers (from VISTA) in all families. Sequence data in over 14,000 unaffected parents was used to quantify the strength of purifying selection acting across different coding and non-coding elements and within DNase hypersensitivity sites and transcription factor binding sites. We assessed a number of methods for predicting coding and non-coding functional effects including VEP, SIFT, PolyPhen, CADD, and fathmm-MKL using concordance between predicted variant deleteriousness and strength of purifying selection (quantified by the mutability adjusted proportion of singletons metric). In both non-coding and coding regions, CADD outperforms all other metrics. Preliminary analysis on 4,294 probands has identified over 400 de novo mutations in putative enhancers and ultra-conserved elements. De novo SNVs are enriched in enhancers and ultra-conserved elements compared to control non-coding elements and a null model based on tri-nucleotide mutation rates. Using DNase hypersensitivity data from 111 different tissues, we find further enrichment in non-coding elements predicted to be active in fetal brain and other fetal tissues. The work presented here will include an updated analysis with more than 8,000 cases. This findings allow us to estimate what fraction of cases have a pathogenic de novo mutation in one of these non-coding elements and set an upper bound on the proportion of mutations in these elements that are likely to be pathogenic. These analyses have generated hypotheses relating to specific variants in individual non-coding elements that have been engineered using the CRISPR/Cas9 system into mouse and zebrafish models, and preliminary results on the phenotypic consequences of these alleles will also be presented.

235

Unveiling microcephaly mechanisms associated with DNA repair disorders using induced pluripotent stem cells and cerebral organoids.

F. Pirozzi[1], J. Ngo[1], T.H. Kim[1], K. Plona[1], Y. Chen[1], E. Gilmore[2], A. Wynshaw-Boris[1]. 1) Department of Genetics and Genome Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH 44106; 2) Department of Pediatrics Division of Neurology, University Hospitals Case Medical Center, Cleveland, OH 44106.

Genome stability is crucial for proper brain development and growth; in fact, loss of function of DNA repair genes such as Ligase 4 (LIG4) and Polynucleotide kinase 3'-phosphatase (PNKP) lead to severe microcephaly with intellectual disability and increased cancer risk. However, the pathogenesis of DNA repair deficiency-related microcephaly is poorly understood. Our hypothesis is that mutations in DNA repair genes cause microcephaly due to lack of an ideal number of Neuronal Progenitor Cells (NPCs) or neurons during the early stages of brain development. This effect could be due to three non-exclusive mechanisms: (1) reduced generation/proliferation of NPCs, (2) premature/abnormal differentiation of NPCs into neurons, or (3) increased apoptosis in one or both populations. To test our hypothesis, we generated induced Pluripotent Stem Cells (iPSCs) from healthy controls as well as microcephalic patients with known mutations in LIG4 or PNKP. We differentiated these iPSCs into NPCs, cortical neurons, and 3D cerebral organoids, which allowed us to study proliferation, differentiation, apoptosis and early self-organizing neuronal structures. Our results support our hypothesis that causative mechanisms of DNA-repair-related microcephaly arise during early stages of brain development. Specifically, we found that NPCs derived from LIG4-iPSCs do not have different proliferation rates compared to controls, but rather they display increased apoptosis. In addition, when directly induced into neurons, LIG4-iPSCs differentiated more rapidly than controls. However, 2 weeks after neuronal induction, the LIG4 neurons displayed increased cell death. Finally, LIG4 cerebral organoids were at least 2 times smaller than control organoids. Immunostaining of LIG4 organoid sections showed increase in cleaved Caspase3 in the outer edge of the sub-ventricular-like zone, further supporting a role for apoptosis in microcephaly. Organoids generated from PNKP mutant iPSCs displayed a more severe phenotype, with reduced size (up to 5 times smaller than controls) followed by early and complete degeneration. We used CRISPR/Cas9 techniques to generate isogenic cell lines to rule out effects due to the genomic background. In summary, we were able to recapitulate human microcephaly caused by mutations in DNA repair genes *in vitro*. Our models suggest that premature differentiation of NPCs followed by apoptosis of specific neuronal populations might play a significant role in microcephaly.

236

Early biomarkers and mechanisms of retinal degeneration identified in a mouse model of human mitochondrial dysfunction: The *harlequin* apoptosis-inducing factor hypomorph. *K.A. Hill[1], T.C. MacPherson[1], A.M. Laliberte[1], A. Prtenjaca[1], E.A. Dolinar[1], J. Mayers[1], T. Privorozky[1], Y. Balboul[1], A. Bentley-De Sousa[1], A. Li[1], M. Edwards[1], I. Kisilevsky[1], S. Rajkarnikar[1], J.R.J. Thompson[1], B. Rubin[1], M.A. Bernards[1], C.M.L. Hutnik[2].* 1) Dept Biol,Biol & Geol Sci Bldg, Univ Western Ontario, London, ON, Canada; 2) Ivey Eye Research Institute, London, ON, Canada.

Retinal degenerative disorders (RDDs) are among the most common ophthalmological impairments causing blindness in the world today, affecting 13 million people and costing nearly $8.7 billion annually in the U.S. alone. A growing body of evidence suggests a central role for mitochondrial dysfunction as an underlying etiology of numerous human RDDs. The *harlequin* (X*hq*Y; *hq*) mouse is a genetic model of human retinal and cerebellar degeneration providing a framework to identify early diagnostic biomarkers and disease mechanisms of neurodegeneration associated with mitochondrial dysfunction. This mitochondrial dysfunction results from a spontaneous proviral insertion in intron 1 of the *Programmed cell death 8 (Pdcd8)* gene, causing downregulation of the corresponding apoptosis-inducing factor protein and resulting in compromised complex 1 function. Early indication of metabolic deficiency in *hq* mice manifests as elevated food consumption with lower body mass compared to wild-type (WT) mice. LC-MS of primary metabolites confirmed decreased oxidative phosphorylation and increased glycolysis and glycerol phosphate shuttling in the *hq* retina as early as 2 months of age (moa). Microglial activation was observed in *hq* retinas *in situ* at 2 moa and retinal transcriptome changes at 4 moa are consistent with a para-inflammatory response. Microglial phagocytosis of ATP-starved photoreceptors is evident by 4 moa. Monthly electroretinography (ERG) of *hq* and WT retinal function from 2 to 10 moa revealed reduced b-wave amplitude and supernormal a-wave amplitude in *hq* mice at 2 moa. At 3 moa, *hq* a-wave amplitude is subnormal and *hq* a- and b-wave amplitude reduction progresses with age. The frequency of the *hq* ERG oscillatory potentials (OP) decreases at 2 moa while OP latencies increase at 3 moa. The 10-90% rise time of the *hq* ERG a-wave, i.e., the photoreceptor response time, is subnormal only at 10 moa. Postmortem histological analysis of *hq* retinas revealed retinal thinning with photoreceptor losses by 4 moa. Collectively, our results indicate functional deficits precede structural losses. The earliest functional deficiencies are detectable in the inner retina rather than the photoreceptors. There is a previously unrecognized therapeutic window prior to retinal neuron loss and an even larger window prior to cerebellar neurodegeneration (7 moa). ERG is a sensitive tool for assessment of retinal health directly and as a proxy for cerebellar neurodegeneration. .

237

Issues important to families considering exome sequencing. *S. Adam[1], P.H. Birch[1], R.R. Coe[1], N. Bansback[2], M.B. Connolly[3], E. Toyota[3], M.J. Farrer[4], M.K. Demos[3], J.M. Friedman[1].* 1) Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada; 2) School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada; 3) Division of Neurology, Department of Pediatrics, University of British Columbia, Vancouver, BC, Canada; 4) Centre for Applied Neurogenetics, Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, BC, Canada.

Purpose: Our goal was to determine which issues parents believe are important to consider when choosing whether or not to have whole exome sequencing (WES) for their child. **Method:** As part of a study of WES for children with early-onset epilepsy of unknown cause, parents used DECIDE, an interactive online decision support tool that presents a list of issues which may influence WES decisions for some families. These issues include the desire for knowledge/to get an answer, wanting information for reproductive planning, and worry about insurability of the child and family members. Parents were asked to select at least 4 issues that were of personal value and then to weigh the relative importance of their chosen issues using a pie chart. **Results:** One parent from each of 109 families was asked to participate. 84% of were female; 79% had at least a college diploma. Parents were positive about the process of selecting and weighing issues, many stating that it helped to clarify their values and define why they did or did not want WES. The issues selected most frequently were knowledge (chosen by 92% of families), guiding management (85%), advancing science/helping others (72%), and obtaining resources (68%). The least frequently selected issues were easing parental guilt (24%) and enabling grieving/acceptance (31%). The issues ranked as most important were management (in 63% of families), knowledge (62%), resources (36%), and advancing science (30%). Reproductive issues were selected by 37%, but 72% of participants stated that their families were complete. **Conclusion:** While the issues of parental guilt and grieving/acceptance are often mentioned in the literature as important genetic counselling topics surrounding testing and diagnosis, these were not prominent in our group. Unexpectedly, altruism (advancing science) was valued by most of our families and was the most important issue for 1/3 of families. Families undergoing WES found DECIDE helpful for considering the various issues and implications around WES and for facilitating values clarification. Genetic counsellors may find DECIDE useful for identifying patients' values and stimulating discussion. DECIDE is described here: doi:10.1007/s10897-016-9971-8; and can be viewed here: http://bit.ly/DECIDE-GWS.

238

Parental expectations and attitudes towards receiving genomic results in healthy children. *J.L. Williams, F.D. Davis, A.L. Fan, L. Bailey, K. Fultz, M.S. Williams, M.F. Murray, A.K. Rahm.* Genomic Medicine Institute, Geisinger Health System, Danville, PA.

Background: The Geisinger Health System has begun to return actionable genomic results from 76 genes (ACMG 56 + 20) to adult participants in the MyCode® biobank. We sought to explore parental attitudes concerning the return of genomic results in healthy children associated with conditions that are medically actionable in childhood vs. adult onset. **Methods**: Four focus groups were conducted with adult MyCode® participants who had at least one child enrolled in the biorepository. Participants were selected on child age (0-8 or 9-17) and geographic location. A deliberative engagement format was used and included education about the American College of Medical Genetics (ACMG) and American Academy of Pediatrics recommendations and the related risks, benefits, and ethical principles. Parents were presented two scenarios for discussion: 1) their healthy child has a pathogenic change for Marfan Syndrome that is medically actionable in childhood; 2) their healthy child has an adult-onset condition, Lynch syndrome. Thematic analysis was conducted on verbatim transcripts of each group. **Results**: All parents identified the desire to contribute to research as a main reason for consenting to the biobank participation of their children. Regardless of scenario, parents stated that the genetic information was important for their child's future health, the information was similar to any other unexpected medical information, results should be placed in their child's EMR, parents should first be told the result without the child present, and disclosure to the child should be tailored over time based on age and need. Thematic differences between scenarios revealed that parents considered the importance of Lynch results to their children's future health outweighed preservation of decision making autonomy at the age of majority. Participants suggested additional counseling about the adult-onset conditions to help prepare parents for such results. **Conclusions**: The majority of parents endorsed return of results irrespective of whether or not there is actionability in childhood. This contrasts with professional society recommendations that favor preservation of autonomy for conditions with actionability limited to adults. Review of DNA Day essays from adolescents about return of adult onset conditions shows that many prefer deferral of return until adulthood. Efforts to reconcile these three perspectives are needed.

239

Healthcare outcomes and costs after genome sequencing among healthy adults: Results of a randomized controlled trial. *J.L. Vassy[1,2], K.D. Christensen[1], E.F. Schonman[1], D. Dukhovny[3], P.M. Diamond[4,5], C.L. Blout[1], J. Oliver Robinson[4], J.B. Krier[1], M.F. Murray[6], A.L. McGuire[4], R.C. Green[1,7,8] for the MedSeq Project.* 1) Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 2) VA Boston Healthcare System, Boston, MA; 3) Oregon Health & Science University, Portland, OR; 4) Baylor College of Medicine, Houston, TX; 5) UTHealth School of Public Health, Houston, TX; 6) Geisinger Health System, Danville, PA; 7) Broad Institute of MIT and Harvard, Cambridge, MA; 8) Partners Healthcare Personalized Medicine, Cambridge, MA.

Background: Genome sequencing (GS) in healthy adults is controversial. It might enable disease prevention for patients and their families. However, it could trigger a cascade of costly medical care that does not improve health outcomes and might cause harm. We aimed to quantify the impact of GS on the immediate and 6-month healthcare and costs in healthy adults. **Methods:** In the MedSeq Project, 9 primary care physicians (PCPs) were enrolled to participate with their patients (n=100) in a randomized trial of GS in primary care. Eligible patients were 40-65 years old and deemed generally healthy by their PCPs. Patients were randomized to receive a family history report only (FH arm) or a FH report plus an interpreted GS report (GS arm) that included potentially clinically relevant variants in 4600 genes associated with dominant and recessive monogenic disease, polygenic risk estimates for 8 cardiometabolic traits, and 5 pharmacogenetic markers. After discussing the reports with the patient, each PCP completed a survey about clinical actions taken based on the patient's results (immediate care). Six months after each disclosure visit, healthcare was abstracted from electronic health record (EHR) data (6-month care). Immediate and 6-month healthcare costs were determined from billing codes and Medicare price weights. We compared outcomes in the 2 arms with non-parametric tests and regression modeling. **Results:** Immediately post-disclosure, PCPs were more likely to take clinical actions for GS (34%) vs. FH patients (16%, p=0.04). For GS patients, PCPs ordered more cardiac tests (7 vs. 0, *p*=0.01), lab tests (12 vs. 5, *p*= 0.28), imaging studies (3 vs. 0, *p*=0.08), and referrals (7 vs. 6, *p*=0.97). Mean immediate costs were $77 (GS) and $67 (FH; *p*=0.046). In 6-month care, compared to the FH arm, the GS arm underwent a greater number of cardiology tests (20 vs. 7, *p*=0.04), lab tests (196 vs. 143, *p*=0.33), imaging tests (57 vs. 44, *p*=0.35), and specialty visits (123 vs. 106, *p*=0.67), but not PCP visits (35 vs. 37, *p*>0.99). Six-month costs were higher in the GS arm (median $616 vs. $456; mean $1324 vs. $1115) but not significantly so (*p*=0.56). Six-month EHR review found supporting phenotypic evidence in only 4 of 13 patients with pathogenic variants. **Conclusion:** Adding GS to the care of healthy adults may increase healthcare utilization and costs, but the per-person increase may be modest. Clinical outcomes are needed to determine the value of this additional healthcare.

240

Beyond diagnostic yield – An economic perspective on the impact of whole exome sequencing (WES) in medical decision-making. *T. Vrijenhoek[1,2], E.M. Middelburg[1,2,3], G.R. Monroe[1,2], K.L.I. van Gassen[1], A.M. Hövels[3], N.V. Knoers[1,2], G.W. Frederix[1,2,4].* 1) Department of Genetics, Utrecht University Medical Center, Utrecht, The Netherlands; 2) Center for Molecular Medicine, Utrecht University Medical Center, Utrecht, The Netherlands; 3) Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands; 4) Department of Health Technology Assessment , Julius Center, University Medical Center Utrecht, Utrecht, the Netherlands.

A comprehensive estimate of the utility of clinical whole-exome sequencing (WES) entails more than calculating diagnostic yields; the consequences of a diagnostic result should also be considered. We estimated the impact of a WES-based diagnosis on other types of healthcare activities and associated costs. We performed retrospective cost analyses on the per-patient healthcare activities of 371 patients with suspected genetic disorders, who were referred for WES at the end of their diagnostic odyssey. The average daily per-patient costs before and after diagnosis were $33.28 and $18.97 for patients with a positive diagnosis (35%), $27.02 and $24.56 for those with a negative diagnosis (24%) and $29.35 and $21.60 for those with an uncertain diagnosis (41%). We thus confirm the high diagnostic yield for WES-based diagnostics. Moreover, we show that healthcare costs are substantially reduced for patients with a positive diagnosis ($p = 0.07$) compared to those with a negative or an uncertain diagnosis Subsequently, we assessed the potential for WES to impact healthcare costs, by evaluating the scenario in which WES would be applied earlier in the diagnostic trajectory (‘WES-first’). Therefore we considered the effect of replacing all genetic testing activities with WES. Average costs would be reduced in all patients, with the highest reduction in patients receiving a positive diagnosis ($2,506) followed by those with an uncertain diagnosis (VUS; $2,428) and a negative diagnosis ($2,203). Whereas replacing all other diagnostic tools with WES or WGS is merely a theoretical option at this point in time, our data indicate that both patients and society may benefit from early implementation of these technologies. We estimate that cost-effective replacement (i.e. no increase of costs) by whole-genome sequencing (WGS) in such ‘genetics-first’ scenario would be possible, but requires a substantial decrease of sequencing prices. Our study confirms WES to be a cost-effective replacement of traditional diagnostic technologies for patients with intellectual disability. Moreover, WES has a considerable impact on the medical decision-making process; having a positive genetic diagnosis reduces subsequent healthcare costs, while patients with a negative or uncertain diagnosis likely continue their diagnostic odyssey. With our study we open up the debate on efficient use of constrained resources in an increasingly genetics-oriented healthcare system.

241

Assessing the acceptability of genomic services for community health centers that provide care for under-served populations. *K.A.B. Goddard[1], J.V. Davis[1], L. Jacob[2], C. McMullen[1], J.L. Holup[1], P. Foley[2], E.K. Cottrell[2], J.L. Schneider[1], B. Wilfond[3].* 1) Center for Health Research, Kaiser Permanente Northwest, 3800 North Interstate Avenue, Portland, OR 97227; 2) OCHIN, 1881 SW Naito Pkwy, Portland, OR 97201; 3) Treuman Katz Center for Pediatric Bioethics, Seattle Children’s Hospital, 1900 Ninth Avenue, Seattle, WA 98101.

Access to health care is inequitably distributed and contributes to poorer health outcomes for racial/ethnic minorities, individuals with lower socio-economic status, and the underinsured. The introduction of exome and genome sequencing may exacerbate existing health disparities by concentrating use of these newer technologies among wealthier, better-insured, and more medically informed patients. Minority populations may also hold different attitudes toward genetic services including concerns about misuse of results and different expectations of health benefits and how results should be managed. Other factors include limited health literacy and mistrust of the medical system. Within care settings that serve diverse populations, we aimed to identify patient and provider priorities for genomic services, the perceived utility of these services, and implementation barriers and facilitators. In 2015, the OCHIN Practice-based Research Network served a population with 15% African American and 24% Hispanic adult patients, and 45% of adult patients had incomes below the federal poverty level. Thirty-one % of adult patients were uninsured and 48% were insured through Medicaid/Medicare. We conducted interviews with 14 OCHIN providers, and had group and individual interviews with 14 OCHIN patients (urban, rural, and Spanish-speaking). We developed three scenarios to facilitate discussions: hereditary cancer; carrier screening; and developmental delay. Genomic services were of relatively low priority for providers, but they were open to providing these services if patients wanted them. Genetic services are accessed mostly through outside referrals to specialists. Actionability, cost, and access to downstream care were key concerns that would limit expanded use of genetic services. Providers questioned the relevance of genetic services for patients and worried about anxiety that could result from expanded use. Providers identified patient health literacy as a potentially significant barrier for adoption and successful delivery of genetic services. Patients expressed that access to genetic services was important, even if the particular service focus, e.g., developmental delay, was not relevant to them. Patient preferences were mixed for receiving genetic services by telehealth and for results disclosure by primary care providers vs. genetic counselors. This work will inform the direction of future research to broaden the evidence base for emerging genetic services.

242

Enhancing diversity in genomic medicine research: Factors influencing enrollment and retention. *J. Berg*[1], *E. Moore*[2], *E. Corty*[1], *M. Roche*[3], *Z. Girnary*[1], *B. Ania*[2], *J. O'Daniel*[1], *F-C. Lin*[2], *C. Rini*[2], *J. Evans*[1], *G. Henderson*[4]. 1) Genetics, UNC Chapel Hill, Chapel Hill, NC; 2) School of Public Health, UNC Chapel Hill, Chapel Hill, NC; 3) Pediatrics, UNC Chapel Hill, Chapel Hill, NC; 4) Social Medicine, UNC Chapel Hill, Chapel Hill, NC.

Inclusion of patients from diverse backgrounds in genomic medicine research is critical to enhance understanding of genomic data and to identify best practices for implementation. However, disparities persist despite efforts to increase diversity. It is likely that a range of factors influence participation, yet few studies have systematically assessed disparities in enrollment and retention of genomic research participants. The North Carolina Clinical Genomic Evaluation by Next-generation Exome Sequencing (NCGENES) study, which included 416 adult and 229 child participants, employed specific strategies to enhance enrollment of under-represented minority candidates. Spanish-speaking team members and Spanish language materials facilitated interactions with Hispanic patients, and a satellite clinic was established in a community cardiomyopathy practice with a high proportion of African American patients. We systematically analyzed the cascade of recruitment, enrollment, and retention in the study. 1228 patients were nominated as potential candidates, in categories proportional to the demographics of North Carolina. Of these candidates, 1063 were approached for study participation, 645 enrolled in the study, and 538 were retained to complete psychosocial survey assessments. Sociodemographic characteristics and related variables were analyzed as potential factors associated with enrollment and attrition throughout the study. While successful in enrolling 30% of participants from under-represented groups, the study still experienced disproportionate enrollment and retention. Multivariate analyses indicated that minorities and adults who lived farther away from the clinic were significantly less likely to enroll. Adults who were African American and had less education, as well as parents whose children were Hispanic, were significantly less likely to be retained. Attrition occurred across all time points, and different factors were associated with enrollment compared to retention. These results suggest that multiple approaches will be needed to increase diversity in genomic medicine research. Broad inclusion in such research is necessary to ensure that results are generalizable across society; systematic loss of the most vulnerable members of society from research will exacerbate disparities in translation of genomic sequencing to medical care. Our findings highlight contributing factors and will guide efforts to address these problems.

243

Aneuploidy and cancer predisposition caused by mutations in genes controlling chromosome segregation. *N. Rahman*[1], *S. Yost*[1], *B. de Wolf*[2], *S. Hanks*[1], *M. Clarke*[1], *A. Zachariou*[1], *E. Ramsay*[1], *H. Wylie*[1], *S. Seal*[1], *E. Ruark*[1], *M.U. Rashid*[3], *G.J.P. Kops*[2]. 1) Institute of Cancer Research, London, London, United Kingdom; 2) Hubrecht Institute KNAW (Royal Netherlands Academy of Arts and Sciences), Uppsalalaan 8, 3584 CT Utrecht, The Netherlands; 3) Shaukat Khanum Memorial Cancer Hospital & Research Centre, Lahore, Punjab 54000, Pakistan.

Accurate chromosome segregation during cell division is essential to maintain the correct number of chromosomes in cells. Errors of chromosome segregation can lead to aneuploidy, which is an important cause of human disease, implicated in recurrent miscarriage, infertility, developmental disorders and cancer. Many biological processes, including spindle assembly, chromatid-spindle attachment, attachment error-correction, and the spindle assembly checkpoint (SAC) are involved in ensuring chromosome segregation proceeds flawlessly and that aneuploidy is prevented. Rare individuals with constitutional mosaic aneuploidies involving varying chromosomes are well documented. We have been studying this condition, which is sometimes referred to as mosaic variegated aneuploidy (MVA), for the last decade. We previously reported biallelic mutations in the spindle assembly checkpoint gene *BUB1B* as a major cause of MVA. More recently we identified biallelic mutations in *CEP57*, which encodes a centrosomal protein involved in kinetochore attachment, in individuals with MVA. Together these two genes only account for a proportion of MVA cases. Through exome sequencing and functional studies of 43 individuals from 20 families with MVA we now report mutations in *TRIP13*, *KNL1*, *ZWINT* and *CCDC84* as causes of MVA syndrome. *BUB1B* and *TRIP13* mutations result in substantial impairment of the spindle assembly checkpoint and are associated with a very high cancer risk, particularly for Wilms tumor. We also show that *TRIP13* mutations can cause Wilms tumor in the absence of constitutional aneuploidy and a founder *TRIP13* mutation is a major cause of Wilms tumor in Pakistan. By contrast the other genes are primarily implicated in kinetochore function and have not, to date, resulted in cancer. The stratification of cancer risk in MVA individuals by genotype is very important for clinical management. MVA cases due to *BUB1B* and *KNL1* are targeted by one truncating mutation and one hypomorphic mutation; biallelic truncating mutations appear to be embryonic lethal in these genes. By contrast, *TRIP13*, *ZWINT* and *CCDC84* patients have biallelic loss-of-function mutations. These data identify new aneuploidy predisposition genes and uncover complex relationships between mutational spectra, clinical phenotypes and functional mechanisms regulating chromosome segregation.

244

Discovery and dissection of regulatory elements of the Mendelian disease gene *HPRT1* using programmed CRISPR/Cas9 guide pairs for multiplexed deletion scanning. *M. Gasperini[1], G. Findlay[1], A. McKenna[1], C. Lee[1], J. Milbank[1], J. Shendure[1,2].* 1) Dept of Genome Sciences, University of Washington, Seattle, WA; 2) Howard Hughes Medical Institute, Seattle, WA.

Loss-of-function mutations to *HPRT1*, the gene encoding for hypoxanthine-guanine phosphoribosyltransferase, result in Lesch-Nyhan syndrome, a severe, early-onset disease marked by motor dysfunction, cognitive and behavioral disorders, and hyperuricemia. Although hundreds of causal mutations in *HPRT1* coding sequence have been identified, a minority of patients present with reduced HPRT1 activity despite the absence of coding mutations. Many of these patients may carry unidentified noncoding mutations that result in reduced *HPRT1* expression, but finding noncoding elements critical for gene function remains challenging. CRISPR/Cas9 genome editing has previously been used to scan noncoding regions for functional elements. Current approaches use the short insertions and deletions (indels) that result from non-homologous end joining to disrupt regulatory elements. Consequently, these screens are fundamentally limited by low and incomplete coverage. We sought to overcome this weakness by leveraging pairs of single-guide RNAs (sgRNAs) to generate large deletions of intervening DNA between the target sites of each pair of sgRNA. To identify *cis*-regulatory sequences that are essential for *HPRT1* function, we induced thousands of highly overlapping 1- and 2-kb deletions across the full 40 kb gene and its surrounding 160 kb. 11,365 programmed sgRNA pairs were synthesized on a microarray and lentivirally delivered to pools of cells via low multiplicity of infection. After editing, we selected for sgRNA pairs that caused loss of *HPRT1* function by dosing cells with 6-thioguanine. This approach enabled comparison of full *HPRT1* exon dropout to deletions in the promoter, introns, UTRs, and uncharacterized nearby intergenic sequence. We defined the length of *HPRT1*'s putative promoter, and also observed putatively critical non-coding sequences in the first 2 kb of intron 1, the middle of intron 3, and 15 kb downstream. Our ongoing experiments aim to shed additional light on how noncoding mutations can lead to Lesch-Nyhan syndrome by further dissecting these newly implicated candidate regulatory sequences. Our unpublished method for dense, redundant, and large-scale CRISPR/Cas9 deletion scanning facilitated a comprehensive search for noncoding sequence essential to *HPRT1* function. More broadly, we envision that this method will be valuable for functional characterization of large regions of the genome that are inaccessible to robust perturbation using individual sgRNAs.

245

Discovery and challenges in low-pass whole genome sequencing of over 140,000 individuals from throughout China. *X. Jin[1], on behalf of the Chinese Millionome Consortium[1,2,3,4].* 1) BGI, Shenzhen, 518083, China; 2) Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; 3) Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark; 4) Department of Integrative Biology, University of California, Berkeley, Berkeley, California 97420, USA.

The past decade has witnessed the launch of several ultra-large-scale whole genome sequencing (WGS) studies, though these vary greatly with respect to cohort composition, sequencing strategy and sample size. A major challenge in scaling such studies is the recruitment and acquisition of samples and phenotypes from hundreds-of-thousands to millions of individuals. At the same time, non-invasive prenatal testing (NIPT) for fetal trisomies – by sequencing of maternal plasma cell-free DNA (cfDNA) - has become the fastest adopted molecular test in history. Here we show that the low-pass WGS data generated for NIPT testing can be repurposed for human genetics, including for population genetics and for genome-wide association studies. As a proof-of-concept, we analyzed WGS data from >140,000 individuals throughout China, generated for clinical NIPT. Despite the low coverage per individual (~0.1x), we are able to identify population-specific rare variants, to estimate allele frequencies down to ~0.2%, and to discover multi-allelic polymorphisms. As we have high-resolution geographic information on each individual, we are able to explore fine-scale patterns of population differentiation across the country's 30 provinces. We show that the allele frequencies of a few disease or drug response-related variations contradicts existing knowledge, confirming the importance of population-specific information for accurate variant interpretation. In mappable regions of the genome, we achieve high imputation accuracy, which in turn suggests the potential of this data to be used for genome-wide association studies (GWAS). Preliminary GWAS on phenotypes including height, weight, fertility, and maternal age confirm known associations and reveal novel signals. We also explored additional uses for NIPT data, empowered by the scale of this study. For example, we find that individuals who developed cancer during pregnancy display copy number variation in cfDNA consistent with aneuploidy. By analysis of the non-human sequences in cfDNA, we demonstrate the detection, geographical distribution and subtyping of several pathogenic viruses in China. Overall, these data demonstrate the power of repurposing NIPT data, which has already been generated by us on >1 million individuals, for human genetics. Our aggregate analysis of low-pass WGS of >140,000 individuals is a pilot for the Chinese Million-ome Project, which will include genomic analysis of at least one million Chinese individuals.

246

Functional studies using *Drosophila melanogaster* reveal pervasive epistasis and common mechanisms for copy-number variant genes. *S. Girirajan, L. Pizzo, M. Jensen, E. Huber, P. Patel, K. Vadodaria, A. Kubina, S. Yennawar, J. Iyer, M.D. Singh.* Biochem & Molecular Biol, Pennsylvania State Univ, University Park, PA.

Recent high-throughput genomics studies have identified several candidate genomic regions and genes for complex human disorders. Systematic functional analysis of each one of these candidate genes is limited due to a lack of high throughput and quantitative assays. We developed a battery of highly sensitive functional assays in *Drosophila melanogaster* for testing orthologs of human neurodevelopmental genes. We took advantage of the tissue-specific expression system conferred by the UAS-Gal4 system, and used RNA interference strategies to achieve eye-specific (*GMR*-Gal4), neuron-specific (*Elav*-Gal4), and ubiquitous (*da*-Gal4) knockdown of conserved genes in flies. Combining data from gene expression using RNA sequencing with novel methods for measuring neuronal phenotypes in multiple fly RNAi lines allowed us to correlate the effect of gene disruption and dosage alterations to severity. We performed detailed phenotyping on >80 fly lines representing eight human copy-number variant regions, including 16p11.2, 16p12.1, 3q29, 16p13.11, 1q21.1, 15q11.2, 15q13.3, and distal 16p11.2, for dosage sensitivity of fly orthologs of human genes. We find differential effects of gene dosage for several genes within CNV regions. For example, 11 out of 13 fly orthologs within 3q29 showed dosage-dependent change in severity. We also identified some key genes such as *DLG* and *NCBP2* in 3q29, *KCTD13* and *MAPK3* in 16p11.2, and *POLR3E* in 16p12.1, that have an essential role towards a phenotype. We prioritized a subset of 15 genes, and tested ~100 *cis* interactions (with other genes within the CNV) and ~360 *trans* interactions interrogating known disease genes in conserved pathways (such as *PTEN, CHD8, SHANK3*). Using this approach, we found several modulators of phenotypes interacting within autophagy and insulin signaling pathways. Using RNA sequencing experiments on ten top candidate genes and empirical data from functional studies, we also created a network of CNV genes interacting within common pathways. This allowed us to identify several novel candidate genes (such as *KIF23* and *SLC6A15*), including those with associations with biological functions (such as FMRP-related proteins and genes involved in post-synaptic density). Our results suggest a complex and pervasive epistasis of genes that contribute to the observed phenotypic variability in affected individuals with the CNVs.

