Distribution Agreement

Signature:

Trenell J. Mosley   Date

Investigating rare genetic disorders to gain insight into human biology
By

Trenell J. Mosley
Doctor of Philosophy
Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

Michael E. Zwick, Ph.D.
Advisor

Jennifer G. Mulle, MHS, Ph.D.
Advisor

Karen Conneely, Ph.D.
Committee Member

David J. Cutler, Ph.D.
Committee Member

Michael P. Epstein, Ph.D.
Committee Member

Michael J. Gambello, M.D., Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Investigating rare genetic diseases to gain insight into human biology

By

Trenell J. Mosley
B.S., The University of Texas at Austin, 2015

Advisor: Michael E. Zwick, Ph.D.
Advisor: Jennifer G. Mulle, MHS, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology
2021

Abstract

Investigating rare genetic diseases to gain insight into human biology

By Trenell J. Mosley

Discerning how genetic variation contributes to phenotypes is a critical part of understanding biology. Historically, scientists have contributed to our comprehension of variation by observing exceptional phenotypes. In humans, this can translate to the investigation of rare genetic diseases (RGDs), which offer unique insights into human biology. There are over 7,000 defined rare genetic diseases that affect more than 350 million people worldwide. By studying RGDs diseases, we have gained insights into essential biological mechanisms that underlie both rare and common diseases and have led to the development and improvement of interventions for them. The advent of next-generation sequencing (NGS) technologies has increased our ability to detect and interpret the genetic variation that underlies rare genetic diseases and has accelerated essential discoveries. Thus, our continued study of RGDs will only increase our understanding of RGDs and human biology. Through my dissertation work, I sought to improve our understanding of human biology by investigating two classes of rare genetic diseases and their underlying variation: a rare monogenic disorder caused by a single nucleotide variant (SNV) and rare, genomic disorders caused by repeat-mediated copy-number variants (CNVs). First, we ascertained two siblings of Middle Eastern descent presenting with a rare syndrome consisting of short stature and insulin resistance. Then, using whole-genome sequencing (WGS), genetic analysis, and functional testing, I identified the underlying genetic cause as an intronic splicing variant in *POC1A*, thereby giving insights into the allelic spectrum of *POC1A*-related primordial dwarfism disorders and insulin resistance. For my genomic disorders project, I used a systematic and comprehensive literature search, single nucleotide polymorphism (SNP) genotyping, and WGS, to determine the parent of origin data for multiple pathogenic CNV loci. I demonstrated a significant association between sex-specific patterns in meiotic recombination and parental origin at these loci, which has implications for assessing risks for forming pathogenic CNVs. Taken together, my dissertation work advances our knowledge of the genetic causes underlying rare genetic diseases, has future implications for prospective interventions and counseling for individuals with rare genetic disorders, and gives insight into essential biological processes in human beings.

Investigating rare genetic diseases to gain insight into human biology

By

Trenell J. Mosley
B.S., The University of Texas at Austin, 2015

Advisor: Michael E. Zwick, Ph.D.
Advisor: Jennifer G. Mulle, MHS, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology
2021

Acknowledgments

First and foremost, I am deeply grateful to my parents, Darneta S. Smith, Brent M. Mosley, and Arenda C. Mosley for loving and supporting me through all my endeavors, and the rest of my family for their encouragement throughout my life. I would like to thank my many friends for their support and for giving me confidence when I could not find it myself. Lastly, I would like to thank my mentors Dr. Amanda Marie James, Dr. Jennifer G. Mulle, and Dr. Michael E. Zwick for their guidance, advocacy, and encouragement throughout my studies.

Dedication

I dedicate this dissertation to my g-pa, John G. Smith Jr., my grandpa Maurice B. Mosley, and my great-grandma Myrtis L. Causey-Hicks. Although they are no longer here, they are my inspiration and constant source of motivation. I stand on their shoulders.

Table of Contents

Table of Tables

Table of Supplemental Tables

Table of Figures

Table of Supplemental Figures

CHAPTER I. Introduction

Trenell J. Mosley, Jennifer G. Mulle, Michael E. Zwick

History of Genetic Diseases and Variation

Genetic variation in populations of sexually reproducing organisms is essential for the science of genetics. Understanding how this genetic variation influences phenotype underlies research programs focused on answering critical questions in biology and evolution. Combining genetic variation with the principles of genetics enables us to dissect biological mechanisms and reveal processes underlying the physiology and development of organisms. Dr. William Gahl, former Clinical Director of the National Human Genome Research Institute and Director of the Undiagnosed Diseases Program, summarized this thought perfectly: “Evolution requires imperfect fidelity of replication, that is mutations, and these mistakes ultimately reveal the exquisite functionality of Nature” [1]. Scientists have historically investigated genetic variation through the observation and characterization of exceptional phenotypes in animals and plants. The number of breakthroughs is genuinely breathtaking. From Calvin Bridges’ use of the X-linked *white* locus to demonstrate the chromosome theory of heredity in 1916 [2], to Victor McKusick’s application of genetics to study human diseases [3], from Barbara McClintock’s discovery of jumping genes in maize [4], to the discovery of LDL receptor from studies of familial hypercholesterolemia [5], and Steve Warren cloning and identifying a new mutational mechanism underlying Fragile X [6,7], the genetic research methodology works.

From a medical standpoint, geneticists aim to identify the full spectrum of DNA variation that influences phenotypes creating a comprehensive genotype-phenotype map of all observable genetic variation. This endeavor requires understanding the underlying characteristics of genetic variation that contribute to a disease or trait — the genetic architecture [8]. Genetic variation contributing to traits is understood through a framework that includes a spectrum of population allele frequencies (AF) and effect sizes. Common variants (AF >5% in the population) often have small effect sizes and are typically associated with common complex traits like height or cardiovascular disease [8,9]. On the other hand, rare Mendelian or monogenic disorders are caused by alleles that have a large effect on phenotype and are rare in the general population. Investigations at each end of this spectrum (rare and large effects vs common and small effects) offer different implications and insights into disease structure and are investigated with different methodologies [8].

Common Genetic Diseases and the Utility of Genome-wide Association Studies

Since 2005, geneticists have employed genome-wide association methods (GWAS), which exploit linkage disequilibrium and correlative structure of the human genome, in order to investigate common variation and disease susceptibility [10]. These study designs benefit from the "commonness" of common traits, and thus, are able to ascertain that large number of participants (typically in the thousands) needed to yield results. Insights into disorders like acute macular degeneration and inflammatory bowel disease are examples of the early rewards of the establishment of GWAS as a means to uncover genetic factors contributing to diseases [11,12]. In the ~16 years since the first GWAS study, we have seen major insights into additional disorders like type 2 diabetes, bipolar disorder, coronary artery disease, and schizophrenia [13,14]. Because variants in a GWAS are usually of modest effect size and are *associated* with a trait and not demonstrated to be causal, insights have been limited. The reliance on linkage-disequilibrium also creates frustration in the accurate identification of the specific variants responsible due to the correlation between adjacent variants [10]. Even if the specific variants are identified, it is difficult to interpret their potential effects on function as most variants map to non-coding regions of the genome, inferences are limited to their potential effects on transcriptional regulation of genes.

Thus, while the study of common variation in common diseases offer a direct route to understanding diseases that affect large numbers of people and biology, the complexity of the genetic etiology has limited applications to clinical translation. The GWAS study design, in attempting to ascertain large numbers of cases and controls, also will generally ignore the potential impact of environmental exposures on the phenotype of interest.

Investigating Rare Genetic Diseases

Early investigations in the 80s and 90s of human diseases focused on rare genetic diseases (affect <200,000 people) as the tractable pathway to unraveling human biology due to technology limitations. Researchers mapped rare genetic variants via linkage studies involving numerous unrelated pedigree structures or multiplex familial studies [10]. Since the completion of the draft human genome reference in 2001 [15,16], and the advent of sequencing technologies, the timeline to discovering mutations in genes responsible for genetic disorders has vastly accelerated. The Online Mendelian Inheritance in Man currently catalogs over 6,800 genetic phenotypes with a known molecular basis, 5,801 of which are monogenic disorders and traits (OMIM: https://www.omim.org/statistics/geneMap). An estimated 50-60 new genetic diseases are added per year [17]. The large effect size of rare variation enables geneticists to interpret and functionally validate the pathogenicity of putative variants on a relatively shortened timeline than common variants, thereby offering relatively faster clinical and/or pharmaceutical implications [8]. While additional development of variant and functional database resources, such as gnomAD, ENCODE, and DECIPHER have aided prioritizing and interpreting rare variants, challenges remain [18-20]. Ascertaining multiple patients and families with the same rare disorder can be challenging to find and cost-prohibitive to recruit. Moreover, interpreting and validating some classes of variation, particularly non-coding and incompletely penetrant variants, pose additional challenges. So why does the field continue to investigate rare genetic diseases if they seem so intractable?

Even though rare genetic diseases (RGDs) are operationally defined in the U.S. as diseases affecting less than 200,000 individuals, in aggregate, they affect between 25-30 million U.S. citizens. Furthermore, it is estimated that approximately 7,000 RGDs affect more than 350 million people worldwide—greater than the population of the world's third most populated country [19,21,22]. Thus, despite their rarity on an individual level, they collectively represent a significant burden on the health and economy of patients and a rich source of biological insight.

Insight into biological mechanisms. In the 1970s, Goldstein and Brown studied a rare homozygous form of familial hypercholesterolemia (FH) (MIM: 143890) (frequency 1 in 1,000,000) and made Nobel-worthy discoveries in cholesterol metabolism. Patients with FH exhibit 6- to 10-fold increases in serum levels of low-density lipoprotein (LDL), a primary cholesterol carrier, and have heart attacks early in childhood. Investigations of the cells cultured from FH patients resulted in discovering the LDL receptor in 1974, which transports LDL particles into a cell [5]. Subsequent studies further revealed fundamental and previously unknown biology about cholesterol metabolism and general concepts of receptor-mediated endocytosis, receptor recycling, and feedback regulation of receptors [23-26]. This final principle was the basis for the development of now widely used statins in lowering cholesterol [26,27].

More recently, investigations of Niemann-Pick Disease Type C (MIM: 257220) syndrome revealed aspects of the biology of the Ebola virus cellular entry mechanisms. Niemann-Pick Disease Type C is a rare (frequency 1 in 120,000) lipid storage disorder characterized by accumulation of LDL-derived cholesterol in the lysosomes [28,29]. In 1997, through studies of individuals with Niemann-Pick Disease Type C, Carstea and others discovered mutations in the endo/lysosomal cholesterol transporter protein, NPC1, was responsible for this rare [29]. However, it was not until almost 14 years later that cells from patients with Niemann-Pick Disease Type C and mutation carriers would be demonstrated to be resistant against Ebola virus infection, and the mechanism of entry would be elucidated. The Ebola virus can infect nearly every cell type it encounters through the universality of its glycoprotein-coated spike protein [30]. Once in the cell, it requires trafficking in and out of the lysosome via NPC1 to replicate itself. Niemann-Pick Disease Type C cells lack a proper functioning NPC1 protein, and in these cells, the Ebola virus is unable to exit the lysosome and copy itself [31,32]. Understanding this biology led to the development of therapeutic bispecific-antibodies that show promise in treating the Ebola virus and other related viruses [33].

Insights into common diseases. Along with novel insights into biology, RGDs and variation can aid in the understanding of common diseases, as rare disease pathology can overlap with those investigated in common diseases [34,35]. The transforming growth factor-β (TGFβ) signaling pathway is highly involved in the maintenance of tissue homeostasis [36], and overexpression of TGFβs is associated with diseases including cancer, fibrosis, and inflammation [37]. The signaling pathway's role in disease progression has made it a standard target for drug development [37]. Studies of Marfan syndrome (MFS) (MIM: 154700), caused by mutations in the fibrillin 1 gene *FBN1*, offer a new avenue into regulating vascular symptoms in TGFβ-related diseases such as fibrosis [37,38]. Patients with MFS exhibit increased levels of TGFβ signaling. Disrupted *FBN1* fails to sequester TGFβ, resulting in increased levels of unbound TGFβ and excessive activation of the pathway [38]. These insights demonstrated the importance of microfibrils in regulating TGFβ levels.

An estimated 38 million people worldwide are infected with Human immunodeficiency virus, which has killed approximately 32.7 million since the start of the epidemic in the 80s [39]. HIV-1 related viruses require co-receptors to infect target cells, and the CCR5 chemokine receptor is the major co-receptor for macrophage-tropic HIV-1 strains. The CCR5 receptor has turned out to be crucially important in understanding how HIV enters cells and has been the target for drug development efforts [40]. Individuals that remained unaffected after multiple exposure HIV-1 were discovered to contain a rare genetic variant, a 32-base pair deletion in the CCR5 gene [41,42]. The deletion results in a non-functional co-receptor, which does not allow for HIV-1 viruses to fuse to the membrane and infect cells. Thus, individuals homozygous for the CCR5 Δ32 allele are highly resistant to HIV infection [40,43,44]. Since the discovery of these individuals, accelerated development of CCR5 inhibitors has been pursued as highly effective antiretroviral therapeutics [40].

Clinical Implications. The increased understanding of the etiology of rare diseases ultimately benefits those affected by them. Individuals with RGDs face a variety of physical, mental health, social, and economic burdens [19]. Currently, the diagnostic rate for Mendelian disorders is less than 50% [45]. Clinicians are often unfamiliar with the symptoms or presentations of rare diseases. As such, patients frequently suffer from a delay in diagnosis. In a survey of eight RGDs, including Fragile X syndrome and cystic fibrosis, 25% of patients experience between 5-30 years between the emergence of their first symptoms and final diagnosis. Even when a diagnosis was made, it was incorrect in 40% of the cases [46]. This diagnostic odyssey imparts financial, health, and psychological burdens on patients and their families. Studies have found that on average the cost of care for individuals with RGDs is $305,428, with hospital charges costing between $17,000 and $77,000 more than charges for non-genetic-related discharges [47,48]. Patients and their families experience stressors such as having to attend multiple appointments, missing work, loss of income, hopelessness, and uncertainty about the future, which can all predispose to anxiety and depression in 86% and 75% of patients, respectively [49,50]. Continued investigation of RGDs offers to continue hope for patients, especially with the usage of genome sequencing. Discovery and collaborative models such as the NIH Undiagnosed Disease Network (UDN) and Centers for Mendelian Genomics have already demonstrated an ease in financial burden. Cost of care in the UDN evaluation framework averages $18,903— representing an approximate 94% decrease in cost compared to outside the UDN [48]. As of 2017, the Centers for Mendelian Genomics uncovered 327 novel genes linked to RGDs and maintains a continued rate of 263 novel discoveries per year [18,35]. As of 2019, the UDN has provided diagnosis to 231 of 791 evaluated individuals (diagnostic rate = 29%) and revealed 17 new disease-gene associations [51]. As the underpinnings of mendelian disorders are discovered, this gives insight into other rare disease pathologies, potentially decreasing the time to diagnosis and treatment for patients [52].

Contributions of Variation to Rare Genetic Diseases

The first step to unraveling the etiology behind RGDs, is understanding the variation that underlies them. Too add complexity to this there is also variation in the types of variants that causes these diseases. The scale of rare disease genetic variation ranges from single nucleotide variants to large-scale structural variants and each class of variation can cause disease through different mechanisms.

Single Nucleotide Variants. Single nucleotide variants (SNVs) are the most common and well understood type of genetic variation in the human genome. Historically SNVs were investigated using single gene approaches, however with the advent of sequencing, interrogations of entire genomes have greatly contributed to our understanding of this variation class and its contribution to disease. SNVs occur as substitutions, insertions, or deletions and outnumber other classes of variation 7 to 1 [53]. There is an approximate 0.1% difference between any two humans amounting to ~3 million single nucleotide differences [53,54]. The estimated rate of mutation for SNVs is approximately 10^{-8} per base pair per generation [55]. SNVs are linked to a large amount of both common and RGDs; it is estimated they account for 85% of disease associations [56]. SNVs can cause RGDs through multiple effects. Rare coding SNVs particularly can be pathogenic through a large impact on protein function, structure, or even post-translational modifications [57]. Additionally, SNVs within non-coding regions of the genome, such as splice sites or promoter sequences, can lead to disease by affecting regulation of gene expression [57]. Much of our understanding of human mutation rates and evolution has come from investigation of SNVs and single nucleotide polymorphisms (SNPs; SNV $\geq$ 1% AF in population) and their links to disease as well as their distribution within the genome and across populations [53,54,58]. Our persistent quest to understand SNVs has revealed the potential the contribution of other types of variation to disease and genome evolutions.