247

Novel genome-wide sequence variants influence antibody response to Epstein-Barr Virus in an African Population. *N. Sallah[1,2], T. Carstensen[1,3], K. Wakeham[4], R. Bagni[5], N. Labo[6], M. Pollard[1,3], D. Gurdasani[1,3], K. Ekoru[1,3], C. Pomilla[1,3], E. Young[1,3], S. Fatumo[1,3], G. Asiki[4], A. Kamali[4], M. Sandhu[1,3], P. Kellam[2], D. Whitby[6], R. Newton[4], I. Barroso[1].* 1) Human Genetics, Wellcome Trust Sanger Institute, Cambridge, UK; 2) Virus Genomics, Wellcome Trust Sanger Institute, Cambridge, UK; 3) Department of Medicine, University of Cambridge, Cambridge, UK; 4) MRC/Uganda Virus Research Institute, Uganda Research Unit on AIDS, Uganda; 5) Protein Expression Lab, Frederick National Laboratory for Cancer Research, Frederick, MD, USA; 6) Viral Oncology Section, Aids and Cancer Program, Frederick National Laboratory for Cancer Research, Frederick, MD, USA.

Globally, 95% of the adult population are infected with Epstein Barr Virus (EBV), a common human herpesvirus. While infection is lifelong and generally asymptomatic, EBV is associated with about 200,000 new cases of cancer and more than 140,000 deaths annually. How host genetic variation influences infectious disease traits such as EBV is largely unknown, particularly in Africa. As Immunoglobulin G (IgG) antibody levels to EBV have been shown to be heritable and associated with developing malignancies, we use it as a proxy for infection and potential disease risk. We therefore perform the first genome-wide association analysis of anti-EBV IgG traits in an African population, using a combined approach including array genotyping, whole-genome sequencing and imputation to a panel with African sequence data to extensively capture genetic variation and aid locus discovery. In 1562 Ugandans, we identify two novel African-specific loci associated with anti-VCA IgG responses, an intergenic variant on chromosome 7 ($p=4.0x10^{-10}$) and an intronic variant in *GALC* ($p=6.8x10^{-10}$). We also identify a variant in *HLA-DQA1* ($p=2.6x10^{-17}$) associated with anti-EBNA-1 responses. Trans-ancestry meta-analysis and fine-mapping with European-ancestry individuals suggest the presence of distinct *HLA* class II variants driving associations in Uganda. Our study reinforces the importance of studying diverse populations to uncover population specific differences and bridges the gap in our understanding of the host genetic contribution to the immune control of EBV infection in Africa.

248

A *CPT1A* missense mutation associated with fatty acid metabolism and reduced height in Greenlanders.

L. Skotte[1], A. Koch[1], V. Yakimov[1], S. Zhou[2,3], B. Søborg[1], M. Andersson[1], S.W. Michelsen[1], J.E. Navne[1], J.M. Mistry[1], P.A. Dion[2,4], M.L. Petersen[5], M.L. Børresen[1], G.A. Rouleau[2,4], F. Geller[1], M. Melbye[1,6,7], B. Feenstra[1]. 1) Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark; 2) Montreal Neurological Institute and Hospital, McGill University, Montréal (Que), Canada; 3) Département de médecine, Faculté de médecine, Université de Montréal, Montréal (Que), Canada; 4) Department of Neurology and Neurosurgery, McGill University, Montréal (Que), Canada; 5) Queen Ingrid Health Care Center, Nuuk, Greenland; 6) Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; 7) Department of Medicine, Stanford University School of Medicine, Stanford, California, USA.

Inuit have lived for thousands of years in an extremely cold environment on a diet dominated by marine-derived fat. To investigate how this selective pressure has affected the genetic regulation of fatty acid metabolism, we assessed 232 serum metabolic phenotypes by NMR spectroscopy in a population-based sample of 1570 Greenlanders. Using array-based and targeted genotyping, we found that rs80356779, a p.Pro479Leu variant in *CPT1A*, was strongly associated with markers of fatty acid metabolism, including degree of unsaturation ($p = 1.25 \times 10^{-32}$), levels of polyunsaturated fatty acids, *n*-3 fatty acids, and DHA relative to total fatty acid levels ($p = 2.20 \times 10^{-14}$, $p = 2.42 \times 10^{-18}$, $p = 8.61 \times 10^{-26}$). Further, we found that each copy of the derived allele of rs80356779 reduced height by an average of 2.1 cm ($p = 1.80 \times 10^{-9}$). We found strong signatures of positive selection at the locus, most extremely for rs80356779, where the derived allele was fixed in individuals of pure Inuit ancestry within our sample, and absent in the CEU and CHB HapMap populations. In exome sequencing data from 104 individuals from a sister population, the Nunavik Inuit, we found no other likely causal candidate variant than rs80356779. *CPT1A* encodes the liver isoform of carnitine palmitoyltransferase I, a key regulator of mitochondrial long-chained fatty acid oxidation and carriers of the p.Pro479Leu mutation may have had a distinct survival advantage by being able to rely on fatty acids and ketone bodies for energy even in the non-fasting state. Our findings illustrate how carefully designed studies of populations adapted to extreme dietary or environmental conditions can provide important knowledge about basic human physiology.

249

Evolution of hypertension: Understanding population differences - Insights from population genomic data.

H. Schaschl, T. Göllner, M. Fieder. University of Vienna, Vienna, Austria.

Long term hypertension in humans is a major risk factor of cardiovascular disease which is one of the leading causes for heart disease. Human hypertension is strongly influenced by genes, environment and ecology. It appears that the heritable component of hypertension is about 30%-50%. Genome-wide association studies have identified more than 50 genetic loci potentially involved in blood pressure regulation. Research data suggest that cardiovascular disease varies among ethnic groups, for instance there are higher stroke rates in South Asians than in Europeans. In particular, the combination of hyperglycemia and hypertension appears to be highly detrimental for South Asians. It is not fully understood why hypertension prevalence differs considerably between human populations. In this first study, we tested the hypothesis that genetic loci associated with hypertension are targets of natural selection, such as recent positive selection and/or balancing selection. We found several genetic loci across the human genome that are associated with hypertension are targets of recent selection. In particular, genetic variants in the solute carrier *SLC35F3* gene (1q42.2) appear to be under strong diversifying selection (*P<0.001*). This locus has been recently identified as a risk locus for hypertension (Zhang et al. 2014). The study showed that *SLC35F3* acts as thiamine (Vitamin B_1) transporter. The risk-allele homozygotes (rs17514104; T/T) at this locus displayed decreased erythrocyte thiamine content. This variant predicted heritable cardiovascular traits previously associated with thiamine deficiency, including elevated cardiac stroke volume with decreased vascular resistance. Our study revealed several SNPs within this locus to be under strong positive selection, characterised by very high *Fst*-values ranging from 0.40 to 0.65. Average *Fst* at this locus across worldwide populations is 0.33. Interestingly, some SNPs (e.g., rs34546351, rs12401568) are highly differentiated (*Fst >0.4*) between geographically relative close populations from Southeast Asia and East Asia. The risk allele rs17514104 (in intron 2) is in high LD with several SNPs that are under selection. Thus, we suggest that population genetics allows to determine the 'true' causal variants. Future study should include transcriptomic data and protein interaction networks to elucidate the molecular networks underlying hypertension and to understand the natural history of hypertension predisposition.

250

Whole-genome sequencing study of serum peptide levels: The Atherosclerosis Risk in Communities study. *P.S. de Vries*[1], *B. Yu*[1], *E.V. Feofanova*[1], *G.A. Metcalf*[2], *M.R. Brown*[1], *A.L. Zeigham*[1], *R.A. Gibbs*[2], *E. Boerwinkle*[1,2], *A.C. Morrison*[1]. 1) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX.

Short peptides have a biological role as precursors of other biomolecules and as intermediates in metabolism, as they are cleaved from larger polypeptides and proteins. Blood levels of peptides in humans potentially represent biomarkers of disease. Because of their molecular proximity to gene action, it is expected that effects of DNA sequence variation on peptide levels may be large. In the Atherosclerosis Risk in Communities (ARIC) study, we measured 25 serum peptides using untargeted gas chromatography and liquid chromatography mass spectrometry, and performed whole-genome sequencing (WGS) in a sample of 1,458 European Americans (EAs) and 1,679 African Americans (AAs) and whole-exome sequencing (WES) in a largely overlapping sample of 1,330 EAs and 1,850 AAs. Common WGS variants (MAF ≥ 5%) were analyzed individually, and an agnostic genome-wide sliding window analysis strategy (4kb window length with 2kb skip length) was applied across the genome in EAs and AAs, respectively, to aggregate low-frequency WGS variants (MAF < 5%) within a window using the Sequence Kernel Association Test (SKAT) or a T5 burden test. Additionally, low-frequency WGS variants were grouped according to annotated functional motifs, and WES variants were grouped by gene. Only variants or genetic regions which were significant in the meta-analysis across both ethnicities are interpreted here. In the evaluation of single common variants, we identified 23 associations with serum peptides levels ($p<1\times10^{-9}$, accounting for 2 million independent common variants and 25 peptides), and among those, 22 were novel gene-metabolite pairs. Notably, variants at the *KLKB1* locus were associated with six peptides, and variants at the *GRK6* locus were associated with five peptides. In three cases, low-frequency variants in sliding windows near associated common variants were significantly associated with peptide levels ($p<1.4\times10^{-9}$, accounting for ~700,000 regions, 25 peptides, and 2 tests). Gene-based analysis of low-frequency WES variants identified a further three genes associated with peptide levels, including *CPN1*, a metallo-protease that cleaves proteins and peptides. By integrating –omic technologies into deeply phenotyped populations, we showed that sequencing variants consistently affect multiple human peptide levels in two ethnicities. These data and results highlight new avenues of gene function and novel molecular mechanisms for regulating serum peptide levels.

251

A massively scalable phenotyping approach using social media for genetic studies. *J. Yuan*[1,2], *A. Gordon*[1], *D. Speyer*[1,2], *D. Zielinski*[1], *R. Aufrichtig*[1], *J. Pickrell*[1,3], *Y. Erlich*[1,2]. 1) New York Genome Center, New York, NY; 2) Computer Science, Columbia University, New York, NY; 3) Biological Sciences, Columbia University, New York, NY.

While DNA sequencing is largely a tractable problem, massive phenotyping is still a challenge, especially for Internet-based studies. Traditional methods, such as physical exams, scale poorly for large numbers of individuals. Questionnaires are easier to collect, but administering lengthy or frequent questionnaires creates a negative experience for participants, leading to lower completion rates. Electronic health records are a great resource for phenotypes, but they exhibit large heterogeneity when collected from various resources and are subject to an array of confidentiality restrictions that complicate their collection. Recent studies have highlighted the value of obtaining digital phenotypes by interpreting the interactions of users with digital outlets as a reflection of underlying traits. In particular, these studies have shown that social media data enables the collection of various phenotypes including big five personality traits, sexual orientation, sleeping patterns, and even heart rate from regular user videos. The ubiquity of the data and its ease of collection through standard APIs enable a new methodology for large scale phenotypic collection. Here, we report our ongoing efforts to enable participants to donate their social-media data along with their genomes in order to understand the genetics of digital phenotypes. In our previous work, we developed DNA.Land (https://dna.land), an online platform where users may register and securely contribute their Direct to Consumer genomic data, as well as receive reports of ancestry and shared relatives with other DNA.Land users. Since our launch in ASHG2015, we have obtained over 20,000 users, many of whom have been eager to share personal information such as family history. We are now building a new component in DNA.Land in which users can contribute their Facebook data for scientific studies. We will present our IBM Watson-based system to predict traits from social media data and will describe the type of information DNA.Land users will receive. In addition, we will discuss the particular challenges in collecting this data with respect to both computational efforts and privacy concerns. Our approach is applicable for other types of large scale efforts such as the Precision Medicine Initiative and can easily scale to millions of people.

252

Partitioning phenotype-ancestry correlations. *D.S. Park[1], J. Jeff[2], B. Glicksberg[2], G. Belbin[2], N. Abul-Husn[2], R.J. Loos[2], J.H. Cho[2], E. Kenny[2], N. Zaitlen[1].* 1) University of California San Francisco, San Francisco, CA; 2) The Charles Bronfman Institute for Personalized Medicine, The Icahn School of Medicine at Mount Sinai, New York.

The prevalence of common complex diseases differs between world-wide populations as a result of differences in both genetic and environmental factors. In admixed populations, which contain genetic information from multiple ancestral populations, the extent to which genetics drives phenotypic differences between ancestral populations will induce a correlation between phenotype and genetic ancestry. Genetic ancestry has also been shown to be correlated with environmental covariates. Thus, it is unclear how much of a given phenotype-ancestry correlation will be driven by genetics. Partitioning the correlation into genetic and environmental components provides insight into what drives differences between world-wide phenotypic distributions and has important implications for global health and precision medicine. In this work, we present a novel statistical method that partitions the phenotype-ancestry correlation into genetic and environmental components. We show analytically and via extensive simulation that our approach provides unbiased estimates of the genetic and environmental contributions to the correlation between ancestry and phenotype. We further show that existing methods for estimating heritability in admixed populations are biased when ancestry is correlated with phenotype and our approach corrects this bias. We apply our method to admixed populations and phenotypes in the Bio*Me* Biobank at the Icahn School of Medicine at Mount Sinai to investigate the relationship between genetic ancestry and disease risk. We analyzed self-reported African American (AA; n=3,705) and Hispanic/Latino (HL; n=5,104) participants in Bio*Me* using phenotype data from electronic health records (EHRs) linked to genotype data. EHR data is comprised of over 14,000 medical billing codes (ICD-9), which classify diseases, injuries, and health encounters. Analyzing the full medical phenome, we found 16, known and novel, significant correlations between ancestry and ICD-9 based phenotypes including asthma ($p<2.0x10^{-3}$; HL), hypertension ($p<1.0x10^{-4}$; HL), cardiac dysrhythmias (p<0.02; HL), and anemia ($p<1.0x10^{-4}$; AA). For each of these associations we estimate the contribution of genetics and environment to the phenotype-ancestry correlation and show that the contributions vary between phenotypes. We discuss the implications of these for medical research and clinical practice.

253

Marked high rate of multiple pathogenic variants in patients with Mendelian traits that show phenotypic expansion. *E. Karaca[1], T. Harel[1], Z. Coban Akdemir[1], S.N. Jhangiani[2], Y. Bayram[1], D. Muzny[2], E. Boerwinkle[2], R.A. Gibbs[2], J.R. Lupski[1,2,3,4].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA; 3) Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA; 4) Texas Children's Hospital, Houston, TX, USA.

Contrary to the widely held paradigm that a genetic syndrome is associated with a singular unifying molecular diagnosis, recent studies reported that in~5% of patients with a molecular diagnosis, the phenotype is attributed to mutations in two distinct disease loci. Moreover, the contribution of oligogenic inheritance or even the combinatorial effect of rare variants to disease burden and variable expressivity is now being uncovered by the increasing number of next generation sequencing studies on genetically and clinically heterogeneous disease cohorts. In this respect, in order to identify potential additional contributors to the patient phenotypes, we deeply re-analyzed the whole exom sequencing data from 19 families in which we found novel variants in known genes and they all represented with additional phenotypic features, also known as phenotypic expansion. As a result, in addition to previously described variants, we found additional potentially deleterious variants in at least one different known disease gene in 6 families (31%), while in 2 families (10%) we identified copy number variations (CNVs) and in other 2 families (10%) we found potentially deleterious single nucleotide variants (SNVs) in novel candidate genes, that could explain expanded phenotypic features in the probands. In conclusion, our findings, compatibly with many recent studies, underscore the role of oligogenic inheritance and mutation burden in the etiology of genetically and clinically heterogeneous disease cohorts. Moreover, several families in this study were found to have rare and potentially deleterious variants in either 3 or 4 known disease genes that each can interestingly explain distinct phenotypic features seen in the probands. Our data also emphasize the high possibility of the presence of multiple pathogenic variants in individuals with phenotypic expansion. We will also discuss potential mechanisms that play role in determining the phenotype in the existence of multiple pathogenic and/or potentially pathogenic variants in the same individual.

254

Translational deep phenotyping for clinical genomics. *T. Groza*[1,2], *T. Roscioli*[1,2], *M. Cowley*[1,2], *G. Baynam*[3,4,5,6,7], *H. Dawkins*[5], *M. Haendel*[8], *C. Mungall*[9], *D. Smedley*[10,11], *P.N. Robinson*[12,13,14], *M.E. Dinger*[1], *A. Zankl*[1,15,16]. 1) Kinghorn Center for Clinical Genomics, Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia; 2) St Vincent's Clinical School, Faculty of Medicine, UNSW Australia; 3) School of Paediatrics and Child Health, University of Western Australia, Perth, WA 6840, Australia; 4) Institute for Immunology and Infectious Diseases, Murdoch University, Perth, WA 6150, Australia; 5) Office of Population Health Genomics, Public Health and Clinical Services Division, Department of Health, Perth, WA 6004, Australia; 6) Genetic Services of Western Australia, King Edward Memorial Hospital, Perth, WA 6008, Australia; 7) Telethon Kids Institute, Perth, WA 6008, Australia; 8) Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon 97239, USA; 9) Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; 10) Queen Mary University of London, London E1 4NS, United Kingdom; 11) Genomics England Ltd., London EC1M 6BQ, United Kingdom; 12) Institute for Medical and Human Genetics, Charité–Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany; 13) Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany; 14) Berlin Center for Regenerative Therapies (BCRT), Charité–Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany; 15) Academic Department of Medical Genetics, The Children's Hospital at Westmead, Sydney, NSW 2145, Australia; 16) Discipline of Genetic Medicine, Sydney Medical School, University of Sydney, Sydney, NSW 2145, Australia.

The genomic era is transforming genetic diagnostics and patient management. We now have the capacity to generate 1000s of whole genome sequences per annum which has brought about the need for secure clinical and genomic data storage and the linking of genomic pathology services to families and clinicians. Although marked improvements have occurred in WES/WGS diagnostic yield, a proportion of cases remain unsolved. One of the main challenges is the interpretation and prioritisation of variants with the demand for genomic testing now rapidly outstripping the ability of the clinical community to interpret these results. Phenotype has long been used to determine underlying genetic etiologies, as well as to substantially reducing the search-space for genomic variation. The incomplete linking of detailed phenotypic terms to genomic variants is now the major factor limiting rapid diagnostics. Patient Archive (PA) is a platform developed as part of the Monarch Initiative that translates deep clinical phenotyping and genome-scale biology to patient-centered human disease pathogenesis. The accurate and detailed phenotype profile created in PA, combined with clinical genomic data accelerates the identification of disease aetiology and facilitates disease stratification and prognosis. PA is deployed at the Kinghorn Centre for Clinical Genomics and has facilitated a 50% diagnostic rate for whole genome sequencing referrals. PA's innovative features are rooted in the use of the Human Phenotype Ontology (HPO). Unlike other existing phenotyping platforms, a patient clinical phenotype profile is created using an automatic extraction of HPO concepts from free text clinical records or the labels of uploaded images. HPO-driven semantic similarity matching algorithms are then used for patient matchmaking, disorder prediction or gene list generation. Clinical data exchange is protected with secure access control on an individual or group context, with full integration into the Global Alliance for Genomics and Health - MatchMaker Exchange Program. Finally, cross-platform data interoperability is enabled via the Monarch Phenotype Exchange Format. Patient Archive provides a seamless solution for the harmonization of phenomic information to enable the acceleration of translational and clinical applications. The platform is freely available and it is currently serving both clinical (Dept. of Health, Govt. of Western Australia) and research (UDP Japan) environments.

255

A whole exome sequencing (WES) rare variant association study of Multiple Sclerosis (MS) subtypes identifies a significant association between *LIX1L* and primary progressive MS (PPMS). *P. Bronson*[1], *K.D. Nguyen*[1], *K. Estrada*[1], *K. Gutwin*[2], *I. Kockum*[2], *J. Hillert*[3], *S. John*[1], *T. Harris*[4,5], *A. Day-Williams*[1]. 1) Computational Biology & Genomics, Biogen, Inc., Cambridge, MA; 2) Scientific Computing, Biogen, Inc., Cambridge, MA; 3) Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden; 4) Precision Medicine, Biogen, Inc., Cambridge MA; 5) SV Life Sciences, Boston, MA.

Background: Genetic studies of MS have identified >100 non-MHC loci. The strongest genetic associations are *HLA-DRB1*15:01* (OR = 3.1) and *HLA-A*02:01* (OR = 0.73). However, the genetics of PPMS, a severe type of MS that affects ~10% of patients, has yet to be interrogated, so there is no anchor to focus drug target discovery and validation efforts to treat PPMS. **Objectives:** To test whether there is a larger burden of rare protein coding variation in any gene in the genomes of PPMS patients compared to controls. **Methodology:** We WES 1,589 Swedish samples (509 PPMS, 552 relapsing remitting MS (RRMS) and 528 controls) in collaboration with the Karolinska Institute. We included 2,996 Swedish WES from dbGap (phs000473.v1.p1) (total N = 4,585). Stringent QC was applied. EPACTS was used for single variant tests that compared 509 PPMS cases to 3,524 controls (total N = 4,033) and also to 552 RRMS cases (total N = 1,061). We conducted gene-based association tests in R SKAT-O with resampling. Gene-based tests were restricted to genes with >1 rare (MAF < 0.005) stop gain, splice donor, splice acceptor, or missense variant, where missense variants were required to be predicted damaging (CADD score > 20 or PolyPhen2 "probably damaging"). The significance threshold (Bonferroni correction for 12,750 gene-based tests) was $P <$ 3.92e-6. **Results:** We observed a putative gene-based signal in *LIX1L* ($P =$ 3.8e-6) in the PPMS vs. controls analysis. Four rare, damaging *LIX1L* variants were analyzed, and were carried by four different PPMS cases and no RRMS patients or controls. In the single variant analysis, the *HLA-DRB1*15:01* signal was similar in PPMS and RRMS but *HLA-A*02:01* (tagged by rs1143146, $r^2 =$ 0.95, controls MAF = 0.55) reached genome-wide significance in RRMS (OR = 0.61, $P =$ 2.8e-12) and was suggestive of association in PPMS (OR = 0.79, $P =$ 9.6e-4). Rs1143146 was under-represented in RRMS (MAF = 0.43) vs. PPMS (MAF = 0.50) ($P =$ 5.8e-3). **Discussion:** This is the first large WES study of MS subtypes (PPMS), and we identified a novel risk locus at *LIX1L*. *LIX1L* is involved in autophagosome maturation, is differentially expressed throughout the brain, and interacts with the Calpain small subunit 2. Calpains are activated by brain trauma, and inhibiting calpains ameliorates neuropathology in animal models. The weaker *HLA-A*02:01* signal observed in PPMS suggests the complex HLA relationship with MS may differ by disease subtype. Further research is warranted to confirm our findings. .

256

Exome sequencing reveals novel candidate genes and potential oligogenic inheritance in patients with hypergonadotropic hypogonadism. *Y. Bayram[1], S. Turan[2], Z. Aycan[3], T. Tos[4], B. Hacihamdioglu[5], Z. Coban Akdemir[1], S. Bas[2], Z. Atay[2], T. Guran[2], S. Abali[2], J.J. White[1], G. Yesil[6], E. Karaca[1], S.N. Jhangiani[7], D.M. Muzny[7], A. Bereket[2], R.A. Gibbs[7], J.R. Lupski[1,7,8,9], Baylor-Johns Hopkins Center for Mendelian Genomics.* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA; 2) Department of Pediatric Endocrinology and Diabetes, Marmara University Hospital, Istanbul, Turkey; 3) Department of Pediatric Endocrinology, Sami Ulus Children's Hospital, Ankara, Turkey; 4) Department of Medical Genetics, Sami Ulus Children's Hospital, Ankara, Turkey; 5) Department of Pediatric Endocrinology, Suleymaniye Training and Research Hospital, Istanbul, Turkey; 6) Department of Medical Genetics, Bezmialem University, Istanbul, Turkey; 7) Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA; 8) Department of Pediatrics, Baylor College of Medicine, Houston, Texas, USA; 9) Texas Children's Hospital, Houston, Texas, USA.

Hypergonadotropic hypogonadism (HH) is characterized by hypogonadism due to an impaired response of the gonads to the gonadotropins (FSH and LH) and in turn a lack of sex steroid production and elevated gonadotropin levels. Most common known causes of HH are chromosomal abnormalities, acquired damage of the ovaries, and congenital disorders that affect normal ovarian development and function. There are also some syndromic and non-syndromic causes of HH described as single gene disorders. Beside these, in most of the cases with HH the etiology of gonadal dysfunction is idiopathic and presumably genetic. To identify novel molecular etiologies in HH we applied whole exome sequencing (WES) to 29 affected female individuals from 27 unrelated families including 21 with reported consanguinity. WES revealed variants in known genes including *AR*, *NOBOX*, *MCM8* (2 families), *PSMC3IP*, and *TG* in 6 families. In 7 families we found pathogenic variants in novel candidate genes including *SOHLH1* (recently published), *PADI6*, *C3*, *PPP2R1A*, *DMRTA2*, and *YY1* with supporting functional data and positive segregation data. Interestingly, in 3 families we identified a potential mutational burden or oligogenic inheritance model. In one patient with HH and deafness we identified homozygous variants in two different genes (*POLG* and *NSMF*) which were previously associated with hypogonadism. In the same patient, we also found compound heterozygous variants in *PCDH15*, a known gene for deafness. In another patient with HH and obesity we identified homozygous variants in two known genes for hypogonadism (*CHD7* and *MCM9*) and another homozygous variant in *PRKD1* that was reported as a candidate gene in patients with obesity. In a third patient potentially explained by a mutational burden model we identified homozygous variants in *GALT* (Associated with galactosemia, a known cause of HH) and in *DNAH6* that was previously reported in a patient with ovarian failure. In this study we described variants in known genes that were previously associated with HH and candidate novel genes using WES in patients with HH. The most interesting finding of our study is to describe oligogenic and mutational burden models in some patients which is one of the advantages of genome wide sequencing approaches. We suggest that increased using of genomic sequencing methods in genetically heterogeneous phenotypes such as HH provide further molecular etiological insights including potential contributions of multilocus variation.

257

Novel analytic approaches used to solve unsolved whole exome sequencing data. *N. Sobreira[1], F. Schiettecatte[2], H. Ling[1,3], E. Pugh[1,3], D. Witmer[1,3], K. Hetrick[1,3], P. Zhang[1,3], K. Doheny[1,3], S.N. Jhangiani[4], Z.H. Coban Akdemir[5], J. Posey[5], V.R. Sutton[5], E. Karac[5], Y. Bayram[5], J.R. Lupski[5], A. Hamosh[1], D. Valle[1].* 1) Human Gen, Johns Hopkins Univ, Baltimore, MD; 2) FS Consulting, Salem, MA; 3) Center for Inherited Disease Research (CIDR), JHUSOM, Baltimore, MD; 4) Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, 77030; 5) Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, 77030.

The Baylor-Hopkins Center for Mendelian Genomics have sequenced more than 6,225 samples and identified more than 222 novel Mendelian disease genes, 231 known genes and 136 phenotypic expansions. However, as in other clinical and research sequencing centers, for more than half of the probands sequenced the responsible gene cannot be determined. Here we present alternative analytic approaches we have used to address these cases. We have re-analyzed WES data from 1063 unsolved samples selecting rare (MAF <1%) functional variants (missense, nonsense, frameshift and splicing) in known imprinted genes, in the genes on pseudoautosomal regions, genes that escape X-inactivation, and genes on the Y chromosome. We found that the genes in the pseudoautosomal regions of the X are not captured by the Agilent SureSelect v4 baits used to sequence these samples. The analysis of variants in the genes on chromosome Y identified 52 rare functional variants and the analysis of variants in the 242 imprinted genes identified 4,337 rare functional variants. These variants are being further evaluated to define causality but we have already identified interesting candidate variants in the imprinted genes ZDBF2 and ABCC9, as well as in the UTY gene on chromosome Y. Our second approach, to solve families in whom we identified 2 or more candidate genes, has focused on the use of GeneMatcher (www.genematcher.org) which facilitates data sharing and improves the search for patients or model organisms with variants in specific candidate genes. In GeneMatcher matches can be made based on gene name, genomic location, OMIM® number and on phenotypic features. As part of the Matchmaker Exchange, we have developed an Application Programing Interface (API) that, as of today, allows the GeneMatcher users to submit their data to query PhenomeCentral and/or DECIPHER. We have also been working with other matchmaker databases like The Monarch Initiative and MyGene2 on the API implementation to connect them to GeneMatcher. A version 2.0 of the API that will allow for more detailed queries is now being developed. As of 1 May 2016 , 4,459 genes were submitted by 1,675 individuals from 55 countries. There have been 5,267 matches involving 1, 216 genes (100 matches with PhenomeCentral and 87 with DECIPHER) that have enabled collaborations and the description of novel Mendelian phenotypes and novel Mendelian disease genes, such as *SPATA5*, *HNRNPK*, *TELO2*, *RSPRY1*, *HIVEP2*, *CHAMP1* and others.

258

Low-pass whole-genome sequencing in clinical cytogenetics: A validated approach. *Z. Dong[1,2,3], F. Chen[3], H. Wang[1,2], H. Jiang[3], K.W. Choy[1,2,4].* 1) Dept. of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong, Hong Kong; 2) Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China; 3) BGI-Shenzhen, Shenzhen, China; 4) 4.The Chinese University of Hong Kong-Baylor College of Medicine Joint Center For Medical Genetics, Hong Kong, China.

Purpose: Chromosomal microarray analysis is the gold standard for copy-number variant (CNV) detection in prenatal and postnatal diagnosis. We aimed to determine whether next-generation sequencing (NGS) technology could be an alternative method for CNV detection in routine clinical application.Methods: Genome-wide CNV analysis (≥50 kb) was performed on a multicenter group of 570 patients using a low-coverage whole-genome sequencing pipeline. These samples were referred for chromosomal analysis; CNVs (i.e., pathogenic CNVs, pCNVs) were classified according to the American College of Medical Genetics and Genomics guidelines.Results: Overall, a total of 198 abortuses, 37 stillbirths, 149 prenatal, and 186 postnatal samples were tested. Our approach yielded results in 549 samples (96.3%) and the other 21 samples were failed because of the poor DNA quality due to fetal demise. In addition to 119 subjects with aneuploidies, 103 pCNVs (74 losses and 29 gains) were identified in 82 samples, giving diagnostic yields of 53.2% (95% confidence interval: 45.8, 60.5), 14.7% (5.0, 31.1), 28.5% (21.1, 36.6), and 30.1% (23.6, 37.3) in each group, respectively. Mosaicism was observed at a level as low as 25%.Conclusions: Patients with chromosomal diseases or microdeletion/ microduplication syndromes were diagnosed using a high-resolution genome-wide method. Our study revealed the potential of NGS to facilitate genetic diagnoses that were not evident in the prenatal and postnatal groups.

259

Medical management differs for children following whole genome sequencing compared to chromosome microarray. *R.Z. Hayeems[1,2], J. Bhawra[1], K. Tsiplova[1], N. Monfared[3], M.S. Meyn[3], S. Bowdin[3], D. Stavropoulos[4], C. Marshall[4], R. Basran[4], C. Shuman[3], S. Ito[5], I. Cohn[5], C. Hum[6], M. Girdea[6], M. Brudno[6,7], R.D. Cohn[3], W.J. Ungar[1,2].* 1) Child Health Evaluative Sciences, Hospital for Sick Children, Toronto, Ontario, Canada; 2) Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Canada; 3) Division of Clinical and Metabolic Genetics, Hospital for Sick Children, Toronto, Canada; 4) Department of Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Canada; 5) Division of Clinical Pharmacology and Toxicology, Hospital for Sick Children, Toronto, Canada; 6) Centre for Computational Medicine, Hospital for Sick Children, Toronto, Canada; 7) Department of Computer Science, University of Toronto, Toronto, Canada.

Objective: Rational integration of whole genome sequencing (WGS) into clinical practice requires knowledge of its downstream clinical and cost consequences. The objective of the study was to compare the volume and type of health services prompted or discontinued following chromosome microarray analysis (CMA) compared to WGS in a pediatric tertiary care setting. **Methods:** Using a prospective observational cohort, this study ascertained the clinical trajectory of care prompted by primary WGS results, compared to CMA for 100 children with developmental delay and/or congenital anomalies. All children pursued both CMA and WGS. Clinical services consumed (specialist visit, laboratory testing, medical imaging, cascade family testing) following the receipt of CMA negative results were compared to clinical services consumed following the receipt of WGS results. Outcomes were captured for the immediate 6 months following result disclosure using chart extraction and clinician verification. **Results:** Ongoing care/characterization of presenting clinical features prompted 87.8% of all care activities for these children; 67.0% of these activities were specialist referrals and 21.0% were medical imaging. The proportions of total services prompted by WGS and CMA were similar (6.7% WGS vs. 5.6% CMA; p>0.05) but WGS prompted more specialist consults (69.8% of WGS-prompted activity vs. 0% of CMA-prompted activity; p<0.01) and CMA prompted more laboratory testing (88.7% of CMA-prompted activity vs. 11.9% of WGS-prompted activity; p<0.01). Importantly, utilization of laboratory testing was decreased for both positive and negative WGS results. Thirteen clinical activities were discontinued following WGS (i.e. 6 specialist consults, 3 lab tests, 2 imaging tests, 2 other) whereas no activities were discontinued following CMA. **Conclusion:** Genetic testing prompted a minority of health service use in this phenotypically complex pediatric population, but the nature of the activity triggered by CMA and WGS differed significantly. Minimal laboratory testing and discontinued services following WGS suggests that WGS may avert ongoing diagnostic investigations and other unnecessary health care use, even when WGS does not lead to a diagnosis. Data reflecting per patient cost of clinical activities triggered by primary CMA/WGS variants will be presented. Early translation of WGS should occur alongside rigorous approaches to measuring its downstream clinical care and cost consequences.

260

The genetics clinic of the future: Towards implementation of whole genome sequencing in diagnostic practice. *I.J. Nijman[1], S. van Lieshout[2], T. Bradley[2], P. van Zon[1], M. van Stralen[3], F. van Ruissen[4], D.M. van Beek[3], P. Neerincx[5], J. ten Hoeve[6], F. Baas[4], R. Sinke[5], B. Sikkema[5], F. Hogervorst[6], J.K. Ploos van Amstel[1], G.H. Schuring-Blom[1], M. Weiss[3], E.C. Cuppen[2], E.A. Sistermans[4].* 1) dept. of Genetics, University Medical Center Utrecht, Utrecht, Netherlands; 2) Hartwig Medical Foundation, Amsterdam; 3) VU University Medical Center Amsterdam; 4) Academic Medical Center, Amsterdam; 5) University Medical Center Groningen; 6) Netherlands Cancer Institute, Amsterdam.

DNA-testing is currently considered relatively late in the process of diagnosis. We envision that in the nearby future the field of clinical genetic diagnostics will shift towards a genome first approach. We have initiated a project to implement a diagnostic track based on whole genome sequencing (WGS) first approach. Apart from cost efficiency, logistics (sample/IT), quality assurance and ethical, legal and social issues, we are validating genome sequencing as a ‘one-test-fits-all’ to replace targeted gene panel- and exome sequencing for SNVs and indels, and array technology for CNV calling. Moreover, detection of translocations or integration of SNV and CNV calls are explored. In this project, a collaboration has been established among four university medical centers in the Netherlands, Groningen (UMCG), Amsterdam (VUmc, AMC) and Utrecht (UMCU) and the comprehensive cancer center (AVL/NKI). Multidisciplinary working groups include laboratory technicians, clinical laboratory geneticists and bioinformaticians. The sequencing and primary bioinformatic processing are to be outsourced to the Hartwig Medical Foundation which runs an Illumina Xten facility for routine high throughput WGS. This approach ensures uniform sequencing data generation and analysis, enabling data sharing and exchange (both technical as interpretation) between the partners. The bioinformatics groups involved have consolidated their local pipelines in a central genome analysis pipeline based on BWA and GATK in a private cloud computing environment. The results of sequencing and analyses have been validated on genome-in-a-bottle samples and are equal or better than targeted analysis, while SV and CNV analysis validation is ongoing. Results show that 40x coverage is sufficient to reach 99.2% sensitivity and 99.7% precision (99.5% and 99.6% respectively for the exome target) for snvs. For indels a 97.2% sensitivity and 90.3% precision (98.6% and 94.1% for the exome target) were obtained. Increased sequence coverage for WGS (range 30-120x) reduces the number of false positives, but it has limited effect to reduce false negatives. However, indel detection improves. We intend to process ~400 samples (postnatal Intellectual disability) on WGS before the end of this year and increase this number to the majority of patients in 2019, which will be >10.000 yearly in the centers involved.

261

Reanalyzing yearly whole-exome sequencing results: An 8% increase in diagnostic yield. *J. Thevenon[1], P. Kuentz[1], S. Nambot[1], A.-L. Bruel[1], D. Lehalle[1], M. Assoum[1], N. Jean-Marçais[1], A. Masurel-Paulet[1], P. Callier[1], A.-L. Mosca-Boidron[1], S. El Chehadeh[1], C. Poé[1], T. Jouan[1], M. Chevarin[1], N. Gigot[1], J.-B. Rivière[1], M. Lefebvre[1], E. Tisserant[1], J.-F. Deleuze[2], Y. Duffourd[1], L. Faivre[1], C. Thauvin[1].* 1) FHU TRANSLAD, University Health Centre, Dijon, France; 2) Centre National de Génotypage, Évry, France.

The World Health Organization estimates the existence of more than 8.000 rare disorders. Although most of these disorders have a suspected genetic origin, only half have their molecular basis identified. The development of next generation sequencing has dramatically changed the pace of gene discovery in such heterogeneous conditions and hundreds of disorders have their molecular basis unraveled each year. Unlike targeted sequencing approaches, whole-exome sequencing offers the possibility of re-assessing the data for recently described disorders after an initial negative interpretation. We present the approach of a regional center performing clinical exome sequencing for the diagnosis of rare disorders with congenital anomalies since 2013. Sequencing was performed in collaboration with the Centre National de Génotypage. Raw data were analyzed on the Computing center of the University of Burgundy. Overall, the mean depth of coverage was 90x, with 93% of coding regions referenced in RefSeq sequenced by at least 10 reads. Clinical interpretation focused on genes referenced in OMIM. On the first analysis of 400 consecutive probands, a diagnostic yield of 30% positive and of 15% non-conclusive results was achieved. Prospectively, the data of 130 negative and non-conclusive cases analyzed in 2013 and 2014 were processed from fastq files on an updated bioinformatic pipeline. Reanalysis time after the initial results ranged from 12 to 18 months. This strategy lead to the identification of seven positive diagnosis caused by disease-causing variations of genes previously unreported in a human disorder, namely *DDX3X* (2 cases)*, DNMT3A, PPP2R5D, PPP2R1A, KCNA2, EEF1A2*; turned two non-conclusive results into positive because of the accumulation of genetic evidence for *PEX6* in Heimler syndrome, and *DNM1L* in a severe mitochondrial disorder; identified a genetic variation that was missed by the previous bioinformatics pipeline with the detection of a recurrent intragenic deletion of *CLN3* in a patient. Overall, the reanalysis of the data of 130 negative probands could identify 11 positive diagnoses (8%) and raised a number of candidate variants that were systematically shared through the MatchMaker exchange initiative after parental consent. These results confirm the usefulness of reanalyzing WES data for the diagnosis of syndromes with congenital anomalies. Such strategy should be undertaken before moving towards whole-genome sequencing.

262

Risk, recommendation, and rationale: How primary care providers interpret and manage genome sequencing results. *I.J. Richardson[1], J.K. Davis[2], C. Kirby[2], R.C. Green[1,3,4,5], P.A. Ubel[2], J.L. Vassy[1,3,6].* 1) Brigham and Women's Hospital, Boston, MA; 2) Fuqua School of Business and Sanford School of Public Policy, Duke University, Durham, NC; 3) Harvard Medical School, Boston, MA; 4) Partners Healthcare Personalized Medicine, Cambridge, MA; 5) Broad Institute of MIT and Harvard, Cambridge, MA; 6) VA Boston Healthcare System.

Background: As genomics plays an increasingly prominent role in clinical medicine, it is unclear how primary care providers (PCPs) will interpret the spectrum of possible genome sequencing (GS) results and make clinical recommendations. **Methods:** We enrolled 9 PCPs and their generally healthy adult patients in the MedSeq Project, a trial of GS in primary care. Each patient's interpreted GS results were returned on a genome report that included monogenic disease variants; carrier status; and polygenic risk estimates for 8 cardiometabolic traits, such as coronary disease and type 2 diabetes. Each report was delivered to the patient's PCP, who then met with the patient to discuss the results. Disclosure sessions were recorded, transcribed, and coded with thematic content analysis to describe how PCPs interpret and manage GS results. **Results:** For each result discussed in 48 PCP-patient sessions, we identified a "take-home" recommendation, categorized as *continuing current management* (*i.e., no change*), *treatment*, *further evaluation*, *behavior change*, *remembering for future care*, or *sharing with family members*. We analyzed how the PCP came to each recommendation by identifying 1) how the PCP described the risk of the given GS result (*high*, *low*, *uncertain*, or *no risk*) and 2) the rationale the PCP gave for translating that risk into a specific recommendation. Quantitative analysis showed that the majority of risk codes for polygenic (271/375, 72%) and carrier results (101/117, 86%) were categorized as *no risk* for the patient. In contrast, only 4/12 (33%) codes about monogenic results were categorized as *no risk* and were more often categorized as *uncertain risk* (5/12, 42%). The most frequent recommendations around polygenic and carrier results were *continuing current management* (326/425, 77%) and *sharing with family members* (66/123, 54%), respectively. Only 5% of recommendation codes for monogenic results referenced the need for *further evaluation.* PCPs' rationales for their recommendations were categorized into *patient contextual factors*, *family contextual factors*, *technical/scientific limitations of GS*, and *reasons explained in genetic terms only* (*e.g., low genetic attribution of disease risk*). Example quotes illustrate these findings. **Conclusion:** Faced with different types of potentially actionable genetic results in their healthy patients, PCPs often use contextual factors and the perceived limitations of GS information to justify continuing current management.

263

HiSeq X performance: Optimization for clinical applications. *K. Walker[1], R. Sanghvi[1], Q. Wang[1], H. Doddapaneni[1], J. Hu[1], Z. Momin[1], J. Santibanez[1], J. Farek[1], A. English[1], W. Salerno[1], Y. Han[1], H. Dinh[1], E. Boerwinkle[1,2], R. Gibbs[1], D.M. Muzny[1].* 1) Baylor College of Medicine HGSC, Houston, TX; 2) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX.

High-throughput parallel nucleotide sequencing has revolutionized genomic research and reshaped applications in clinical health care. The HiSeq X Ten platform further expands these opportunities with unprecedented capacity. The Human Genome Sequencing Center at Baylor College of Medicine adopted the HiSeq X Ten system with eventual deployment in a CAP/CLIA environment. To date, we have analyzed 1,647 HiSeq X flowcells, representing >13,174 30X human genomes. These studies have included primarily common disease cohorts, inherited cancers and mendelian disease. PCR-Free library methods were evaluated and implemented for optimized coverage in GC-rich regions. Metrics related to coverage, sample integrity and variant representation were established to ensure high quality genome sequencing. We have implemented standard metrics including % pass filter, % aligned bases, % error rate, % unique reads and % Q30 bases to achieve > 90 GB unique aligned bases per lane. Genome coverage metrics are also tracked for 90% of genome covered at 20x and 95% at 10x with a minimum of $86x10^9$ mapped, aligned bases with Q20 or higher. Additional metrics such as library insert size, duplicate reads, error rates, % pair reads and mean quality scores are also monitored. Platform sensitivity and precision at 30x coverage was determined to be 97.8% and 99.6% respectively using control sample NA12878. To ensure integrity, we have implemented the Fluidigm SNPTrace assay to confirm sample identity and VerifyBamID to detect sample contamination. All quality control metrics are generated at-scale with HgV, the HGSC workflow management system that integrates our LIMS with existing and novel software including our custom HiSeq X AlignStats program and ERIS, the HGSC array-agnostic sample concordance framework. When considering clinical WGS, costs for required coverage are paramount. High SNP, Indel and SV concordance between 20X and 30X WGS suggests that for variant calling 20X would be sufficient. However, the typical clinical exome benchmark is high with consistent exonic coverage of clinically relevant genes. Our analysis shows that clinical WGS may require ~40X coverage to reach this benchmark. These efforts have provided valuable insight to how sequencing depth and coverage uniformity impact the ability to accurately detect variants. Establishment of robust PCR-Free WGS methods and associated pipeline metrics are essential for applications in both the research and clinical setting.

264

DName barcodes allow absolute quality control of genetic test processes based on NGS. *H. Cuppens[1], S. Dillen[1], M. Jaspers[2], K. Verleysen[1].* 1) DName-iT, Leuven, Belgium; 2) KULeuven, Experimental Otorhinolaryngology Group, Leuven, Belgium.

Genetic tests based on next generation sequencing (NGS) are complex and consist of many handling steps from patient sampling to result, which can lead to errors such as sample switching and/or contamination. Genetic testing labs take extreme measures to overcome these errors and will often process patient samples in duplicate to validate the obtained results, amongst many other QC procedures. However, this still does not guarantee 100% quality assurance and comes with a considerable extra cost. We have developed a powerful technology that provides quality assurance throughout the entire genetic testing process. Because of the parallel properties of NGS, spiking of biological samples with DNA molecules now allows for the first time 100% quality assurance. The technology allows to economically produce over millions of unique spiking DNA molecules (DName barcodes), which are used only once. Our proof of principle experiments illustrate that these barcodes can be universally applied with respect to the used DNA extraction method, NGS template preparation method and NGS platform, and detect sample switching and allows very sensitive detection of contaminations at any level in the genetic NGS test process. Different unique barcodes can be added at different time points in a genetic test process, such as at the time of blood collection (barcodes in the blood collection tube itself), after DNA extraction before NGS sequencing, so that quality assurance problems can be traced to different subprocesses in a genetic test process. In this way NGS testing can be outsourced, while retaining full control of the process. The unique quality assurance properties of these barcodes will boost the reliability of genetic tests while maintaining patient confidentiality. Furthermore, it will help to unleash the unprecedented informative power of NGS by improving accuracy, fidelity and automation of genetic test processes at a reduced cost.

265

A curated database of 130,000 variants from more than 1,200 genomes: Challenges for the clinical laboratory. *D.L. Perry, A.J. Coffey, J. Avecilla, M. Bennett, N. Burns, A. Chawla, B. Juan, J. Kakishita, A. Khouzam, E. Thorpe, R.J. Taft.* Illumina, Inc., San Diego, CA.

The Illumina Clinical Services Laboratory (ICSL) offers a clinical whole genome sequencing test for generally healthy adults, providing a clinical interpretation of variants found in 1,691 genes associated with more than 1,200 genetic disorders. Variant curation is performed in accordance with the American College of Medical Genetics and Genomics guidelines, with the addition of one category, variant of unknown significance-suspicious (VUS-S), which includes variants with insufficient information to qualify as likely pathogenic.Currently the ICSL database contains over 130,000 variants that have been curated from over 1,200 genomes (from 2012 – present). Keeping the curation database current presents challenges. Variants are generally only recurated if they reoccur in another genome and meet inclusion criteria for the clinical report (VUS-S, likely pathogenic and pathogenic). Variants classified as variants of unknown significance (VUS), likely benign or benign, or variants not seen since the initial curation, are not systematically revisited. As this latter group represents the majority of the variants in our database, a systematic strategy for keeping abreast of new information is needed. Here, we present two strategies to address outdated variant-level information in the ICSL database. First, we applied the frequency information from the Exome Aggregation Consortium (ExAC) database to the 75,295 variants curated prior to the ExAC release followed by a systematic search for literature on these variants. Second, we compared ICSL classifications to those also found in ClinVar to determine similarities and differences in variant classifications (n=13,326). Fifty-eight percent (n=7,967) of the variants had an exact match in classification. Thirty-nine percent (n=5,318) differed in exact classification, but matched in the overall direction of the classification (i.e. pathogenic/likely pathogenic/VUS-S vs. benign/likely benign). Three percent (n=341) were truly discrepant in classification, but these were all single entries to ClinVar from OMIM, and thus were not interpreted by another clinical laboratory. For clinical laboratories maintaining large curation databases, such comparisons are valuable checkpoints in assessing the state of the curated database and identifying areas of focused updates to the database. These exercises demonstrate the value of standardization in variant curation and highlight the importance of publically available databases.

266

Improving gene annotation to facilitate identification of missing variants of clinical significance. *A. Frankish[1], C.A. Steward[1], M-M. Suner[1], D. Pervouchine[2], B. Uszczynska[2], J-M. Gonzalez[1], S. Fitzgerald[1], D. Grozeva[3], K. Carss[4], P. Cossette[5], F. Hamdan[6], J. Michaud[6], R. Petryszak[7], E. Tapanari[1,7], M. Diekhans[8], B.A. Minassian[9], F.L. Raymond[3], R. Guigó[2].* 1) Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; 2) Centre for Genomic Regulation, Barcelona. Catalonia. Spain; 3) Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK; 4) Department of Haematology, University of Cambridge, Cambridge, UK; 5) Centre Hospitalier de l'Universite de Montreal, Canada; 6) Hopital Ste Justine, Montreal, Canada; 7) European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Hinxton, UK; 8) Center for Biomolecular Science and Engineering, UCSC, CA, USA; 9) The Hospital for Sick Children, Toronto, Canada.

High quality gene models provide a foundation for the annotation of sequence variation in clinical samples. Complete and accurate annotation of gene structure and function has the potential to reduce both false negative (from missing annotation) and false positive (from incorrect annotation) errors in identification of disease-associated variants in genome and exome sequence. As part of the GENCODE project, our team produces detailed reference annotation of all human and mouse protein-coding genes. We will describe the re-annotation of 70 genes on reference diagnostic panels for Early Infantile Epileptic Encephalopathies (EIEE), a group of disorders characterized by early onset seizures and developmental delay. The re-annotation utilised public transcriptional evidence, focussing on next-generation sequence data from human brain. We used RNAseq, SLRseq and PacBio reads to find new alternative splicing (AS) events such as novel exon inclusion and skipping, and shifted splice sites. Cap analysis gene expression (CAGE) data was used to confirm transcription start sites allowing better functional annotation of transcript models. Although these genes had been annotated previously by GENCODE, with 685 AS transcripts compared to 193 in the RefSeq geneset, our re-annotation added a further 1092 AS transcripts, 706 novel exons, 224 shifted splice sites and more than 141kb of additional genomic coverage of which approximately 15.2kb represented novel CDS. More than 80% of these novel splice features were expressed in foetal brain, albeit generally less highly than those in existing annotation. Similarly, comparing the overlap between existing and novel CDS annotation and Ensembl constrained elements in 39 mammals shows that while the former share an overlap of ~92%, in novel CDS sequence this drops to ~30%. Together, these results suggest that while overall the new annotation has different characteristics to existing annotation, it does capture previously un-annotated features that are both expressed and conserved, suggesting functionality. Using existing and updated GENCODE annotation to interpret variants for 1000 WES and 100 WGS from clinical datasets including patients with EIEE we identified a mean of ~61 novel genic variants and variants with changed consequences per WES and ~985 per WGS. Filtering for those with the most deleterious consequences identified 385 variants in WES and 83 in WGS that are currently undergoing more detailed follow-up analysis.

267

5-hydroxmethylcytosine-mediated epigenetic dysregulation in cerebellar degeneration. *B. Yao[1], H. Bao[1], L. Chen[6], L. Lin[1], M. Poidevin[1], S. Yang[1], X. Li[1], C. Stoyas[2], A. La Spada[2], F. Ayhan[3], L. Ranum[3], L. Duvick[4], H. Orr[4], Z. Zalewski[5], D. Nelson[5], H. Wu[6], P. Jin[1].* 1) Human Genetics, Emory University, Atlanta, GA; 2) Pediatrics, University of California San Diego, San Diego, CA; 3) Molecular Genetics and Microbiology, University of Florida, Gainesville, FL; 4) Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN; 5) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 6) Biostatistics and Bioinformatics, Emory University, Atlanta, GA.

Ataxia is a group of neurological disorders featuring uncontrolled movements and balance due to the common damage of cerebellar motor neurons termed Purkinje cells (PCs). 5-hydroxymethylcytosine (5hmC), which can be oxidized from 5-methylcytosine by Tet proteins, is highly enriched in mammalian brains such as PCs and play critical roles in brain. We previously showed aberrant 5hmC levels and distributions in a mouse model of Fragile X ataxia and tremor syndrome (FXTAS). To further understand the precise epigenetic roles of 5hmC in FXTAS pathogenesis, we crossed mice expressing PC-specific GFP-ribosomal protein with FXTAS mcie to isolate PCs from control and FXTAS mice by cell sorting and profile their genome-wide 5hmC. By integrating these 5hmC profilings with PC-specific translating ribosome affinity purification coupled with sequencing (TRAP-seq), we found that the aberrant 5hmC modification could potentially influence the expression of the gene related to the FXTAS pathogenesis, even before the disease onset. Mechanistically, we found that RNA-binding protein hnRNP A2/B1, which could be sequestered by extensive CGG repeats in FXTAS FMR1 mRNAs, also possesses DNA binding ability and is involved in epigenetic modulations. Chromatin-immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) of hnRNP A2/B1 specifically in PCs revealed their preferential binding to active promoters and coordinated with 5hmC oxidase Tet2. PC-specific overexpression of hnRNP-A2/B1 amliorated the PC degeneration in FXTAS mice. We then systematically characterized the genome-wide 5hmC distributions in several ataxia mouse models, including spinocerebellar ataxia (SCA) type 1, 7, 8 and 17, and found the 5hmC alteration was a common feature among these models. Disease-associated 5hmC alteration “hotspots” were identified, which strongly correlated with specific neuronal pathways. We identified a list of genes associated with ectopic 5hmC alterations across these ataxia models and comfirmed their contributions of ataxia pathogenesis by testing their ability to modify the neuronal toxicity in ataxia fly models. Taken together, our study reveal a novel epigenetic character of RNA-binding protein hnRNP A2/B1 in the pathologenesis of FXTAS, and indicate a general role of 5hmC-mediated epigenetic modulation in ataxia-related diseases.

268

Predicting haploinsufficiency from epigenomics data enables discovery of novel risk genes of developmental disorders.

Y. Shen[1,2], X. Han[1]. 1) Department of Systems Biology, Columbia Univ, New York, NY; 2) Department of Biomedical Informatics, Columbia Univ, New York, NY.

Haploinsufficiency (HIS) makes major contributions to the pathogenesis of various developmental disorders (DD). Precise prediction of haploinsufficient genes is critical for interpreting deleterious *de novo* variants and CNVs detected in DD exome or genome sequencing studies. Currently the main approach is to measure mutation intolerance based on depletion of rare deleterious variants in large population that do not have early onset developmental disorders. While highly effective in predicting HIS of sufficiently large genes, this approach has limited power for genes that are either small or with low background mutation rate. Moreover, there is no direct functional link between general mutation intolerance and a developmental disorder of a particular organ or tissue. Here we present an orthogonal approach utilizing large-scale epigenomic data to predict gene haloinsufficiency. We hypothesize that the expression of haploinsufficient genes during development is precisely regulated by a combination of transcription factors and epigenomic elements to reduce transcription noise, and such regulation can be detected by distinct patterns of epigenomic marks in relevant tissues and cell types. Based on this model, we developed a computational method that predicts haploinsufficiency by machine learning approaches using Roadmap Epigenomics project data as features. Using enrichment of *de novo* truncating mutations as a benchmark, we showed that this method achieved similar prediction accuracy as mutation intolerance metrics for neurodevelopmental disorders, and superior performance for structural birth defects. Furthermore, the predicted HIS genes are much less biased with gene size or background mutation rate. Finally, we show that the epigenomics-feature based predictions and mutation intolerance are complementary, and in combination one can maximize sensitivity and specificity in HIS prediction. Our method will facilitate discovery and interpretation of novel risk genes in developmental disorder genetic studies.

269

The majority of pathogenic copy number variations in congenital limb malformation affect non-coding regulatory elements.

M. Spielmann[1,2,11], R. Flöttmann[1], B. Kragsteen[2], S. Geuer[2], M. Socha[3], L. Allou[2], A. Sowińska-Seidler[3], J. Wagner[1], A. Jamsheer[3], B. Oehl-Jaschkowitz[4], D. de Silva[5], I. Kurth[6], I. Maya[7], F. Santos[8], W. Hülsemann[9], E. Klopocki[10], R. Mountford[12], G. Bork[13], D. Horn[1], P. Lapunzina[14], D. Duboule[15], S. Mundlos[1,2,11]. 1) Institute for Medical and Human Genetics, Charité -University Berlin,, Berlin, Germany; 2) Max Planck Institute for Molecular Genetics, Berlin, Germany; 3) Department of Medical Genetics, Poznan University of Medical Sciences, Poznan, Poland; 4) Gemeinschaftspraxis für Humangenetik Homburg/Saar, Homburg, Germany; 5) Department of Physiology, Faculty of Medicine, University of Kelaniya, Thalagolla Road, Ragama, Sri Lanka; 6) Institute of Human Genetics, Jena University Hospital, Friedrich-Schiller-University Jena, Jena, Germany; 7) Raphael Recanati Genetics Institute, Rabin Medical Center, Beilinson Hospital, Petah Tikva, Israel; 8) Pediatrics, Hospital Universitario Central de Asturias, Oviedo, Asturias, Spain; 9) Handchirurgie Kinderkrankenhaus Wilhelmstift, Hamburg, Germany; 10) Institute for Human Genetics, Biozentrum, Universität Würzburg, Würzburg, Germany; 11) Berlin-Brandenburg School for Regenerative Therapies (BSRT), Berlin, Germany; 12) Merseyside and Cheshire Regional Molecular Genetics Laboratory, Liverpool Women's NHS Foundation Trust, Liverpool, UK; 13) Institute of Human Genetics, University of Ulm, Ulm, Germany; 14) INGEMM (Instituto de Genética Médica y Molecular), Hospital Universitario La Paz-IdiPaz, Universidad Autónoma de Madrid, Madrid, Spain; 15) School of Life Sciences, Federal Institute of Technology, Lausanne, Switzerland; Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland.

Congenital limb malformations can occur as part of a syndrome or as an isolated form and are thought to be largely genetic in origin. The extensive genetic heterogeneity of these anomalies requires a genome-wide detection of all types of genetic variation. Studies in single affected families have previously demonstrated the importance of copy number variations (CNVs) in limb malformations, but no large scale study has been performed so far and the majority of cases remain undiagnosed. Here we applied high resolution copy number analysis to 350 patients with congenital limb malformations. All patients included had not received a molecular diagnosis after candidate gene testing. We found 41 pathogenic CNVs in known disease loci and identified 5 new loci previously not known to be associated with limb malformations. The pathogenic CNVs affected non-coding *cis* regulatory elements more frequently than expected (56%). We performed functional studies in transgenic mice using the CRISPR/Cas9 system and/or segregation studies in these families to investigate the pathogenicity of 5 novel CNVs causing limb defects. We reached a diagnostic yield of 12% in this cohort, which is comparable to copy number studies in other cohorts such as in individuals affected with intellectual disability. However, the majority of the pathogenic CNVs (56%) were likely to result from changes in the non-coding *cis* regulatory landscape, while only 44% were due to gene dosage effects or haploinsufficinency. We show that that CNVs have the potential to alter the topological associated domain (TAD) architecture of the genome by deleting or duplicating enhancer elements or misplacing TAD boundaries causing misexpression and disease. Our results suggest that CNVs affecting non-coding regulatory elements are a major cause of congenital limb malformations.

270

Subcutaneous adipose eQTLs coincident with GWAS loci identify 140 candidate target genes for cardiometabolic traits. *Y. Wu[1], M. Civelek[2,3], C.K. Raulerson[1], L.J. Scott[4], C. Pan[3], A. Ko[3], A. He[5], C. Tilford[5], C. Fuchsberger[4], A.E. Locke[4], H.M. Stringham[4], A.U. Jackson[4], N.K. Saleem[6], N. Narisu[7], P.S. Chines[7], J. Kuusisto[6], P. Gargalovic[5], T. Kirchgessner[5], P. Pajukanta[3], F.S. Collins[7], M. Boehnke[4], M. Laakso[6], A.J. Lusis[3], K.L. Mohlke[1].* 1) Department of Genetics, University of North Carolina, Chapel Hill, Chapel Hill, NC; 2) Department of Biomedical Engineering, University of Virginia, Charlottesville, VA; 3) Department of Medicine, University of California, Los Angeles, CA; 4) Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI; 5) Bristol-Myers Squibb, Pennington, NJ; 6) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland; 7) National Institutes of Health, Bethesda, MD.

Genome-wide association studies (GWAS) have identified hundreds of loci for cardiometabolic traits, but many of the underlying genes and mechanisms remain unclear. Mapping of expression quantitative trait loci (eQTLs) to these loci may prioritize candidate genes for cardiometabolic traits. To identify eQTLs that affect local gene expression and coincide with GWAS loci, we analyzed subcutaneous adipose tissue from 770 individuals from the METabolic Syndrome In Men (METSIM) study. Using 1,221 reported GWAS variants for type 2 diabetes, glycemic traits, lipids, obesity, metabolic syndrome and cardiovascular disease, we tested each variant for association with expression level of genes located <1 Mb. At a significance level of FDR<1% (equivalent $P<2.4\times10^{-4}$), we found that 611 (50%) of these GWAS variants were associated with expression level of ≥1 local gene. After LD-based clumping ($r^2>0.8$), 459 GWAS index variants were associated with 682 genes for a total of 944 GWAS variant-eQTL gene pairs. We next evaluated whether these eQTLs were coincident with the GWAS signals. Among the 944 adipose cis-eQTLs, 140 (15%) were coincident with a GWAS locus based on pairwise LD $r^2>0.8$ between the GWAS SNP and the lead eSNP (that exhibited the strongest association with the gene's expression level). Reciprocal conditional analyses at each GWAS variant and lead eSNP pair further supported the coincident signals (GWAS SNPs: all conditional $P>5.6\times10^{-3}$; lead eSNPs with conditional $P<2.4\times10^{-4}$: all $\Delta\log_{10}(P)>8.5$). Of these coincident eQTLs, 93 (66.4%) had not been previously described in large-scale genome-wide association studies that interrogated available resources for eQTLs. These eQTL genes are potential targets of the GWAS variants at loci for T2D (n=6 novel eQTL genes), glycemic traits (n=11), lipids (n=39), obesity (n=13), metabolic syndrome (n=7) and cardiovascular disease (n=21). Among the GWAS loci with coincident eQTLs, 80 had one eQTL gene while 26 contained ≥2 eQTL genes per GWAS locus. In addition, the expression level of 49 of these 140 eQTL genes was associated with the level of the corresponding GWAS trait at $P<0.05$, suggesting biologically plausible roles of the genes in mediating the variant-trait association. Taken together, our finding highlight 140 plausible target genes at 106 GWAS loci and improve the understanding of the regulatory effects and adipose biology in cardiometabolic risk.

271

Common pan-cancer pathways and gene sets can identify tumors across over 30 cancer types. *F.C. Lamaze[1], M. Agbessi[2], J.C. Grenier[2], V. Bruat[2], P. Awadalla[1,2], PCAWG consortium.* 1) OICR, Toronto, Ontario, Canada; 2) Sainte-Justine Hospital, Montreal, Quebec, Canada.