DNA Repeat Variants. Repetitive DNA is broadly defined as DNA sequences present in multiple copies in the genome [59] and comprises approximately 50% of the human genome [15]. Repetitive variation can be divided into two classes both of which contribute to disease: Tandem repeats that lie in a head-to-tail arrangement and interspersed repeats that are scattered throughout the genome [59]. This class of genetic variation can contribute to genetic disease through several mechanisms. Tandem repeats particularly are highly mutable and can expand or contract via replication slippage [60]. Expansion of short tandem repeats (STRs) are known to cause rare repeat expansion diseases like Fragile X syndrome (FXS) (MIM: 300624). FXS is caused by expansion of the CCG repeat in the *FMR1* gene. In unaffected individuals this triplet repeat is present in <50 copies and *FMR1* is expressed normally [61]. However, in individuals with more than 200 repeats, *FMR1* expression is silenced and no FRM1 protein is produced, resulting in FXS [61]. FXS ad other repeat expansion diseases exhibit a phenomenon known as anticipation. As tandem repeats expand, the longer a repeat tract becomes, the more prone to slippage the tract is. Tandem repeats can expand across generations within a single pedigree, and as the repeat tract lengthens across generations, the associated disease can worsen or age of onset can shorten [59].

In comparison interspersed repeats are remnants of transposable elements (TEs) that have “jumped” around the genome throughout time [62]. Most are inactive; however, active TEs can cause disease by insertion into a gene and inactivate via frameshift or inactivating, such as in X-linked Dystonia-Parkinsonism (XDP) (MIM: 314250). SVA retrotransposition into the intron of the gene *TAF1* leads to abnormal splicing, intron retention, and overall reduction in *TAF1* expression [24]. In addition to insertional inactivation, TEs, specifically ancient and fixed interspersed repeats, can contribute to chromosomal instability and act as substrates for recombination processes that produce structural variants [62]. Recombination between *Alu* repeats located in *HPRT* produces duplication of exons 2 and 3 and leads to Lesch Nyhann syndrome (MIM: 300322), a rare neurological and behavior disorder [63].

Structural Variants. The extent of structural variants (SVs) and their contributions to disease was not appreciated until 2004, when Sebat and colleagues demonstrated the widespread existence of copy number polymorphisms in human populations [64]. Since then, continued scrutiny has demonstrated that SVs contribute more variation to the human genome than SNVs by sheer amount of DNA material involved [65], and occur 1,000 to 10,000-fold more frequently than SNVs, with a mutation rate ranging between 10^{-5} and 10^{-4} [66]. Given the size and frequency of SVs it is no wonder they are estimated to be responsible for 25% of all protein-truncating events in the genome [67]. SVs are broadly classified as unbalanced and balanced. Unbalanced SVs are those that result in the net gain or loss of DNA material, and are often referred to as copy-number variants (CNVs). Balanced SVs, such as inversions and translocations, maintain the same amount of DNA material [67]. SVs most obviously contribute to disease through the alteration of dosage-sensitive genes. We see this with classical CNV disorders, also known as genomic disorders, like Charcot-Marie Tooth disease type 1A (CMT1A) (MIM: 118220) and Hereditary Neuropathy with liability to Pressure Palsies (HNPP) (MIM: 162500) caused by the respective duplication or deletion of *PMP22*. If genes span both breakpoints of an SV, have the same orientation, and have a maintained reading frame, rearrangement can produce gene fusion products, which has been well demonstrated in the red-green opsin genes in X-linked color blindness [53,68,69]. Structural rearrangements can subject genes to position effects, by removing or altering regulatory sequences, as seen with *SOX9* rearrangements and campomelic dysplasia [70,71]. Atypical presentations of genomic disorders, such as 22q11.2 deletion syndrome (MIM: 611867) have been attributed to contribution of recessive SNVs that were unmasked by deletions and inherited *in trans* [72]. Particularly complex rearrangements can produce regions of uniparental disomy (UPD) or absence of heterozygosity (AOH), which can be pathogenic if the region is imprinted or contains a recessive variant [68].

The multiple downstream mechanisms by which SVs cause disease, highlights the large interest in biological processes underlying formation of SVs. Studies so far have demonstrated that SVs can be formed through both replicative and recombination processes, and are largely influenced by genomic architecture [73]. Investigations of SV mechanisms have contributed to our understanding of the dynamics of genomic architecture, which has allowed the subsequent discovery of novel SVs and related disorders [74-76].

Utility of Next-Generation Sequencing in Rare Genetic Diseases

Uncovering disease-gene relationships require the ability to sensitively and comprehensively detect genetic variants of all types. Before next-generation sequencing (NGS) technologies were invented, rare disease genes were mapped using linkage analyses, which required the ascertainment of large multiplex pedigrees or large samples of small pedigrees [10]. Findings then needed to be followed up with segregation analyses and laborious functional studies. While this workflow is still required in the field, advances in genome-wide analysis tools have made the study of RGDs more tractable [34]. Next-generation sequencing with efficient mapping, genotype calling, and variant annotation can lead to the rapid discovery of novel disease-causing variants underlying RGDs. The unbiased nature of NGS bypasses the requirement for extensive knowledge of the disease pathology or previous genetic analysis to pinpoint certain regions in the genome (*i.e.*, linkage analysis), allowing for Mendelian gene discovery that is agnostic to biological hypotheses [77] . Clinical and research settings are increasingly using whole-exome sequencing (WES) and whole-genome sequencing (WGS), but each comes with its pros and cons

[77]. WES captures only the coding regions of the genome—the exons, and thus leverages the ability to interrogate all ~20,000 human protein-coding genes. In contrast, WGS examines the coding and non-coding regions of the genome and, therefore, can capture splicing or regulatory variants. Additionally, WGS has greater capacity to detect SVs such as deletions, insertions and duplications, but limited capacity to detect copy-number neutral SVs, like inversions [21,78]. The capture and amplification process of WES opens the approach to batch effects and biases, and while WGS ameliorates this, the sheer number of mostly intergenic or unannotated variants produced by WGS to interpret (~4 million SNVs/genome) creates a daunting task for clinicians and researchers. While the cost of WES ($500/exome) is lower compared to WGS ($1,000/genome), WGS offers increased sample capacity and reduced labor time [21,78,79]. Nonetheless, the benefits and applicability of NGS technologies are reflected by the 250% increase in gene entries in the Online Mendelian Inheritance in Man (OMIM) database since 2007 [34]. Applying WGS and WES, optionally combined with other omics platforms, and rigorous functional validation, offer an optimal, rapid, and accurate path from disease identification to molecular cause [80]. The dramatic decline in cost of sequencing combined with ubiquitous cloud computing and publicly available genomics resources suggests the impact of this workflow will continue to increase in the future.

As scientists continue to identify and investigate RGDs, our understanding of RGDs and biology will only increase. The following work constitutes an effort to apply genetic principles to two cases of rare diseases in order to understand their underlying causes. In Chapter Two, we leverage consanguinity within a family, WGS analysis, and molecular studies to identify and functionally validate a causal single nucleotide variant for a rare, monogenic primordial dwarfism disorder. In Chapter Three, we investigate rare repeat-mediated CNV disorders and link patterns of meiotic recombination to distributions and biases in parent of origin for pathogenic CNVs.

References

1. Gahl WA. The battlefield of rare diseases: where uncommon insights are common. Sci Transl Med. 2012;4(154):154ed7. Epub 2012/10/05. doi: 10.1126/scitranslmed.3004980. PubMed PMID: 23035044; PMCID: PMC3790314.

2. Bridges CB. Non-Disjunction as Proof of the Chromosome Theory of Heredity (Concluded). Genetics. 1916;1(2):107-63. Epub 1916/03/01. doi: 10.1093/genetics/1.3.309. PubMed PMID: 17245853; PMCID: PMC1193656.

3. McKusick VA. A 60-year tale of spots, maps, and genes. Annu Rev Genomics Hum Genet. 2006;7:1-27. Epub 2006/07/11. doi: 10.1146/annurev.genom.7.080505.115749. PubMed PMID: 16824022.

4. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36(6):344-55. Epub 1950/06/01. doi: 10.1073/pnas.36.6.344. PubMed PMID: 15430309; PMCID: PMC1063197.

5. Goldstein JL, Brown MS. Binding and degradation of low density lipoproteins by cultured human fibroblasts. Comparison of cells from a normal subject and from a patient with homozygous familial hypercholesterolemia. J Biol Chem. 1974;249(16):5153-62. Epub 1974/08/25. doi: 10.1016/S0021-9258(19)42341-7. PubMed PMID: 4368448.

6. Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell. 1991;65(5):905-14. Epub 1991/05/31. doi: 10.1016/0092-8674(91)90397-h. PubMed PMID: 1710175.

7. Warren ST, Zhang F, Licameli GR, Peters JF. The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites. Science. 1987;237(4813):420-3. Epub 1987/07/24. doi: 10.1126/science.3603029. PubMed PMID: 3603029.

8. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet. 2018;19(2):110-24. Epub 2017/12/12. doi: 10.1038/nrg.2017.101. PubMed PMID: 29225335.

9. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832-8. doi: 10.1038/nature09410.

10. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577(7789):179-89. Epub 2020/01/10. doi: 10.1038/s41586-019-1879-7. PubMed PMID: 31915397; PMCID: PMC7405896.

11. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science. 2006;314(5804):1461-3. Epub 2006/10/28. doi: 10.1126/science.1135245. PubMed PMID: 17068223; PMCID: PMC4410764.

12. Klein C, Gahl WA. Patients with rare diseases: from therapeutic orphans to pioneers of personalized treatments. EMBO Mol Med. 2018;10(1):1-3. Epub 2017/11/29. doi: 10.15252/emmm.201708365. PubMed PMID: 29180354; PMCID: PMC5760852.

13. Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet. 2017;49(1):27-35. Epub 2016/11/22. doi: 10.1038/ng.3725. PubMed PMID: 27869829; PMCID: PMC5737772.

14. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-78. Epub 2007/06/08. doi: 10.1038/nature05911. PubMed PMID: 17554300; PMCID: PMC2719288.

15. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921. Epub 2001/03/10. doi: 10.1038/35057062. PubMed PMID: 11237011.

16. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304-51. Epub 2001/02/22. doi: 10.1126/science.1058040. PubMed PMID: 11181995.

17. Lee CE, Singleton KS, Wallin M, Faundez V. Rare Genetic Diseases: Nature's Experiments on Human Development. iScience. 2020;23(5):101123. Epub 2020/05/19. doi: 10.1016/j.isci.2020.101123. PubMed PMID: 32422592; PMCID: PMC7229282.

18. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015;97(2):199-215. Epub 2015/07/15. doi: 10.1016/j.ajhg.2015.06.009. PubMed PMID: 26166479; PMCID: PMC4573249.

19. Stoller JK. The Challenge of Rare Diseases. Chest. 2018;153(6):1309-14. Epub 2018/01/13. doi: 10.1016/j.chest.2017.12.018. PubMed PMID: 29325986.

20. Whicher D, Philbin S, Aronson N. An overview of the impact of rare disease characteristics on research methodology. Orphanet J Rare Dis. 2018;13(1):14. Epub 2018/01/21. doi: 10.1186/s13023-017-0755-5. PubMed PMID: 29351763; PMCID: PMC5775563.

21. Bick D, Jones M, Taylor SL, Taft RJ, Belmont J. Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases. J Med Genet. 2019;56(12):783-91. Epub 2019/04/27. doi: 10.1136/jmedgenet-2019-106111. PubMed PMID: 31023718; PMCID: PMC6929710.

22. Current Population: United States Census Bureau; 2021 [cited 2021 May 3]. Available from: https://www.census.gov/popclock/print.php?component=counter.

23. Anderson RG, Goldstein JL, Brown MS. Localization of low density lipoprotein receptors on plasma membrane of normal human fibroblasts and their absence in cells from a familial hypercholesterolemia homozygote. Proc Natl Acad Sci U S A. 1976;73(7):2434-8. Epub 1976/07/01. doi: 10.1073/pnas.73.7.2434. PubMed PMID: 181751; PMCID: PMC430596.

24. Aneichyk T, Hendriks WT, Yadav R, Shin D, Gao D, Vaine CA, et al. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. Cell. 2018;172(5):897-909 e21. Epub 2018/02/24. doi: 10.1016/j.cell.2018.02.011. PubMed PMID: 29474918; PMCID: PMC5831509.

25. Basu SK, Goldstein JL, Anderson RG, Brown MS. Monensin interrupts the recycling of low density lipoprotein receptors in human fibroblasts. Cell. 1981;24(2):493-502. Epub 1981/05/01. doi: 10.1016/0092-8674(81)90340-8. PubMed PMID: 6263497.

26. Goldstein JL, Brown MS. The LDL receptor. Arterioscler Thromb Vasc Biol. 2009;29(4):431-8. Epub 2009/03/21. doi: 10.1161/ATVBAHA.108.179564. PubMed PMID: 19299327; PMCID: PMC2740366.

27. LaRosa JC, He J, Vupputuri S. Effect of Statins on Risk of Coronary DiseaseA Meta-analysis of Randomized Controlled Trials. JAMA. 1999;282(24):2340-6. doi: 10.1001/jama.282.24.2340.

28. Vanier MT. Niemann-Pick disease type C. Orphanet Journal of Rare Diseases. 2010;5(1):16. doi: 10.1186/1750-1172-5-16.

29. Carstea ED, Morris JA, Coleman KG, Loftus SK, Zhang D, Cummings C, et al. Niemann-Pick C1 disease gene: homology to mediators of cholesterol homeostasis. Science. 1997;277(5323):228-31. Epub 1997/07/11. doi: 10.1126/science.277.5323.228. PubMed PMID: 9211849.

30. Aleksandrowicz P, Marzi A, Biedenkopf N, Beimforde N, Becker S, Hoenen T, et al. Ebola virus enters host cells by macropinocytosis and clathrin-mediated endocytosis. J Infect Dis. 2011;204 Suppl 3(Suppl 3):S957-S67. doi: 10.1093/infdis/jir326. PubMed PMID: 21987776.

31. Carette JE, Raaben M, Wong AC, Herbert AS, Obernosterer G, Mulherkar N, et al. Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. Nature. 2011;477(7364):340-3. Epub 2011/08/26. doi: 10.1038/nature10348. PubMed PMID: 21866103; PMCID: PMC3175325.

32. Cote M, Misasi J, Ren T, Bruchez A, Lee K, Filone CM, et al. Small molecule inhibitors reveal Niemann-Pick C1 is essential for Ebola virus infection. Nature. 2011;477(7364):344-8. Epub 2011/08/26. doi: 10.1038/nature10380. PubMed PMID: 21866101; PMCID: PMC3230319.

33. Wec AZ, Nyakatura EK, Herbert AS, Howell KA, Holtsberg FW, Bakken RR, et al. A "Trojan horse" bispecific-antibody strategy for broad protection against ebolaviruses. Science. 2016;354(6310):350-4. Epub 2016/09/10. doi: 10.1126/science.aag3267. PubMed PMID: 27608667; PMCID: PMC5647781.

34. Fernandez-Marmiesse A, Gouveia S, Couce ML. NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. Curr Med Chem. 2018;25(3):404-32. Epub 2017/07/20. doi: 10.2174/0929867324666170718101946. PubMed PMID: 28721829; PMCID: PMC5815091.