Extensive genetic and phenotypic variations exist within and between cancers representing a challenge for the discovery of biomarkers and personalised oncology. While some loci are now "canonical" or common, and despite some evidence for functional convergence in cancer development (e.g. metabolic pathways and cell cycle), common biomarkers identification have largely been unexplored to define similar phenotypes in lesions that are yet genetically divergent. Paired tumour and normal tissues biopsied from over 1700 patients, including whole transcriptomes and genomes, across 31 different cancers from the Pan Cancer and TCGA consortium were used to define a common set of transcripts and pathways across all of these cancers. The value of these tools is that we can now predict with high confidence the status of any given sample (tumor vs. normal), regardless of cancer type. First we used a 1000x resampling strategy to identify a common set of 322 deregulated genes. These genes are largely functionally enriched for cell cycle, signalling and recombination pathways or processes. We used a machine learning strategy where the entire sample set was split into training (n = 1438) and comparable prediction (n = 958) sets. We were able to further refine our gene-set to only 162 genes which segregate with high predictive accuracy (98%), specificity (99%) and sensitivity (93%) tumors from normal tissue biopsies across all cancers. For each cancer, accuracy, sensitivity and specificity also remained high - all above 97%, 93% and 100% (1st quartile), respectively. We were also able to show that this gene set was highly predictive beyond humans, correctly identifying tumor vs. normal tissues in other mammalian cancers including Tasmanian devils. The selective differential expression of these biomarkers in tumor compared with paired normal tissues in mammals suggests that they resume common differentiation pathways during "mammalian" carcinogenesis. This stable expression in tumor can be used in un-described, or rare cancers, or when the differentiation markers are unreliable. Finally, these biomarkers could be used for increasing the accuracy of the early detection diagnosis and stratification of tumor types in the clinic.

272

Longitudinal next generation sequencing (NGS) of plasma cell-free DNA (cfDNA) during a phase 1 basket trial of targeted AKT inhibitor (AZD5363) reveals clinical correlates and emergent somatic mutations. *J. Reichel[1], L. Smyth[2], J. Tang[1], F. Meng[1], J. Patel[1], S.D. Selcuklu[1], D. You[1], A. Samoila[3], S. Chandarlpaty[4], M. Berger[1,4].* 1) Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY; 2) Breast medicine service and developmental therapeutics service, Memorial Sloan Kettering Cancer Center, New York, NY; 3) Laboratory Medicine, Memorial Sloan Kettering Cancer Center; 4) Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center.

Cancer is a heterogenous disease where somatic mutations evolve over time as tumor cells grow and divide and respond to selective pressure from drug therapies. This evolutionary plasticity poses a significant challenge for the acquisition of complete genomic cancer profiles; tissue biopsies are derived from a limited subset of cells rather than the full tumor, and often are sampled once, rather than throughout a patient's course of therapy. Incomplete genomic assays can stymie the long-term efficacy of treatments, and blind clinicians and researchers to evolutionary dynamics such as acquired drug resistance. Short fragment cell-free DNA (cfDNA) is released by cells undergoing apoptosis, necrosis, and/or vesicular budding, and can be profiled using next-gen sequencing techniques. Compared to tissue sequencing, the profiling of the circulating tumor DNA (ctDNA) component of total cfDNA is less invasive, more readily available, and may provide greater insight to the full heterogeneity of disease. The E17K somatic mutation of *AKT1* is a recurrent, oncogenic event in solid tumors. Little is known about genomic mutations coincident with targeted inhibition of AKT. Therefore, we performed longitudinal profiling of 410 genes (MSK-IMPACT) on 78 plasma cfDNA specimens from a cohort of 25 MSKCC patients harboring E17K-mutant solid tumors receiving AZD5363, an oral catalytic selective pan-AKT inhibitor, as part of the multi-center basket clinical trial NCT01226316. Tumor tissues specimens were also sequenced pre-treatment and upon progression, whenever possible. The variant allele fraction (VAF) of *AKT1* E17K could be correlated with tumor response, with durable radiological responses seen in 5/5 (100%) pts with persistent (>21 days) cfDNA clearance. The most frequently mutated genes in addition to *AKT1* were ESR1 (27%), TP53 (23%) and GATA3 (18%) with an average of 4 muts per patient. Pre-Tx cfDNA NGS revealed muts not observed in the archival specimen in 4/9 (44%) pts. 7 muts emerged on NGS of on-study and progression cfDNA, which were not observed or found at low abundance in pre-Tx cfDNA, including muts in *GRIN2A*, *TP53* and *CEBPA*, among others. Second site muts in *AKT1* were not observed.

273

Identification of biomarkers to personalize treatment after resection surgery for stage I adenocarcinoma. *A. Clemenceau[1], P. Joubert[1], Y. Bossé[1,2].* 1) Institut universitaire de cardiologie et de pneumologie de Québec, Québec, Canada; 2) Department of Molecular Medicine, Laval University, Québec, Québec, Canada.

Background and objective: Lung cancer is the leading cause of cancer-related death. Surgical resection for early stages lung cancer remains the best hope for a cure. However, 30 to 50% of surgically-treated patients will relapse. We have no marker or tool to predict relapse and remission after surgery. **The objective of this study is to identify gene expression biomarkers associated with lung cancer survival after resection for stage 1 adenocarcinoma**.**Methods:** Candidate genes were selected based on literature review and analyses performed in public databases (PRECOG) as well as our own microarray gene expression dataset in lung cancer and non-tumor pulmonary parenchyma collected at 0, 2, 4 and 6 cm from the lesion. Selected genes were measured by quantitative real-time PCR (qPCR) in tumor of 244 patients with stage 1 adenocarcinoma. Kaplan-Meier analyses were performed to establish the discriminatory performance of these biomarkers.**Results:** Based on literature and analyses performed in PRECOG, we have selected 10 candidate genes that showed promising prognostic value. Complementary analyses with our own microarray dataset enabled us to choose three genes associated with poor outcome (*RRM1*, *EZH2* and *FOXM1*) and two genes associated with favourable outcome (*BTG2*, *SELENBP1*). Pathological stages (stage 1A and 1B) were significantly associated with survival in our series of 244 patients (Kaplan-Meier log-rank p=0.013). Preliminary qPCR results for *BTG2* and *RRM1*, in the same series revealed modest, but statistically significant differences in survival curves between patients with low compared to high gene expression. Suggestive differences in survival curves were also observed for *EZH2*. qPCR analysis for *FOXM1* are in progress.**Conclusion:** New prognostic tools are urgently needed to help clinicians guide adjuvant therapy following early-stages lung cancer surgery. Our preliminary results support *BTG2* and *RRM1* as a good predictor of prognosis for stage I adenocarcinoma. Additional gene expression biomarkers must be evaluated (e.g. *FOXM1*). Combinations of genes and molecular phenotypes (mRNA, methylation marks, etc) must also be evaluated to differenciate high-risk patients beyond clinicopathologic staging.

274

Whole exome sequencing of metastatic colorectal and lung cancers: Advances and challenges in interpretation of somatic and germline variants. *A. Ghazani[1,2,3], N. Oliver[1,2], J. St. Pierre[1], A. Garofalo[3], L. Sholl[4], N. Lindeman[4], J. Garber[5], S. Joffe[6], P. Jänne[1], S. Gray[1], L. Garraway[1,2,3], E. Van Allen[1,2,3], N. Wagle[1,2,3].* 1) Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA; 2) Center for Cancer Precision Medicine, Dana-Farber Cancer Institute, Boston, MA; 3) Broad Institute of MIT and Harvard, Cambridge, MA; 4) Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA; 5) Center for Cancer Genetics and Prevention, Dana-Farber Cancer Institute, Boston, MA; 6) Department of Medical Ethics and Health Policy, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA.

Implementing precision medicine in oncology requires an integrated approach for the interpretation of both somatic and germline variants. We describe a systematic approach to integrated genomic interpretation and reporting in the CanSeq program at the Dana-Farber Cancer Institute and the Broad Institute. CanSeq is a prospective study of patients with metastatic colorectal and lung adenocarcinoma. Participants undergo whole exome sequencing (WES) of tumor and germline samples, and clinically relevant results are returned to the clinician and patient. To date, 229 adults with metastatic colorectal and lung adenocarcinomas have enrolled in the study. WES (Illumina, HiSeq 2500) was performed on both formalin-fixed paraffin-embedded (FFPE) tumor biopsy and fresh peripheral blood from each patient. All genome variants were ranked ordered based on clinical and biological relevance by Precision Heuristics for Interpreting the Alteration Landscape (PHIAL). Genome variants were interpreted using an integrated somatic and germline framework. The results were then returned to the clinical team. Initially, the process of review and return of results were made by a molecular tumor board, and ultimately evolved to a protocol-based interpretive approach to allow scalability. Tumor and blood samples for 165 of the 229 patients have, so far, been sequenced, interpreted and returned to the clinical team. Among genome variants, 31% (511/768 somatic) and 5% (43/806 germline) variants were associated with clinical evidence. For the remaining 69% of the somatic variants and 95% of the germline variants, a well-established interpretation framework does not exist. A total of 43 germline findings were reported to 35 patients. Among these 43 reported germline variants, 44% were classified as pathogenic, 17% were categorized as likely pathogenic, and the remaining 39% were considered variants of unknown significance. Overall, 71% of the reported germline variants were found in cancer-related genes and 29% in non-cancer related genes. In total, 91% of patients had an actionable or potentially actionable somatic alteration and 21% had at least one reported germline alteration. Variants of unknown clinical significance constitute a large proportion of the findings both in somatic and germline components of WES. The systematic approach to interpretation of somatic and germline variants developed within CanSeq can help identify potentially actionable genomic alterations in WES. .

275

ADHD risk loci identified by genome-wide association meta-analysis. *D. Demontis, the iPSYCH-Broad ADHD Workgroup and the Psychiatric Genomics Consortium: ADHD Subgroup.* Biomedicine, Aarhus University, Denmark, Aarhus, Denmark.

Attention-deficit hyperactivity disorder (ADHD) is a highly heritable childhood behavioural disorder affecting 3-6% of school-age children and ~4% of adults. Several moderately sized genome-wide association studies (GWASs) have been performed, but until now no single markers have passed the threshold for genome-wide significance. The SNP heritability of ADHD has been estimated to 0.28, indicating that common SNPs contribute substantially to ADHD susceptibility and that increasing GWAS sample sizes is likely to produce significant results. Here we present the first genome-wide significantly associated loci with ADHD from a large-scale meta-analysis of GWASs of ADHD. This is based on collaboration between the Danish *i*PSYCH initiative, the Broad Institute, and the Psychiatric Genomics Consortium (PGC). The study includes nine PGC samples (totalling ~4,400 cases and ~10,300 controls) and the iPSYCH sample (~14,500 cases and ~22,400 controls), resulting in a grand total of ~18,900 cases and ~32,700 controls included in the meta-analysis. The meta-analysis identified 10 independent genome-wide significant loci, revealing new and important information on the underlying biology. Furthermore, gene-based analysis (using MAGMA) identified several significantly associated genes, some of which are located outside the 10 identified loci. Genetic correlation between the nine PGC samples and the iPSYCH sample was very high (rg = 0.77; P =9.25x10-41 (estimated using LD score regression)). The SNP heritability was estimated to 0.23, and when using cell type specific annotations a significant enrichment in the heritability of central nervous system specific annotations was found. Additional information about potential functional effects was obtained by identifying significant differences in gene expression between cases and controls in various brain tissues using PrediXcan and Metaxcan, which was used to impute gene expression for the genotyped individuals. Furthermore, genetic correlation estimates, based on summary statistics from our ADHD meta-analysis and summary statistics from GWASs of childhood intelligence (Benyamin et al. Mol. Psych. 2014) and educational attainment (Okbay et al., Nature 2016), found a strong negative genetic correlation of ADHD with the two measures of cognitive performance. The results presented here represent a substantial progress in our understanding of the genetic architecture of ADHD.

276

Genome- and phenome-wide study of "nail biting": Not just a habit. *C. Tian, J. Tung, D. Hinds.* Research, 23andme, Mountain View, CA.

We describe the discovery of genetic and phenotypic associations with "nail biting," technically known as onychophagia. Over 180,000 participants who consented to research in the 23andMe customer base responded to the question "How often do you bite your nails"; 37% reported biting their nails and 7% said they bite very often. Consistent with the literature, "nail biting" was correlated with "conscientiousness" and "neuroticism" of our five dimensional personality questionnaires. Individuals who become nervous easily or are moody report a higher frequency of nail biting. Our genome-wide scan identified 21 significant associations ($p < 5e-8$) with nail biting. We identified a loss of function variant (rs117612447, p=4.6e-22) in *KRT31*, a keratin gene involved in hair and nail formation, and a variant (rs10876505, p=5.5e-9) near *HOXC13*, a gene linked to nail and hair developmental disorders. Six of the identified loci (rs713843, p=4.2e-26; rs35754740, p=4.8e-11; rs4776970, p=7.4e-11; rs4775313, p=8.4e-11; rs62264775, p=9.4e-9; rs149994299, p=2e-8) were also associated with BMI in the same direction. Five of the identified loci (rs1442883, p=3.8e-19; rs8095324, p=1.7e-13; rs7837754, p=3.3e-12; rs7411445 [*NEGR1*], p=8.6e-10; rs2977694 [*CSMD1*], p=7.2e-8) were also associated with "sweet tooth," but in the different directions. The *NEGR1* and *CSMD1* regions that have been previously implicated in psychiatric disorders. We also identified variants near *GRIN2A* in 16p13.2 (rs2014151, p=6e-19) and near *NRG1* (rs13255543, p=5.7e-13). Mutations in these two regions have previously been linked to diseases such as autism, schizophrenia, and bipolar disorder. We estimated a positive genetic correlation between nail biting and BMI (LD score rg=0.17, p=1.46e-14). We found a near-zero genetic correlation between nail biting and sweet tooth. Although they shared many associations, the effects from those pleiotropic loci are not in the same direction. Overall, our findings revealed genetic contributions to nail biting. They also point to a possible connection between nail biting, BMI, and taste perception, which is interesting in light of prior findings that BMI GWASes implicate neural regulations; personality factors such as anxiety and the ability to cope with stress have been discovered to change hormones and act on taste. Our study may provide molecular evidence for neural mechanisms underlying personality and taste.

277

Dysregulation of RNA splicing in developing brains increases risk for neuropsychiatric disease. *Y. Li*[1]*, T. Raj*[2]*.* 1) Department of Genetics, Stanford University, Stanford, CA; 2) Departments of Neuroscience, Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY.

Most neuropsychiatric risk loci identified by genome-wide association studies (GWAS) localize in non-coding regions of the genome. A popular hypothesis is that risk variants function by altering aspects of gene regulation. We recently showed that genetic variants that affect RNA splicing (splicing QTLs, or sQTLs), alongside those that affect gene expression levels, underlie a substantial fraction of the genetic associations to disease (Li et al., 2016). Motivated by the high diversity of splicing events in human brains compared to other tissues (Barbosa-Morais et al., 2012), we aimed to characterize the impact of aberrant RNA splicing on neuropsychiatric diseases.   There is a strong association between neuropsychiatric diseases and brain development, we therefore identified splicing "switches" between pre- and post-natal brains and searched for disease-associated genes with evidence of switching. To do this we used a novel splicing quantification method, LeafCutter (Li et al., 2016), to analyze RNA-seq data from 36 post-mortem prefrontal cortex (PFC) samples spanning brain development, maturation and aging (Jaffe et al., 2015). This revealed hundreds of splicing events that are developmental-specific, many of which are complex and previously undetected. As expected, these developmentally specific splicing events were enriched in genes associated with autism and schizophrenia. Interestingly, we found that a large number of developmental-specific micro-exons, i.e. exons that are shorter than 51nt, (Li et al., 2015) were highly enriched among disease genes. Indeed, alternative inclusions of micro-exons are believed to alter protein-protein interactions and were previously shown to be dysregulated in autism (Irimia et al., 2014).   To establish causal links between dysregulation of RNA splicing and disease, we analyzed the RNA-seq profiles of 461 PFC from the AMP-AD consortium and identified common genetic variants associated with RNA splicing. We identified over five thousand sQTLs at 1% FDR. Using MetaXcan (Alvero et al., 2016), to associate sQTL loci to neuropsychiatric risk, we identified 39 genes whose splicing patterns are associated with schizophrenia, including the Complement component 4B (C4B) and the fragile X mental retardation gene (FXR1). Several associated splicing events show dynamic changes during brain development, which supports a mechanistic link between dysregulation of RNA splicing in developing brains and neuropsychiatric disease.

278

Patient-derived iPSC model of an *ABCA7* deletion associated with Alzheimer disease. *H.N. Cukier[1,2], S.P. Gross[1], B.W. Kunkle[1], B.N. Vardarajan[3], S. Rolati[1], K.L. Hamilton-Nelson[1], P.L. Whitehead[1], B.A. Dombroski[4], R. Lang[5], L.A. Farrer[6], M.L. Cuccaro[1,7], J.M. Vance[1,2,7], J.R. Gilbert[1,7], G.W. Beecham[1,7], E.R. Martin[1,7], R.M. Carney[1,7], R. Mayeux[3], G.D. Schellenberg[4], G.S. Byrd[5], J.L. Haines[8], M.A. Pericak-Vance[1,2,7], D.M. Dykxhoorn[1,7].* 1) John P. Hussman Institute for Human Genomics, University of Miami, Miller School of Medicine, Miami, FL; 2) Department of Neurology, University of Miami, Miller School of Medicine, Miami, FL; 3) The Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Gertrude H. Sergievsky Center, Departments of Neurology, Psychiatry, and Epidemiology, College of Physicians and Surgeons, Columbia University, New York, NY; 4) Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 5) Department of Biology, North Carolina A&T State University, Greensboro, NC; 6) Departments of Medicine, Neurology, Ophthalmology, Genetics & Genomics, Epidemiology, and Biostatistics, Boston University, Boston, MA; 7) Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, FL; 8) Department of Epidemiology and Biostatistics, Institute for Computational Biology, Case Western Reserve University School of Medicine, Cleveland, OH.

The *ATP-binding cassette, sub-family A (ABC1), member 7* (*ABCA7*) gene has been implicated as a risk factor in Alzheimer disease (AD) across populations including African American (AA), Asian, and non-Hispanic white (NHW). However, the effect is significantly stronger in AA than in the Asian and NHW populations. While some rare loss-of-function *ABCA7* variants have been identified in NHWs, we recently identified a relatively common 44 base pair deletion (rs142076058) in African Americans significantly associated with disease ($p=1.414x10^{-5}$, OR=1.81 [95% CI:1.38-2.37], Cukier, et al, 2016). The deleted allele is predicted to produce a frameshift mutation (p.Arg578Alafs) resulting in a truncated protein. To further understand the mechanism by which the *ABCA7* deletion may be acting, iPSC lines were developed from two unrelated AA AD individuals bearing the deletion, as well as two gender, aged, and ethnical matched control individuals. Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood and reprogrammed by transducing with Sendai virus vectors expressing the Yamanaka factors. Two clones from each individual were generated, validated for pluripotency, and karyotyped to ensure that no gross chromosomal abnormalities were present. The iPSC lines were cultured first as embroid bodies, then neural rosettes, and finally as a monolayer in media supplemented with small molecules to promote cortical neuronal differentiation. RNA purified from both PBMCs and neuronal cells from AD individuals produce a stable RNA transcript from the *ABCA7* allele with the deletion. iPSC-derived cortical neurons from cases and controls were then examined by an enzyme-linked immunosorbent assay (ELISA) to measure levels of pathogenic b40-amyloid at day 45 and 95 of differentiation. While higher levels of b40-amyloid were found in the older neurons (day 95) compared to younger neurons (day 45), the AD cases had higher levels of b40-amyloid compared to controls. In conclusion, this deletion in *ABCA7* may represent an ethnic-specific pathogenic alteration in AD resulting in impaired APP processing and increased production of toxic b-amyloid production.

279

iPS cells-based pathophysiological investigation for large subcutaneous hematomas in Ehlers-Danlos syndrome caused by CHST14/D4ST1 deficiency. *T. Kosho[1], F. Yue[2], T. Era[3], J. Nakayama[4], T. Yamaguchi[1], N. Miyake[5], S. Mizumoto[6], S. Yamada[6], R. Kawamura[1], K. Wakui[1], T. Yoshizawa[7], Y. Takahashi[1], K. Matsumoto[7], T. Hirose[8], J. Minaguchi[8], K. Takehana[8], M. Uehara[9], J. Takahashi[9], M. Ishikawa[1], C. Masuda[10], S. Shimazu[10], Y. Nitahara-Kasahara[10], A. Watanabe[10], T. Okada[10], K. Matsumoto[11], Y. Nomura[12], Y. Kakuta[13], A. Hatamochi[14], Y. Fukushima[1], K. Sasaki[2].* 1) Department of Medical Genetics, Shinshu University School of Medicine, Matsumoto, Japan; 2) Department of Histology and Embryology, Shinshu University School of Medicine, Matsumoto, Japan; 3) Department of Cell Modulation, Institute of Molecular Embryology and Genetics, Kumamoto University, Kumamoto, Japan; 4) Department of Molecular Pathology, Shinshu University Graduate School of Medicine, Matsumoto, Japan; 5) Department of Human Genetics, Yokohama City University Graduate School of Medicine, Yokohama, Japan; 6) Department of Pathobiochemistry, Faculty of Pharmacy, Meijo University, Nagoya, Japan; 7) Division of Laboratory Animal Research, Research Center for Human and Environmental Sciences, Shinshu University, Matsumoto, Japan; 8) Department of Veterinary Pathology, School of Veterinary Medicine, Rakuno Gakuen University; 9) Department of Orthopaedic Surgery, Shinshu University, School of Medicine, Matsumoto, Japan; 10) Department of Biochemistry and Molecular Biology, Nippon Medical School, Tokyo, Japan; 11) Department of Biosignaling and Radioisotope Experiment, Interdisciplinary Center for Science Research, Organization for Research and Academic Information, Shimane University, Izumo, Japan; 12) Scleroprotein and Leather Research Institute, Tokyo University of Agriculture and Technology, Faculty of Agriculture, Tokyo, Japan; 13) Department of Bioscience and Biotechnology, Faculty of Agriculture, Kyushu University, Fukuoka, Japan; 14) Department of Dermatology, Dokkyo Medical University, Mibu, Japan.

Carbohydrate sulfotransferase 14 (CHST14)/dermatan 4-*O*-sulfotransferase-1 (D4ST1) deficiency represents a clinically recognizable form of Ehlers-Danlos syndrome (EDS), also named as musculocontractural type EDS type 1. It is caused by biallelic mutations in *CHST14* and is clinically characterized by multiple congenital malformations (craniofacial features, multiple congenital contractures) and progressive multisystem fragility-related complications (skin hyperextensibility and fragility, recurrent dislocations and progressive talipes or spinal deformities, large subcutaneous hematomas). Pathological and glycobiological studies on affected skin specimens suggested multisystem fragility to be caused by impaired assembly of collagen fibrils resulting from loss of dermatan sulfate (DS) in the decorin glycosaminoglycan side chain that promotes electrostatic binding between collagen fibrils. An ongoing international collaborative study describes that large subcutaneous hematomas are frequent (85%) and one of the most serious complications, typically occurring after minor traumas, spreading in several hours with severe pain, and sometimes accompanying hemorrhagic shock. Intranasal administration of desmopressin, with hemostatic and vasoconstrictive properties, efficiently reduces the progression in some patients. Hypothesizing that large subcutaneous hematomas in the disorder would be attributable to impaired contraction followed by rupture of small-sized arteries caused by the structural fragilities, we performed induced-pluripotent stem cells (iPSCs)-based pathophysiological studies. After validating undifferentiation status and pluripotency of iPSCs derived from cultured skin fibroblasts of three patients and a healthy subject, vascular smooth muscle cells (VSMCs) were induced. Significantly reduced contraction of VSMCs from patient specific iPSCs was observed after stimulation by a muscarinic agonist or a calcium agonist, compared with VSMCs from normal iPSCs. Impaired vascular formation was observed on a Matrigel where patient iPS-derived VSMCs and human endothelial cells (HUVEC) were transplanted together into SCID mouse, compared with normal iPS-derived VSMC transplantation. These results would support the hypothesis, which also implicates an important role of DS in the maintenance of arterial structure in humans.

280

A *Drosophila melanogaster* model of 16p11.2 deletion supports a complex *cis* and *trans* interaction model for neurodevelopmental disorders. *M. Singh, J. Iyer, L. Pizzo, M. Jensen, P. Patel, E. Huber, S. Girirajan.* Department of Biochemistry and Molecular Biology, The Pennsylvania State University, State College, PA.

Rare CNVs such as the 16p11.2 deletion are associated with extensive phenotypic heterogeneity, complicating disease gene discovery and functional evaluation. We used the powerful genetic system of *Drosophila melanogaster* and a series of quantitative methods to assay the phenotype, function, and interactions of fly orthologs of 16p11.2 genes. Using the UAS-Gal4 system and RNA interference, we evaluated phenotypes for 24 fly lines (representing 14 orthologs) in a tissue-specific manner. For example, using the GMR-Gal4 driver, we were able to identify seven orthologs that showed rough eye phenotypes in a dosage-dependent manner. Of these genes, we prioritized four genes (*KCTD13, MAPK3, DOC2A, and PPP4C*) and performed deeper mechanistic experiments to identify cellular phenotypes. Neuronal proliferation assays on the fly imaginal discs suggested increased mitosis for the *KCTD13* line and decreased mitosis for the *MAPK3* fly line. We then generated recombinant lines to investigate the effect of gene interactions on the observed cellular and developmental phenotypes. We performed ~33 *cis* and 240 *trans* interaction studies to identify genes that modify the effect of knockdown of the selected orthologs. For example, we were able to rescue the rough eye phenotypes in knockdown fly models of both *MAPK3* and *KCTD13* by also knocking down *ALDOA*. Similarly, combined knockdown of *MAPK3* and *PTEN* rescued the eye phenotypes due to *MAPK3* knockdown alone. This suggests that reduced dosage of *MAPK3* contributes to a defect in the insulin-signaling pathway, potentially providing a functional basis of phenotypes in 16p11.2 deletion. These observations of rescuing the eye phenotype were also validated by proliferation assays using phosphorylated histone antibodies and bromouridine staining. We also performed RNA sequencing of fly brain samples from six select orthologs and prioritized 15 differentially expressed genes for functional validation. Interaction studies with fly lines of the selected downstream targets identified 10 genes that rescued *MAPK3* phenotypes and three genes that rescued *KCTD13* phenotypes. Integration of functional data and RNA sequencing data also allowed us to identify novel interacting partners in the neurodevelopmental gene network that have potential disease relevance. Our result suggest a complex interplay of *cis* and *trans* interaction contribute to the ultimate phenotype in individuals with 16p11.2 deletion .

281

Lymphoproliferation due to *Fas-ligand* defects in cats. A potential model for autoimmune lymphoproliferative syndrome (ALPS). *L.A. Lyons[1], D. Aberdein[2], J.S. Munday[2], B. Gandolfi[1], K.E. Dittmer[2], R. Malik[3], 99 Lives Consortium.* 1) College of Veterinary Medicine, University of Missouri, Columbia, MO; 2) Pathobiology, Institute of Veterinary, Animal, and Biomedical Sciences, Massey University, Palmerston North, New Zealand; 3) Centre for Veterinary Education, University of Sydney, Sydney, NSW, 2006, Australia.

Autoimmune lymphoproliferative syndrome (ALPS) is a rare non-neoplastic lymphoproliferative disease typically seen in infants or young children with diverse racial background. The lymphoproliferation results in marked lymphadenopathy and splenomegaly with autoimmune cytopenias often present in affected individuals. The majority of patients with ALPS have germline defects in the *FAS* gene, although defects in the *FAS ligand* (*FASL*) or *caspase 10* also cause ALPS. The phenotype is segregating and detected in over 300 families worldwide, however, the majority of the patients remains undiagnosed or misdiagnosed. Patients with ALPS are also predisposed to developing malignancies with 10 - 20% of affected individuals subsequently developing lymphoma. The advancement of cat genomics and the 99 Lives cat genome initiative are demonstrating the feasibility and efficiency of discovering new feline biomedical models. Recent successes have included two inherited blindnesses, osteoarthritis, and Congenital Myasthenia Syndrome models. British shorthair kittens in multiple litters developed massive lymphadenopathy before 8 weeks of age that was similar to human ALPS and an autosomal recessive disorder. Consistent with ALPS, the lymphadenopathy was due to a non-neoplastic proliferation of circulating CD3+/CD4–/CD8– ‘double negative’ T-lymphocytes, which were also abundant in the lymph nodes. The whole genome of two affected kittens was sequenced and compared to 82 cat genomes, including 9 wild felids. Over 3000 variants were unique to the affected individuals and concordant with the phenotype, 15 were predicted to have an effect on protein products. Both BSH kittens had homozygous insertions of adenine on cat chromosome F1 within exon 3 of *FAS-ligand*. The resultant frame-shift and premature stop codon was predicted to result in a severely truncated protein that is unlikely able to activate FAS. Homozygous variants were subsequently identified in three additional affected BSH kittens while heterozygous variants were found in 10 of 15 unaffected, but closely related BSH cats. All BSH cats in the study were from a population with significant inbreeding. The variant was not identified in a survey of 510 non-BSH cats. Identification of a genetic defect in the FAS-mediated apoptosis pathway confirms that the lymphoproliferative disease in BSH cats fulfills the diagnostic criteria for ALPS in humans, suggesting that cats are a potential animal model for ALPS.

282

Knockdown of chromatin remodeling gene *Cecr2* causes subfertility in both male and female mice, but by different mechanisms. *H.E. McDermid, C.B. Weatherill, K. Rowel Lim, V.V. Nguyen, R.C. Humphreys, K.A. Norton.* Biological Sciences, University of Alberta, Edmonton, Alberta, Canada.

Mammalian reproduction requires a complex interplay of genes which must be regulated spatially and temporally. Part of this process depends on chromatin remodellers, which are able to affect nuclear processes through modulation of chromatin structure. Mutations in chromatin remodelling gene *Cecr2* result in reproduction defects in both male and female mice. **Males:** Knockdown of *Cecr2,* which is expressed in spermatogonia, results in an unusual form of male subfertility that is most severe at maturity and improves with age. Over the first 2 months of sexual maturity, mutant male litter sizes improve from 11.7% to 58.3% of their wildtype brothers. Histological analysis reveals severe defects in the seminiferous tubules of newly mature (6 week) males, including tubules that have very few cell layers and are not completing spermatogenesis. These defects become less prevalent and less severe with age, and at 3+ months testes appear close to normal in structure. Testes appear normal at birth and at 2 weeks postnatal, suggesting a defect in progression from spermatogonia to spermatocyte. Additionally, analysis of wildtype eggs, fertilized *in vivo* and collected 5 hours later, indicates that subfertility in mutants is due to fewer oocytes being fertilized. Mutants males aged 42-60 days show a fertilization rate of ~5% compared to their wild type brothers. At 60-100 days of age the mutant males show ~35% of the normal fertilization rate, and after 100 days the fertilization rate is not significantly different between mutants and wild types. This subfertility defect may be alleviated by an increase in *Cecr2* transcript from the mutant allele. **Females:** Mutant females crossed with normal males show litter sizes reduced by ~50% compared to their normal female littermates. However, ovarian histology is normal, as is the number of oocytes fertilized and percent of embryos surviving to the blastocyst stage in culture. There is also no evidence of increased late embryonic death. This suggests that embryo loss in *Cecr2* mutants occurs just before or soon after implantation, and that the mutant female reproductive system may be less hospitable to the obligate heterozygous embryos. **Conclusion:** The knockdown of *Cecr2* results in subfertility in both male and female mice, but the phenotype suggests that the mechanism by which this occurs is different. In males the production of sperm is directly affected, while in females the defect occurs after fertilization in the reproductive tract.

283

Common genetic variation drives molecular heterogeneity in human iPSCs. *H. Kilpinen[1], A. Goncalves[2], D. McCarthy[1], A. Leha[2], D. Bensaddek[3], A.I. Lamond[3], R. Durbin[2], D.J. Gaffney[2], O. Stegle[1] on behalf of the HipSci Consortium.* 1) European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI); Wellcome Genome Campus, Cambridge, UK; 2) Wellcome Trust Sanger Institute; Wellcome Genome Campus, Cambridge, UK; 3) Centre for Gene Regulation & Expression, University of Dundee; Dundee, UK.

Induced pluripotent stem cells (iPSC) are increasingly used to model functional effects of human disease alleles. However, variable characterization of many existing iPSC lines limits their use for research. The Human Induced Pluripotent Stem Cells Initiative (www.hipsci.org) has generated a reference panel of iPSCs to understand the heterogeneity and sources of biological variability in these lines in the context of genetic changes. We analyzed the first 522 iPSC lines from 189 healthy individuals in a framework that captures variability across individuals, lines, and single cells. We show that donor effects are the major driver of cellular heterogeneity in iPSCs, and present the first map of regulatory variants affecting the transcriptome of pluripotent stem cells (PSC). We identified 2169 genes with an eQTL in *cis* (FDR 5%) of which 503 were specific to iPSCs. This proportion of tissue-specific eQTL (23%) was higher than in 44 tissues studied by the GTEx Project (except in the testis). A subset of the iPSC-specific eQTL affected also protein abundance in 50 donors. The majority of iPSC-specific eQTL (77%) were driven by variants distinct from effects in other tissues, with only 6% explained by stem cell specific gene expression. iPSC-specific eQTL were highly enriched in active enhancers and poised promoters in PSCs, as well as in distal binding sites of key regulators of pluripotency. iPSC eQTL tagged 45 loci associated with complex traits at which the eQTL effect size was larger in iPSCs than in somatic tissues. Additional 8 loci were tagged by iPSC-specific eQTL, e.g. rs10069690, lead eQTL variant for the *Telomerase Reverse Transcriptase* (*TERT*) associated with germline predisposition to multiple cancers. We associated rs10069690 with alternative splicing of *TERT* in iPSCs, and hypothesize that the eQTL may lead to genotype-dependent variability in telomerase activity in somatic cells, affecting cancer susceptibility. In summary, we show that variation in iPSC gene regulation is similar to that in somatic tissues, with lineage-specific gene expression driven by distal regulatory elements. We identified eQTL that show regulatory function only in iPSCs and report disease-associated loci tagged by them. These loci may drive disease susceptibility through molecular changes early in development, not captured by adult tissues. Thus, iPSCs can be used to study genetic effects of traits that manifest in transient states during cellular growth and differentiation. .

284

Inter-individual variation in epigenetic marks between human induced pluripotent stem (iPS) cell lines. *A.T. Filimon Goncalves[1], P. Danecek[1], A. Leha[1], H. Kilpinen[2], R. Durbin[1], O. Stegle[2], D. Gaffney[1], HipSci Consortium (http://www.hipsci.org/).* 1) Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 2) EMBL EBI, Cambridge, United Kingdom.

Human induced pluripotent stem cells (hIPSCs) vary substantially at molecular and functional levels, particularly in their ability to differentiate into alternative cell lineages. Here we describe the initial analysis of 522 hIPSCs derived from 189 healthy individuals by the Human Induced Pluripotent Stem Cell Initiative. We used variance component analysis to estimate the sources of variation across a range of hIPSC molecular and functional phenotypes. Most of the non-technical variance we observed, including the ability of an hIPSC line to differentiate into the three germ layers, was explained by differences between donor individuals. This result suggests that common genetic variation may subtly alter core components of the regulatory networks controlling cellular differentiation. We also generated the most extensive map to date of recurrent genetic abnormalities in iPSCs, linked a subset of these to downstream transcriptional changes and identified putative targets of selection. Finally, we used methylation data in 285 lines (201 donors) and ChIP-seq for three histone modifications in 118 lines (86 donors) to map quantitative trait loci (QTLs) that alter hIPSC epigenomes. We mapped methylation QTLs for over 7000 genes at an FDR of 1%, including many members of the pluripotency regulatory network. We also show that some instances of suspected aberrant methylation in iPSCs are likely to be driven by common genetic variation. For the histone modification data we employed a probabilistic model recently developed by our group that combines allele-specific and population level sequencing data, to boost the power for association detection. Despite the smaller sample size, we also detected thousands of genome-wide significant associations. Our study provides a first comprehensive picture of the genetic and phenotypic variability in human pluripotent stem cells, including the major drivers of this variation. Further analysis will reveal how these genetically mediated changes to iPSCs epigenomes drive variation in important stem cell properties including maintenance of pluripotency, cell growth and differentiation.

285

Gene expression and splicing differences across 250 cellular environments. *F. Luca[1], A. Richards[1], A. Pai[2], D. Kurtz[1], G. Moyerbrailean[1], G. Davis[1], A. Alazizi[1], C. Harvey[1], N. Hauff[1], Y. Sorokin[1], X. Wen[3], R. Pique-Regi[1].* 1) Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI; 2) Massachusetts Institute of Technology, Cambridge, MA; 3) University of Michigan, Ann Arbor, MI.

Most human traits result from a complex interaction between environmental exposures and an individual's genotypes. While genome-wide association studies (GWAS) have identified a large number of loci associated with complex traits variation, these loci are generally located in non-coding regions and explain a small portion of the variation. We hypothesized that an in depth characterization of the transcriptional response to environmental perturbations can elucidate the genetic architecture and molecular mechanisms underlying complex trait variation. We exposed 5 cell types to a panel of 50 treatments that represent common environmental exposures. On average 3041 genes are differentially expressed. We used a high throughput approach to identify 89 environmental conditions resulting in large gene expression changes and we deeply sequenced the RNA (130M reads/sample on average) from each of these conditions in three individuals. We used Weighted Gene Coexpression Analysis (WGCNA) to construct a coexpression network for 14,527 genes, comprised of 87 modules. Modules associated with >12 treatment conditions contained genes involved in N-linked glycosylation, cellular homeostasis, and response to endoplasmic reticulum stress. These shared response patterns may represent common response mechanisms across several treatments. We found that changes in gene expression are correlated with differences in splicing. We identified 9,221 alternative splicing sites (FDR=15%) in 4,680 unique genes using Mixture of Isoforms (MISO). The highest proportion of changes were retained intron (RI) and alternative first and last exons, with 70% of genes with RI retaining the intron after exposure to treatments. Genes with RI are enriched for functions related to RNA binding, suggesting a mechanism for the regulation of the overall cellular response to environmental stimuli through widespread changes in splicing. Across the different treatments, Caffeine, Iron, Selenium, Tunicamycin, Vitamin D and the anti-allergy drug Loratadine produced the largest effects on RNA processing. We found strong evidence that gene expression response to environmental perturbations is a major mechanism in complex traits. 22% of differentially expressed genes overlap with those identified in GWAS analyses compared to 4% of non-differentially expressed genes expressed in our samples. This overlap corresponds to a 7-fold enrichment ($p < 2.2 \times 10^{-16}$).

286

Human induced pluripotent stem cells: A powerful model to investigate inter-individual regulatory variation across cell types. *N.E. Banovich[1], Y.I. Li[2], A. Raj[2], M.C. Ward[1], P.Y. Tung[1], J.E. Burnett[1], M. Myrthil[1], S.M. Thomas[1], C.L. Burrows[1], I. Gallego Romero[1,3], B.J. Pavlovic[1], J.K. Pritchard[2,4], Y. Gilad[1].* 1) Human Genetics, University of Chicago, Chicago, IL; 2) Genetics, Stanford University, Stanford, CA; 3) School of Biological Sciences, Nanyang Technological University, Singapore; 4) HHMI, Stanford University, Stanford, CA.

Human induced pluripotent stem cells (iPSCs) provide a powerful system to study complex human traits. To investigate inter-individual variation in gene regulation across multiple cell types from the same individuals, we established and validated a panel of 59 iPSCs from lymphoblastioid cell lines (LCLs) of Yoruba individuals, which we have extensively studied in the past. The genome sequences of all individuals were also available to us. We collected RNA sequencing, chromatin accessibility, and DNA methylation data from the LCLs, the iPSCs, as well as from iPSC-derived cardiomyocytes (iPSC-CMs) from 15 of the same individuals. Using these gene regulatory data, we identified thousands of genetic associations with inter-individual variation in gene expression levels (eQTLs), methylation levels (meQTLs), and chromatin accessibility (caQTLs), in each cell type. We found that regulatory variation is lower in iPSCs compared with the differentiated cell types, consistent with the intuition that developmental processes are generally canalized. By considering transcription factor footprints and inferred chromatin states, we were able to provide putative mechanistic explanations for many differences in regulatory QTL associations across cell types. In particular, we identified a large number of cell-type specific regulatory QTLs in distal enhancers, which are likely to regulate tissue-specific gene expression patterns. This study demonstrates the power of the iPS cellular model to dynamically study inter-individual variation in gene regulation.

287

***In vitro* human neuronal differentiation validated as genomic model system to study major psychiatric illnesses.** *A.P.S. Ori[1], M.H.M. Bot[1], R.T. Molenhuis[1], L.M. Olde Loohuis[1], R.A. Ophoff[1,2].* 1) Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, CA; 2) Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.

Large-scale genetic studies of neuropsychiatric disorders have identified hundreds of susceptibility alleles. An important next challenge is to understand how identified genetic risk loci impact biological pathways and disease. For this, model systems are needed that recapitulate biological pathways in cell types that align with etiological mechanisms underlying the disease. We therefore aim to investigate an *in vitro* model of neuronal differentiation using human neural stem cells (hNSC) and evaluate its potential for studying molecular pathways important for major psychiatric illnesses based on genome-wide disease risk identified through GWAS. We differentiated WA09, a widely used hNSC line with standardized lab protocols, to a neuronal lineage across 30 days and assayed genome-wide gene expression profiles at seven time points in replicates. Using a transition mapping approach, we demonstrate that the observed *in vitro* gene expression profiles significantly match *in vivo* human neurodevelopmental stages and cellular laminae of human cortical regions. Using an empirical Bayes approach, we identified >25% of genes to be differentially expressed across 30 days of neuronal differentiation. By subsequent fuzzy c-means clustering with bootstrapping specifically tailored to time course data, we identified 10 high confidence clusters of genes with distinct expression patterns. Biological annotations of these clusters highlights pathways and mechanisms important for neuronal differentiation, such as transcription factor activity, regulation of translation, and synaptic transmission. We next used stratified LD score regression, a statistical method that partitions heritability from GWAS summary statistics, to estimate how genes important for neuronal differentiation contribute to heritability of major psychiatric illnesses. We show that differentially expressed genes are significantly enriched for schizophrenia, bipolar disorder, and major depressive disorder. More specifically, we show that the heritability falls into specific clusters that can be distinct or shared across these diseases. These findings validate WA09 neuronal differentiation as an *in vitro* genomic tool to study major psychiatric illnesses and highlight disease-specific signatures across clusters. This work provides directions for GWAS functional follow-up studies in a model that is robust and simple and that allows for genomic manipulations across an isogenic background in a controlled environment.

288

Saturation genome editing to characterize thousands of mutations at the *HPRT1* locus. *G.M. Findlay[1], M.J. Gasperini[1], J. Shendure[1,2].* 1) Dept. of Genome Sciences, School of Medicine, University of Washington, Seattle, WA; 2) Howard Hughes Medical Institute, Seattle, WA.

The vast majority of genetic variants encountered clinically cannot be confidently interpreted. While many computational approaches exist to predict phenotypic consequences of mutations, these methods fall short of providing clinically actionable information for even a small fraction of variants. To address this challenge, we developed an experimental method, saturation genome editing, in which hundreds to thousands of programmed mutations are introduced to their endogenous genomic context in cultured cells. By subjecting these cells to appropriate assays, massively parallel DNA sequencing can be used to measure the effects of each mutation in a multiplex fashion. To date, we have shown this method can be applied to robustly measure mutations' impact on gene splicing and cellular fitness. To expand the applicability of this method and its relevance to medical genetics, we performed new saturation genome editing experiments at the *HPRT1* locus. Clinically, hundreds of different *HPRT1* mutations can cause Lesch-Nyhan syndrome, and loss of HPRT1 enzymatic activity confers cancerous cells with resistance to the chemotherapeutic 6-thioguanine (6-TG). To deeply mutagenize this locus, we designed saturation genome editing libraries containing all possible single nucleotide variants (SNVs) for every coding exon of *HPRT1* as well as parts of the promoter, the 5'- and 3'-untranslated regions, and a few highly conserved intronic regions. Performing 6-TG selection on edited populations of cells allowed us to test whether the mutations we introduced resulted in loss-of-function of the gene. Our preliminary results, which provide data for the mutational effects of over 1,000 coding variants, correlate well with biological expectation. Additionally, saturation genome editing performed within an intronic region of unknown functional importance shows that while most SNVs do not produce loss-of-function alleles, a small fraction (~3%) have strong effects, enabling cells to survive 6-TG treatment. Additional experiments are ongoing towards the goals of testing every possible coding SNV within all exons of the gene, creating thousands of additional non-coding mutations in putative regulatory regions, and characterizing the mechanisms by which mutations in non-coding sequences lead to loss-of-function. Although preliminary, our results to date illustrate the vast potential of genome editing for characterizing both coding and non-coding variants within human disease genes.

289

Using multiplex non-coding CRISPR deletion libraries to individually perturb every CTCF binding site in the human genome. *S.K. Reilly[1,2], R. Tewhey[1,2], E.A. Brown[1,2], P.C. Sabeti[1,2].* 1) Broad Institute, Cambridge, MA; 2) Harvard University, Center for SyCenter for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA.

The vast majority of sequence variants associated with human traits, disease susceptibility, and evolutionary adaptation are located in non-coding regions of the human genome. While many of these variants are associated with *cis*-regulatory regions such transcription factor (TF), histone domains, and chromatin looping elements, the regulatory components often occur at over 100,000 locations in the genome, making a precise and comprehensive understanding of the importance of individual TF sites difficult. The ubiquitously expressed CCCTC-binding factor (CTCF) is of particular regulatory interest because it organizes overall 3-D organization of the genome, defines chromatin boundaries, and directs enhancer-promoter looping events. 8157 CTCF motifs contain common variants, and 965 of these are located in regions implicated in recent human selection. Furthermore, variants in CTCF sites have been shown to alter TF binding, chromatin looping, gene expression, and cancer progression making them prime candidates in the search for disease-causing and adaptive human variation. Most previous studies of non-coding regulatory regions have investigated single genomic loci, sacrificing throughput, or have knocked down a specific underlying factor genome-wide, sacrificing locus specificity. Also, while comprehensive CRISPR deletion libraries to target genes have been generated, similar approaches for non-coding regions of the genome have not been developed. Here we describe a novel non-coding CRISPR screening method and its use in investigating all CTCF sites individually. Targeting this factor is of primary importance due to its fundamental role in gene regulation and genome architecture. We synthesized 100,000 guides to target the PAM site in the core CTCF motif of loci directly bound by CTCF or in a CTCF-directed loop in LCL or K562 cell lines. A lentiviral library of these guides and cas9 was used to infect both cell lines. We measured guide abundance by sequencing at the time of infection as well as after positive and negative selection to assess each individual CTCF site's role in viability. Our results describe global trends and individual dissections of CTCF importance to gene expression, chromatin looping, and *cis*-regulatory logic. Furthermore, we find allelic variation in CTCF's contribution to evolution and disease. Lastly, we describe how the system can be rapidly applied to other non-coding targets and provide paradigms for experimental design and analysis.

290

Personalized medicine for neurometabolic disorders via an integrated '-omics' approach. *C. van Karnebeek*[1,5], *M. Tarailo-Graovac*[2,5], *A. Matthews*[2,5], *C. Shyr*[2,5], *J. Lee*[2,5], *D. Wishart*[6], *C. Ross*[1,2], *H. Vallance*[3], *G. Sinclair*[3], *R. Salvarinova*[1], *S. Stockler*[1], *L. Kluijtmans*[4], *G. Horvath*[1], *R. Wevers*[4], *W. Wasserman*[2,5]. 1) Pediatrics, University of British Columbia, Vancouver, British Columbia, Canada; 2) Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada; 3) Pathology & Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada; 4) Radboud University Medical Centre, Nijmegen, The Netherlands; 5) Centre for Molecular Medicine and Therapeutics / CFRI; 6) The Metabolomics Innovation Centre, University of Alberta, Edmonton, Canada.

Introduction: Translation of a genomic diagnosis into disease-modifying treatments is challenging, particularly for intellectual developmental disorder (IDD). However, the exception is inborn errors of metabolism, since many of these are responsive to therapy that targets pathophysiological features at the molecular or cellular level. In this study we aimed to unravel the genetic basis of unexplained neurometabolic phenotypes and initiate targeted treatments. **Methods**: We combined phenomics (comprehensive characterization a patient's clinical and biochemical phenotype) with exome / genome sequencing analysis through a semiautomated bioinformatics pipeline in consecutively enrolled patients with IDD and unexplained metabolic phenotypes. In parallel, metabolomics analysis using QTOF (untargeted) and NMRspectroscopy (targeted) was performed to further characterize the affected metabolic pathway and discover biomarkers. **Results**: Exome/genome sequencing and metabolomics was performed in 77 probands (predominantly nonconsanguineous parents, European descent). In 58 probands, a diagnosis was established: variants in 15 genes newly implicated in disease, 27 known genes with newly identified phenotypes, and 16 genes with expected phenotypes (majority variants classified as pathogenic or probably pathogenic)In 7 patients, complex phenotypes of 7 patients were explained by coexisting monogenic conditions. Novel biomarkers were identified in 3 patients. In 29 patients the diagnosis changed the clinical management, including preventive measures such as regular screening for malignancy and avoidance of disease triggers in 7, immune-modulating therapies such as chemotherapy or stem cell transplantation in 4, more precise symptomatic management (supplementation with neurotransmitters, medications) in 9, and treatments targeting the identified abnormality at a cellular or molecular level in 9 patients. These changes in clinical management improved or stabilized IDD related outcomes (epilepsy, autism, movement) to different degrees. **Conclusions:** Deep phenotyping combined with genomic and metabolomics analysis in 77 probands with IDD and unexplained metabolic abnormalities led to a diagnosis in 75%, the discovery of 15 genes newly implicated in neurometabolic disease, and enabled personalized medicine -change in treatment beyond genetic counseling- in 37%. Combining large data sets is effective in pinpointing the candidate gene, revealing new biomarkers and validating causality.

291

Semi-automated bioinformatics approach to exome/genome data analysis for diagnosis and discovery of neurometabolic disease. *M. Tarailo-Graovac*[1-3], *A. Matthews*[1-3], *J.Y.A. Zhu*[4], *C. Shyr*[1-3], *G.A. Horvath*[3,5,6], *R. Salvarinova*[5,6], *C.J. Ross*[2,3,5], *J.J.Y Lee*[1,3,4], *S. Stockler-Ipsiroglu*[3,5,6], *R. Wevers*[7], *H. Vallance*[3,8], *G. Sinclair*[3,8], *C.D. van Karnebeek*[1,3,5,6], *W.W. Wasserman*[1-3]. 1) Centre for Molecular Medicine and Therapeutics, Vancouver, BC, Canada; 2) Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada; 3) Child & Family Research Institute, Vancouver, BC, Canada; 4) Genome Science and Technology Graduate Program, University of British Columbia, Vancouver, BC, Canada; 5) Division of Biochemical Diseases, Vancouver, BC, Canada; 6) Department of Pediatrics, BC Children's Hospital, Vancouver, BC, Canada; 7) Department of Laboratory Medicine, Radboud University Medical Centre, Nijmegen, Netherlands; 8) Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada.

Introduction: Next generation sequencing (NGS) has revolutionized gene discovery and diagnosis of rare Mendelian disorders, including Intellectual Developmental Disorders (IDDs). Our TIDE-BC (Treatable Intellectual Disability Endeavor) project combines neurometabolic phenotype and NGS to identify potentially treatable IDD genes. **Methods:** NGS sequencing was performed on the consecutively enrolled patients with unexplained neurometabolic phenotype, the parents, and any affected siblings. We developed a semi-automated gene-discovery pipeline, which involves manual inspection of data quality and collaborative interactions between clinicians and bioinformaticians. The pipeline allows for analysis of both nuclear and mitochondrial genomes, as well as gene/variant-centric classification, including assessment of variant prevalence in presumably unaffected individuals. One of the pipeline development steps included analysis of The Exome Aggregation Consortium (ExAC) dataset for presence of individuals with pathogenic variants described in severe Mendelian pediatric disorders. **Results:** Exome/genome sequencing analysis was performed in >150 probands meeting selection criteria. Diagnosis was established in ~70% of the probands; the majority is due to genes newly implicated in disease or known genes with newly identified phenotypes. Coexisting monogenic conditions were found to contribute to complex clinical phenotype in >15% of the probands. The identified inheritance patterns included: recessive (homozygous, compound heterozygous, X-linked), dominant (*de novo*: autosomal, X-linked and mosaic; and heterozygous inherited), as well as mitochondrial. The ExAC dataset for variant classification based on population frequency proved to be an important resource; however, four of our patients had confirmed pathogenic variants with the same genotype as 44 ExAC individuals. Analysis of > 900 genes believed to cause Mendelian pediatric disorders (from variable to highly penetrant severe) revealed presence of individuals in the ExAC dataset with pathogenic 'genotypes'. **Conclusions:** The success of our study emphasizes the strength of a collaborative semi-automated approach for accurate data analysis and diagnosis using NGS technology. Better understanding of the population datasets is warranted as presence of a pathogenic variant in a presumably unaffected population does not rule out the possibility that the variant is pathogenic (implicated in a severe Mendelian pediatric disorder).

292

Massively parallel digital transcriptional profiling of single cells.

G. Zheng[1], J. Terry[1], P. Belgrader[1], P. Ryvkin[1], Z. Bent[1], R. Wilson[1], S. Ziraldo[1], T. Wheeler[1], G. McDermott[1], M. Gregory[2], J. Shuga[1], L. Montesclaros[1], D. Masquelier[1], S. Nishimura[1], M. Schnall-Levin[1], C. Hindson[1], R. Bharadwaj[1], A. Wong[1], K. Ness[1], L. Beppu[7], J. Deeg[7], C. McFarland[8], K. Loeb[5,7], W. Valente[2,3,4], N. Ericson[2], E. Stevens[7], J. Radich[7], T. Mikkelsen[1], B. Hindson[1], J. Bielas[2,4,5,6]. 1) 10x Genomics Inc, Pleasanton, CA; 2) Translational Research Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA; 3) Medical Scientist Training Program, University of Washington School of Medicine, Seattle, WA; 4) Molecular and Cellular Biology Graduate Program, University of Washington, Seattle, WA; 5) Department of Pathology, University of Washington, Seattle, WA; 6) Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, WA; 7) Clinical Research Division, Fred Hutchinson Cancer Research Centre, Seattle, WA; 8) Seattle Cancer Care Alliance Clinical Immunogenetics Laboratory, Seattle, WA.

Characterizing the transcriptome of individual cells is fundamental to understanding complex biological systems. Current methods for single cell RNA-sequencing (scRNA-seq) face practical challenges due to throughput limitations, custom set-up requirements, or complicated experimental protocols. We describe a fully-integrated, droplet-based system that enables 3' mRNA digital counting of up to tens of thousands of single cells per sample with ~50% cell capture efficiency. We first validated the sensitivity of the system with scRNA-seq of human (Jurkat) and mouse (3T3) cell lines. At 100k reads/cell, we detected a median of ~4,500 genes in each human and mouse cell, indicating comparable sensitivity to existing droplet-based platforms. We then demonstrated the ability of the system to detect heterogeneous populations by profiling >100,000 peripheral blood mononuclear cells from healthy donors. We successfully detected all major subpopulations at expected proportions from fresh and cryopreserved cells. Lastly, we developed a method to characterize both immune cell subtypes and genotypes by integrating single cell digital RNA profiling with de novo single nucleotide variant (SNV) calling. Our method enables the study of host and donor cell chimerism in an allogeneic hematopoietic stem cell transplant (HSCT) setting. We applied our method on bone marrow mononuclear cells from patients who underwent HSCT for acute myeloid leukemia (AML). The SNV analysis detected the proportion of donor and host genotypes that was consistent with independent clinical chimerism assay. In one post-transplant AML patient, we detected 87% host cells and 13% donor cells. In addition, scRNA-seq analysis revealed an expansion of myeloid progenitors in the host cells where the donor cells consisted mostly of lymphocytes. This observation is consistent with the relapse of the patient's AML with a return of the malignant host AML at the expense of donor hematopoiesis. Our method provided insight into the disease pathology that would normally require multiple clinical assays. We envision that our technology will enable widespread adoption of high throughput single cell mRNA analysis, and accelerate the characterization of diverse developmental systems as well as tumor samples in basic and clinical research.

293

Single nucleotide mutagenesis starts from two-cell cleavage-stage: Direct evidence from genome-wide analysis.

K. Choy[1,2,4], Z. Dong[1,2,3], Y. Cao[1,2], F. Chen[3], H. Jiang[3]. 1) Department of Obstetrics and Gynecology, The Chinese University of Hong Kong, Hong Kong, China; 2) Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China; 3) BGI-Shenzhen, Shenzhen 518083, China; 4) The Chinese University of Hong Kong-Baylor College of Medicine Joint Center For Medical Genetics, Hong Kong, China.

Emerging family-based studies infer a high mutation rate of single nucleotide variants (SNVs) per cell division during embryogenesis, but there is a lack of direct evidence of single nucleotide mutagenesis from the human cleavage-stage embryos. Here, by utilizing intra-embryonic single-cell transcriptome data from 15 human and mouse embryos, we observed high SNVs mutation rates in protein-coding region among the embryos in different cleavage-stages. To further investigate the earliest timing of pre-morula mutagenesis, we conducted over 150× whole genome sequencing within an individual's five different organs, including placenta tissue from two un-related fetuses. Genome-wide analysis revealed a subset of all-organs shared SNVs in approximately 50% mosaic level, directly proofed that pre-morula mutagenesis started from two-cell stage. The pre-morula mutation rate per individual is 14.23±2.50 times of the somatic mutation rate documented in early organogenesis, indicating a high DNA replication error rate with random occurrence in cleavage-stage.

294

PIK3CA-Related Overgrowth Spectrum (PROS) molecular spectrum and recommendations for testing among a novel series of 205 patients.

P. Kuentz[1], Y. Duffourd[2], J. St-Onge[2], A. Sorlin[2], V. Carmignac[2], T. Jouan[2], J. Amiel[3], N. Bahi-Buisson[3], D. Bessis[4], O. Boute[4], A.-C. Bursztejn[5], C. Chiaverini[6], C. Coubes[4], A. Goldenberg[7], B. Isidor[8], L. Martin[9], A. Maruani[10], J. Mazereeuw-Hautier[11], C. Mignot[3], F. Morice-Picard[12], F. Petit[4], A. Phan[13], R. Touraine[14], M. Vincent[8], M. Willems[4], N. Marle[2], S. Hadj-Rabia[3], P. Vabres[2], J.-B. Rivière[2], L. Faivre[2]. 1) CHU BESANCON, BESANCON, France; 2) CHU DIJON, DIJON, France; 3) APHP, PARIS, France; 4) CHU MONTPELLIER, MONTPELLIER, France; 5) CHU NANCY, NANCY, France; 6) CHU NICE, NICE, France; 7) CHU ROUEN, ROUEN, France; 8) CHU NANTES, NANTES, France; 9) CHU ANGERS, ANGERS, France; 10) CHU TOURS, TOURS, France; 11) CHU TOULOUSE, TOULOUSE, France; 12) CHU BORDEAUX, BORDEAUX, France; 13) CHU LYON, LYON, France; 14) CHU ST ETIENNE, ST ETIENNE, France.

Vascular malformations are a hallmark of the *PIK3CA* related overgrowth spectrum (PROS) which includes CLOVES (congenital lipomatous overgrowth, vascular and epidermal nevi, skeletal anomalies) and MCAP (macrocephaly-capillary malformation) syndromes. We sought to determine the diagnostic yield of *PIK3CA* mutation search on affected tissue, saliva or blood depending on clinical presentation. We analyzed results of next generation sequencing *PIK3CA* mutation detection in a series of 205 patients who were classified either without (group A) or with (group B) cerebral involvement, roughly assimilated with CLOVES or MCAP spectra, respectively. Cerebral involvement consisted at least of macrocephaly or hemimegalencephaly. In most patients, *PIK3CA* was analyzed on at least one affected tissue sample, usually a skin biopsy (149 samples). There were altogether 128 group A (62.8%) and 77 group B cases (37.2%). Overall, testing was positive on 114 patients (55.6%) and the diagnostic yield was 57.8% for group A and 51.9% for group B. The highest diagnostic yield was obtained testing fresh affected tissue with rates between 55.6% (affected skin - group B) and 81.3% (other affected tissue - group A). Affected skin cultured fibroblasts showed more random results (34.8% - group A; 66.7% - group B). Interesting rates were obtained with saliva (64.3%) and blood samples (44.2%) in group B. As previously reported, very low yield was confirmed with blood samples in group A (8.5%). *PIK3CA* mutations were distributed over the whole gene sequence. 12 mutations had previously been reported in group A, and 28 in group B. We identified 23 novel mutations in group A and 10 in group B. Both groups shared seventeen common mutations, including the three major hotspot mutations (p.Glu542Lys, p.Glu545Lys and p.His1047Arg) accounting for only 26% of positive cases and never found in blood or saliva. Molecular diagnosis of PROS should preferably be performed on affected tissue. It is also possible on blood or saliva whenever central nervous system involvement is present, but with a lower diagnostic yield. Cultured fibroblast testing is not recommended, as mutations were missed in 12.5% of cases. The mutational spectrum extends far beyond known hot spots, hence complete *PIK3CA* sequencing is warranted. Absence of *PIK3CA* mutation in 44.4% of patients may be due either to inappropriate clinical diagnosis or tissue sampling, undetectable mosaicism rate (<1%), or involvement of additional unknown genes.

295

Frequency of mosaicism in parents of children with early-onset epilepsy.

C. Myers[1], Z. Thuesmunn[1], A. Muir[1], G. Hollingsworth[2], A. Schneider[2], G. Carvill[1], L. Sadleir[5], I. Scheffer[2,3,4], H. Mefford[1]. 1) Department of Pediatrics, Division of Genetic Medicine, University of Washington, Seattle, Washington 98195, USA; 2) Epilepsy Research Centre, Department of Medicine, Austin Health, The University of Melbourne, Heidelberg, Victoria, 3084, Australia; 3) Florey Institute of Neuroscience and Mental Health, The University of Melbourne, VIC 3010, Australia; 4) Department of Paediatrics, Royal Children's Hospital, The University of Melbourne, Parkville, Victoria, 3050, Australia; 5) Department of Paediatrics and Child Health, University of Otago, Wellington, New Zealand.