35. Posey JE, O'Donnell-Luria AH, Chong JX, Harel T, Jhangiani SN, Coban Akdemir ZH, et al. Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet Med. 2019;21(4):798-812. Epub 2019/01/19. doi: 10.1038/s41436-018-0408-7. PubMed PMID: 30655598; PMCID: PMC6691975.

36. Morikawa M, Derynck R, Miyazono K. TGF-beta and the TGF-beta Family: Context-Dependent Roles in Cell and Tissue Physiology. Cold Spring Harb Perspect Biol. 2016;8(5).

Epub 2016/05/04. doi: 10.1101/cshperspect.a021873. PubMed PMID: 27141051; PMCID: PMC4852809.

37. Akhurst RJ, Hata A. Targeting the TGFbeta signalling pathway in disease. Nat Rev Drug Discov. 2012;11(10):790-811. Epub 2012/09/25. doi: 10.1038/nrd3810. PubMed PMID: 23000686; PMCID: PMC3520610.

38. Neptune ER, Frischmeyer PA, Arking DE, Myers L, Bunton TE, Gayraud B, et al. Dysregulation of TGF-beta activation contributes to pathogenesis in Marfan syndrome. Nat Genet. 2003;33(3):407-11. Epub 2003/02/25. doi: 10.1038/ng1116. PubMed PMID: 12598898.

39. Fact Sheet – World AIDS Day 2020: United Nations Programme on HIV/AIDS; 2020 [cited 2021 May 3]. Available from: https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf.

40. Latinovic OS, Reitz M, Heredia A. CCR5 Inhibitors and HIV-1 Infection. J AIDS HIV Treat. 2019;1(1):1-5. Epub 2019/08/16. doi: 10.33696/AIDS.1.001. PubMed PMID: 31414081; PMCID: PMC6693856.

41. Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, et al. Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection. Cell. 1996;86(3):367-77. doi: 10.1016/s0092-8674(00)80110-5.

42. Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber CM, et al. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature. 1996;382(6593):722-5. Epub 1996/08/22. doi: 10.1038/382722a0. PubMed PMID: 8751444.

43. Chung BHY, Chau JFT, Wong GK. Rare versus common diseases: a false dichotomy in precision medicine. NPJ Genom Med. 2021;6(1):19. Epub 2021/02/26. doi: 10.1038/s41525-021-00176-x. PubMed PMID: 33627657; PMCID: PMC7904920.

44. Hutter G, Nowak D, Mossner M, Ganepola S, Mussig A, Allers K, et al. Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. N Engl J Med. 2009;360(7):692-8. Epub 2009/02/14. doi: 10.1056/NEJMoa0802905. PubMed PMID: 19213682.

45. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biol.

2019;20(1):58. Epub 2019/03/21. doi: 10.1186/s13059-019-1667-6. PubMed PMID: 30890163; PMCID: PMC6425644.

46. Survey of the Delay in Dagnosis for 8 Rare Diseases in Europe: Eurordis - Rare Diseases Europe; 2007 [cited 2021 May 3]. Available from: https://www.eurordis.org/sites/default/files/publications/Fact_Sheet_Eurordiscare2.pdf.

47. Gonzaludo N, Belmont JW, Gainullin VG, Taft RJ. Estimating the burden and economic impact of pediatric genetic disease. Genet Med. 2019;21(8):1781-9. Epub 2018/12/21. doi: 10.1038/s41436-018-0398-5. PubMed PMID: 30568310; PMCID: PMC6752475.

48. Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheatle-Jarvela AM, et al. Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. N Engl J Med. 2018;379(22):2131-9. Epub 2018/10/12. doi: 10.1056/NEJMoa1714458. PubMed PMID: 30304647; PMCID: PMC6481166.

49. Shire. Rare disease impact report: insights from patients and the medical community. 2013.

50. Uhlenbusch N, Lowe B, Depping MK. Perceived burden in dealing with different rare diseases: a qualitative focus group study. BMJ Open. 2019;9(12):e033353. Epub 2020/01/01. doi: 10.1136/bmjopen-2019-033353. PubMed PMID: 31888936; PMCID: PMC6937088.

51. Schoch K, Esteves C, Bican A, Spillmann R, Cope H, McConkie-Rosell A, et al. Clinical sites of the Undiagnosed Diseases Network: unique contributions to genomic medicine and science. Genet Med. 2021;23(2):259-71. Epub 2020/10/24. doi: 10.1038/s41436-020-00984-z. PubMed PMID: 33093671; PMCID: PMC7867619.

52. Yates J, Gutiérrez-Sacristán A, Jouhet V, LeBlanc K, Esteves C, DeSain TN, et al. Finding commonalities in rare diseases through the undiagnosed diseases network. Journal of the American Medical Informatics Association. 2021;00:1-9. doi: 10.1093/jamia/ocab050.

53. Eichler EE. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. N Engl J Med. 2019;381(1):64-74. Epub 2019/07/04. doi: 10.1056/NEJMra1809315. PubMed PMID: 31269367; PMCID: PMC6681822.

54. Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. J Hum Genet. 2010;55(7):403-15. Epub 2010/05/21. doi: 10.1038/jhg.2010.55. PubMed PMID: 20485443.

55. Shendure J, Akey JM. The origins, determinants, and consequences of human mutations. Science. 2015;349(6255):1478-83. Epub 2015/09/26. doi: 10.1126/science.aaa9119. PubMed PMID: 26404824.

56. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009;106(45):19096-101. Epub 2009/10/29. doi: 10.1073/pnas.0910672106. PubMed PMID: 19861545; PMCID: PMC2768590.

57. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43. Epub 2020/05/29. doi: 10.1038/s41586-020-2308-7. PubMed PMID: 32461654; PMCID: PMC7334197.

58. Roses AD, Akkari PA, Chiba-Falek O, Lutz MW, Gottschalk WK, Saunders AM, et al. Structural variants can be more informative for disease diagnostics, prognostics and translation than current SNP mapping and exon sequencing. Expert Opinion on Drug Metabolism & Toxicology. 2016;12(2):135-47. doi: 10.1517/17425255.2016.1133586.

59. Pavlicek A, Kapitonov VV, Jurka J. Human Repetitive DNA. Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 822-31.

60. Ryan CP. Tandem repeat disorders. Evol Med Public Health. 2019;2019(1):17. Epub 2019/02/26. doi: 10.1093/emph/eoz005. PubMed PMID: 30800316; PMCID: PMC6379701.

61. Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, et al. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res. 2013;23(1):121-8. Epub 2012/10/16. doi: 10.1101/gr.141705.112. PubMed PMID: 23064752; PMCID: PMC3530672.

62. Payer LM, Burns KH. Transposable elements in human genetic disease. Nat Rev Genet. 2019;20(12):760-72. Epub 2019/09/14. doi: 10.1038/s41576-019-0165-8. PubMed PMID: 31515540.

63. Brooks AJ, Waters MJ. The growth hormone receptor: mechanism of activation and clinical implications. Nat Rev Endocrinol. 2010;6(9):515-25. Epub 2010/07/29. doi: 10.1038/nrendo.2010.123. PubMed PMID: 20664532.

64. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-8. Epub 2004/07/27. doi: 10.1126/science.1098918. PubMed PMID: 15273396.

65. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81. Epub 2015/10/04. doi: 10.1038/nature15394. PubMed PMID: 26432246; PMCID: PMC4617611.

66. Lupski JR. Genomic rearrangements and sporadic disease. Nat Genet. 2007;39(7 Suppl):S43-7. Epub 2007/09/05. doi: 10.1038/ng2084. PubMed PMID: 17597781.

67. Collins RL, Brand H, Karczewski KJ, Zhao X, Alfoldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature. 2020;581(7809):444-51. Epub 2020/05/29. doi: 10.1038/s41586-020-2287-8. PubMed PMID: 32461652; PMCID: PMC7334194.

68. Harel T, Lupski JR. Genomic disorders 20 years on-mechanisms for clinical manifestations. Clin Genet. 2018;93(3):439-49. Epub 2017/09/28. doi: 10.1111/cge.13146. PubMed PMID: 28950406.

69. Palau F, Lofgren A, De Jonghe P, Bort S, Nelis E, Sevilla T, et al. Origin of the de novo duplication in Charcot-Marie-Tooth disease type 1A: unequal nonsister chromatid exchange during spermatogenesis. Hum Mol Genet. 1993;2(12):2031-5. Epub 1993/12/01. doi: 10.1093/hmg/2.12.2031. PubMed PMID: 8111370.

70. Velagaleti GV, Bien-Willner GA, Northup JK, Lockhart LH, Hawkins JC, Jalal SM, et al. Position effects due to chromosome breakpoints that map approximately 900 Kb upstream and approximately 1.3 Mb downstream of SOX9 in two patients with campomelic dysplasia. Am J Hum Genet. 2005;76(4):652-62. Epub 2005/02/24. doi: 10.1086/429252. PubMed PMID: 15726498; PMCID: PMC1199302.

71. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451-81. Epub 2009/09/01. doi: 10.1146/annurev.genom.9.081307.164217. PubMed PMID: 19715442; PMCID: PMC4472309.

72. McDonald-McGinn DM, Fahiminiya S, Revil T, Nowakowska BA, Suhl J, Bailey A, et al. Hemizygous mutations in SNAP29 unmask autosomal recessive conditions and contribute to

atypical findings in patients with 22q11.2DS. J Med Genet. 2013;50(2):80-90. Epub 2012/12/13. doi: 10.1136/jmedgenet-2012-101320. PubMed PMID: 23231787; PMCID: PMC3585484.

73. Lupski JR. Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environ Mol Mutagen. 2015;56(5):419-36. Epub 2015/04/22. doi: 10.1002/em.21943. PubMed PMID: 25892534; PMCID: PMC4609214.

74. Dittwald P, Gambin T, Gonzaga-Jauregui C, Carvalho CM, Lupski JR, Stankiewicz P, et al. Inverted low-copy repeats and genome instability--a genome-wide analysis. Hum Mutat. 2013;34(1):210-20. Epub 2012/09/12. doi: 10.1002/humu.22217. PubMed PMID: 22965494; PMCID: PMC3738003.

75. Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, et al. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. Genome Res. 2013;23(9):1395-409. Epub 2013/05/10. doi: 10.1101/gr.152454.112. PubMed PMID: 23657883; PMCID: PMC3759717.

76. Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, et al. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet. 2006;38(9):1038-42. Epub 2006/08/15. doi: 10.1038/ng1862. PubMed PMID: 16906162.

77. Posey JE. Genome sequencing and implications for rare disorders. Orphanet J Rare Dis. 2019;14(1):153. Epub 2019/06/27. doi: 10.1186/s13023-019-1127-0. PubMed PMID: 31234920; PMCID: PMC6591893.

78. Chiara M, Pavesi G. Evaluation of Quality Assessment Protocols for High Throughput Genome Resequencing Data. Front Genet. 2017;8:94. Epub 2017/07/25. doi: 10.3389/fgene.2017.00094. PubMed PMID: 28736571; PMCID: PMC5500642.

79. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) [cited 2021 May 5]. Available from: www.genome.gov/sequencingcostsdata.

80. Rodenburg RJ. The functional genomics laboratory: functional validation of genetic variants. J Inherit Metab Dis. 2018;41(3):297-307. Epub 2018/02/16. doi: 10.1007/s10545-018-0146-7. PubMed PMID: 29445992; PMCID: PMC5959958.

CHAPTER II. A novel intronic mutation causes aberrant splicing in variant POC1A-related syndrome

Trenell J. Mosley, Chuan-En Wang, Weiya He, Yue Feng, Kun Qian, Jennifer G. Mulle, Michael E. Zwick, William R. Wilcox

Introduction

Primordial dwarfism (PD) is a group of clinically and genetically heterogeneous disorders characterized by severe intrauterine growth retardation (IUGR) and post-natal growth delay and abnormalities [1,2]. Several subtypes of PD exist broadly categorized by the presence or absence of microcephaly and additional phenotypic elements [1]. Seckel syndrome, Microcephalic Osteodysplastic Primordial Dwarfism (MOPD) types I, II, and III, and Meier-Gorlin syndrome are classified as microcephalic PD disorders. In contrast, Silver-Russell syndrome (SRS), 3M syndrome, and **S**hort stature, **H**yperextensibility of joints or hernia, **O**cular depression, **R**ieger anomaly, delayed **T**eething (SHORT) syndrome are classified as normocephalic PD disorders [2-5]. Historically, PD subtypes have been further differentiated via specific physical presentations, made complex by the variation within a single subtype, and even more difficult by overlapping features between subtypes. Fortunately, the advent and decrease in the cost of next-generation sequencing technologies have facilitated the rapid discovery of genes causal for PD, and increased the possibility of an accurate molecular and clinical diagnosis.

Following the initial discovery of *ATR*, in 2003, mutations in *BRCA2*, *CENPJ*, *CEP152*, *XRCC4*, *ATRIP*, *POC1A*, and *PCNT* were shown to be associated with PD [3,5-11]. Most causal genes are essential for fundamental cellular processes, such as DNA-damage response, mitosis, and DNA replication initiation [1]. One interpretation of these data attributes the pathophysiology of PD to an overall imbalance of cell proliferation and apoptosis [12]. In addition to increased rates of accurate PD diagnosis, NGS studies and discoveries highlight the genetic overlap and pleiotropy within PD and PD genes. Recent studies have uncovered allele-specific pleiotropy in *POC1A*-related PD. Mutations in *POC1A* are causal for both **S**hort stature, **O**nychodysplsia, **F**acial dysmorphisms and Hyper**T**richosis (SOFT) syndrome and a related but distinct variant POC1A-related (vPOC1A) syndrome [11,13]. While patients with vPOC1A syndrome present with a milder overlap of SOFT syndrome, they *exclusively* present with dyslipidemia with insulin resistance, and acanthosis nigricans as additional symptoms [11,13-15]. To-date, three mutations in three patients have been reported for vPOC1A syndrome and nine for SOFT syndrome. All nine SOFT syndrome mutations are isolated to exons 2-6, and perturb the encoded WD40 domain encoded protein [4,11,16-22]. In contrast, vPOC1A syndrome mutations solely affect proper translation of exon 10, leading to defects in the C-terminal spacer sequence and Poc1 domain in the encoded protein [13-15].

Here we report on two siblings born to consanguineous parents (2nd cousins), presenting with vPOC1A syndrome (Figure 2-1). Whole-genome sequencing analysis revealed they carry a novel variant in *POC1A* intron 9 found to affect splicing of *POC1A* transcripts. This mutation contributes to the allelic spectrum of PD-related *POC1A* variants and further delineates the allele-specific boundary between vPOC1A syndrome and SOFT syndrome.

Materials and Methods

DNA Library Prep

HudsonAlpha Discovery (Huntsville, AL) prepared Illumina sequencing libraries for samples V-8 and V-9 using their standard protocols. Briefly, a Qubit (ThermoFisher Scientific, Waltham, MA) fluorometric assay was used to measure each DNA sample’s concentration. DNA integrity was verified via agarose gel electrophoresis. After quality control, all samples with passing metrics were processed to create a sequencing library. DNA samples were normalized to 1,000 ng of DNA in 50 ul of water. Following normalization, samples were acoustically sheared via Covaris LE-220 instrument to a final fragment size of ~350-400 bp. The sheared DNA was then transformed into a standard Illumina paired-end sequencing library via standard methods. The sheared DNA was end-repaired and A-tailed using New England Biolabs End-Repair and A-Tailing kits, respectively, using the manufacturer’s recommended conditions. Following each step, the library was purified via Agencourt AMPure XP beads and eluted in water. Standard Illumina paired-end adaptors were ligated to the A-tailed DNA via New England BioLabs Rapid Ligation kit. Following ligation, the reactions were purified using AMPure XP beads. The purified ligated DNA was amplified via PCR using KAPA Biosystems HIFI PCR kit using 6 cycles of PCR. The primers were standard Illumina primers with a custom 7-base sample barcode in the i7 position to allow sample identification/de-multiplexing following sequencing. The final library was quality controlled using size verification via PerkinElmer LabChip GX and real-time PCR using the KAPA SYBR FAST qPCR Master Mix, primers and standards according to the manufacturer’s directions. Libraries were normalized to 2.5 nM stocks for use in clustering and sequencing.