The epileptic encephalopathies (EEs) are devastating epilepsies characterized by refractory seizures that often begin in infancy or childhood, frequent epileptic activity on EEG and developmental slowing or regression. Due to the severity of the condition, cases are usually sporadic with no other affected family members. Massively parallel sequencing, or next-generation sequencing, has revealed the prominent role of *de novo* pathogenic variants in causing EEs. *De novo* variants may arise post-zygotically in the developing fetus, resulting in a mosaic distribution of the mutation in the affected child. Alternatively a post-zygotic mutation may arise in the parental germline, resulting in a constitutional mutation in the affected child. Mosaic parents may not have any noticeable signs of the disorder depending on the cell type and percentage of cells that carry the mutation, but germline mosaicism leads to a critical increase in recurrence risk in offspring. We have used a highly sensitive, improved version of the molecular inversion probe technology (single *m*olecule *m*olecular *i*nversion *p*robes) to investigate the frequency of somatic (and germline) mosaicism in parents of children with heterozygous *de novo* mutations. We screened 142 families where the affected child's EE was attributed to a substitution or small indel in one of 31 established epilepsy gene and reported as 'de novo' by either clinical or research analysis of parental DNA. To date, we have identified 10 cases of low-level somatic mosaicism, detecting the causative variant in DNA isolated from parental blood or saliva. The fraction of mutant alleles identified ranged from 0.1 – 11%, levels that are likely to go undetected by traditional Sanger sequencing methods. Six of the pathogenic variants originated maternally and four originated paternally. In two families, the pathogenic variant was also transmitted to an affected sibling. The rate of parental mosaicism has been greatly underestimated in a sporadic patient with an EE and an apparent *de novo* mutation. This is because low levels (<11%) of mosaicism are found in ~10% of cases and escape detection by traditional methodologies. Our findings highlight the critical importance of more sensitive tools for detecting low frequency mosaicism. Identifying of a mosaic mutation in parents is crucial for reproductive counseling and family planning.

296

The contribution of post-zygotic mutations resulting in sequence and structural mosaicism to dominant developmental disorders. *M.E. Hurles, C. Wright, J. Kaplanis, D. Rajan, A. Sifrim, D. King, DDD Study.* Human Genetics, Wellcome Trust Sanger Institute, Cambridge, United Kingdom.

New mutations can arise at any stage within the cellular lineage of the germline, from the zygote to the mature gamete. This timing of mutation is of direct clinical relevance: both in terms of missed diagnoses and for determining the risk of recurrence in siblings. We undertook a comprehensive analysis of both parental and child mosaicism in 4,293 parent-child trios exome sequenced within the Deciphering Developmental Disorders study. We identified candidate post-zygotic pathogenic mutations in children, considering both sequence and structural variants, and experimentally validated them in DNA derived from both saliva and blood, for example, by re-sequencing all members of a trio to 100,000X depth. We found that ~3% of pathogenic mutations detectable using exome sequencing were post-zygotic mutations within children. Moreover, we identified 9 mosaic pathogenic mutations that were present within saliva-derived DNA, but absent or at much lower frequency in blood-derived DNA. These mutations are likely under stronger negative selection in blood, and were enriched for large structural variants. These mutations represent pathogenic variants that are likely to be missed if only using blood-derived DNA for genetic testing.We also identified and experimentally validated (again using ultra-deep targeted sequencing) 20 pathogenic mutations in children (representing ~2% of all pathogenic mutations) that were mosaic within parental somatic tissues (saliva-derived DNA), at frequencies from 6-60% of cells. Using standard variant calling pipelines, half of these mutations were called in the parent, and initially were falsely interpreted to be benign variants inherited from an unaffected parent. These families are at considerably greater risk of recurrence in siblings, and analysis of multi-sibling human and mouse pedigrees suggests that this risk may be as high as 10-20%. Correspondingly, families exhibiting parental somatic mosaicism for a pathogenic de novo mutation were enriched for siblings affected by the same disorder, compared to other families with pathogenic de novo mutations. Clinicians were informed of the likely increased recurrence risk in their families to assist in management of subsequent pregnancies. Our results highlight and quantify the clinical importance of identifying parental and child mosaicism, and provide insights into the timing of pathogenic mutations.

297

***GNA11* and *GNAQ* post-zygotic mosaicism cause an overlapping spectrum of neurocutaneous disorders.** *V.A. Kinsler[1and2], A.C. Thomas[1], Z. Zeng[3], J-B. Rivière[4], R. O'Shaughnessy[5], L. Al-Olabi[1], J. St-Onge[4], D.J. Atherton[2], H. Aubert[6], L. Bagazgoitia[7], S. Barbarot[6], E. Bourrat[8and9], C. Chiaverini[10], W.K. Chong[11], Y. Duffourd[4], K. Forde[2], M. Glover[2], L. Groesser[12], S. Hadj-Rabia[13], H. Hamm[14], R. Happle[15], P. Kuentz[4], J-P. Lacour[10], I. Mushtaq[16], S. Polubothu[1and2], R. Waelchli[2], M. Wobser[14], E.E. Patton[3], P. Vabres[4and17].* 1) Genetics and Genomic Medicine , UCL Institute of Child Health, London, London, United Kingdom; 2) Paediatric Dermatology, Great Ormond St Hospital for Children, London, United Kingdom; 3) 1.MRC Institute of Genetics and Molecular Medicine, MRC Human Genetics Unit & Edinburgh Cancer Research UK Centre, Edinburgh, UK; 4) 1.Equipe d'Accueil 4271, Génétique des Anomalies du Développement, University of Burgundy, Dijon, France; 5) Infection, Immunity, Inflammation, UCL Institute of Child Health, London, UK; 6) Department of Dermatology, Nantes University Hospital, Nantes, France; 7) Dermatology, Hospital Universitario Ramón y Cajal, Madrid, Spain; 8) Dermatology, Saint-Louis Hospital, Paris, France; 9) General Paediatrics, Robert-Debré Hospital, Paris, France; 10) Dermatology, University Hospital of Nice, Nice, France; 11) Neuroradiology, Great Ormond St Hospital for Children, London, UK; 12) Dermatology, Regensburg University Clinic, Regensburg, Germany; 13) Paediatric Dermatology, Necker Enfants-Malades Hospital, Paris, France; 14) Dermatology, University Hospital Wuerzburg, Wuerzburg, Germany; 15) Dermatology, Freiburg University Medical Center, University of Freiburg, Germany; 16) Paediatric Urology, Great Ormond Street Hospital for Children, London, UK; 17) Dermatology, Dijon University Hospital, Dijon, France.

Common birthmarks can be an indicator of underlying genetic disease, but are often overlooked. Mongolian Blue Spots (dermal melanocytosis) are usually localized and transient, but can be extensive, permanent and associated with extra-cutaneous abnormalities. Co-occurrence with vascular birthmarks defines a subtype of Phakomatosis Pigmentovascularis (PPV), a group of syndromes associated with neuro-vascular, ophthalmological, overgrowth and malignant complications. In a cohort of children with extensive dermal melanocytosis and PPV we discover here that these phenotypes are caused by activating mutations in *GNA11* and *GNAQ*, genes that encode Gα subunits of heterotrimeric G-proteins. The mutations were found within the affected skin but not the blood, indicating that these conditions are post-zygotic mosaic disorders. We also describe for the first time Sturge-Weber syndrome caused by *GNA11* mosaicism. *In vitro* expression of mutant *GNA11*^{R183C} in human cell lines demonstrated activation of the downstream p38 MAPK signalling pathway. Novel transgenic mosaic zebrafish models expressing mutant *GNA11*^{R183C} under promoter *mitfa* developed extensive dermal melanocytosis recapitulating the human phenotype. PPV and extensive dermal melanocytosis are new diagnoses in the group of mosaic heterotrimeric G-protein disorders, joining McCune-Albright and Sturge-Weber syndromes. These findings will allow accurate clinical and molecular diagnosis of this subset of common birthmarks, thereby identifying infants at risk of serious complications, and will provide novel therapeutic opportunities.

298

Characterizing functional regulatory variants in iPSC-derived human cardiomyocytes.

P. Benaglio[1], C. DeBoever[2], H. Li[3], F. Drees[3], M. D'Antonio[4], H. Matsui[3], A. Arias[3], A. DAntonio-Chronowska[1], E. Smith[1], K. Frazer[1,3]. 1) Department of Pediatrics and Rady Children's Hosp, University of California San Diego, La Jolla, CA; 2) Bioinformatics and Systems Biology, University of California San Diego, La Jolla, CA; 3) Institute for Genomic Medicine, University of California San Diego, La Jolla, CA; 4) Moores Cancer Center, University of California San Diego, La Jolla, CA.

Our goal is to demonstrate the utility of using human induced pluripotent stem cell derived cardiomyocytes (iPSC-CM) to understand how non-coding DNA variants influence cardiomyocyte molecular phenotypes. Genetic variants that have been associated with common phenotypes and diseases are enriched in non-coding regulatory regions of trait-relevant cell types. Cellular systems such as iPSCs that can be used to model virtually any human tissue and can be profiled in depth are a promising strategy to characterize the function of regulatory variants in cell type-specific contexts. We have generated a large collection of human iPSCs from 222 different individuals genotyped by whole-genome sequencing and are differentiating each line into cardiomyocytes. We are profiling each line by RNA-Seq, ATAC-Seq, ChIP-Seq and DNA methylation to identify common and rare genetic variants associated with these molecular phenotypes (QTLs). Here, I will focus on the characterization of iPSC-CMs from 7 related individuals by RNA-Seq, ChIP-Seq of two transcription factors (TF) (NKX2-5 and SRF) and one histone modification (H3K27ac). Across 14 differentiations (2 replicates per person), we observe cardiac-specific gene expression and epigenetic profiles that are more similar to those obtained from fetal heart than any other tissue in the Roadmap project. Across individual iPSC-CM lines, we observe that 2-10% molecular traits show quantitative differences. SNPs are enriched in variant ChIP-Seq peaks and show allelic specific effects (ASE) consistent between individuals, suggesting that DNA variants underlie epigenetic differences between iPSC-CMs derived from different people. SNPs that showed ASE in NKX2-5 and SRF peaks were enriched in TF motif-modifying sequences, and were consistent with the direction of the ASE, suggesting a functional role for these SNPs in determining differential TF binding. A subset of these variant peaks overlap with loci previously associated with cardiovascular traits or eQTLs and can be explained by differences in NKX2-5 and SRF binding in our model system, suggesting that TF binding may underlie these previous associations. Overall these findings highlight the extraordinary potential of iPSCs and derived cell types as a new model to dissect the impact of human genetic variants on intermediate molecular phenotypes and offer a powerful, yet unexplored tool to functionally characterize human regulatory variants.

299

Discovery of adipose-specific GxBMI interactions on the regulation of gene expression.

K.S. Small[1], C.A. Glastonbury[1], A. Vinuela[1,2], A. Buil[2], G.H. Halldorsson[3], G. Thorleifsson[3], H. Helgason[3,4], U. Thorsteindottir[3,5], K. Stefansson[3,5], E.T. Dermitzakis[2,6,7], T.D. Spector[1]. 1) Department of Twin Research and Genetic Epidemiology, ,King's College London, London, United Kingdom; 2) Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, CH-1211, Switzerland; 3) deCODE Genetics, Sturlugata 8, Reykjavik, IS-101, Iceland; 4) School of Engineering and Natural Sciences, University of Iceland, Reykjavik, IS-107, Iceland; 5) Faculty of Medicine, University of Iceland, Reykjavik, IS-101, Iceland; 6) Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, CH–1211, Switzerland; 7) Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.

Obesity is a global epidemic that is causally associated at the population-level to a range of diseases including type 2 diabetes and cardiovascular disease. However, there is marked heterogeneity in obesity-related outcomes amongst individuals. This may reflect genotype-dependent responses to adiposity. As adiposity, measured by body mass index (BMI), is associated to widespread changes in gene expression and regulatory variants mediate the majority of known complex trait *loci*, we sought to identify gene-by BMI interactions (G×BMI) on the regulation of gene expression in a multi-tissue RNA-Seq dataset from the TwinsUK cohort (N = 856). We first characterized the direct association between BMI and gene expression and find that 16,818, 9,216, 6,640 genes in adipose, skin and whole blood respectively had at least one exon associated to BMI (FDR 5%). Approximately half of the associations detected in adipose were observed in the other two primary tissues (skin π_1 = 0.53, blood π_1 = 0.54) whereas adipose captured over 75% of the associations detected in skin (π_1 = 0.78) and blood (π_1=0.76). We next performed a global *cis*-scan for BMI-dependent regulatory variants and identified sixteen *cis*-G×BMI interactions at an FDR of 5% (top *cis* interaction: *CHURC1* - rs7143432, $P = 2.0\times10^{-12}$) and one variant regulating 53 genes in *trans* (top *trans* G×BMI interaction: *ZNF423* - rs3851570*,* $P = 8.2\times10^{-13}$), all in adipose tissue. The interactions were adipose-specific, enriched for variants overlapping adipocyte enhancers, and regulated genes enriched for metabolic and inflammatory processes. We demonstrate the robustness of our findings by replicating a subset of G x BMI regulatory interactions in an independent adipose RNAseq dataset (deCODE genetics, N =754). We also confirmed the interactions with an alternate measure of obesity, dual-energy X-ray absorptiometry (DXA) derived visceral fat volume measurements, in a subset of TwinsUK individuals (N = 564). The identified G×BMI regulatory effects demonstrate the dynamic nature of gene regulation and reveal a functional mechanism underlying heterogeneous response to obesity.

300

Functional fine-mapping of coronary artery disease risk variants. *B. Liu[1,2], M. Pjanic[3], T. Nguyen[3], S. Montgomery[2,4], C. Miller[3], T. Quertermous[3].* 1) Biology, Stanford University, Palo Alto, CA; 2) Pathology, Stanford University, Palo Alto, CA; 3) Medicine - Division of Cardiovascular Medicine, Stanford University, Palo Alto, CA; 4) Genetics, Stanford University, Palo Alto, CA.

Coronary artery disease (CAD) is the leading cause of death globally. Its complex etiology is influenced by both genetic and environment risk factors. Large-scale population association studies have identified over 150 risk loci, explaining ~10% of the disease heritability. Functional interpretation of the risk associated variants has been challenging as the majority of the variants lie outside of coding regions and are predicted to be regulatory. One of the key arbiters in coronary artery disease is the human coronary artery smooth muscle cell (HCASMC). These cells constitute the majority of the vessel wall and have been shown to give rise to ~30% of the total cells in the atherosclerotic lesion. While these cells remain differentiated in healthy tissue, they become highly proliferative and invasive in response to vessel injury and lesion expansion. Global changes in the HCASMC transcriptome and the underlying regulatory mechanisms responsible for these disease state transitions have not been fully characterized due to limited sample availability. Here, we have generated RNA and whole-genome sequencing datasets on 52 serum-stimulated HCASMC (mimicking the disease state), as well as ten serum-free HCASMC cell lines (mimicking the healthy state), representing one of the largest collections of these valuable samples to date. By comparing the transcriptomes of HCASMC donors and closely related tissue types in the GTEx project, we identify gene expression and splicing quantitative trait loci (eQTL and sQTL) and unique regulatory features of the HCASMC transcriptome. Additionally, comparisons between the serum-perturbed and serum-free HCASMC donors reveal transcriptomic differences between healthy and disease states related to the phenotypic modulation of these cells. Through QTL mapping, we identify HCASMC-specific eQTLs and sQTLs as compared to related GTEx tissues. Most of these QTLs overlap chromatin-accessible regions as determined by the Assay for Transposase-Accessible Chromatin (ATAC-seq). As expected, CAD GWAS variants are enriched in HCASMC QTLs as compared to other tissues, further emphasizing the disease relevance of this critical vascular cell type. By combining eQTL, ATAC-seq and whole genome data, we report on fine-mapping of multiple CAD-associated alleles. This study represents the first comprehensive characterization of the HCASMC transcriptome from multiple donors, ultimately revealing insights into the genetic underpinnings of CAD.

301

Cardiometabolic variants in adipose tissue ATAC-seq peaks at GWAS loci with coincident adipose tissue eQTLs. *M.E. Cannon[1], K.W. Currin[1], A. Safi[2], L. Song[2], Y. Wu[1], M. Civelek[3], M. Laakso[4], G.E. Crawford[2], K.L. Mohlke[1].* 1) Department of Genetics, University of North Carolina at Chapel Hill, NC, USA; 2) Department of Pediatrics, Division of Medical Genetics, Duke University, NC, USA; 3) Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, USA; 4) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.

Many genome-wide association study (GWAS) signals for cardiometabolic phenotypes consist of tens or hundreds of trait-associated variants. To identify candidate variants responsible for cardiometabolic GWAS signals, we examined expression quantitative trait loci (eQTLs) and assay for transposase-accessible chromatin (ATAC-seq) data from subcutaneous adipose tissue of METabolic Syndrome in Men (METSIM) study participants. We hypothesized that regulatory variants may alter chromatin accessibility or transcription factor binding to alter gene expression. Based on gene expression levels and genotypes in 770 subcutaneous adipose samples, 107 GWAS loci harbor coincident eQTLs (r^2>.8; conditional analysis). We generated ATAC-seq data in 3 subcutaneous adipose samples (mean 71,512,258 uniquely mapped reads) and called a mean of 46,291 regions of chromatin accessibility (peaks) using MACS2. In total, we identify 150 variants exhibiting pairwise LD of r^2>.8 with the eQTL lead variant that overlap ATAC-seq peaks at GWAS-eQTL loci. Of these variants, 43% overlap ATAC-seq peaks in all three samples, 24% in two samples, and 33% in only one sample. These 150 variants represent 60 of the 107 (56%) GWAS-eQTL loci. Of the 60 GWAS-eQTL loci that contain variants in ATAC-seq peaks, 25 contain only one variant overlapping an ATAC-seq peak, 31 contain 2-5 variants in one or more peaks, and 4 contain 6-10 variants in one or more peaks. Among these GWAS and eQTL loci overlapping an ATAC-seq peak is a signal within *MLXIPL* associated with increased HDL cholesterol, decreased triglycerides, and increased *MLXIPL* expression level (*P*=8E-106). One of 32 candidate variants at this signal, rs55747707, is located in an ATAC-seq peak in intron 1 of *MLXIPL* and may act as a regulatory variant. Another GWAS and eQTL locus overlapping an ATAC-seq peak is a signal within *DOCK6* and upstream of *ANGPTL8* associated with increased HDL cholesterol and increased *ANGPTL8* expression level (*P*=1E-9). One of 10 candidate variants, rs12463177, is located in an ATAC-seq peak and may act as a regulatory variant. We confirmed that rs12463177 exhibits allelic differences consistent with the direction of the eQTL in transcriptional reporter assays and electrophoretic mobility shift assays in adipose and liver cell types. Taken together, these data demonstrate the utility of combining regulatory datasets, such as eQTLs and ATAC-seq, to aid in the identification of functional regulatory variants at GWAS loci.

302

Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. *C.L. Miller[1], M. Pjanic[1], T. Wang[1], T. Nguyen[1], A. Cohain[2], J.D. Lee[1,3], L. Perisic[4], U. Hedin[4], R.K. Kundu[1], D. Majmudar[1], J.B. Kim[1], O. Wang[1], C. Betsholtz[5,6], A. Ruusalepp[7,8], O. Franzen[2,8], T.L. Assimes[1], S.B. Montgomery[3,9], E.E. Schadt[2], J.L.M. Bjorkegren[2,6,10], T. Quertermous[1].* 1) Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, CA, USA; 2) Department of Genetics and Genomic Sciences, Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, NY, USA; 3) Department of Genetics, Stanford University School of Medicine, CA, USA; 4) Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; 5) Department of Immunology, Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden; 6) Department of Medical Biochemistry and Biophysics, Vascular Biology Unit, Karolinska Institutet, Stockholm, Sweden; 7) Department of Cardiac Surgery, Tartu University Hospital, Tartu, Estonia; 8) Clinical Gene Networks AB, Stockholm, Sweden; 9) Department of Pathology, Stanford University School of Medicine, CA, USA; 10) Department of Physiology, Institute of Biomedicine and Translation Medicine, University of Tartu, Estonia.

Coronary artery disease (CAD) is the leading cause of mortality and morbidity worldwide, and is driven by both genetic and environmental risk factors. Meta-analyses of genome-wide association studies in 184,305 cases and controls have identified 163 loci associated with CAD and myocardial infarction susceptibility in humans. A majority of these variants reside in non-coding regions and are co-inherited with hundreds of candidate regulatory variants, presenting a challenge to elucidate their functions. Herein, we use integrative genomic, epigenomic, and transcriptomic profiling of perturbed human coronary artery smooth muscle cells and normal and diseased tissues to begin to identify causal regulatory variation and mechanisms responsible for CAD associations. Using these genome-wide maps we prioritize 64 candidate variants and perform allele-specific binding and expression analyses at 7 top candidate loci: 9p21.3, SMAD3, PDGFD, IL6R, BMP1, CCDC97/TGFB1, and LMOD1. We also perform gain-of-function studies using site-specific integrase-mediated transgenesis in mice to measure the impact of a candidate regulatory variant on enhancer function in vivo, as well as loss-of-function CRISPR/Cas9-mediated genome editing in human cells. We further validate our findings in large external expression quantitative trait loci (eQTL) cohorts, which together reveal new links between CAD associations and regulatory function in the appropriate disease context.

303

***GNB5* variants cause a novel multisystem syndrome associated with sinus bradycardia and intellectual disability.** *P. De Nittis[1,2], E.M. Lodder[3,4], C.D. Koopman[3,4], W. Wiszniewski[5], C. Fishinger Moura de Souza[6], N. Lahrouchi[7], N. Guex[1,8], V. Napolioni[9], F. Tessadori[4], T. de Boer[3], L. Beekman[7], E.A. Nannenberg[10], I. Ratbi[11], A.A.M Wilde[7], W.F. Simonds[12], M. Neerman-Arbez[13], V.R. Sutton[5,14], F. Kok[15], J.R. Lupski[5,14,16], G. Merla[2], C.R. Bezzina[7], J. Bakkers[3,4], A. Reymond[1].* 1) Centre for Integrative Genomics, University of Lausanne, CH-1015 Lausanne, Switzerland; 2) Medical Genetics Unit, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Foggia, Italy; 3) Department of Medical Physiology, Division of Heart and Lungs, University Medical Center Utrecht, Utrecht, The Netherlands; 4) Hubrecht Institute-KNAW, University Medical Centre Utrecht, Utrecht, The Netherlands ; 5) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA; 6) Medical Genetics Service, Hospital de Clinicas de Porto Alegre, Porto Alegre, Brazil; 7) Department of Clinical and Experimental Cardiology, Heart Center, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; 8) SIB-Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland; 9) Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Palo Alto, CA, USA; 10) Department of Clinical Genetics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; 11) Département de Génétique Médicale, Mohammed V University of Rabat, Morocco; 12) Metabolic Diseases Branch/ NIDDK, National Institutes of Health, Bethesda, MD, USA; 13) Department of Genetic Medicine and Development, University Medical Centre (CMU), 1211 Geneva, Switzerland; 14) Texas Childrens' Hospital, Houston, TX, USA; 15) Child Neurology Division, Department of Neurology of University of Sao Paulo School of Medicine, Sao Paulo, Brazil; 16) Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA.

We describe a novel causal locus for an autosomal recessive genetic syndrome manifesting with neurological and cardiac features. We propose the acronym HEIGHTS for the combination of unique finding of early-onset Heart rate disturbance with Eye disease, Intellectual disability, Gastric problems, HypoTonia and Seizures. Whole exome sequencing of 9 individuals from 6 unrelated families, 4 different continents, and with overlapping clinical manifestations, identified bi-allelic loss-of-function and missense variants in the *GNB5* gene. It encodes the G-protein β-subunit 5. G-protein coupled signaling plays a crucial role in neuronal communication, including regulation of the antagonistic effects of the parasympathetic and sympathetic branches of the autonomic nervous system. A striking association between the nature of the variants and the clinical severity of HEIGHTS syndrome was observed, as individuals with loss-of-function alleles had more severe symptoms than carriers of a missense variant. Consistent with the possible involvement of *GNB5* in HEIGHTS syndrome, ablation of the orthologous genes in mouse and zebrafish recapitulated the phenotypic spectrum of affected individuals. Knocked-out mice presented with neurobehavioral developmental delay, learning deficiency, impaired gross motor coordination, defective visual adaptation and perturbed development and functioning of retinal bipolar cells. Knock-out of both *GNB5* paralogs, *gnb5a* and *gnb5b,* in fish larvae resulted in an abnormal escape response, a symptom of neurological dysfunction and possibly muscle hypotonia. Heart rate measurements showed that g*nb5a*/*gnb5b* signaling was crucial for its parasympathetic control, suggesting that the lack of *GNB5* is associated with extreme bradycardia at rest. Indicative of the dependency of proper eye-movement control on *GNB5* function, optokinetic response was severely affected in mutant larvae. Collectively, our work provides for the first time evidence for a direct role of *GNB5* in the control of heart rate, motor capacity, and vision.

304

Biallelic hypomorphic mutations in a linear deubiquitinase define otulipenia, an early-onset systemic autoinflammatory disease. *Q. Zhou[1], X. Yu[2], E. Demirkaya[3], N. Deuitch[1], D. Stone[1], W. Tsai[4], H. Wang[1], Y. Park[1], A. Ombrello[1], T. Romeo[1], E. Remmers[1], J. Chae[1], M. Gadina[4], S. Ozen[5], M. Abinun[6], D. Kastner[1], I. Aksentijevich[1].* 1) Inflammatory Disease Section, National Human Genome Research Institute, Bethesda, USA; 2) Genetics and Pathogenesis of Allergy Section, National Institute of Allergy and Infectious Diseases, Laboratory of Allergic Diseases, Bethesda, USA; 3) FAVOR, Institute of Health Sciences, R&D Center, Ankara, Turkey; 4) Translational Immunology Section, National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda, USA; 5) Department of Pediatric Nephrology and Rheumatology, Hacettepe University Faculty of Medicine, Ankara, Turkey; 6) Institute of Cellular Medicine, Newcastle University, Newcastle, United Kingdom.

Background: Autoinflammatory diseases are caused by mutations in genes regulating innate immune responses. More than 20 genes have been associated with various monogenic autoinflammatory disorders. **Methods:** We performed whole-exome and candidate gene sequencing in the patients and their unaffected family members. We used an NF-κB luciferase assay and overexpression experiments in 293 cells to confirm the causality of mutations. Patient samples were analyzed using immunoprecipitation, immunoblotting, gene expression and cytokine profiling. **Results:** We studied 3 unrelated families, one of Pakistani and two of Turkish ancestry, with neonatal-onset systemic inflammation, neutrophilic dermatitis/rash, and lipodystrophy. We identified three novel homozygous mutations in the *FAM105B* gene, which encodes OTULIN, the only deubiquitinase that specifically hydrolyzes Met-1 linked ubiquitin chains. The p.Leu272Pro and p.Tyr244Cys mutations are located near the linear ubiquitination binding region, while the p.Gly174Aspfs*2 mutation truncates the protein. Transfected OTULIN mutant plasmids showed decreased enzyme activity and a substantial defect in deubiquitination of target molecules NEMO, RIPK1, TNFR1 and ASC. This defect could be partially rescued by cotransfecting the mutation proteins with wild type OTULIN. Stimulated patients' fibroblasts and PBMCs showed increased phosphorylation of IKKα/β, IκBα, JNK, P38 and a higher linear-ubiquitinaton level of NEMO, RIPK1, TNFR1, and ASC. These results indicate that inefficient deubiquitination of OTULIN target proteins might explain increased NF-κB activity in mutant cells. Levels of IL-1β, TNF, IL-6, IL-18, IL-12, and IFN-γ were substantially increased in patient serum samples and stimulated cells. **Conclusion:** A new disorder caused by loss-of-function mutations in OTULIN expands the spectrum of autoinflammatory diseases caused by defects in deubiquitination and proteasomal degradation.

305

Characterization and successful treatment of a novel autosomal dominant immune dysregulatory syndrome caused by a *JAK1* gain-of-function mutation. *M.L. McKinnon[1,3], R.J. Ragotte[2,3], K.L. Del Bel[2,3], A. Saferali[2,3], S. Turvey[2,3].* 1) Department of Medical Genetics, University of British Columbia. C234 - 4500 Oak St, Vancouver, British Columbia, Canada. V6H 3N1; 2) Department of Pediatrics, Rm 2D19, 4480 Oak Street, BC Children's Hospital, Vancouver, BC V6H 3V4; 3) Child and Family Research Institute (CFRI), University of British Columbia, 950 West 28th Ave, Vancouver, BC, V5Z 4H4.

Introduction: Janus Kinase 1 (JAK1) plays an essential, non-redundant role in the JAK/STAT signaling cascade, a key pathway in the control of hematopoiesis and immune function. Significant progress has been made in elucidating the role of JAK1, but gaps in our knowledge still persist. Loss-of-function mutations in *JAK1* are perinatal lethal in mice, while somatic gain-of-function mutations have been linked to T-cell acute lymphoblastic leukemia. **Results:** We describe the first known patients carrying a germ-line gain-of-function mutation in *JAK1*. Two young children presented with a clinical phenotype that included severe atopic dermatitis, markedly elevated peripheral blood eosinophil counts with eosinophilic infiltration of the liver and gastrointestinal tract, hepatosplenomegaly, autoimmunity, and failure to thrive. Their mother, shown to possess a *de novo,* somatic mosaic mutation in *JAK1* presented with an intermediate phenotype and mosaic pattern of atopic dermatitis. Functional analysis established the gain of function phenotype caused by the mutation and in vitro studies demonstrated that the enhanced signaling could be controlled by ruxolitinib, an approved JAK1/2 inhibitor. Informed by these experimental data, the patients were treated with ruxolitinib with remarkable improvement in a variety of clinical end-points, including hematological profiles and growth parameters. **Conclusion:** This characterization of a human *JAK1* gain-of-function mutation expands ourcurrent understanding of the role of JAK1 in eosinophil biology, hematopoiesis and immune function. This case highlights successful drug-repurposing and pathway-specific targeted therapy through the use of ruxolitinib, an approved JAK1/2 inhibitor, in this novel genetic disease.

306

Mutation spectrum of *NOD2* in an early-onset inflammatory bowel disease cohort reveals recessive Mendelian inheritance as a main driver of Crohn's Disease. *J.E. Horowitz*[1]*, N. Warner*[2]*, J. Staples*[1]*, R. Murchie*[2]*, C. Van Hout*[1]*, A. King*[1]*, K. Fiedler*[2]*, J.G. Reid*[1]*, J.D. Overton*[1]*, A.R. Shuldiner*[1]*, A. Baras*[1]*, F. Dewey*[1]*, A. Griffiths*[2]*, O. Gottesman*[1]*, A. Muise*[3]*, C. Gonzaga-Jauregui*[1]*, Geisinger-Regeneron DiscovEHR Collaboration.* 1) Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY, USA; 2) SickKids IBD Centre, Hospital for Sick Children, University of Toronto, Toronto, Canada; 3) Departments of Pediatrics and Biochemistry, University of Toronto, Toronto, Canada.

Inflammatory bowel disease (IBD), clinically defined as Crohn's Disease (CD) or ulcerative colitis, results in chronic inflammation of the gastrointestinal tract in genetically susceptible hosts. IBD is typically diagnosed in the 3rd decade of life but can arise earlier. Pediatric-onset IBD is especially severe, with greater likelihood of intestinal stricturing, perianal disease, impaired developmental growth, and poor treatment response. GWAS have implicated 163 loci with IBD susceptibility and progression in adults. Of these, the nucleotide binding and oligomerization domain containing 2 (*NOD2*) gene is the most replicated locus associated with adult CD, to date. However, its role in pediatric-onset IBD is not well understood. We performed whole exome sequencing on a cohort of 1,183 probands with pediatric-onset IBD (ages 0-18y) and their affected or unaffected parents and siblings, where available. We first conducted trio-based analysis on 492 complete trios for gene identification and utilized the remaining 691 probands for replication of candidate genes. In our initial analyses, we identified 12 families with recessive compound heterozygous or homozygous variants in *NOD2* (MAF<2%). Our observation that some of these rare variants occur in *trans* with more common, previously-reported CD risk alleles (2%<MAF>5%) led us to survey additional probands for recessive inheritance of *NOD2* variants. From this, we identified 103 probands with recessive *NOD2* variants, carrying a known *NOD2* CD-risk allele in addition to either another known *NOD2* CD-risk allele or a completely novel *NOD2* variant. To determine the contribution of recessive inheritance of rare and low frequency *NOD2* alleles to IBD in the general population, we surveyed 1,146 adult IBD patients from the Regeneron Genetics Center-Geisinger Health System DiscovEHR study that links whole exome data to electronic health records. Here, we found that ~7% of cases in this cohort could be attributed to recessive inheritance of *NOD2* variants, including 18% of CD cases. Of these, 1% had a diagnosis before 18y, consistent with early-onset CD. In sum, ~8% of the probands in our pediatric-onset IBD cohort conform to a recessive, Mendelian mode of inheritance for rare and low frequency (MAF<5%) deleterious variants in *NOD2*. We confirmed this recessive inheritance in an adult IBD cohort and identified several early-onset CD cases. Collectively, our findings implicate *NOD2* as a Mendelian disease gene for early-onset IBD.

307

Homozygous *BRCA1* truncation causes Fanconi Anemia. *A. Seo*[1]*, T. Walsh*[1,2]*, N. Rosenfeld*[3]*, M.K. Lee*[1,2]*, O. Dgany*[4]*, E. Levy-Lahad*[3]*, A. Shimamura*[5]*, M.-C. King*[1,2]*, H. Tamary*[4,6]*.* 1) Department of Genome Sciences, University of Washington, Seattle, WA; 2) Division of Medical Genetics, University of Washington, Seattle, WA; 3) Medical Genetics, Shaare Zedek Medical Center, Hebrew University-Hadassah School of Medicine, Jerusalem, Israel; 4) Hematology-Oncology, Felsenstein Medical Research Center, Petah Tikva, Israel; 5) Bone Marrow Failure and Myelodysplastic Syndrome Programs, Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Boston, MA; 6) Hematology Unit, Schneider Children's Medical Center of Israel, Petah Tikva, Israel.

Inherited loss-of-function mutations in *BRCA1* significantly increase breast and ovarian cancer risk, and recent evidence suggests that inherited biallelic *BRCA1* mutations lead to the inherited bone marrow failure syndrome Fanconi Anemia, subtype S (FANCS). The two FANCS patients described to date were each compound heterozygotes for a truncation and a hypomorphic missense allele in *BRCA1*. Here, we report the first evidence of a homozygous *BRCA1* truncation (p.W372X, c.1115G>A, exon 11; NM_007294) found in two affected sisters with clinical features of Fanconi Anemia. The affected sisters showed microcephaly, bird-like face, developmental delay, café au lait spots, and optic nerve hypoplasia. Although CBCs were normal, chromosome breakage testing on blood was positive with increased chromosomal breaks and radial formation in response to either mitomycin C or diexpoxybutane. The older sister was diagnosed with T-cell ALL at age 5 years and subsequently died. The younger sister is still alive at age 4 years. Family history included a great uncle with gastrointestinal cancer at age 55 years, but no history of breast or ovarian cancer. The parents of the two sisters are younger than 40 years. Sequencing of all other known Fanconi Anemia genes was negative for mutations. Deep sequencing of *BRCA1* did not show any evidence of somatic reversion to wildtype. RT-PCR from peripheral blood RNA from the father, a heterozygous carrier, suggests that the mutant transcript is expressed in lower abundance than the wildtype transcript. We speculate that although full length BRCA1 protein is not present in the affected sisters, the naturally occurring splice variant lacking exon 11 provided enough BRCA1 function to escape embryonic lethality, but not enough for normal development. This study suggests that certain severely damaging inherited mutations in *BRCA1* are compatible with life as homozygotes, and further strengthens the argument for *BRCA1* being not only a cancer susceptibility gene but also a Fanconi Anemia gene. Sequencing *BRCA1* in patients with pediatric bone marrow failure, particularly patients from consanguineous families, may reveal other homozygous truncating alleles. Supported by NIH grants R24DK099808 and R35CA197458.

308

Whole genome sequencing and imputation in inflammatory bowel disease uncovers mechanisms for multiple therapeutically-relevant associations.

K.M. de Lange[1], Y. Luo[1,2,3], L. Moutsianas[1], J.C. Lee[4], L. Jostins[5,6], C. Lamb[7], N. Kennedy[8], J.C. Mansfield[7], M. Parkes[4], C.A. Anderson[1], J.C. Barrett[1], UK Inflammatory Bowel Disease Genetics Consortium. 1) Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK; 2) Division of Genetics and Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 3) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 4) Inflammatory Bowel Disease Research Group, Addenbrooke's Hospital, Cambridge, UK; 5) Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, UK; 6) Christ Church, University of Oxford, St Aldates, UK; 7) Newcastle University, Newcastle upon Tyne, UK; 8) Gastrointestinal Unit, Wester General Hospital University of Edinburgh, Edinburgh, UK.

GWAS have identified 210 risk loci for inflammatory bowel disease (IBD), nearly all of which are driven by common variants. To investigate lower frequency variants (MAF<5%), we sequenced the whole genomes of 4280 IBD patients at low coverage, and compared them to 3652 population controls at 76 million sites. A significant burden of rare, damaging missense variants was detected in genes previously implicated by IBD GWAS ($P<1x10^{-5}$), but only the well-known gene *NOD2* reached individual significance. This suggests that while informative allelic series may exist at many IBD genes, it will require sequencing tens of thousands of samples to identify them. We imputed into new and existing GWAS cohorts using a reference panel augmented with our sequence data, and tested for association at 12 million sites in 35,275 samples. We identified a 0.6% risk-increasing missense variant in *ADCY7*, which encodes an adenylate cyclase. These enzymes convert ATP to cAMP, a modulator of innate and adaptive immune function. Notably, this was the only such variant (MAF<1%,OR~2) detected, suggesting such alleles make a relatively minor contribution to disease risk. Indeed, while low-frequency SNPs represent 81% of the variants included in this analysis, they explain only 1.5% of variation in disease liability. We also meta-analyzed our data with published summary statistics (59,957 samples total) and identified 26 novel risk loci. Three of our new loci and 2 published loci contain integrin genes, which encode IBD therapeutic targets. At 4 of these 5 loci (*ITGA4*,*ITGB8*,*ITGAL*,*ICAM1*) the signal co-localizes with monocyte LPS-response eQTLs, with upregulation consistently associated with IBD risk. This reveals the specific mechanism linking genetic association to targets of existing therapeutics. We then fine-mapped 65 loci, and identified likely causal missense variants in the primary immunodeficiency gene *PLCG2* and the negative regulator of inflammation, *SLAMF8*. We have performed one of the largest whole genome sequence-based studies for a complex disease to date, coupled with a large new GWAS in the same disease, to greatly extend the allele frequency spectrum tested for association to IBD. We find multiple associations of high therapeutic relevance, underscoring the promise of extending current GWAS approaches to integrate sequencing and context-specific expression data. We caution that fully exploring the role of rare variation will be an expensive and logistically complicated endeavor.

309

A functional genome-wide genomics approach identifies genetic variations that contribute to variability in innate immune responses.

V.K. Magadi Gopalaiah[1], Y. Li[1], M. Oosting[2], S. Smeekens[2], M. Jaeger[2], R. Aguirre-Gamboa[1], T. Schoffelen[2], A.F.M. Jansen[2], M. van Deuren[2], J.W.M. van der Meer[2], R.J. Xavier[3,4], M.G. Netea[2], C. Wijmenga[1], Human Functional Genomics Project. 1) University of Groningen, University Medical Center Groningen, department of genetics, Groningen, the 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.

The variability of host immune responses determines individual susceptibility to immune-mediated diseases such as infections, autoimmune diseases, inflammatory diseases and the severity of disease experienced. Cytokine production is a key process that maintains the immune balance by regulating both innate and adaptive arms of immunity. There have been no genome-wide assessments of the genetics of cytokine production in humans so far. We therefore used two independent cohorts of 500 and 200 healthy individuals from the Human Functional Genomics Project to study the impact of genetic variation on cytokine production in response to major human pathogens. We performed the stimulation of three different cellular systems (whole-blood, peripheral blood mononuclear cells (PBMC), and macrophages) with a broad panel of bacterial, fungal, viral, and non-microbial stimuli to induce cytokine production, which was correlated with approximately 8.0 million genetic variants (SNPs). The discovery was performed in the 500FG cohort and the validation in the 200FG cohort. We identified 17 novel genome-wide significant ($P < 5 x 10^{-8}$) cytokine production QTLs (cQTLs), some of which map to genes that encode proteins known to influence immune responses, like the *TLR1-6-10* cluster, cytokine and complement inhibitors, and the kallicrein system. The genetic heritability of cytokine production capacity was strongest for the IL-1b/IL-6 pathway. We also observed a significant epistasis between loci that regulate cytokine production in response to various microbial stimuli. The cQTLs are more often located in regions under positive selection, and are significantly enriched among SNPs associated with infections and immune-mediated diseases. In conclusion, we present the first comprehensive analysis of how genetic variation affects cytokine production capacity in humans. Genetic variation is shown to be a crucial factor determining variability in innate immune responses, and this study demonstrates the power of a systems biology approach in identifying new regulatory pathways in immunity.

310

Assessment of the genetic basis of rosacea severity by genome-wide association study and expression analysis highlights immuno-inflammation and skin pigmentation as key mechanisms.

J.L. Aponte[1], M.N. Chiano[3], L.M. Yerges-Armstrong[2], D.A. Hinds[5], C. Tian[5], A. Gupta[4], D. Rajpal[2], J. Freudenberg[2], M.G. Ehm[2], D.M. Waterworth[2]. 1) Genomic Medicine, PAREXEL International, Research Triangle Park, NC; 2) Target Sciences, GlaxoSmithKline, King of Prussia, PA; 3) Target Sciences, GlaxoSmithKline, Stevenage, UK; 4) Stiefel, GlaxoSmithKline, Research Triangle Park, NC; 5) 23andMe Inc., Mountain View, CA.

Rosacea is a common, chronic skin disease of variable severity with limited treatment options. The exact pathogenesis of rosacea is unknown. Insight into what drives the severity of this disease and novel therapeutic interventions for rosacea may be gained from a genome-wide investigation in a well-powered study. Therefore, we performed a genome-wide association study (GWAS) of self-reported rosacea severity data in a large cohort of 73,265 individuals of European ancestry who were participants of the genetics company 23andMe. This is the largest GWAS study of rosacea to date and by including a severity score we were able to detect many additional loci over and above a simple binary definition of disease. Seven loci with variants significant at the genome-wide level ($p \leq 5 \times 10^{-8}$) were identified. Fine mapping analyses highlighted likely effector or flanking genes for these loci, which included: *IRF4* (rs12203592, $p=8.6\times10^{-17}$), a HLA region flanked by *PSMB9* and *HLA-DMB* (rs57390839, 9.7×10^{-15}), *HERC2* (rs1129038, 1.3×10^{-11}), *SLC45A2* (rs16891982, 4.5×10^{-10}), *IL13* (rs847, 6.6×10^{-9}), a region flanked by *NRXN3* and *DIO2* (rs149851565, 9.4×10^{-9}), and a region flanked by *OVOL1* and *SNX32* (rs77779142, 2.7×10^{-8}). All associations except the HLA locus association have not been reported previously. Of these seven associated loci and another precedented variant with a p-value just below the significance threshold, rs1805007 ($p \leq 2.5\times10^{-7}$, *MC1R*), three have previously been associated with skin phenotypes/pigmentation (*HERC2*, *SLC45A2* and *MC1R*), two are associated with genes previously associated with immuno-inflammation phenotypes (*IL13*, *HLA-DMA/B*), and one gene (*IRF4*) is associated with both categories. Genes located in three of the loci (*PSMB9-HLA-DMA*, *HERC2*, and *NRX3-DIO2*) were differentially expressed in a previously published clinical rosacea transcriptomics study comparing lesional to non-lesional samples. Together, these loci provide insight into the underlying disease processes in rosacea, suggesting a combination of skin type and susceptibility to inflammation, though none are themselves obvious drug targets.

311

Joint omics analysis connects human IgG *N*-glycosylation to multiple immunological loci.

X. Shen[1,2,3], L. Klarić[1,3,4], M. Mangino[5,6], I. Trbojević-Akmačić[4], M. Pučić-Baković[4], I. Rudan[1], O. Polašek[7], C. Hayward[3], T.D. Spector[5], J.F. Wilson[1,3], G. Lauc[4,8], Y.S. Aulchenko[1,9,10,11]. 1) Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, United Kingdom; 2) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; 3) MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; 4) Genos Glycoscience Research Laboratory, Zagreb, Croatia; 5) Department for Twin Research, King's College London, London, United Kingdom; 6) National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St. Thomas' Foundation Trust, London, United Kingdom; 7) Faculty of Medicine, University of Split, Split, Croatia; 8) University of Zagreb faculty of Pharmacy and Biochemistry, Zagreb, Croatia; 9) PolyOmica, Groningen, The Netherlands; 10) Novosibirsk State University, Novosibirsk, Russia; 11) Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia.

Jointly modeling of a number of phenotypes using multivariate methods has been often neglected in genome-wide association studies (GWAS). Modern omics techniques allow characterization of functional phenomena with a large amount of related phenotypes, which can gain from joint analysis. Here, we develop a multi-phenotype GWAS workflow in structured populations, perform association analysis using 23 immunoglobulin G (IgG) *N*-glycosylation phenotypes in 1,960 individuals, and conduct two sets of replication analyses in three independent cohorts (with total size of 6,169). Our multi-phenotype method has identified five known (*ST6GAL1, B4GALT1, FUT8, SMARCB1-DERL3, SYNGR1-TAB1-MGAT3*) and allowed for discovery and replication of five novel loci (*IGH, ELL2, HLA-B-C, AZI1-TMEM105, FUT6-FUT3*) associated with the human IgG *N*-glycome. Three of the novel loci included a strong immune system related candidate gene: the immunoglobulin heavy locus (*IGH*), IgG transcription factor *ELL2,* and the human leukocyte antigen (*HLA-B-C*). The IgG *N*-glycosylation loci share pleiotropy with multiple eQTL and complex disease. By subsequent functional analysis, we found the established IgG *N*-glycosylation loci to date show strong enrichment in hemic and immune system, antibody producing and B cell lines. We emphasize the use of multivariate analysis in large-scale association studies, for which we foresee a great potential in revealing complex biology underlying high-throughput omics data.

312

Causal effects of protein biomarkers on immune diseases. *A. Johansson, U. Gyllensten, S. Enroth.* Uppsala Univeristy, Uppsala, Sweden.

Protein biomarkers are associated with development of diseases or disease outcomes and might therefore serve as potential drug targets. However, most biomarkers are increased as a response to disease, rather than being directly casual for disease pathogenesis. In this study we have used Mendelian randomization to evaluate the causal effect of protein biomarkers that are associated with cardiovascular or immune diseases. A total of 144 protein biomarkers were measured and cis-regulatory genetic variants influencing the protein levels identified in a Swedish cohort (N=961). Using cis-regulatory genetic variants we constructed models to predict the genetically determined protein levels. The models were then applied to compute genetic scores, in data from the UK biobank (N=152249), and evaluate the causal effects of the biomarkers. A total of 13 immune or cardiovascular phenotypes were investigated. Results are reported as odds ratio (OR) per one standard unit increase in biomarker levels for all associations that were significant after adjusting for multiple testing (FDR q-value < 0.05). We identified cis-regulatory genetic variants for 29 biomarkers of which at least one SNP meet the threshold for genome-wide significance ($p<5\times10^{-9}$). We show that many inflammatory biomarkers directly increase the risk of inflammatory diseases including IL6RA levels with risk of the composite phenotype hay fever, allergic rhinitis or eczema (OR =1.045, $p=9.6\times10^{-7}$), both CCL24 and IL6RA with risk of eczema (OR = 1.026, p= 0.00084 and OR= 1.092, $p= 9.9\times10^{-5}$ respectively), and MIC-A with asthma (OR = 1.035, $p= 3.3\times10^{-6}$) whereas IL-12 decrease risk of psoriasis (OR= 0.72, $p =1.0\ 10^{-5}$). We also identified an association between CXCL10 and allergy or anaphylactic reaction to foods (OR= 1.77, $p = 4.9\times10^{-6}$), and to drugs (OR= 1.44, P= 0.00071), an association that is partly driven by a deleterious missense variant (rs11548618). No biomarkers were associated with any cardiovascular phenotype after adjusting for multiple testing. We have shown that many biomarkers that are associated with immune diseases are directly causal for a disease phenotype. However, most biomarkers for cardiovascular disease showed weak evidence for being causal and are therefore more likely to be increased as a response of disease progression. Acknowledgments: This research has been conducted using the UK Biobank Resource.

313

RUFUS: Accurate and sensitive reference free variant detection. *A. Farrell[1,2], D. Lee[1,2], G. Marth[1,2].* 1) Human Genetics, University of Utah, Salt Lake City, UT; 2) USTAR Center for Genetic Discovery, Salt Lake City, UT.

We developed a novel k-mer based variant detection tool, RUFUS, that vastly improves specificity and sensitivity for germline and somatic/tumor *de novo* mutations, and may reveal some of the missing heritability in many genetic diseases. RUFUS is based on direct k-mer comparison, removing the reference from variant detection, and any associated reference bias. This vastly improves the detection of medium sized (20-500bp) insertions/deletions (INDELs) that current methods are unable to reliably detect: small variant detectors (e.g. GATK, FreeBayes) are effective at finding 1-20bp events in read alignments; and structural variant callers (LUMPY, WHAM, etc.) are effective at >500bp events where insert size variations and read coverage anomalies can be confidently detected. As a result, medium length, INDELs have been missed by most sequencing studies. We are currently applying RUFUS to 519 family quartets (mother, father, autistic child, unaffected sibling), a total of 2,076 samples sequenced by the Simon's Foundation Autism Research Initiative to 30x whole genome coverage. This will be the largest *de novo* variation study to date, and combined with RUFUS's sensitivity for all variant types and sizes will provide the most complete picture of *de novo* variation ever constructed. Our preliminary data on a 40 family pilot study has shown that the rate of medium length *de novo* events is twice that of structural events (12 medium-length vs 6 SV events), suggesting that these events may be more common than previously thought.In addition to increased sensitivity for variants of all sizes, RUFUS also shows far higher specificity over mapping based approaches. Previous research has suggested that the human per-nucleotide *de novo* SNV mutation rate is ~1.25×10^{-8}, or roughly 75 mutations per generation. In our analysis of numerous disease family trio data sets at the University of Utah, RUFUS finds between 77 and 116 *de novo* mutations per child genome, including 74-101 SNV mutations; 98% also seen in mapping-based variant calls. Conversely, traditional mapping based methods call on average 150,000 *de novo* calls per child, dominated by mapping and reference errors, drowning out true variation, and post-processing and genome masking is necessary to improve these, still leaving thousands of *de novo* calls. RUFUS requires no filtering or masking of the genome, enabling true genome wide variant detection of all mutation types, at uniquely high specificity.

314

A hybrid approach for *de novo* human genome sequence assembly, phasing, and detection of complex structural variation. *Y. Mostovoy[1], M. Levy-Sakin[1], J. Lam[1], E.T. Lam[2], A.R. Hastie[2], P. Marks[3], J. Lee[2], C. Chu[1], C. Lin[1], Z. Džakula[2], H. Cao[2], S.A. Schlebusch[4], K. Giorda[3], M. Schnall-Levin[3], J.D. Wall[5], N.J.L. Meeks[6], K.C. Chatfield[6], C.R. Coughlin II[6], T.H. Shaikh[6], P. Kwok[1,5,7].* 1) Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA; 2) BioNano Genomics, Inc., San Diego, CA; 3) 10X Genomics, Inc., Pleasanton, CA; 4) Department of Molecular and Cell Biology, University of Cape Town, Cape Town, South Africa; 5) Institute for Human Genetics, University of California, San Francisco, San Francisco, CA; 6) Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO; 7) Department of Dermatology, University of California, San Francisco, San Francisco, CA.

Despite tremendous progress in genome sequencing, the basic goal of producing a phased (haplotype-resolved) genome sequence with end-to-end contiguity for each chromosome at reasonable cost and effort is still unrealized. In this study, we describe an approach to performing *de novo* genome assembly and experimental phasing by integrating data from Illumina short-read sequencing, 10X Genomics linked-read sequencing, and BioNano Genomics genome mapping to yield high-quality, phased, *de novo* assembled human genomes. Our approach is well-suited to detecting classes of structural variants that are often inaccessible to Illumina short-read data alone, such as large, complex inversions, insertions, and copy-number variants. Our pilot project using human sample NA12878 achieved an assembly with an N50 of over 30 Mb, with scaffolds that in some cases spanned entire chromosomal arms. The accuracy of the assembly matched or exceeded that of previously published assemblies for this sample, with phase block lengths far exceeding those in previous publications. We were readily able to assemble inversion polymorphisms that were known to be present in the sample, as well as to detect thousands of other inversions, insertions, and deletions. Expanding on our pilot project, we applied our approach to additional human genomes from diverse populations, permitting us to characterize population-specific genomic architecture. To take advantage of our ability to characterize complex structural variants, we further applied our technique to the genomes of multiple patients with microdeletion syndromes, including the 22q11.2, 7q11.23, 15q13.3 and 16p12.2 microdeletion syndromes. We leveraged our hybrid datasets to characterize and phase the structural variants at these highly complex microdeletion-prone loci, permitting detailed analysis of the architecture associated with pathogenic deletions in affected samples. In this study, we demonstrate the power of our hybrid approach to achieve *de novo* genome assemblies with high contiguity, long phased haplotype blocks, and detection of complex structural variation.

315

A hybrid approach combining next and third generation sequencing data for powerful structural variant detection. *X. Fan[1,2], Z. Chong[1], L. Nakhleh[2], K. Chen[1,2], Human Genome SV Consortium.* 1) Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, Houston, TX; 2) Department of Computer Science, Rice University, Houston, TX.

Structural variation (SV) detection, although essential in understanding genetic diseases and population diversity, has not yet been fully resolved due to complex sequence context and limited read length of the Next Generation Sequencing (NGS) data. The Third Generation Sequencing (TGS) technology produces long reads, facilitating structural variation detection. Nevertheless, their high error rate (>15%) challenges accurate reference alignment, leading to potential false negative and false positive calls. We propose a reference-free, hybrid approach that combines reads from NGS and TGS to produce accurate SV call set. By identifying the variant-containing NGS reads and aligning them to the TGS reads, the approach extracts the TGS reads that span SV breakpoints and assembles them into long contigs. It then aligns the assembled contigs to the reference, performs variant calling, and confirms the detected breakpoints using the NGS reads. The approach takes the advantage of the high accuracy of NGS and the long read length of TGS. We applied the approach to a well-studied haploid hydatidiform mole genome CHM1 and the three trios in the Human Genome Structural Variation Consortium (formerly the 1000 Genomes SV group). The result showed that our method, while having a low false discovery rate, can effectively detect structural variations that have weak signals in either NGS or TGS data, which are challenging to existing NGS- or TGS-specific methods. Particularly, our approach is advantageous in detecting large novel insertions, small INDELs in the twilight zone (10-50bp) including those with alternate copies of short tandem repeats. In all, by combining NGS and TGS reads synergistically, our proposed approach helps to achieve high quality personal genome analysis.

316

De novo assembly of individual human haplotypes from diploid samples. *D. Church[1], R. Abbas[1], A. Fehr[1], B. Galvin[1], S. Garcia[1], H. Heaton[1], P. Hardenbol[1], J. Herschleb[1], C. Jabara[1], M. Imielinski[2], S. Kyriazopoulou-Panagiotopoulou[1], V. Kumar[1], P. Marks[1], H. Ordonez[1], M. Pratt[1], P. Shah[1], A. Xu[1], N. Weisenfeld[1], I. Yousif[1], G. Zheng[1], M. Schnall-Levin[1], D. Jaffe[1].* 1) 10x Genomics, Pleasanton, CA; 2) Weill Cornell Medical College, Cornell University, New York, NY.

Traditional genome analysis involves sequencing a genome, aligning the reads to a reference assembly and then identifying differences between the sequenced reads and the reference sequence. While this approach has expanded our knowledge of population variation, our picture of both global and individual genomic variation remains far from complete. Importantly, the ability to perform long-range haplotype reconstruction is severely limited for individual samples using traditional approaches. A significant challenge is that reads from each haplotype are indistinguishable from each other in the alignment phase. If the haplotypes in the sample are significantly different from each other and from the reference, correct reconstruction of genotypes and long-range haplotypes becomes impossible. We describe a novel approach for the *de novo* assembly of individual human genomes, requiring only 1ng of input DNA. We have developed a microfluidic system that allows for the high-throughput partitioning of high-molecular weight DNA. Unique barcodes are applied within each partition, allowing for the retention of long-range information even when using short read sequencing, creating a data type called Linked-Reads. The Supernova Assembler™ takes advantage of Linked-Reads to perform *de novo* diploid assembly. Analysis of several individual human genomes demonstrates that we can obtain accurate long assemblies, typically with scaffold N50s greater than 10 Mb, containing multi-megabase phased haplotypes, with N50 phase blocks of 2 Mb in Caucasian samples. Molecule length and sample heterozygosity have the biggest impact on the length of the phased haplotypes. Initial analysis suggests gene representation is good, with >99% of protein coding genes aligning to the assembly, with >94% of these alignments covering 95% or more of the coding sequence and less than 3% contain frameshifting errors. Preliminary analysis also suggests improved power for identifying novel insertions and other complex variants. This approach provides additional power to discriminate heterozygous haplotype differences, especially for more complex variant types or in regions where both haplotypes differ from the reference. We will describe the *de novo* assemblies of three related individuals, and the assembly-based analysis of their genomes. We will contrast this with reference-based approaches and describe approaches for integrating data from both reference and assembly-based approaches.

317

Monovar: Single-nucleotide variant detection in single cells. *H. Zafar[1,2], Y. Wang[3], L. Nakhleh[1], N. Navin[2,3], K. Chen[2].* 1) Department of Computer Science, Rice University, Houston, TX, USA; 2) Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA; 3) Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

Single cell sequencing (SCS) has emerged as a powerful new method for resolving genomic heterogeneity in complex cell populations and profiling mutations in rare subpopulations. SCS tools have had a major impact on diverse fields of biology, including cancer research, neurobiology, microbiology, immunology and development. However, inherent technical errors in SCS datasets, including false-positive (FP) errors, allelic dropout (ADO) events and coverage non-uniformity plague the bioinformatics analysis of such data. Existing DNA variant detection algorithms were designed for bulk tissue samples and therefore make many assumptions regarding the underlying properties of the sequencing data. When existing variant callers are applied on SCS datasets, a huge number of sequencing artifacts are incorrectly reported as real biological variations. To address these problems, we developed Monovar (https://bitbucket.org/hamimzafar/monovar), a novel statistical method for detecting and genotyping single nucleotide variants (SNVs) in SCS data. Monovar leverages data from multiple cells to mitigate non-uniform coverage distribution across single cells. The underlying probabilistic model of Monovar accounts for false-positive errors and allelic dropout events. A candidate site is characterized as SNV based on the posterior probability calculated using Bayes' rule. Efficient variant calling and genotyping are performed using a dynamic programming approach. Monovar exhibited superior performance over standard algorithms like GATK and Samtools in three simulated benchmark datasets. An isogenic fibroblast cell line was used to validate the sensitivity and specificity of Monovar on real SCS datasets. Next, we applied Monovar to delineate clonal substructure and identify driver mutations in three human tumor datasets. Our experiments suggest that Monovar shows substantial improvements over standard SNV calling algorithms for the analysis of SCS datasets produced using different whole genome amplification (WGA) protocols. Monovar being capable of analyzing large-scale datasets and handling different WGA protocols is well suited for addressing the growing need for accurate single-cell DNA variant detection. As SCS methods move into the clinic, we expect that Monovar will be used for important applications in cancer diagnosis and treatment, personalized medicine and pre-natal genetic diagnosis, where the accurate detection of SNVs is critical for patient care.

318

RNA-seq analysis of Parkinson disease patient-derived dopaminergic neurons reveals disease specific genes and pathways across multiple time-points. *K. Belle*[1,2,3], *S. Sivasankaran*[1,2], *A. Mehta*[1], *D. Van Booven*[1,3], *M. Seignon*[1,3], *J. Vance*[1,2,3], *D. Dykxhoorn*[1,2,3]. 1) John P. Hussman Institute for Human Genomics, University of Miami, Miami, FL; 2) Dr. John T. MacDonald Foundation Department of Human Genetics, Miami, FL 33136; 3) University of Miami Miller School of Medicine, Miami, FL 33136.

Parkinson disease (PD) is a progressive neurodegenerative disease characterized by the loss of dopaminergic neurons (DAns) in the substantia nigra. However, the molecular and cellular mechanisms that drive pathogenesis are not well characterized. It has been suggested that impairment in cellular functionality may predate the clinical diagnosis by many years. One of the main difficulties in understanding PD pathogenesis has been a lack of model systems that recapitulate the complex genetic architecture found in the disease. Induced pluripotent stem cell (iPSC) based models of neuronal differentiation have increasingly been utilized to help understand the pathophysiology of many disorders including PD. To determine if there are convergent molecular mechanisms that drive PD pathogenesis, we derived a cohort of iPSC lines from individuals with PD, as well as age and ethnicity matched control lines, including lines from 17 PD patients and six unaffected individuals. These lines were differentiated into DAns and RNA-seq analysis was performed at multiple time-points during the differentiation process. RNA-seq data was analyzed at each time-point, generating a list of differentially expressed (DE) genes which were analyzed for enrichment in biological pathways or functions through gene set enrichment analysis (GSEA). EdgeR was used to identify sets of genes that were differentially expressed in PD cases compared to control neuronal cultures. We identified approximately 30, 60, and 50 genes distinguishing these two groups at 21, 45, and 120 days post initiation of neuronal differentiation, respectively. Of these genes, six were seen across multiple time-points, including TUBA4A, a neuronal microtubule protein used for cell stability and mitochondrial trafficking, and DDX43, shown to be downregulated during neurodegenerative stress. Additionally, PDGFRA, a gene with known roles in neuronal development and degeneration and CRYAB, which is highly expressed in the substantia nigra of PD patients. Gene set enrichment analysis showed overrepresentation of several pathways between the cases and controls, including focal adhesion (p=0.009), extracellular matrix receptor interactions (p=0.02), and cell adhesion molecules (p=0.009). Our results suggest that initial PD dysfunction may take place earlier than expected during neuronal differentiation.

319

RNA-seq analysis identifies phenotypic heterogeneity among *ex vivo* purified dopamine neurons and highlights their progressive temporal diversification. *P.W. Hook, S.A. McClymont, L.A. Goff, A.S. McCallion.* McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD.

Dopamine (DA) neuronal degeneration is a central pathognomonic feature of Parkinson disease (PD) and related disorders. However, not all DA neurons are functionally equal, nor are they uniformly sensitive to mutations underlying Mendelian Parkinsonian syndromes. To help illuminate the disease etiology of these conditions and the shared and discrete biology of these neurons, we set out to describe their molecular signatures. Using a transgenic mouse model (Tg(Th-EGFP)DJ76Gsat), expressing GFP in DA neurons, we can microdissect, dissociate and retrieve via FACS the DA neurons from regions of interest (ventral midbrain, MB; forebrain, FB; olfactory bulb) for RNA-seq-based analyses. To determine the baseline transcriptional signatures, we subjected sorted terminally differentiated FB and MB DA neurons (E15.5) to bulk RNA-seq, generating approximately 740 million reads (four biological replicates per region, ≤92e6 reads per library). By principal component analysis (PCA), the MB and FB populations clearly separate along PC1, indicating the largest source of variation in gene expression lies in the different transcriptional signatures of these DA populations. Preliminary differential expression analysis reveals a subset (n≤1379, FDR=0.01) of genes whose transcription levels differ significantly between these neuronal populations, offering potential insight into their discrete cellular functions and their relative robustness/sensitivity. To assay the cellular heterogeneity of these populations of neurons, we employed single-cell RNA-seq at multiple time points, selected to follow terminal differentiation (E15.5) and remodeling of DA populations (P7, P28). At E15.5 and P7, our data reveals distinct populations of DA neurons between and within neuroanatomical isolates through the use of dimensionality reduction algorithms, PCA and t-distributed stochastic neighbor embedding. Additionally, cellular heterogeneity increases over time, consistent with progressive functional specification of these populations. Analysis at P28 is ongoing. Validation of the spatial/temporal expression of selected genes via RNAscope and functional validation of findings are ongoing. However, our data is already illuminating investigations into PD-associated intervals. For 12 of 26 replicated PD GWAS loci, the gene closest to the lead SNP is not expressed in MB DA neurons at E15.5, although neighboring genes are. Results of these analyses and preliminary functional studies will be presented.