Whole-Genome Sequencing

DNA samples for individuals V-8 and V-9 were sequenced at the Hudson–Alpha Institute of Biotechnology (Birmingham, AL) using their published protocols. Sequencing was performed to approximately 30X coverage per genome on the Illumina HiSeq X platform. Following sequencing, all base-calling was performed using standard Illumina software to generate the final FASTQ files for each sample.

Sequencing Quality Control

Read length, per base sequencing quality, per base pair sequence content, and per sequence GC content and distribution for each FASTQ file generated from WGS were assessed using FastQC [23]. A read length of 150 bp is expected for all Illumina HiSeq X paired-end reads. For both samples all reads were 150 bp in length. All read positions had an average quality score $\geq$ 20 and there was no bias of nucleotide content by read position. For both samples the average %GC content of all reads were normally distributed with an average %GC content equal to 38%.

Sequence Alignment: PEMapper

FASTQ files were aligned on a per sample basis with PEMapper [24] using default parameters and a Smith-Waterman alignment threshold of 95%, as recommended for 150-bp paired-end reads. Alignment was performed relative to the human Hg38 reference as reported by the University of California at Santa Cruz (UCSC) Genome Browser on July 1, 2015. The output from PEMapper, pileup and indel files, were used as input for variant calling with PECaller [24]. Alignment performance was checked before moving to variant calling. All samples had $\geq$ 65% of sequence reads uniquely mapped and an average depth of coverage $\geq$ 20X.

Variant Calling: PECaller

Variant calling was performed in a single batch using PECaller [24], which assumes multiple samples sequenced with the same technology will be available. Optimal PECaller performance is achieved when at least 50 genomes are called in batch; 59 control genomes, were included with the genomes from individuals V-8 and V-9 (61 genomes total). Variants were called with the default theta value of 0.001 and a 95% posterior probability for a genotype to be considered called variant. Calls were produced for repeat-masked (unique) subset of the human

Hg38 reference as reported by the University of California at Santa Cruz (UCSC) Genome Browser on July 1, 2015. The initial .snp file output from PECaller was used in a subsequent step to merge SNP variant calls with INDEL variant calls, producing a final "merged" .snp file. This raw file was used for site and sample quality control. A sum total of n=7,049,674 variants were called for all 61 samples.

Whole-genome Sequence Quality Control

Quality control (QC) was performed on per-site and per-sample basis. The following metrics were used to flag and/or exclude samples and variant sites from QC and analysis, and were calculated using a custom QC pipeline consisting of multiple in-house-developed scripts, PLINK 1.9 [25], R [26], and Bystro [27]:

1. *Per-site QC: Missing call rate and unlocalized contigs:* Variants with a missing call rate greater than or equal to 10% were removed from subsequent QC *and* variant analysis (n = 264,086). Variants located on unlocalized or random contigs were also removed from subsequent QC and variant analysis, resulting in a final sum total n = 6,740,383 variants.
2. *Sample Mixture Check:* Possible sample mixture was checked by calculating the ratios of minor allele homozygous calls to heterozygous calls. While this number varies between call batches, and thus cannot be compared across different calling experiments, non-mixed samples within the same calling batch should exhibit the similar ratios. A sample's homozygous:heterozygous call ratio that falls three standard deviations outside the batch mean, provides evidence for sample mixture. The ratios were calculated by generating a

.PED file including only autosomes from PLINKv1.9 using the --autosome and --recode flags. No samples were removed on the basis of possible sample mixture.

3. *Per-sample QC: Transition:Transversion Ratio:* Transition:transversion (Ti:Tv) ratios were calculated for each genome in the variant calling batch using a Bystro. The Ti:Tv ratio for an individual genome is expected to be approximately 2.00, with a ratio of 2.04 representing a quality genome. The 59 control genomes used in batch calling were previously validated for quality calling performance, therefore a *mean* Ti:Tv ratio less than 2.00 suggested a failed variant-calling experiment. The mean Ti:Tv ratio for the entire batch indicated a successful variant-calling experiment ($Ti:Tv_{batch}$ = 2.05 ± 0.004). Individuals V-8 and V-9 had Ti:Tv ratio >2.00 (Ti:Tv = 2.04; Ti:Tv = 2.04, respectively), and no other samples were removed from analysis on the basis of Ti:Tv ratio.
4. *Per-sample QC: Silent:Replacement Ratio:* The silent:replacement (sil:rep) ratios were calculated for each genome using a custom python script. The expected sil:rep ratio for a single genome is expected for fall between 1.05 and 1.15, with 1.15 indicating a quality genome. A mean sil:rep less than 1.05 suggested a failed variant-calling experiment. The mean sil:rep ratio for the entire batch indicated a successful variant-calling experiment ($sil:rep_{batch}$ = 1.142 ± 0.011). Individuals V-8 and V-9 had sil:rep ratio 1.05 < sil:rep < 1.15 (sil:rep = 1.13; sil:rep = 1.12, respectively), and no other samples were removed from analysis on the basis of sil:rep ratio.
5. *Per-sample QC: Missing call rate:* The missing call rates for samples were calculated using PLINKv1.9 [25]. Briefly, the “merged” .snp file was converted to BCF format with a custom script (snp_to_vcf2), multiallelic variants were split into single variants using BCFtools 1.3 [28], and loaded into PLINK. The following flags were used during

loading: --*bcf*, and --*keep-allele-order*. Per sample missing call rates were calculated using the –*missing* flag in PLINK, and the batch mean missing call rate and standard deviation was calculated; a mean missing call rate greater than or equal to 3% indicated a failed variant-calling experiment. The mean batch missing rate was 1.00 ± 0.1%, indicating a successful variant-calling experiment. Individuals V-8 and V-9 had missing call rate of 1.1% and 1.2%, respectively, and no other samples in the current analysis were removed on the basis of low call rate.

6. *Sex Check:* PLINK was used to calculate the F coefficient estimates for the X chromosome and impute sex assignment for each sample in the batch. Before sex was inferred, the --*split-x* flag with the *hg38* modifier was used to identify and remove pseudoautosomal regions of the X chromosome. The default parameters for the --check-sex flag were used to infer sex. All samples with an F coefficient less than or equal to 0.2 was assigned as female, and a sample with an F coefficient greater than or equal to 0.9 was assigned as male. All samples were assigned a sex that coincided with our expectation based on pedigree information.

7. *Relationship Inference:* Expected relationships between related samples of the batch were verified with PLINK. Variants were pruned to remove SNPs likely to be linkage disequilibrium using the --*indep-pairwise* flag and 50, 5, and 0.2 for the variant count window size, variant count step size and r^2, respectively. The --*genome* flag was used to infer relationships (coefficient of relatedness; *r*) on this set of pruned SNPs. Among the control genomes there was a known parent-offspring relationship, which was used as a positive control, while the remaining control genomes were known to be unrelated. All

samples' inferred relationships matched our expectations based on information provided by the families.

Variant Annotation

After QC the finalized dataset was submitted to Bystro, an online variant annotator and filterer [27]. The data was submitted according to Bystro instructions.

Regions of Homozygosity (RoH) Analysis

Regions of homozygosity (RoHs) were found using PLINK's --*homzyg* flag. The *group* modifier was used to identify RoHs shared by both case samples where both individuals were also homozygous for the same allele. The default parameters for --*homozyg* were used, and a set of LD-pruned variants were used to identify RoHs. Once allelically shared regions were identified the consensus coordinates for all regions were submitted as a search query to Bystro [27]. As these inferred regions could potentially include heterozygous calls, variants in these regions were then refined by filtering for variants for which both case samples were homozygous. The resulting variants were then sorted by Genome Aggregation Database (gnomAD) total population minor allele frequency and CADD score. All variants with and MAF < 0.001 *and* a CADD score $\geq$ 15 were considered candidate causal variants (Table S2-1).

Genome-wide Homozygous Variant Analysis

Variants homozygous in both case samples were isolated using Bystro's natural language search function. The search query, *homs: (SL106253 SL106254)* was used to find all variants for which both cases were homozygous. The resulting variants were sorted in by MAF and CADD score. All variants with and MAF < 0.001 *and* a CADD score ≥ 15 were considered candidate causal variants (Table S2-1).

In silico Splice Analysis

The potential splicing effects of NM_015426.5:c.981+5G>C were examined *in silico* with four independent splice prediction algorithms: NNSplice (http://www.fruitfly.org/seq_tools/splice.html), SD-Score (https://www.med.nagoya-u.ac.jp/neurogenetics/SD_Score/sd_score.html), and Human Splice Finder 3.1 and MaxEntScan (https://www.genomnis.com/access-hsf).

Plasmids and Site-Directed Mutagenesis

Exons 8-11 of human *POC1A* plus fused partial intronic regions of exons 8-10 were synthesized as a 2.2 kb minigene sequence. The minigene sequence was ligated into the pEGFP-C1 mammalian expression vector between the BsrGI and EcoRI sites at the C-terminus of the eGFP open reading frame (pEGFP-C1-wtPOC1A). The mutant minigene construct (pEGFP-C1-mutPOC1A) containing the NM_015426.5:c.981+5G>C variant was generated with site-directed mutagenesis (SDM). A PCR product was synthesized using pEGFP-C1-wtPOC1A as a template and primers: 12.pEGFP-C1.seq.S 5’-TGAGCAAAGACCCCAACGAGAAGC-3’ and MZ28.g2102c.A 5’-AGTCACAGAGCTCAGGACATGCCTGCCAAGCCAgTTACCAGATTC-3’. Primer MZ28.g2102c.A contains a mismatched basepair (lowercase) in order to substitute g.2102 G>C. The mutated PCR product was ligated into pEGFP-C1-wtPOC1A between the BsrGI and SacI sites. Similarly, a sequence containing a FLAG-tag fused to the C-terminal of the full coding sequence for *DUSP7* (NM_001947) was synthesized (2.3 kb sequence). The sequence was ligated into the pcDNA5/TO mammalian expression vector between. Constructs containing the DUSP7 Y401C variant and the C331S, cat. dead variant was separately generated via SDM with primers: SDMY401C.F 5'-CTAGCGAACAGCTGTGCTTCTCTACCCCTAC-3' & SDMY401C.R 5'- GTAGGGGTAGAGAAGCACAGCTGTTCGCTAG-3' and SDMC331SC.F 5'- CGTGCTGGTCCACTCTCTGGCCGGC-3' & SDMC331SC.R 5'- GCCGGCCAGAGAGTGGACCAGCACG -3', respectively.

HEK293 Cell Transfection

HEK293 cells were transfected with plasmid DNA (WT, Mutant, and empty plasmid, separately). Cells transfected with no DNA were also included as a negative control. Transfection was performed with TransfeX™ Transfection reagent (ATCC, Cat. No. ACS-4005) according to the standard protocol. Briefly, the day prior to transfection, 1x 10^6- 1.5 x 10^6 cells were seeded in T25 flasks with 6 mL of growth medium (DMEM +10% FBS) and incubated overnight at 37°C with 5% CO_2. The next day, old growth media was replaced with 6 mL of fresh growth media and DNA transfection complexes were prepared with 3 ug of each plasmid DNA and 6.0 uL of TransfeX™ reagent. Complexes were then incubated at room temperature for 15 minutes, and then distributed to cells via pipetting and rocking. Transfected cells were incubated for 48 hours at 37°C with 5% CO_2. For DUSP7 experiments, HEK293 cells were transfected with an overexpression vector encoding either FLAG-tagged wildtype DUSP7 (WT), catalytically dead DUSP7 (C331S), or Y401C variant DUSP7. Transfection was performed with Lipofectamine LTX™ with Plus Reagent (ThermoFisher, Cat. No. 15338030) according to the standard protocol. Briefly, the day prior to transfection, 5.0 x 10^6 cells were seeded in 100-mm culture dishes with 30 mL of growth medium (DMEM +10% FBS) and incubated overnight at 37°C. The next day, old growth media was replaced with 30 mL of fresh growth media and DNA transfection complexes were prepared with 45 ug of each plasmid DNA and 90 uL of Lipofectamine LTX™ reagent. Complexes were then incubated at room temperature for 30 minutes and then distributed to cells via pipetting and rocking. Transfected cells were incubated for 24 hours at 37°C with 5% CO_2.

RT-PCR

Total RNA was extracted from transfected cells with the Qiagen RNeasy® Mini Kit (Qiagen, Cat. No. 74104) according to the standard protocol. Briefly, 1x 10^7 transfected cells were harvested and lysed via centrifugation with 600 uL of Buffer RLT. 600 uL of 70% ethanol was added to lysates. Total RNA was then applied to RNeasy silica membrane columns and cleaned in a series of 3 spin washes. RNA was eluted with 35 uL of RNase-free water. cDNA synthesis was performed according to the SuperScript® III First-Strand Synthesis System (Invitrogen, Cat. No. 18080-051) standard protocol. Six reactions were included in the protocol with RNA from the following sources: (1) Total RNA extracted from cells transfected with pEGFP-C1-wtPOC1A plasmid, (2)Total RNA extracted from cells transfected with pEGFP-C1-mutPOC1A vector, (3) Total RNA extracted from cells transfected with empty pEGFP-C1 plasmid, (4)Total RNA extracted from cells transfected with no DNA, and (5) HeLa RNA included with SuperScript® III kit. cDNA was synthesized from total RNA in 20-uL reactions using Oligo(dT)$_{20}$ primers. Approximately 5 ug of RNA was mixed with 1 uL of 50 uM Oligo(dT)$_{20}$ primers, 1 uL of 10mM dNTP mix and X uL of DEPC-treated water up to 10 uL. RNA mixtures were incubated at 65°C for 5 minutes, then placed on ice for $\geq$ 1 minute. 12 uL of

10X RT Buffer, 24 uL of 25 mM MgCl2, 12 uL of 0.1M DTT, and 6 uL of RNaseOUT were mixed to create 6X cDNA synthesis mixture. 9 uL of the cDNA synthesis mix and 1 uL of SuperScript® III reverse transcriptase (RT) was added to the 10-uL RNA mix. For the negative control 1 uL of DEPC-treated water was added instead of RT. cDNA synthesis reactions were incubated at 50°C for 50 minutes, terminated at 85°C for 5 minutes, and then chilled on ice. Reactions were collected by brief centrifugation. 1 uL of RNase H was added to each reaction and reactions were incubated at 37°C for 20 minutes to digest template RNA. PCR amplification of target cDNA was performed with the KAPA HiFi HotStart ReadyMix PCR Kit (Roche, Cat. No. KK2601) and the following gene-specific primers: pEGFP-C1-POC1A-specific primers, eGFP primer 5'-TCTATATCATGGCCGACAAGC-3' (forward) and Ex11 primer 5'-TGCTGGTTCTCCAGACACTG-3' (reverse); pEGFP-C1-specific primers, CMV promoter primer 5'-AGGCGTGTACGGTGGGAGGTCTA-3' (forward) and polyA signal primer 5'-GTTCAGGGGGAGGTGTGGGAGGTT-3' (reverse); and human β-actin-specific primers, 5'-GCTCGTCGTCGACAACGGCTC-3' (forward) and 5'-CAAACATGATCTGGGTCATCATCTTCTC-3' (reverse). A summary of PCR templates and associated primers is described in Table S2. For each reaction, 10 uL of PCR-grade water, 12.5 uL of 2X KAPA HiFi mix, 0.75 uL each of 10uM reverse and forward primer and 1 uL of cDNA were mixed to create a 25-uL PCR reaction mix. PCR was performed with the standard cycling protocol and a 71°C annealing temperature. PCR products were separated on 2% agarose gels at 80v for 4 hours. Splicing isoforms were identified by size.