320

Identification of genetic modifiers of the age onset of amyotrophic lateral sclerosis associated with the expanded GGGGCC repeats. *H. Kim[1], H. Bao[1], B. Jiao[1], J.D. Glass[2], T. Wingo[2], P. Jin[1].* 1) Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA; 2) Department of Neurology, School of Medicine, Emory University, Atlanta, GA.

Amyotrophic lateral sclerosis (ALS), the most common form of neuromuscular diseases, is highly progressive and leads to death from respiratory failure within 2-5 years of symptom presentation. The time of diagnosis is between the ages of 40 and 70, with an average age of 55; therefore, age is one critical risk factor for ALS. In addition, genome-wide association studies elucidate that this disorder is mediated by genetic mutations. Among several ALS-associated mutations, the pathogenic expansion of GGGGCC (G_4C_2) repeats in the first intron of C9orf72 has been shown to explain the etiology of both familial ALS (~40%) and sporadic ALS (~25%). Despite the significance of G_4C_2 repeats in disease development, the age of onset in ALS patients with G_4C_2 repeats is variable, and any genetic variants responsible for the age of disease development have yet to be characterized. Given there is no effective treatment, the identification of such genetic factors will play a significant role in managing ALS. Here, we performed whole-genome sequencing of both early-onset (younger than 40-year old) and late-onset (later than 70-year old) ALS patients with expanded G_4C_2 repeats. We identified 135 and 54 genetic mutations for early-onset and late-onset cases, respectively. To further test the potential roles of these genes in expanded G_4C_2 repeat-mediated toxicity, we employed a Drosophila model expressing G_4C_2 repeats using the Gmr-GAL4 driver (G_4C_2 fly), which exhibits neurodegeneration. Of the 189 genes detected in ALS patients, we identified 105 fly orthologs, and then selected the corresponding RNAi transgenic lines (302), crossing each with transgenic flies expressing $r(G_4C_2)_{30}$. Notably, 38 RNAi lines displayed either enhanced or suppressed neuronal toxicity associated with G_4C_2 repeats. Further, targeted sequencing of 19 genes, including PLEKHG2, MYH15, KIF27, HK3, and, PDK3, among a larger cohort of ALS patients with G_4C_2 repeats identified potential genetic modifiers associated with the age of ALS onset. Taken together, our analyses suggest the existence of multiple genetic modifiers that could modulate the age-of-onset in ALS patients associated with G_4C_2 repeats.

321

Biallelic mutations in the nuclear 3’ exonuclease, *TOE1*, cause Pontocerebellar Hypoplasia Type 7 and result in snRNA processing defects. *A. Schaffer[1,2*], R. Markmiller[3*], V. Eggans[4*], M. Zaki[5], S. Sathe[2], S. Grainger[2], B. Rosti[1], E. Van Nostrand[2], Z. Schlachetzki[1], E. Scott[1], L. Heckman[1], E. Dikoglu[1], R. Rosti[1], N. Akizu[1], A. Gregor[1], A. Guemez-Gamboa[1], N. Foulds[6], W. Dobyns[7], N. Chi[8], D. Traver[2], L. Spaccini[9], S. Bova[10], S. Gabriel[11], M. Gunel[12], E.M. Valente[13], E. Bennett[3], G. Yeo[2,14], F. Baas[4], J. Lykke-Andersen[3**], J. Gleeson[1**].* 1) Laboratory of Pediatric Brain Disease, Howard Hughes Medical Institute, The Rockefeller University, New York, NY 10065; 2) Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA; 3) Division of Biological Sciences, University of California, San Diego, La Jolla, CA; 4) Department of Genome Analysis, Academic Medical Center, Meibergdreef 9,1105AZ Amsterdam, the Netherlands; 5) Clinical Genetics Department, Human Genetics and Genome Research Division, National Research Centre, Cairo 12311, Egypt; 6) Southampton University Hospitals Trust, Southampton, Hampshire, UK; 7) Seattle Children's Research Institute, Centre for Integrative Brain Research, Seattle, WA 98195, USA; 8) UCSD Cardiology, University of California San Diego, La Jolla, CA 92093, USA; 9) Clinical Genetic Unit, Dept. of Women, Mother and Neonates, "Vittore Buzzi" Children’s Hospital, Istituti Clinici di Perfezionamento, Milan, Italy; 10) Child Neurology Unit, Dept. of Pediatrics, "Vittore Buzzi" Children Hospital, Istituti Clinici di Perfezionamento, Milan, Italy; 11) Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02142, USA; 12) Yale Program on Neurogenetics, Departments of Neurosurgery, Neurobiology, and Genetics, Yale University School of Medicine, New Haven, CT 06510, USA; 13) IRCCS Casa Sollievo della Sofferenza, CSS-Mendel Institute, San Giovanni Rotondo, Italy; Dept. of Medicine and Surgery, University of Salerno, Salerno, Italy; 14) Department of Physiology, National University of Singapore and Molecular Engineering Laboratory, A*STAR, Singapore.

Pontocerebellar neurodegeneration has its onset so early as to overlap with neurodevelopment, and is thus alternatively referred to as pontocerebellar hypoplasia (PCH). We sought to determine the cause of PCH type 7, a genetically undefined unique recessive syndromic constellation of pontocerebellar atrophy with ambiguous genitalia. We recruited 12 families meeting phenotypic criteria, including the index family that defined the condition. Whole exome sequencing uncovered biallelic-inactivating mutations in *TOE1*, which encodes a conserved unconventional deadenylase. Patient mutations occurred in highly conserved amino acids and were found to destabilize TOE1 protein. Genetic ablation of Toe1 was embryonic lethal, and partial knockdown of Toe1 in zebrafish showed severe degeneration of the mid- and hind-brain regions of the morphants by 48 hours post fertilization as well as reduced and ectopically localized germ cells, phenocopying patients with PCH7. To determine how mutations in TOE1 might cause disease, we looked at bound proteins and RNAs by immunoprecipitation and sequencing. TOE1 was tightly associated with small nuclear RNA (snRNA)-protein (snRNP) complex proteins as well as incompletely processed spliceosomal pre-snRNAs containing 3’ extensions of post-transcriptionally added U/A nucleotides. Using an snRNA-processing assay, we found TOE1 activity was required in immune-isolated protein complexes for processing the 3’ extended snRNAs *in vitro*. Primary fibroblast and neural progenitor cells derived from patients, as well as TOE1 depleted cells, showed accumulation of U1, U2, U4 and U5 snRNAs containing these 3’ extensions, indicating snRNAs are abnormal in the disease state. Our findings reveal the cause of a unique brain malformation and uncover the long-sought 3’ exonuclease required for snRNA processing.

322

Identification of a 1.35 Mb insertion within the distal hereditary motor neuropathy locus (DHMN1) on chromosome 7q34-q36.2. *M.L. Kennerson*[1,2,3]*, A.P. Drew*[1]*, A.N. Cutrupi*[1,2]*, M.H. Brewer*[1,2]*, G.A. Nicholson*[1,2,3]*.* 1) Northcott Neuroscience, ANZAC Res Inst, Concord, Australia; 2) Sydney Medical School, University of Sydney, Camperdown, Australia; 3) Molecular Medicine Laboratory, Concord Hospital, Concord, Australia.

Rationale: Distal hereditary motor neuropathies (dHMNs) are a group of diseases predominantly affecting the motor neurons of the peripheral nervous system leading to chronic disability. We previously mapped the disease locus for a dHMN family (F-54) to a 12 Mb interval on chromosome 7q34-q36.2. Extensive mutation analysis of the region excluded all genes for coding, splice site and untranslated region mutations. We therefore hypothesised that either non-coding DNA or structural variation mutations are likely to be the cause of DHMN1 in family F-54. **Objective:** To identify the genetic mutation causing DHMN1 using whole genome sequencing (WGS). **Methods and results:** Using WGS we identified a novel structural variation (SV) within the DHMN1 locus on chromosome 7q34-q36.2. Split and discordant paired-end reads identified an intra-chromosomal translocation resulting in the insertion of a 1.35 Mb DNA fragment into the DHMN1 locus. The source of the inserted sequence was located 2.3 Mb distal to the disease locus at chromosome 7q36.3 and inserted in the reverse orientation. The SV segregated with the disease phenotype in family F-54 and was absent from 1000 neurologically normal control chromosomes. Five intact genes (*LOC389602*, *RNF32*, *LMBR1*, *NOM1*, *MNX1*) and one partial gene (*UBE3C*) are located within the inserted 1.35 Mb DNA fragment. This region has been duplicated from its original location and results in trisomy of the genes within the inserted DNA. The insertion site is located at the distal end of the DHMN1 locus between the genes *ACTR3B* and *DPP6* and does not disrupt any coding sequences. Twenty one genes are candidates for gene dysregulation caused by the SV based on proximity to the insertion site and trisomy of the introduced genes. In the absence of patient neural tissue, lymphoblasts are being used to investigate if the 1.35 Mb insertion alters expression of the candidate genes in DHMN1 patients. Preliminary data shows a significant decrease in expression of the *RNF32* gene. **Conclusion:** Our data suggests this novel SV insertion is the likely DNA mutation disrupting the DHMN1 locus. Our finding represents a new disease mechanism for hereditary motor neuropathies and highlights the growing importance of interrogating the non-coding genome for SV mutations in families which have been excluded for genome wide coding mutations. M.L.K, A.P.D and A.N.C contributed equally to this work.

323

Direct measurement of the mutagenic impact of recombination through deep genome sequencing of 519 families. *A. Quinlan*[1]*, T. Sasani*[1]*, B. Pedersen*[1]*, R. Layer*[1]*, A. Farrell*[1]*, R. Collins*[2]*, M. Stone*[3]*, H. Brand*[4]*, J. Glessner*[4]*, J. An*[5]*, D. Werling*[5]*, S. Dong*[5]*, M. Gilson*[5]*, L. Smith*[5]*, M. State*[5]*, A. Willsey*[5]*, X. He*[6]*, J. Buxbaum*[7]*, B. Devlin*[8]*, K. Roeder*[9]*, M. Daly*[3]*, H. Coon*[10]*, G. Marth*[1]*, M. Talkowski*[11]*, S. Sanders*[5]*.* 1) Department of Human Genetics, University of Utah, Salt Lake City, UT; 2) Program in Bioinformatics and Integrative Genomics, Division of Medical Sciences & Center for Human Genetics Research, Harvard Medical School & Massachusetts General Hospital, Boston, MA; 3) Center for Human Genetics Research, Massachusetts General Hospital, Boston, MA; 4) Center for Human Genetics Research & Department of Neurology, Massachusetts General Hospital & Harvard Medical School, Boston, MA; 5) Department of Psychiatry, University of California San Francisco, San Francisco, CA; 6) Department of Human Genetics, University of Chicago, Chicago, IL; 7) Department of Psychiatry, Mount Sinai, New York, NY; 8) Department of Psychiatry, University of Pittsburgh Medical Center, Pittsburgh, PA; 9) Department of Statistics, Carnegie Mellon University, Pittsburgh, PA; 10) Departments of Biomedical Informatics, Internal Medicine, and Psychiatry , University of Utah, Salt Lake City, UT; 11) Center for Human Genetics Research, Department of Neurology and Program in Bioinformatics and Integrative Genomics, and Program of Medical and Population Genetics, Massachusetts General Hospital, Harvard Medical School, and The Broad Institute, Boston, MA.

Meiotic recombination rearranges parental alleles, and the process of resolving these crossovers can be mutagenic. Multiple studies have shown that increased rates of allelic diversity are correlated with increased rates of recombination. More recent work has also demonstrated a link between double-strand breaks and increased SNP density in the surrounding genomic region (Pratto et al, 2014). However, direct investigations of the mutagenic impact of meiotic recombination in the human genome have not yet been undertaken, primarily owing to the cost of whole genome sequencing in families. Therefore, we sought to conduct the first direct measurement of the relationship between recombination and de novo mutation via deep (~40X) genome sequencing of 519 families of four (quartets) from the Simons Foundation Autism Research Initiative (SFARI). Following the method described by Coop et al (2008), we first identified recombination events by marking genomic sites where the inheritance state observed in one offspring changes relative to its sibling. We then developed a Hidden Markov Model to eliminate spurious crossovers owing to genotyping error. High confidence de novo mutations were detected in each of 1,038 offspring following a strategy inspired by McRae et al, 2016. Leveraging the high resolution maps of crossovers and de novo mutations from 519 families, we will quantify the mutagenic impact of recombination. We will further dissect whether specific genomic contexts (e.g. repeat, GC, and CpG content) are enriched for recombination and mutation hotspots. In addition, since our maps are ascertained directly from deep DNA sequencing data, we are able to leverage read-backed phasing to determine the parental germline origin of a subset of de novo mutations. By integrating detailed phenotypic information available for the SFARI collection, we will also investigate the impact of parent of origin and ancestry on the patterns and rates of crossover and mutation. Lastly, crossover hotspots have been shown to be enriched at the breakpoints of copy number variants arising from non-allelic homologous recombination (Pratto et al, 2014). We will therefore investigate and compare the interplay between recombination and spontaneous mutation (both single-nucleotide and copy-number) in affected and unaffected offspring.

324

A greater meiotic gene conversion rate in females increases with age. *B.V. Halldorsson[12], M.T. Hardarson[1], B. Kehr[1], U. Styrkarsdottir[1], A. Gylfason[1], G. Thorleifsson[1], F. Zink[1], Ad. Jonasdottir[1], As. Jonasdottir[1], P. Sulem[1], G. Masson[1], U. Thorsteinsdottir[13], A. Helgason[14], A. Kong[1], D. Gudbjartsson[15], K. Stefansson[13].* 1) Research Scientist, deCODE genetics, Reykjavik, IS, Iceland; 2) School of Science and Engineering, Reykjavik University, Reykjavík, Iceland; 3) Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland; 4) Department of Anthropology, University of Iceland, Reykjavik, Iceland; 5) School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland.

Meiotic gene conversion and crossover events, along with mutations, produce the germline genetic diversity that is subsequently shaped by evolution. Many properties of meiotic gene conversion in man are still poorly understood, despite its key role in shaping the genome's GC content and its potential confounding impact on mutation rate inferences and evolutionary divergence time estimates. Recently, the sex-averaged non-crossover (NCO) gene conversion rate (G) of single nucleotide polymorphisms (SNPs) was estimated in humans as 5.9 per million base-pairs (Mb) per generation, with a GC-bias of 68%. Here, we report the discovery of 3,176 SNP and 61 indel gene conversions, based on SNP microarray data from 37,981 Icelanders and 30x whole genome sequence (WGS) data from 530 Icelanders. Our estimate of G is 7.0 (95% CI 6.0-8.0) for SNPs and 5.8 (95% CI 4.1-7.9) for indels per Mb per generation and the GC bias is 67.6% (95% CI 65.7-69.8). For indels we demonstrate a 65.6% (95% CI 53.2-77.4) preference for the shorter allele. G is highly elevated in crossover recombination hotspots and male double strand break (DSB) regions. NCO gene conversions from mothers are longer and G is 2.17 (95% CI 1.94-2.45) times greater than from fathers. Notably, while no age effect was seen in fathers, G increases with the age of mothers, at a rate of 0.58 (95% CI 0.38-0.78) per Mb per year. Similarly, crossover (CO) gene conversion rate is higher in mothers and increases with maternal age. A disproportionate number of NCO gene conversions in older mothers occur outside DSB regions and in regions of comparatively low GC rate. Overall, our results suggest that a different mechanism controls meiotic gene conversions in mothers than fathers, where the mechanism in mothers is age-related.

325

Parent-of-origin specific signatures of *de novo* mutations. *C. Gilissen[5], J.M. Goldmann[1], W.S.W. Wong[2], M. Pinelli[3], T. Farrah[4], D. Bodian[4], A.B. Stittrich[4], L.E.L.M. Vissers[5], A. Hoischen[5], J.C. Roach[4], J.G. Vockley[2,6], J.A. Veltman[5,7], B.D. Solomon[2,8,9], J.E. Niederhuber[2,9].* 1) Human Genetics, Radboud University Medical Center, Nijmegen, Netherlands; 2) Inova Translational Medicine Institute (ITMI), Inova Health Systems, Falls Church, VA, USA; 3) Telethon Institute of Genetics and Medicine (TIGEM), Pozzuoli, 80078 Naples, Italy; 4) Institute for Systems Biology, Seattle, WA, USA; 5) Department of Human Genetics, Donders Centre for Neuroscience, Radboud University Medical Center; 6) Department of Pediatrics, Virginia Commonwealth University School of Medicine, 1201 E Marshall St, Richmond, VA, USA; 7) Department of Clinical Genetics, GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre; 8) Department of Pediatrics, Inova Children's Hospital, Inova Health System, Falls Church, VA, USA; 9) Johns Hopkins University School of Medicine, 733 North Broadway Street, Baltimore, MD, USA.

De novo mutations (DNMs) originating in gametogenesis are an important source of genetic variation. We use a dataset of 7,216 autosomal DNMs with resolved parent-of-origin from whole-genome sequencing of 816 parent-offspring trios to investigate differences between maternally and paternally derived DNMs and study the underlying mutational mechanisms. Our results show that the number of DNMs in offspring increases not only with paternal age, but also with maternal age and that some genome regions show enrichment for maternally derived DNMs. We identify parent-of-origin-specific mutation signatures that become more pronounced with increased parental age, pointing to different mutational mechanisms in spermatogenesis and oogenesis. Moreover, we find DNMs that are spatially clustered to have a unique mutational signature with no significant differences between parental alleles, suggesting a different mutational mechanism. Our findings provide insights into the molecular mechanisms that underlie mutagenesis and are relevant to disease and evolution in humans.

326

Characterization of the STR mutation process at every locus in the genome.

M. Gymrek[1,2], T. Willems[3,4], N. Patterson[5], D. Reich[2,5,6,8], Y. Erlich[3,7,8]. 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 3) New York Genome Center, New York, NY, USA; 4) Computational and Systems Biology Program, MIT, Cambridge, MA USA; 5) Department of Genetics, Harvard Medical School, Boston, MA USA; 6) Howard Hughes Medical Institute, Harvard Medical School, Boston, MA USA; 7) Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY, USA; 8) Equally supervised this work.

Mutations form the substrate of evolution, and thus understanding mutation processes is critical to a wide range of applications, including medical genetics, population genetics, and forensics. Short tandem repeats (STRs) are among the largest contributors of de novo mutations in humans. They are comprised of repeating motifs of 1-6bp that are implicated in over 40 Mendelian disorders and contribute to phenotypic variation in complex traits. STRs are prone to polymerase slippage during replication, resulting in mutation rates that are orders of magnitude higher than those of point mutations. Most studies of STR polymorphism have focused on a highly ascertained set of loci that are extraordinarily polymorphic and easy to genotype, providing biased mutation models and limited application to genome-wide STRs. Here, we harnessed novel bioinformatics tools and an analytical framework to estimate mutation parameters at each STR in the genome. First, we used our lobSTR algorithm to generate the most comprehensive STR polymorphism dataset to date, consisting of nearly 1.5 million loci across 300 samples sequenced to high coverage by the Simons Genome Diversity Project. These samples originate from diverse genetic backgrounds and maximize our power to observe STR evolution across a wide range of time scales. Next, we developed an analytical model of the STR mutation process that shows greater consistency than traditional stepwise models with empirical data. Our model allows us to obtain maximum likelihood estimates of mutation parameters at individual STRs or jointly across loci by correlating genotypes with local sequence heterozygosity. We tested this model on simulated and observed genotypes and accurately recovered mutation rates for markers with published de novo rates. We applied our method to jointly estimate mutation rates for each STR in the genome and detected rates ranging from 10^{-7} to 10^{-2} mutations per generation. Local sequence features are highly predictive of mutation rate but less so for parameters describing other features of STR evolution. We used this call-set to assess patterns of variation at known pathogenic loci, identify potential modifiers of mutation rate, and scan for microsatellites under selective pressure. Our mutation model and per-locus estimates of mutation parameters will enable better assessment of the role of STRs in human traits and will inform future work incorporating STRs into complex trait analysis.

327

Identification of recurrent copy number variants associated with developmental brain disorders from whole exome sequencing of 47,859 participants in the DiscovEHR study.

A.E. Hare-Harris[1], A. Moreno-De-Luca[1], E. Maxwell[2], L. Habegger[2], C. O'Dushlaine[2], S. McCarthy[2], J.D. Overton[2], J. Reid[2], A. Luncas[1], D. Kim[1,3], T.H. Nelson[1], S.A. Pendergrass[1], M. Ritchie[1], D.H. Ledbetter[1], C.L. Martin[1]. 1) Autism & Developmental Medicine Institute, Geisinger Health System, Lewisburg, PA; 2) Regeneron Genetics Center, Tarrytown, NJ; 3) The Center for Systems Genomics, The Pennsylvania State University, State College, PA.

Recurrent copy number variants (CNVs), mediated by segmental duplications, have been identified as a cause of many developmental brain disorders (DBDs), such as autism, intellectual disability, and schizophrenia. To date, most large CNV studies have been conducted in clinical cohorts of individuals ascertained for DBD phenotypes. However, many of the CNVs identified in affected individuals have also been reported in apparently healthy subjects, demonstrating the variable expressivity of CNVs. In order to assess the prevalence and phenotypic spectrum conferred by these CNVs in an unbiased sample, we evaluated 45 recurrent regions previously associated with DBDs in 47,859 exomes from the Geisinger Health System - Regeneron Genetics Center (RGC) DiscovEHR project using the MyCode® Community Health Initiative biorepository and electronic health records (EHR) of Geisinger Health System. To date, MyCode® has mostly adult patient-participants of predominantly European ancestry. Using the CLAMMS method for CNV detection from exome data, we identified 1,947 patients (4.1% of DiscovEHR participants) with at least one CNV in any of the 45 recurrent regions previously associated with DBDs. Twenty-four percent of CNV carriers (n=472) come from 196 pedigrees with segregating CNVs, including at least five pedigrees with 3+ generations. The most common CNV region observed was 2q13 with a frequency of 1 in 149 patients for deletions (n=322) and duplications (n=321). The next most common CNV region was 15q11.2 (BP1-BP2) with a frequency of 1 in 163 for deletions (n=293) and 1 in 199 (n=240) for duplications. Using 1,190 International Classification of Disease 9 (ICD-9) codes extracted from the EHR of each patient-participant, we conducted permutation-based enrichment analyses to identify ICD-9 codes that were significantly enriched in patients with a CNV. Overall, 17 of the 25 CNVs with >2 cases in DiscovEHR were enriched for ICD-9 codes that represent DBD phenotypes ($p<0.01$), such as schizophrenia, bipolar disorder, and seizures. Interestingly, <5% of all CNV carriers were previously identified as having a clinically relevant CNV requiring specialized care by Geisinger Health System's Medical Genetics providers. Therefore, we have captured individuals with subclinical manifestations of these CNVs. Overall, this study provides the first insight into the prevalence and clinical manifestations of recurrent CNVs associated with DBDs in an unselected health-system based cohort.

3398

Epigenomic profiling at high resolution reveals genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *A. Varshney[1], L.J. Scott[2], R. Welch[2], M.R. Erdos[3], P.S. Chines[3], N. Narisu[3], R.D.O. Albanus[4], P. Orchard[4], B.N. Wolford[4], R. Kursawe[5], S. Vadlamudi[6], M.E. Cannon[6], J. Didion[3], J. Hensley[4], A. Kirilusha[3], L.L. Bonnycastle[3], D.L. Taylor[3,7], R.M. Watanabe[8,9], K.L. Mohlke[6], M. Boehnke[2], F.S. Collins[3], S.C.J. Parker[1,4], M.L. Stitzel[5], NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program.* 1) Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA; 2) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA; 3) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA; 4) Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA; 5) The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA; 6) Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA; 7) European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK; 8) Department of Preventive Medicine, University of Southern California Keck School of Medicine, Los Angeles, CA 90089, USA; 9) Department of Physiology and Biophysics, University of Southern California Keck School of Medicine, Los Angeles, CA 90089, USA.

Genome wide association studies (GWAS) have identified >100 single nucleotide polymorphisms (SNPs) that encode type 2 diabetes (T2D) and related trait susceptibility. However, the pathogenic mechanisms for most of these SNPs remain elusive. Here, we examined genomic, epigenomic, and transcriptomic profiles in disease-relevant human pancreatic islets to understand the links between genetic variation, chromatin landscape, and gene expression in the context of T2D. We first integrated genome and transcriptome (RNA-seq) variation across 112 islet samples to produce dense cis-expression quantitative trait loci (cis-eQTL) maps. Further integration with chromatin state maps for islets and other diverse tissue types revealed that cis-eQTLs for islet specific genes are specifically and significantly enriched in islet stretch enhancers. High-depth (>1.4B reads) chromatin accessibility profiling using ATAC-seq in two islet samples enabled us to identify specific transcription factor (TF) footprints embedded in active regulatory elements, which are highly enriched for islet cis-eQTL. Aggregate allelic bias signatures in TF footprints enabled us de novo to reconstruct TF binding affinities genetically, which support the high-quality nature of the TF footprints. Interestingly, we found that T2D GWAS loci were specifically and significantly enriched ($P = 4.4\times10^{-07}$, fold enrichment = 30.1) in islet Regulatory Factor X (RFX) footprints. Remarkably, within and across independent T2D GWAS loci, risk alleles that overlap with RFX footprints uniformly disrupt the RFX motifs at high information content positions. Among the RFX TFs, RFX6 is expressed in islets with high specificity, is involved in maintaining beta cell functional identity, and controls glucose homeostasis. Studies have shown that beta cell specific deletion of RFX6 results in impaired insulin secretion. Rare autosomal recessive mutations in the DNA binding domain of RFX6 result in Mitchell-Riley syndrome, which is characterized by neonatal diabetes. Our findings may represent a novel connection between rare coding variation in the islet master regulatory TF RFX6 and common non-coding variation in multiple target sites for this TF. Together, these results suggest that common regulatory variations impact islet TF footprints and the transcriptome, and that a confluent RFX regulatory grammar plays a significant role in the genetic component of T2D predisposition..

3399

TOPMed: Early insights from sequencing and analysis of 45,934 deep human genomes. *G. Abecasis[1], P. Natarajan[2], G. Peloso[2], S. Lee[1], NHLBI Trans-Omics for Precision Medicine and TOPMed Anthropometry and Lipids Working Groups.* 1) Ctr Statistical Gen, Univ Michigan, Ann Arbor, MI; 2) The Broad Institute, Cambridge, MA.

TOPMed aims to discover mechanisms underlying heart, lung, blood and sleep disorders by whole genome sequencing high-value samples. In a collaboration of diverse scientists, genome sequencing, data coordination and informatics resource centers, and staff at the National Heart Lung and Blood Institute, we deeply sequenced 45,934 genomes (majority non-European, including large numbers of African Americans and Latinos). These 45,934 samples from 26 studies are sequenced at mean depth 37.8X and passed high quality standards. The rate of newly discovered variants in sequenced individuals has decreased only slowly throughout the project, resulting in a current estimate of ~300 million SNPs and indels (>1 variant per 10 basepairs). Analysis of 18,877 individuals in the first 10 studies yielded 183 million SNPs and 10 million indels. Among discovered variants, 43.5% were singletons present in a single individual. Deviations from this fraction can identify functional regions targeted by natural selection. Regions >100kb with an extreme low fraction of singletons (<38%) include loci encoding *HLA* class I and class II genes and *ABO*. Conversely, large fractions of singletons are observed in >100kb windows around *TP53BP1* (>48% singletons), among missense variants and inframe indels (48%) and among truncating mutations and frameshift indels (57%). Early results identify association signals across the frequency spectrum. For example, we recapitulate association with LDL cholesterol at $p<5x10^{-8}$ for a common non-coding variant at *SORT1* and for rare coding variants in *PCSK9* (p.R46L, p.C679X). For LDL, additional signals include a haplotype of two African ancestry-specific variants (both frequency 1.1%, $r^2 = 1$) associated with a 28 mg/dl decrease in LDL (rs17249141 in *LDLR* promoter and rs114197570 in enhancer 4-kb upstream). For BMI, we recapitulate association at $p<5x10^{-8}$ with common variants near *FTO* and identify new signals at low-frequency variants near *DNAH5* (rs76221701). In October 2016, TOPMed will release ~10,000 genomes through dbGaP. A new public variant server will allow easy browsing of all catalogued variants. Preliminary analyses suggest that haplotype imputation will exceed all current panels in accuracy, in individuals of European, African-American or Latino ancestry. In summary, we illustrate the importance of deep genome sequencing to understand human genetic variation and its association with disease, and announce release of the first TOPMed results and data.

3400

Exome sequencing of infant dried blood spots identifies three-quarters of metabolic disorders found by newborn screening, indicating limits to exomes in both newborn screening and diagnostic testing. *A.N. Adhikari[1], Y. Wang[1], R. Gallagher[2], Y. Zou[1], U. Sunderam[3], J. Shieh[2], A. Chellappan[3], L. Bassaganyas[2], B. Cai[4], F. Chen[2], G. Freedman[2], B.A. Koenig[2], M. Kvale[2], D. Lee[4], D. Vaka[2], B. Zerbe[2], S.D. Mooney[4], R. Srinivasan[3], P.-Y. Kwok[2], J.M. Puck[2], S.E. Brenner[1], The NBSeq Project.* 1) University of California, Berkeley, CA, USA; 2) University of California, San Francisco, CA, USA; 3) Tata Consultancy Services, Hyderabad, TS, India; 4) University of Washington, Seattle, WA, USA.

Public health newborn screening (NBS) identifies newborns with rare treatable conditions, permitting early intervention. Tandem mass spectrometry (MS/MS) detects many metabolic disorders with excellent sensitivity, but yields false positives—particularly from preterm and ill infants—as well as false negatives and often does not identify the precise disorder. The NBSeq project is evaluating the potential of whole exome sequencing (WES) in NBS using de-identified, archived dried blood spots (DBS) under an IRB-approved protocol with the California Dept. of Public Health. One aim explores feasibility of WES to replace or augment MS/MS for metabolic disorders. DBS of all California newborns from Jul 2005–Dec 2013 with disorders diagnosed by MS/MS and a selection of false positives were made available (1600 samples). To date we have sequenced 600 samples, and analyzed 184. DNA from DBS samples yielded exomes comparable to those from fresh blood DNA. Initial analysis revealed that WES yields vastly more false negatives than MS/MS. In one quarter of newborns with metabolic disorders, we did not identify two rare potentially damaging alleles for genes responsible for their recessive Mendelian disorder. We systematically explored how different WES interpretation protocols impacted the prediction of metabolic disorders in newborns. Tuned analysis pipelines shifted the balance of false negatives vs. false positives but did not yield acceptable specificity and sensitivity for NBS. While apparently unsuitable for first-line NBS, WES might yet improve NBS in NICU settings or help specify a diagnosis following positive MS/MS screens. These results have implications beyond NBS. Diagnostic WES, which has been reported to offer a clinical breakthrough in 25-50% of cases, is applied to individuals with an indication, whereas NBS is performed without any phenotypic information and must be scalable to the entire newborn population. While NBS looks for a very narrow set of disorders when compared to diagnostic testing, it identifies >97% of affected individuals at >99.5% specificity. The success of diagnostic WES is limited in part because patients' disorders may be genetically complex or not genetic. Our study suggests limits to identifying even monogenic disease as well; our work focused on some of the best-characterized Mendelian disorders, yet recognized only three-quarters of the cases. This therefore suggests a ceiling to the potential of current diagnostic WES.