PCR Purification and Sanger Sequencing

Purified PCR product sequences were verified by Sanger sequencing. PCR reactions were purified using the Qiagen QIAquick® PCR Purification Kit (Qiagen, Cat. No. 28104) according to the standard protocol. Briefly, reactions in 5 volumes of binding buffer were applied to QIAquick spin column by centrifugation. Bound DNA was washed with an ethanol buffer and then eluted in 30 uL of water. Forward and reverse Sanger reactions confirmed PCR sequences.

Recombinant DUSP7 Immunoprecipitation

Recombinant DUSP7 protein was pull downed from transfected HEK293 cells using anti-FLAG resin, according to the FLAG® Immunoprecipitation Kit (Sigma-Aldrich; Cat. No. FLAGIPT1). Briefly, cells were washed with PBS buffer and lysed with custom lysis buffer (50 mM HEPES, pH 7.4; 10% Glycerol; 2 mM EGTA; 2mM MgCl2; 1% Nonidet P-40; 1 mM PMSF; 1:100 protease inhibitor cocktail (Sigma-Aldrich, Cat No. P8340-1ML) on ice for 10 minutes. Cells were then collected and sonicated at 3W for 2 rounds of 20-30 seconds each, and clear lysates were stored at -80°C or used immediately. Before immunoprecipitation Anti-FLAG M2 affinity resin washed two with 1X Wash Buffer, packed and then washed an additional three times via centrifugation (7,000 x g for 30s each). Cell lysates were applied to the washed resin and rocked overnight at 4°C. Lysate-resin solutions were then washed three times with 1X Wash Buffer via centrifugation. Recombinant DUSP7 was eluted with 3X FLAG Elution Buffer (150 ng/uL 3X FLAG peptide, 1X Wash Buffer) via rocking overnight at 4°C followed by centrifugation (7,000 x g for 30s).

DUSP7 End-point *In Vitro* Phosphatase Assay

5 ug of purified recombinant DUSP7 protein were incubated in activation buffer (50 mM HEPES, 5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl) for 30 minutes at 30°C. 1 uL of 100 ng/uL

of active recombinant human ERK2 (prERK2; R&D Systems, Cat. No. 1230-KS) was then added to the solution and incubated for 2 hours at 37°C. Final end-point phosphatase solutions were separated via SDS-PAGE, and final results were analyzed via western blot using anti-phosphoErk1/2 (Cell Signaling Technology; Cat. No. 9101) and anti-Erk1/2 (Cell Signaling Technology; Cat. No. 9102).

Results

Family presenting with short stature and insulin resistance

The younger male sibling (individual V-9) was referred to a clinic at age 12 and presented with short stature (Z = -5.1), spinal stenosis, hip dysplasia, insulin resistance, and acanthosis nigricans. He was diagnosed with intrauterine growth restriction (IUGR) between 36 and 37 weeks and weighed 2 pounds 14 ounces at birth. He displayed strabismus, chronic otitis media, and conductive hearing loss. CT revealed an enlarged internal canal. MRI revealed possible gliosis of the subcortical white matter. Other clinical features included brachydactyly, mild dysmorphic features, mild microcephaly, and a history of episodes of hypothermia. At 15 years of age, the female sibling (individual V-8) presented with similar clinical features: short stature (Z = -5.4), scoliosis, hyperthyroidism, insulin resistance, and acanthosis nigricans. She was also diagnosed with IUGR and weighed 2 pounds 14 ounces at birth. She received surgery for strabismus and scoliosis. She had a history of recurrent otitis media and hearing loss. Upon examination, her clinical features included brachydactyly, mild dysmorphic features, and mild microcephaly (Figure 2-2).

POC1A is a strong causal candidate

To identify candidate causal variants for the phenotypes, we performed whole-genome sequencing followed by variant analysis. We conducted two orthogonal variant analyses under the hypothesis that the disorder is a novel autosomal recessive disorder caused by an identical-by-descent (IBD) mutation shared by both probands. We first identified homozygous regions likely to be inherited IBD from a common ancestor. Chromosomes 3, 6, and 12 contained candidate homozygous regions where both siblings were homozygous for the same allele (Figure 2-3A). Within these regions, we filtered for variants with a minor allele frequency less than 0.001 and a CADD score greater than 15. Next, we conducted a separate genome-wide search for rare, deleterious (MAF < 0.001; CADD > 15 variants homozygous in both siblings. Both analyses converged on nine variants, all located within a homozygous haplotype shared by the two affected probands on chromosome 3 (Table 2-1). Three variants were selected as strong causal candidates: c.981+5C>G in *POC1A* (NM_015426.4:c.981+5C>G), c.1202A>G in *DUSP7* (NM_001947.3:c.1202A>G), and c.2513G>A in *USP19* (NM_001199160.1:c.2513G>A).

Mutations in *POC1A* have previously been implicated in short stature, onychodysplasia, facial dysmorphism, and hypotrichosis (SOFT) syndrome and a related variant POC1A (vPOC1A) syndrome (vPOC1A). The c.981+5C>G variant is located in the ninth intron of *POC1A*, close to the 3’ end of exon 9 (Figure 2-3C). Sanger sequencing confirmed the variant co-segregates with the disorder as expected for an autosomal recessive disorder: Both probands are homozygous for the variant, both parents are heterozygous, and all unaffected siblings are either heterozygous or homozygous for the reference allele (Figure 2-3B). *In silico* analysis revealed that the variant is predicted by four independent splice prediction algorithms to weaken a canonical mammalian splice donor motif located in exon 9 and intron 9 [29-32]. The combined evidence suggested that c.981+5C>G in *POC1A* is a strong candidate causal variant for the disorder.

POC1A variant causes exon skipping of exon 9

We interrogated the potential effect of c.981+5C>G on splicing using a mutant *POC1A* minigene construct. The construct contained exons 8-11 of *POC1A* with intervening portions of introns 8-10 fused to eGFP (Figure S2-1). To determine the variant's effect on splicing, we transfected the construct into HEK293 cells. PCR amplification of cDNA using primers spanning eGFP and exon 11 demonstrated that c.981+5C>G induces skipping of exon 9 in the minigene paradigm when compared to a wildtype construct (Figure 2-4A-B). Notably, the mutation appears to allow some retention of exon 9. Sanger sequencing of the PCR products confirmed this result (Figure 2-4C). Transcripts that skip exon 9 would lack 33 amino acids, which are predicted to encode a portion of the seventh WD40 repeat motif and the C-terminal linker sequence between the final WD40 motif and the Poc1 coiled-coil domain in the POC1A protein (Figure 2-3C).

DUSP7 c.1202A>G variant does not contribute to the probands' phenotypes

There are increasing reports of individuals with a disorder caused by rare mutations in more than one known cause of a Mendelian disease [33,34]. While our functional experiments provide strong evidence for the action of the c.981+5C>G variant, we sought to test the hypothesis that other variants might contribute to the clinical presentation in the probands. The *DUSP7* c.1202A>G variant was an attractive candidate for contributing to the observed probands' phenotypes. *DUSP7* encodes the dual-specificity phosphatase 7 protein, which interacts with the mitogen-activated protein kinase (MAPK) pathway [35,36]. Specifically, DUSP7's canonical substrate is active ERK2, an essential kinase in the MAPK and a secondary branch of the insulin signaling pathway [37,38]. The DUSP7 protein has also been shown to interact with the growth hormone receptor *in vitro* [39]. The variant substitutes a tyrosine for a cysteine residue in the protein's C-terminus structure (p.Y401C). We tested the hypothesis that the DUSP7 mutation abolishes activity towards ERK2 using an endpoint *in vitro* phosphatase assay followed by western blot analysis. This assay indicated that the DUSP7 p.Y401C variant does not affect its activity towards active ERK2. We concluded that this variant does not contribute to the disorder through an ERK2 mechanism (Figure S2-2).

Discussion

We used whole-genome sequencing to identify candidate causal variants for two siblings with consanguineous parents who presented with vPOC1A syndrome. Our study revealed nine rare and putative deleterious homozygous variants on a shared haplotype on chromosome 3. One was a variant in intron 9 of the gene *POC1A,* which encodes Protein of Centriole 1A. Mutations in exon 10 of *POC1A* cause variant POC1A-related (vPOC1A) syndrome which includes short stature and insulin resistance as clinical presentations. We did not observe any point mutations in exon 10. Instead, our variant falls in the +5 position of the 5' consensus sequence, the second most common site of disease-causing splice point mutations [40], and was predicted to cause skipping of exon 9. In line with expectations [41], we showed the mutation predominantly results in a messenger RNA (mRNA) that skips exon 9, with a minor expression of an mRNA that retains exon 9. To the best of our knowledge our variant (NM_015426.4:c.981+5C>G) is the first pathogenic intronic mutation causal for vPOC1A syndrome reported in the literature.

In addition to vPOC1A syndrome, mutations in *POC1A* are also known to be causal for SOFT syndrome. While both syndromes are classified as a type of primordial dwarfism syndromes and have significant overlap in associated features, patients with vPOC1A syndrome distinctively present with dyslipidemia, insulin resistance, and acanthosis nigricans [13-15]. In contrast, SOFT syndrome patients have more severe facial dysmorphia and hypoplastic nails [11,21,22]. Mutations observed in patients distinguish these two syndromes. Mutations causing SOFT syndrome have been reported in exons 2-6 and cause defects in the WD40 domain in the POC1A protein, which is critical for localization at the centrioles [42]. Mutations causing vPOC1A syndrome have only been observed in exon 10, and it is hypothesized that vPOC1A syndrome is the result of the skewing of ratios between 10- and 10+ mRNA [13,14]. Two reported mutations (c.1048delC and c.1047_1048dupC) truncate the POC1A protein at exon 10, which encodes the C-terminal Poc1 coiled-coil domain [13,14] and may be involved in protein-protein interactions [43]. More recently, a mutation in exon 9 (c.884delT) is reported to delete a single nucleotide in exon 9, causing the creation of a premature termination codon (PTC) downstream in exon 10. Interestingly, c.884delT also disrupted the AG 3' acceptor splice site of exon 9 and is predicted to cause abnormal splicing. RT-PCR experiments have shown that the mutation does indeed affect proper splicing, increasing the relative amount of mRNA isoforms that skip exon 10 [15].

Our study yields a surprising result. Our patients present with vPOC1A syndrome yet are homozygous for a mutation that is predicted to remove a portion of the WD40 domain, a molecular lesion exclusively associated with SOFT syndrome. Variant c.981+5C>G results in the deletion of the last 14 protein residues of the seventh WD40 repeat. Within the WD40 domain structure, the last and first blade motif interlock to form a stable propeller structure [44];

therefore, it is possible that deletion of the last 14 residues could destabilize the structure. The WD40 domain is vital for protein localization to the centrosome [42]. Additionally, mice with a *Poc1a* mutation that causes skipping of the in-frame exon 8 and leads to the deletion of 23 amino acids of the final seventh WD40 repeat exhibit a skeletal dysplasia like that seen in SOFT syndrome patients [45,46]. Skipping of exon 9 would affect all three validated transcripts of *POC1A*, therefore potentially disrupting the WD40 domain structure of all three protein isoforms. However, the trace amounts of full-length transcript produced by variant c.981+5C>G could partially compensate for the transcripts missing exon 9, therefore allowing some production of functional POC1A protein.

In addition to the three validated transcripts, *POC1A* is predicted to produce seven additional transcripts via alternative splicing, two of which are expected to skip exon 9. Recently, Majore et al., have reported *POC1A* mRNA species that skip exon 9 in healthy controls [15]. We were unable to assess mRNA species in the patients presented here. However, the existence of an endogenous mRNA species that skips exon 9 could explain the disconnect between the siblings' presentation and the nature of variant c.981+5C>G. Species that skip exon 9 may be expressed in a tissue-specific manner and could partially compensate for the lack of full-length protein isoforms. Thus, the vPOC1A syndrome presented here could result from an alteration in relative mRNA ratios rather than the absence of exon 9 in all isoforms. It is not an uncommon occurrence for changes in transcript ratios to result in diseases. In Frontotemporal dementia with Parkinsonism, chromosome 17 type (FTDP-17), mutations in exon 10 of *MAPT* adjust ratios of tau protein 3R and 4R isoforms via altered regulation of splicing [47,48]. According to the GTEx database, the full-length POC1A transcript is the most abundant in all tissues examined [49]. The splicing effects of c.981+5C>G would decrease this prevalence and could be the mechanism underlying vPOC1A syndrome in our patients. Additional RNA and protein studies investigating the full range of *POC1A* mRNA and protein isoforms are needed to determine whether the causal mechanism for vPOC1a syndrome in the current patients is solely the result of the deletion of exon 9 or is caused by altering the mRNA ratios.

Primordial dwarfism (PD) is considered the result of a net imbalance between cell proliferation and cell death [1]. Like other genes causal for PD, *POC1A* encodes a centriolar protein which is critical for cellular functions, including centrosome organization, centriole integrity, and ciliogenesis [3,50,51]. Proper functioning of the centrosome and centrioles is critical for successful mitosis and cell cycle progression [52]. Defects in the centrosomes can lead to the accumulation of DNA damage, cells with aneuploidy, and an overall decrease in cell cycle efficiency. As cell defects increase, overall cell death increases and outweigh proliferation leading to a decrease in organism cell number and ultimately a reduction in organism size [2,12]. In line with this, previously reported patients with vPOC1A syndrome showed an increased presence of anaphase bridges during cytokinesis due to abnormal mitotic spindles and increased DNA damage, leading to cell cycle arrest and apoptosis [13].

Proteins that localize to centrosomes and cilia tightly link the lifecycles of these vital cellular structures [53-55]. Knockdown of *Poc1*, a *Tetrahymena* ortholog of *POC1A*, results in instability of the basal body that supports both the development of the cilia and centrosomes [42], and *Poc1a*-null mouse embryonic fibroblasts show defects in primary cilia [46]. Thus, mutations in structural proteins within the centrioles could also indicate defects in cilia structure or function. Cilia plays a role in many cell-signaling pathways, including those necessary for growth and development and metabolic insulin regulation [55,56]. Wnt signaling is essential for planar cell polarity in which defects could also contribute to the growth deficiency seen in vPOC1A syndrome patients, particularly in the chondrocytes. During growth, planar cell polarity signals are required for proper division and integration of new chondrocytes in the proliferative zone [57]. Mice with mutations in *Poc1a* are shown to have defects in long bone growth due to disorganized growth plate morphology that becomes more disorganized with age. Specifically, mouse chondrocytes within the proliferative zone exhibit abnormal shape, improper division direction, and fail to integrate into the existing columns of proliferating chondrocytes. Primary cilia also play a role in insulin signaling, as evidenced by several ciliopathies with insulin resistance, obesity, or type-II diabetes as symptoms [58-60]. While previous vPOC1A patients have not shown obvious defects in cilia, a growing body of evidence points to a link between centrosome biology and primary cilia as the mechanism underlying the growth and metabolic phenotypes seen in multiple PD disorders that feature insulin resistance [58-62].

Recent WES and WGS studies in consanguineous families support the possibility of multiple molecular diagnoses for individuals seemingly presenting with one disorder [63,64]. We identified two additional candidate variants based on their predicted and known functions and protein interactions. The NM_001947.3:c.1202A>G variant in *DUSP7* is predicted to substitute a tyrosine residue for cytosine (p.Y401C) within the C-terminal of the DUSP7 protein. DUSP7 is a regulator of ERK1/2, which is a component of the mitogen-activated protein kinase (MAPK) pathway, an essential regulator of growth [37,65,66]. Using an *in vitro* phosphatase assay, we determined that the DUSP7 variant is likely not causal through its role in dephosphorylating ERK1/2. We have not ruled out the contribution of the *USP19* variant, NM_001199160.1:c.2513G>A, to the phenotype. *Usp19*-null mice show increased insulin sensitivity in muscle and liver tissues [67], suggesting some interaction with metabolic insulin signaling.

We demonstrated the splicing effects of NM_015426.4:c.981+5C>G using an *in vitro* minigene splice assay, allowing investigation of single nucleotide changes in the absence of patient cells [68,69]. However, this method has limitations. Minigene assays do not completely replicate the genomic context of genes. In our construct, we truncated the long introns of *POC1A*, which may remove or affect *cis*-regulator*y* elements located more than 300 bp away from the exon-intron boundaries [70]. Also, cell- or tissue-specific conditions may influence the regulation of alternative splicing, splicing efficiency, and dynamics of the spliceosome, which is also not captured by an *in vitro* system [41,68]. Nevertheless, minigene splicing assays are a cost-effective and widely used approach to evaluate splicing effects of genetic variants of unknown significance and provide essential information for the interpretation of putatively causal variants [68-74].

Conclusions

In conclusion, we identify a novel splicing mutation in *POC1A* associated with vPOC1A syndrome. In an *in vitro* system, the mutation removes exon 9 from all relevant *POC1A* transcripts. It is predicted to impact the structure of the WD40 domain in POC1A, potentially making this the first case of partial disruption of the WD40 domain in vPOC1A syndrome. We further demonstrate that the study of rare and heterogenous genetic disorders like PD, can increase the accurate classification of patients via molecular and clinical findings. Overall, this evidence adds to the mutational spectrum of vPOC1A syndrome and *POC1A*-related PD and better defines the allelic series contributing to vPOC1A syndrome and SOFT syndrome.

Tables

Table 2-1. Nine candidate causal variants. Variant analyses converged on nine candidate variants located in a shared haplotype on the p arm of chromosome 3 (chr3:34476778-52317729, hg38). Allele frequency reported from gnomADv2 database [75].

Genomic Variant	Gene Name[a]	Variant Site Type	Codon Change	CADD	Allele Frequency[b]
chr3:g.37545413:A>G	*ITGA9*	Intronic	N/A	17.3	0
chr3:g.41195771:C>T	*CTNNB1*	Intergenic	N/A	17.7	3.9×10^{-4}
chr3:g.41320084:T>C	*ULK4*	Intronic	N/A	15	0
chr3:g.47891083:A>G	*MAP4*	Intronic	N/A	16.2	1.3×10^{-4}
chr3:g.49112616:C>T	*USP19*	Exonic	p.Arg737His	28.8	7.4×10^{-4}
chr3:g.49654750:G>A	*BSN*	Exonic	p.Val1732Met	16.7	3.6×10^{-4}
chr3:g.49968570:C>T	*RBM6*	Exonic/Intronic/5' UTR	p.Ala382Val	19	5.5×10^{-4}
chr3:g.52020873:T>C	*DUSP7*	Exonic	p.Tyr401Cys	18.8	0
chr3:g.52122374:C>G	*POC1A*	Splice site	N/A	16.3	0

[a] For intergenic variants the nearest gene is reported

[b] Overall allele frequency as reported by gnomad v2

Figures

Figure 2-1. Family Pedigree. Pedigree of family. Affected individuals indicated in solid black.

WGS analysis was performed on individuals V-8 and V-9

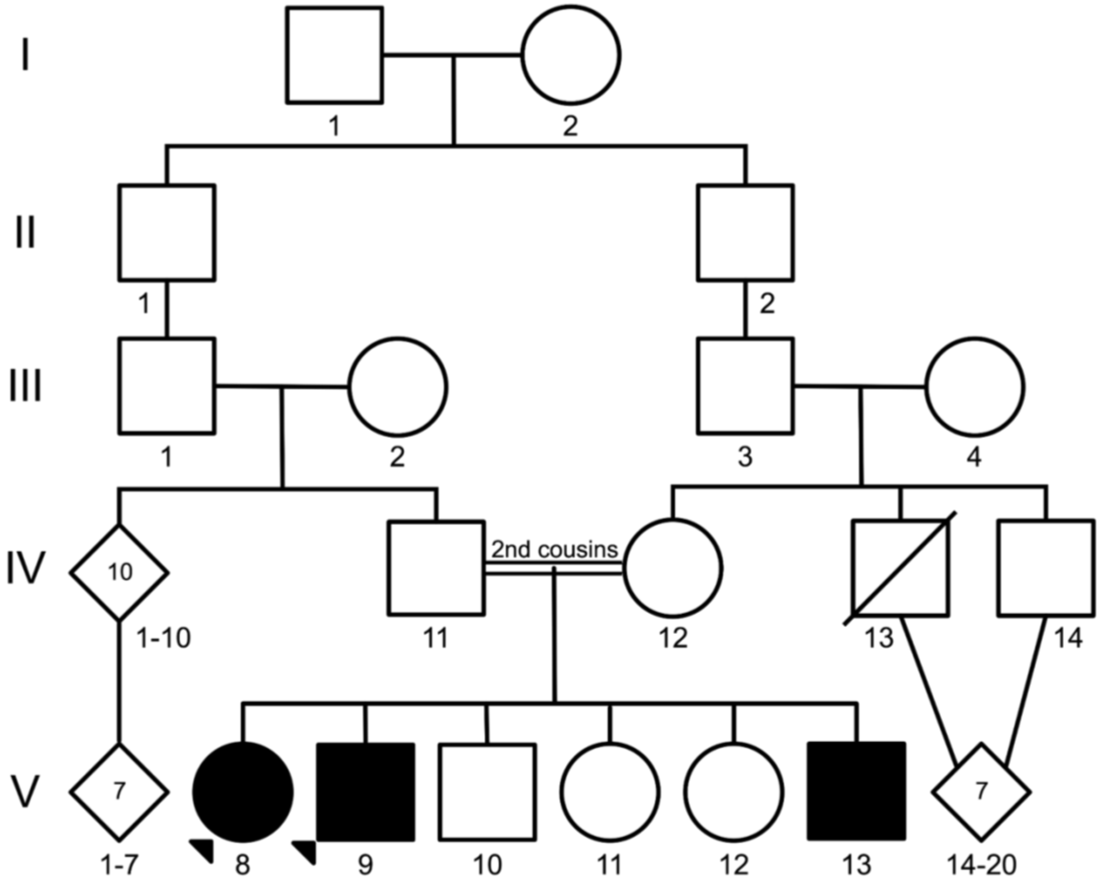


Figure 2-2. Clinical X-rays of affected individuals. Hand and spinal x-rays of V-8 at ages 11 (A) and 12 (B-E) and V-9 (F-H) at age 14.

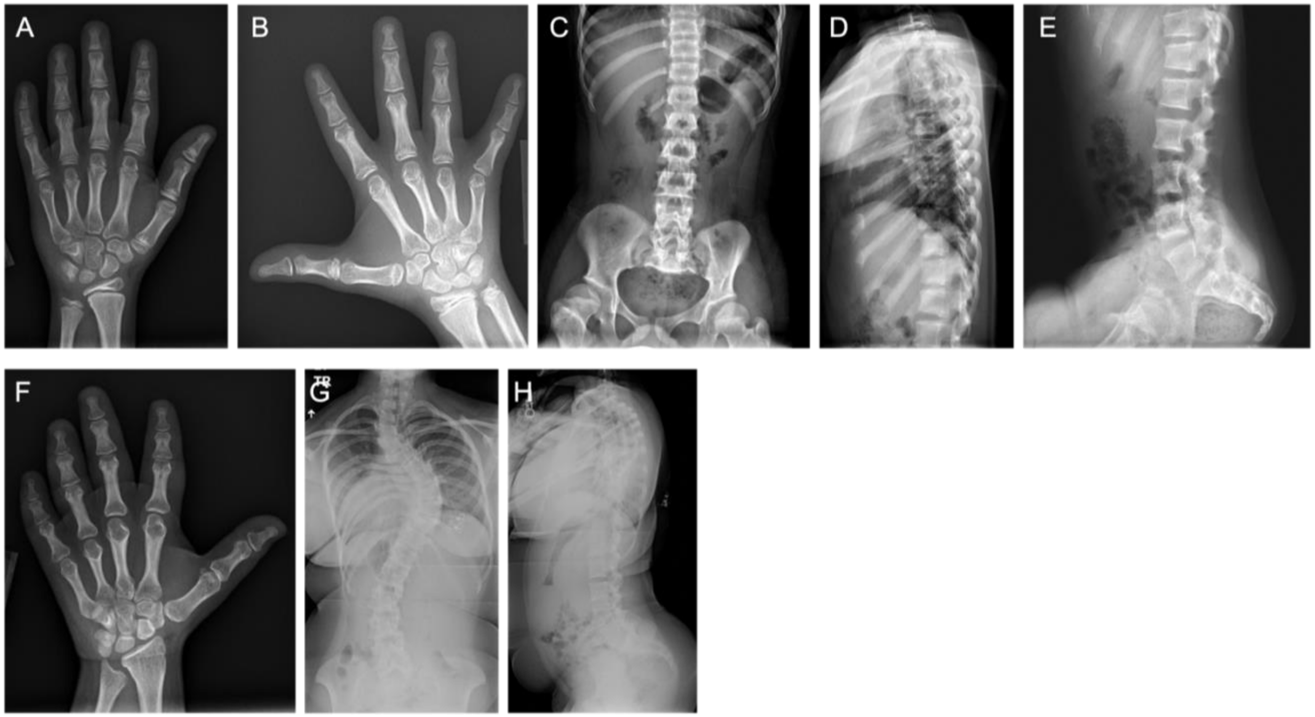


Figure 2-3. Candidate ROHs, sanger validation, and POC1A. (A) Three regions of homozygosity shared between individuals V-8 and V-9 on chromosomes 3, 6, and 12. Candidate variants on chromosome 3 (box) are shown as red and yellow stars, with strong candidates in *POC1A*, *DUSP7*, and *USP19*. (B) Familial segregation of *POC1A*, *DUSP7*, and *USP19* variants was validated using PCR and Sanger sequencing. Sanger traces for POC1A are shown. Both siblings (V-8 & V-9) are homozygous for the alternate allele (G, yellow highlight). The father and mother (IV-11& IV-12), are heterozygous for the alternate allele and reference C allele. (C) Illustration of POC1A experimentally validated mRNA transcripts and POC1A protein isoform 1 (longest protein) with corresponding NCBI RefSeq accession numbers are to the left (NCBI RefSeq). Exons are color-coded and a representation of intron 9 is indicated with a blue dotted-line. Location of reported pathogenic POC1A variants are shown and separated as causal for causal SOFT Syndrome or vPOC1A syndrome. Our reported mutation, NM_015426.4:c.981+5C>G, is shown in red.

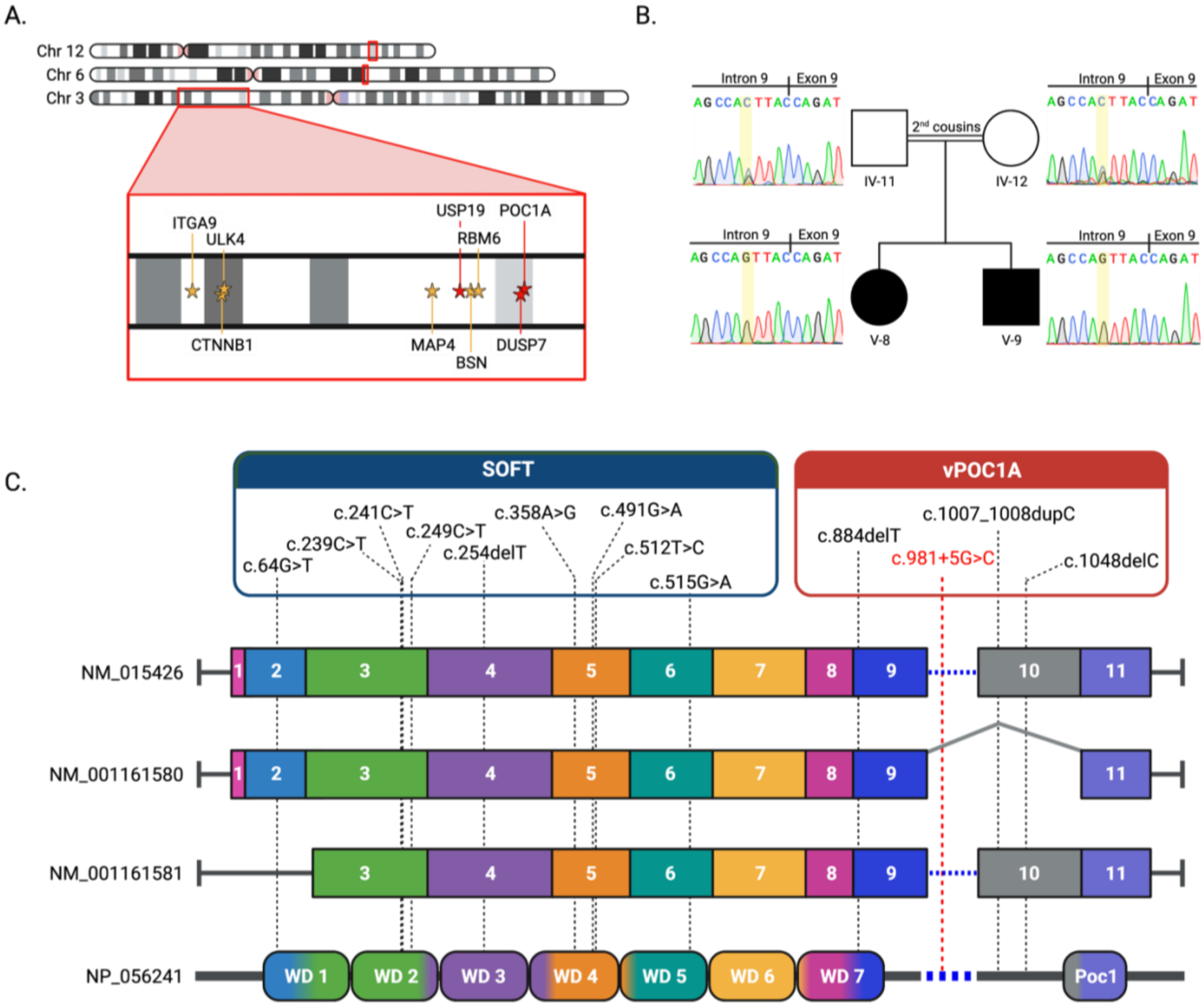
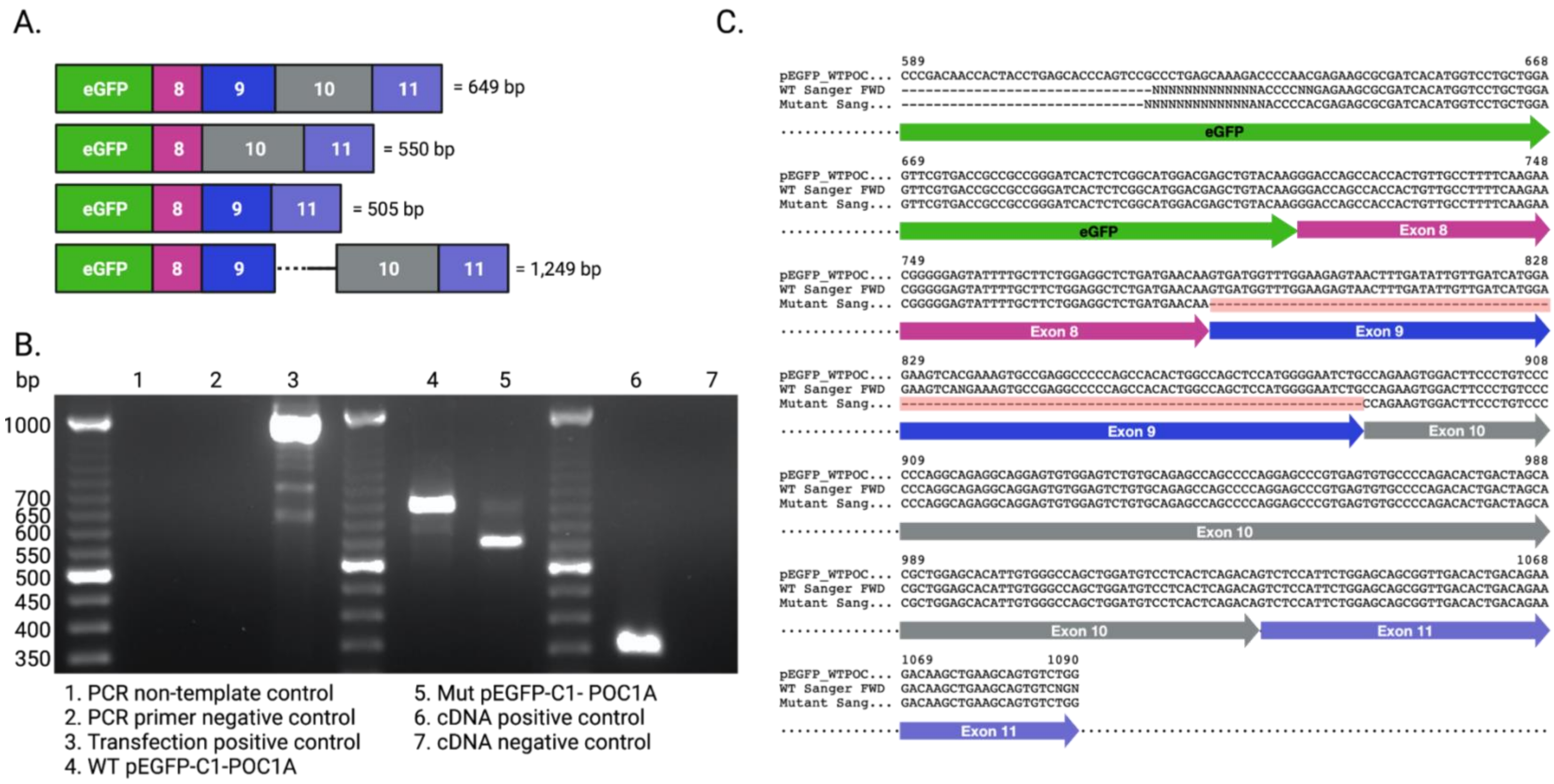

A.
Chr 12
Chr 6
Chr 3
ITGA9
ULK4
USP19
POC1A
RBM6
CTNNB1
MAP4
BSN
DUSP7
B.
Intron 9
Exon 9
AGCCACTTACCAGAT
AGCCAGTTACCAGAT
2nd cousins
IV-11
IV-12
V-8
V-9
C.
SOFT
c.241C>T
c.358A>G
c.491G>A
c.239C>T
c.249C>T
c.254delT
c.512T>C
c.64G>T
c.515G>A
vPOC1A
c.1007_1008dupC
c.884delT
c.981+5G>C
c.1048delC
NM_015426
NM_001161580
NM_001161581
NP_056241
1
2
3
4
5
6
7
8
9
10
11
WD 1
WD 2
WD 3
WD 4
WD 5
WD 6
WD 7
Poc1

Figure 2-4. Minigene splice assay results. (A) Expected product sizes of mutant wand wild type RT-PCR products based on predicted splice effects. From top to bottom: Wild type splicing, skip exon 9, skip exon 10, retain intron 9. (B) PCR products were resolved by gel electrophoresis. Lane 5 shows a band of 550 bp, in line with the expected product size for an isoform that skips exon 9. A band of 650 bp is also present, but fainter relative to the skip 9 fragment. (C) Alignment of wild type and mutant PCR sequences to wild type cDNA sequence. Sanger sequencing and alignment confirms skipping of exon 9 in *POC1A* minigene transcripts (shaded box).


A.
eGFP
8
9
10
11
= 649 bp
eGFP
8
10
11
= 550 bp
eGFP
8
9
11
= 505 bp
eGFP
8
9
10
11
= 1,249 bp
B.
bp
1
2
3
4
5
6
7
1000
700
650
600
550
500
450
400
350
1. PCR non-template control
2. PCR primer negative control
3. Transfection positive control
4. WT pEGFP-C1-POC1A
5. Mut pEGFP-C1- POC1A
6. cDNA positive control
7. cDNA negative control
C.
589 668
pEGFP_WTPOC... CCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGA
WT Sanger FWD -------------------------------NNNNNNNNNNNNNACCCCNNGAGAAGCGCGATCACATGGTCCTGCTGGA
Mutant Sang... ------------------------------NNNNNNNNNNNNNNANACCCCACGAGAGCGCGATCACATGGTCCTGCTGGA
eGFP
669 748
pEGFP_WTPOC... GTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGGGACCAGCCACCACTGTTGCCTTTTCAAGAA
WT Sanger FWD GTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGGGACCAGCCACCACTGTTGCCTTTTCAAGAA
Mutant Sang... GTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGGGACCAGCCACCACTGTTGCCTTTTCAAGAA
eGFP
Exon 8
749 828
pEGFP_WTPOC... CGGGGGAGTATTTTGCTTCTGGAGGCTCTGATGAACAAGTGATGGTTTGGAAGAGTAACTTTGATATTGTTGATCATGGA
WT Sanger FWD CGGGGGAGTATTTTGCTTCTGGAGGCTCTGATGAACAAGTGATGGTTTGGAAGAGTAACTTTGATATTGTTGATCATGGA
Mutant Sang... CGGGGGAGTATTTTGCTTCTGGAGGCTCTGATGAACAA--
Exon 8
Exon 9
829 908
pEGFP_WTPOC... GAAGTCACGAAAGTGCCGAGGCCCCCAGCCACACTGGCCAGCTCCATGGGGAATCTGCCAGAAGTGGACTTCCCTGTCCC
WT Sanger FWD GAAGTCANGAAAGTGCCGAGGCCCCCAGCCACACTGGCCAGCTCCATGGGGAATCTGCCAGAAGTGGACTTCCCTGTCCC
Mutant Sang... ---CCAGAAGTGGACTTCCCTGTCCC
Exon 9
Exon 10
909 988
pEGFP_WTPOC... CCCAGGCAGAGGCAGGAGTGTGGAGTCTGTGCAGAGCCAGCCCCAGGAGCCCGTGAGTGTGCCCCAGACACTGACTAGCA
WT Sanger FWD CCCAGGCAGAGGCAGGAGTGTGGAGTCTGTGCAGAGCCAGCCCCAGGAGCCCGTGAGTGTGCCCCAGACACTGACTAGCA
Mutant Sang... CCCAGGCAGAGGCAGGAGTGTGGAGTCTGTGCAGAGCCAGCCCCAGGAGCCCGTGAGTGTGCCCCAGACACTGACTAGCA
Exon 10
989 1068
pEGFP_WTPOC... CGCTGGAGCACATTGTGGGCCAGCTGGATGTCCTCACTCAGACAGTCTCCATTCTGGAGCAGCGGTTGACACTGACAGAA
WT Sanger FWD CGCTGGAGCACATTGTGGGCCAGCTGGATGTCCTCACTCAGACAGTCTCCATTCTGGAGCAGCGGTTGACACTGACAGAA
Mutant Sang... CGCTGGAGCACATTGTGGGCCAGCTGGATGTCCTCACTCAGACAGTCTCCATTCTGGAGCAGCGGTTGACACTGACAGAA
Exon 10
Exon 11
1069 1090
pEGFP_WTPOC... GACAAGCTGAAGCAGTGTCTGG
WT Sanger FWD GACAAGCTGAAGCAGTGTCNGN
Mutant Sang... GACAAGCTGAAGCAGTGTCTGG
Exon 11

Supplemental Tables

Table S2-1. Rare variants located in ROHs. List of rare (MAF<0.001) variants captured in ROH analysis.

Can be accessed online at: https://emory-my.sharepoint.com/:x:/r/personal/tmosle3_emory_edu/Documents/TMosley_Dissertation%20Links/TableS2-1_RareVariants_ROH.xlsx?d=w41fae0df2469429abb7a8d91863e2e23&csf=1&web=1&e=jU1ybI

Table S2-2. Genome-wide rare homozygous variants. List of rare (MAF<0.001) variants captured in genome-wide homozygous analysis.

Can be accessed online at: https://emory-my.sharepoint.com/:x:/r/personal/tmosle3_emory_edu/Documents/TMosley_Dissertation%20Links/TableS2-2_RareVariants_genomewide.xlsx?d=w3c14264205874d56b3c0a6fa7c7691d7&csf=1&web=1&e=budlql

Table S2-3. RT-PCR Primer and Templates

Lane	Reaction	Template	FWD Primer	REV Primer
1	PCR non-template control	None	028 FWD: 5’-TCTATATCATGGCCGACAAGC-3’	028 REV: 5’-TGCTGGTTCTCCAGACACTG-3’
2	PCR primer negative control	HEK 293 only cDNA	028 FWD: 5’-TCTATATCATGGCCGACAAGC-3’	028 REV: 5’-TGCTGGTTCTCCAGACACTG-3’
3	Transfection positive control	pEGFP-C1 cDNA	pcDNA3.seq.S: 5’-AGGCGTGTACGGTGGGAGGTCTA-3’	pEGFP-c1.seq.A: 5’-GTTCAGGGGGAGGTGTGGGAGGTT-3’
4	WT pEGFP-C1-POC1A	WT pEGFP-C1-POC1A cDNA	028 FWD: 5’-TCTATATCATGGCCGACAAGC-3’	028 REV: 5’-TGCTGGTTCTCCAGACACTG-3’
5	Mut pEGFP-C1-POC1A	Mutant pEGFP-C1-POC1A cDNA	028 FWD: 5’-TCTATATCATGGCCGACAAGC-3’	028 REV: 5’-TGCTGGTTCTCCAGACACTG-3’
6	cDNA positive control	HeLa control cDNA	5’-GCTCGTCGTCGACAACGGCTC-3’	5’-CAAACATGATCTGGGTCATCATCTTCTC-3’
7	cDNA negative control	HeLa control cDNA	5’-GCTCGTCGTCGACAACGGCTC-3’	5’-CAAACATGATCTGGGTCATCATCTTCTC-3’

Supplemental Figures

Figure S2-1. Schematic of POC1A minigene construct, pEGFP-C1-POC1A. *POC1A* exons 8-11 are inserted at C-terminus of eGFP. 300 base pairs of 5' and 3' intronic sequences separate each exon. Location of NM_015426.5:c.981+5G>C is indicated with red star. Placement of primers used in RT-PCR analysis are show as half arrowheads.

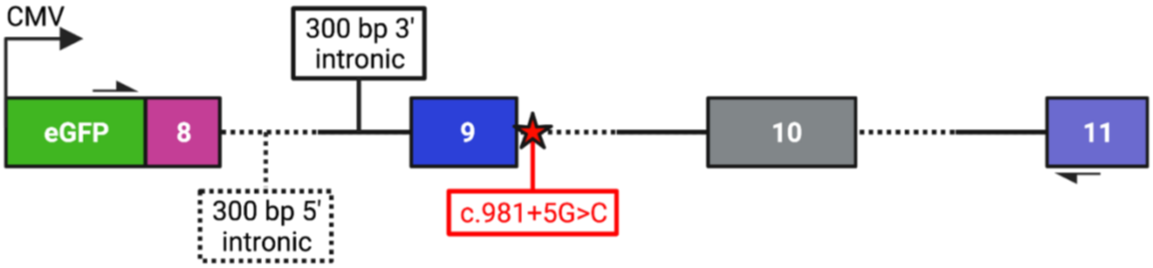
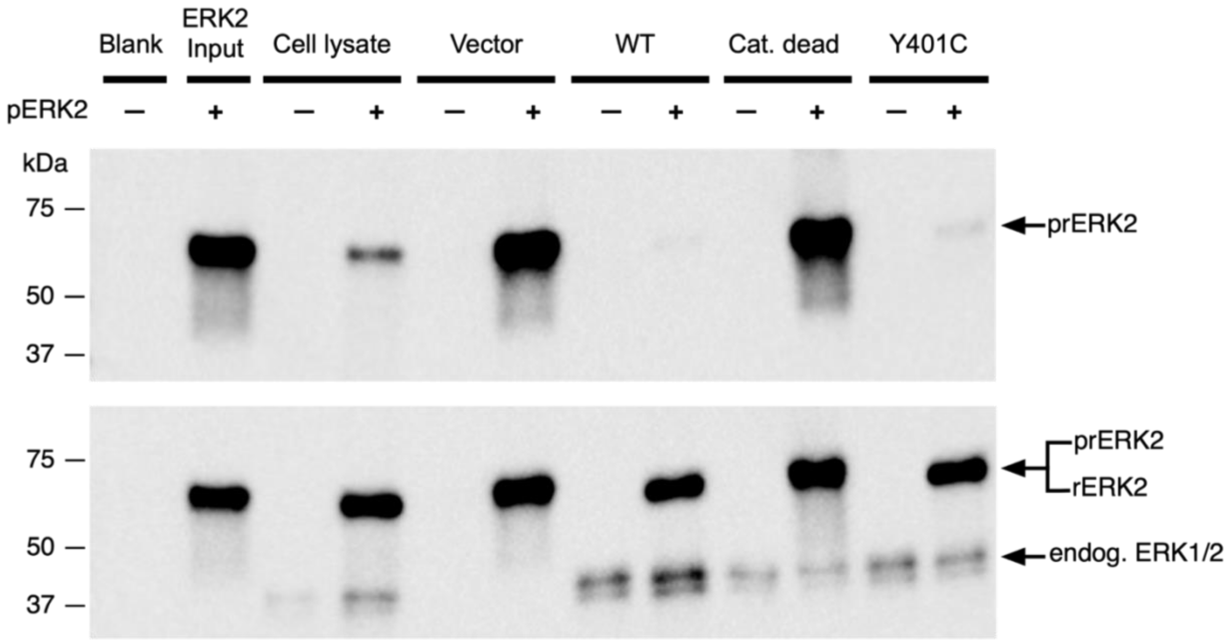


Figure S2-2. DUSP7 in vitro phosphatase assay. In vitro ERK2 phosphatase assay. DUSP7's canonical substrate, prERK2, was incubated alone, with cell lysate from non-transfected cells, eluate from cells transfected with empty vector, purified wildtype DUSP7 protein, purified catalytically dead DUSP7 protein, or with Y401C DUSP7 protein. Samples from assay were blotted with an antibody against prERK2 (upper) or an antibody against total ERK (lower). Absence of a prERK2 band suggest phosphatase activity is present. Y401C maintains wildtype phosphatase activity towards ERK2. Presence of endogenous ERK1/2 bands indicates Y401C DUSP7 maintains binding activity towards ERK2 similar to wildtype DUSP7.



References

1. Alkuraya FS. Primordial dwarfism: an update. Curr Opin Endocrinol Diabetes Obes. 2015;22(1):55-64. Epub 2014/12/10. doi: 10.1097/MED.0000000000000121. PubMed PMID: 25490023.

2. Khetarpal P, Das S, Panigrahi I, Munshi A. Primordial dwarfism: overview of clinical and genetic aspects. Mol Genet Genomics. 2016;291(1):1-15. Epub 2015/09/02. doi: 10.1007/s00438-015-1110-y. PubMed PMID: 26323792.

3. Al-Dosari MS, Shaheen R, Colak D, Alkuraya FS. Novel CENPJ mutation causes Seckel syndrome. J Med Genet. 2010;47(6):411-4. Epub 2010/06/05. doi: 10.1136/jmg.2009.076646. PubMed PMID: 20522431.

4. Koparir A, Karatas OF, Yuceturk B, Yuksel B, Bayrak AO, Gerdan OF, et al. Novel POC1A mutation in primordial dwarfism reveals new insights for centriole biogenesis. Hum Mol Genet. 2015;24(19):5378-87. Epub 2015/07/15. doi: 10.1093/hmg/ddv261. PubMed PMID: 26162852.

5. Shaheen R, Faqeih E, Ansari S, Abdel-Salam G, Al-Hassnan ZN, Al-Shidi T, et al. Genomic analysis of primordial dwarfism reveals novel disease genes. Genome Res. 2014;24(2):291-9. Epub 2014/01/07. doi: 10.1101/gr.160572.113. PubMed PMID: 24389050; PMCID: PMC3912419.

6. Griffith E, Walker S, Martin CA, Vagnarelli P, Stiff T, Vernay B, et al. Mutations in pericentrin cause Seckel syndrome with defective ATR-dependent DNA damage signaling. Nat Genet. 2008;40(2):232-6. Epub 2007/12/25. doi: 10.1038/ng.2007.80. PubMed PMID: 18157127; PMCID: PMC2397541.

7. Kalay E, Yigit G, Aslan Y, Brown KE, Pohl E, Bicknell LS, et al. CEP152 is a genome maintenance protein disrupted in Seckel syndrome. Nat Genet. 2011;43(1):23-6. Epub 2010/12/07. doi: 10.1038/ng.725. PubMed PMID: 21131973; PMCID: PMC3430850.

8. O'Driscoll M, Ruiz-Perez VL, Woods CG, Jeggo PA, Goodship JA. A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome. Nat Genet. 2003;33(4):497-501. Epub 2003/03/18. doi: 10.1038/ng1129. PubMed PMID: 12640452.

9. Ogi T, Walker S, Stiff T, Hobson E, Limsirichaikul S, Carpenter G, et al. Identification of the first ATRIP-deficient patient and novel mutations in ATR define a clinical spectrum for ATR-ATRIP Seckel Syndrome. PLoS Genet. 2012;8(11):e1002945. Epub 2012/11/13. doi: 10.1371/journal.pgen.1002945. PubMed PMID: 23144622; PMCID: PMC3493446.

10. Rauch A. The shortest of the short: pericentrin mutations and beyond. Best Pract Res Clin Endocrinol Metab. 2011;25(1):125-30. Epub 2011/03/15. doi: 10.1016/j.beem.2010.10.015. PubMed PMID: 21396579.

11. Shaheen R, Faqeih E, Shamseldin HE, Noche RR, Sunker A, Alshammari MJ, et al. POC1A truncation mutation causes a ciliopathy in humans characterized by primordial dwarfism. Am J Hum Genet. 2012;91(2):330-6. Epub 2012/07/31. doi: 10.1016/j.ajhg.2012.05.025. PubMed PMID: 22840364; PMCID: PMC3415549.

12. Klingseisen A, Jackson AP. Mechanisms and pathways of growth failure in primordial dwarfism. Genes Dev. 2011;25(19):2011-24. Epub 2011/10/08. doi: 10.1101/gad.169037. PubMed PMID: 21979914; PMCID: PMC3197200.

13. Chen JH, Segni M, Payne F, Huang-Doran I, Sleigh A, Adams C, et al. Truncation of POC1A associated with short stature and extreme insulin resistance. J Mol Endocrinol. 2015;55(2):147-58. Epub 2015/09/04. doi: 10.1530/JME-15-0090. PubMed PMID: 26336158; PMCID: PMC4722288.

14. Giorgio E, Rubino E, Bruselles A, Pizzi S, Rainero I, Duca S, et al. A syndromic extreme insulin resistance caused by biallelic POC1A mutations in exon 10. Eur J Endocrinol. 2017;177(5):K21-K7. Epub 2017/08/19. doi: 10.1530/EJE-17-0431. PubMed PMID: 28819016.

15. Majore S, Agolini E, Micale L, Pascolini G, Zuppi P, Cocciadiferro D, et al. Clinical presentation and molecular characterization of a novel patient with variant POC1A-related syndrome. Clin Genet. 2021;99(4):540-6. Epub 2020/12/30. doi: 10.1111/cge.13911. PubMed PMID: 33372278.

16. Al-Kindi A, Al-Shehhi M, Westenberger A, Beetz C, Scott P, Brandau O, et al. A novel POC1A variant in an alternatively spliced exon causes classic SOFT syndrome: clinical presentation of seven patients. J Hum Genet. 2020;65(2):193-7. Epub 2019/11/27. doi: 10.1038/s10038-019-0693-2. PubMed PMID: 31767933.

17. Barraza-Garcia J, Ivan Rivera-Pedroza C, Salamanca L, Belinchon A, Lopez-Gonzalez V, Sentchordi-Montane L, et al. Two novel POC1A mutations in the primordial dwarfism, SOFT

syndrome: Clinical homogeneity but also unreported malformations. Am J Med Genet A. 2016;170A(1):210-6. Epub 2015/09/17. doi: 10.1002/ajmg.a.37393. PubMed PMID: 26374189.

18. Ko JM, Jung S, Seo J, Shin CH, Cheong HI, Choi M, et al. SOFT syndrome caused by compound heterozygous mutations of POC1A and its skeletal manifestation. J Hum Genet. 2016;61(6):561-4. Epub 2016/01/23. doi: 10.1038/jhg.2015.174. PubMed PMID: 26791357.

19. Mostofizadeh N, Gheidarloo M, Hashemipour M, Dehkordi EH. SOFT Syndrome: The First Case in Iran. Adv Biomed Res. 2018;7:128. Epub 2018/10/13. doi: 10.4103/abr.abr_13_18. PubMed PMID: 30310776; PMCID: PMC6159314.

20. Saida K, Silva S, Solar B, Fujita A, Hamanaka K, Mitsuhashi S, et al. SOFT syndrome in a patient from Chile. Am J Med Genet A. 2019;179(3):338-40. Epub 2018/12/21. doi: 10.1002/ajmg.a.61015. PubMed PMID: 30569574.

21. Sarig O, Nahum S, Rapaport D, Ishida-Yamamoto A, Fuchs-Telem D, Qiaoli L, et al. Short stature, onychodysplasia, facial dysmorphism, and hypotrichosis syndrome is caused by a POC1A mutation. Am J Hum Genet. 2012;91(2):337-42. Epub 2012/07/31. doi: 10.1016/j.ajhg.2012.06.003. PubMed PMID: 22840363; PMCID: PMC3415554.

22. Shalev SA, Spiegel R, Borochowitz ZU. A distinctive autosomal recessive syndrome of severe disproportionate short stature with short long bones, brachydactyly, and hypotrichosis in two consanguineous Arab families. Eur J Med Genet. 2012;55(4):256-64. Epub 2012/03/24. doi: 10.1016/j.ejmg.2012.02.011. PubMed PMID: 22440536.

23. Simon A. FastQC: a quality control tool for high throughput sequence data. 2010.

24. Johnston HR, Chopra P, Wingo TS, Patel V, International Consortium on B, Behavior in 22q11.2 Deletion S, et al. PEMapper and PECaller provide a simplified approach to whole-genome sequencing. Proc Natl Acad Sci U S A. 2017;114(10):E1923-E32. Epub 2017/02/23. doi: 10.1073/pnas.1618065114. PubMed PMID: 28223510; PMCID: PMC5347547.

25. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. Epub 2015/02/28. doi: 10.1186/s13742-015-0047-8. PubMed PMID: 25722852; PMCID: PMC4342193.

26. Team RC. R: A Language and Environment for Statistical Computing. Vienna, Austria2014.

27. Kotlar AV, Trevino CE, Zwick ME, Cutler DJ, Wingo TS. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. Genome Biol. 2018;19(1):14. Epub 2018/02/08. doi: 10.1186/s13059-018-1387-3. PubMed PMID: 29409527; PMCID: PMC5801807.

28. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93. Epub 2011/09/10. doi: 10.1093/bioinformatics/btr509. PubMed PMID: 21903627; PMCID: PMC3198575.

29. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009;37(9):e67. Epub 2009/04/03. doi: 10.1093/nar/gkp215. PubMed PMID: 19339519; PMCID: PMC2685110.

30. Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. J Comput Biol. 1997;4(3):311-23. Epub 1997/10/01. doi: 10.1089/cmb.1997.4.311. PubMed PMID: 9278062.

31. Sahashi K, Masuda A, Matsuura T, Shinmi J, Zhang Z, Takeshima Y, et al. In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5' splice sites. Nucleic Acids Res. 2007;35(18):5995-6003. Epub 2007/08/30. doi: 10.1093/nar/gkm647. PubMed PMID: 17726045; PMCID: PMC2094079.

32. Wang M, Marin A. Characterization and prediction of alternative splice sites. Gene. 2006;366(2):219-27. Epub 2005/10/18. doi: 10.1016/j.gene.2005.07.015. PubMed PMID: 16226402.

33. Eldomery MK, Coban-Akdemir Z, Harel T, Rosenfeld JA, Gambin T, Stray-Pedersen A, et al. Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med. 2017;9(1):26. Epub 2017/03/23. doi: 10.1186/s13073-017-0412-6. PubMed PMID: 28327206; PMCID: PMC5361813.

34. Posey JE, Harel T, Liu P, Rosenfeld JA, James RA, Coban Akdemir ZH, et al. Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. N Engl J Med. 2017;376(1):21-31. Epub 2016/12/14. doi: 10.1056/NEJMoa1516767. PubMed PMID: 27959697; PMCID: PMC5335876.

35. Caunt CJ, Keyse SM. Dual-specificity MAP kinase phosphatases (MKPs): shaping the outcome of MAP kinase signalling. FEBS J. 2013;280(2):489-504. Epub 2012/07/21. doi: 10.1111/j.1742-4658.2012.08716.x. PubMed PMID: 22812510; PMCID: PMC3594966.

36. Orlev LN, Ehud B, Tamar BG, Orit SA, Yoel K, Witz IP. Does the dual-specificity MAPK phosphatase Pyst2-L lead a monogamous relationship with the Erk2 protein? Immunol Lett. 2004;92(1-2):149-56. Epub 2004/04/15. doi: 10.1016/j.imlet.2003.11.024. PubMed PMID: 15081539.

37. Guntur AR, Rosen CJ. IGF-1 regulation of key signaling pathways in bone. Bonekey Rep. 2013;2:437. Epub 2014/01/15. doi: 10.1038/bonekey.2013.171. PubMed PMID: 24422135; PMCID: PMC3818534.

38. Owens DM, Keyse SM. Differential regulation of MAP kinase signalling by dual-specificity protein phosphatases. Oncogene. 2007;26(22):3203-13. Epub 2007/05/15. doi: 10.1038/sj.onc.1210412. PubMed PMID: 17496916.

39. Pasquali C, Curchod ML, Walchli S, Espanel X, Guerrier M, Arigoni F, et al. Identification of protein tyrosine phosphatases with specificity for the ligand-activated growth hormone receptor. Mol Endocrinol. 2003;17(11):2228-39. Epub 2003/08/09. doi: 10.1210/me.2003-0011. PubMed PMID: 12907755.

40. Attanasio C, David A, Neerman-Arbez M. Outcome of donor splice site mutations accounting for congenital afibrinogenemia reflects order of intron removal in the fibrinogen alpha gene (FGA). Blood. 2003;101(5):1851-6. Epub 2002/10/31. doi: 10.1182/blood-2002-03-0853. PubMed PMID: 12406899.

41. Singh RK, Cooper TA. Pre-mRNA splicing in disease and therapeutics. Trends Mol Med. 2012;18(8):472-82. Epub 2012/07/24. doi: 10.1016/j.molmed.2012.06.006. PubMed PMID: 22819011; PMCID: PMC3411911.

42. Pearson CG, Osborn DP, Giddings TH, Jr., Beales PL, Winey M. Basal body stability and ciliogenesis requires the conserved component Poc1. J Cell Biol. 2009;187(6):905-20. Epub 2009/12/17. doi: 10.1083/jcb.200908019. PubMed PMID: 20008567; PMCID: PMC2806327.

43. Wang Y, Zhang X, Zhang H, Lu Y, Huang H, Dong X, et al. Coiled-coil networking shapes cell molecular machinery. Mol Biol Cell. 2012;23(19):3911-22. Epub 2012/08/10. doi: 10.1091/mbc.E12-05-0396. PubMed PMID: 22875988; PMCID: PMC3459866.

44. Xu C, Min J. Structure and function of WD40 domain proteins. Protein Cell. 2011;2(3):202-14. Epub 2011/04/07. doi: 10.1007/s13238-011-1018-1. PubMed PMID: 21468892; PMCID: PMC4875305.

45. Cha KB, Karolyi IJ, Hunt A, Wenglikowski AM, Wilkinson JE, Dolan DF, et al. Skeletal dysplasia and male infertility locus on mouse chromosome 9. Genomics. 2004;83(6):951-60. Epub 2004/06/05. doi: 10.1016/j.ygeno.2003.12.020. PubMed PMID: 15177549.

46. Geister KA, Brinkmeier ML, Cheung LY, Wendt J, Oatley MJ, Burgess DL, et al. LINE-1 Mediated Insertion into Poc1a (Protein of Centriole 1 A) Causes Growth Insufficiency and Male Infertility in Mice. PLoS Genet. 2015;11(10):e1005569. Epub 2015/10/27. doi: 10.1371/journal.pgen.1005569. PubMed PMID: 26496357; PMCID: PMC4619696.

47. D'Souza I, Poorkaj P, Hong M, Nochlin D, Lee VM, Bird TD, et al. Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements. Proc Natl Acad Sci U S A. 1999;96(10):5598-603. Epub 1999/05/13. doi: 10.1073/pnas.96.10.5598. PubMed PMID: 10318930; PMCID: PMC21906.

48. Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev. 2003;17(4):419-37. Epub 2003/02/26. doi: 10.1101/gad.1048803. PubMed PMID: 12600935.

49. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5. Epub 2013/05/30. doi: 10.1038/ng.2653. PubMed PMID: 23715323; PMCID: PMC4010069.

50. Hames RS, Hames R, Prosser SL, Euteneuer U, Lopes CA, Moore W, et al. Pix1 and Pix2 are novel WD40 microtubule-associated proteins that colocalize with mitochondria in Xenopus germ plasm and centrosomes in human cells. Exp Cell Res. 2008;314(3):574-89. Epub 2007/12/11. doi: 10.1016/j.yexcr.2007.10.019. PubMed PMID: 18068700.

51. Keller LC, Geimer S, Romijn E, Yates J, 3rd, Zamora I, Marshall WF. Molecular architecture of the centriole proteome: the conserved WD40 domain protein POC1 is required for centriole duplication and length control. Mol Biol Cell. 2009;20(4):1150-66. Epub 2008/12/26. doi: 10.1091/mbc.E08-06-0619. PubMed PMID: 19109428; PMCID: PMC2642750.

52. Doxsey S, Zimmerman W, Mikule K. Centrosome control of the cell cycle. Trends Cell Biol. 2005;15(6):303-11. Epub 2005/06/15. doi: 10.1016/j.tcb.2005.04.008. PubMed PMID: 15953548.

53. Gupta GD, Coyaud E, Goncalves J, Mojarad BA, Liu Y, Wu Q, et al. A Dynamic Protein Interaction Landscape of the Human Centrosome-Cilium Interface. Cell. 2015;163(6):1484-99. Epub 2015/12/08. doi: 10.1016/j.cell.2015.10.065. PubMed PMID: 26638075; PMCID: PMC5089374.

54. Lee H, Song J, Jung JH, Ko HW. Primary cilia in energy balance signaling and metabolic disorder. BMB Rep. 2015;48(12):647-54. Epub 2015/11/06. doi: 10.5483/bmbrep.2015.48.12.229. PubMed PMID: 26538252; PMCID: PMC4791320.

55. Song DK, Choi JH, Kim MS. Primary Cilia as a Signaling Platform for Control of Energy Metabolism. Diabetes Metab J. 2018;42(2):117-27. Epub 2018/04/21. doi: 10.4093/dmj.2018.42.2.117. PubMed PMID: 29676541; PMCID: PMC5911514.

56. Benzler J, Andrews ZB, Pracht C, Stohr S, Shepherd PR, Grattan DR, et al. Hypothalamic WNT signalling is impaired during obesity and reinstated by leptin treatment in male mice. Endocrinology. 2013;154(12):4737-45. Epub 2013/10/10. doi: 10.1210/en.2013-1746. PubMed PMID: 24105484.

57. Karsenty G, Kronenberg HM, Settembre C. Genetic control of bone formation. Annu Rev Cell Dev Biol. 2009;25:629-48. Epub 2009/07/07. doi: 10.1146/annurev.cellbio.042308.113308. PubMed PMID: 19575648.

58. DiIorio P, Rittenhouse AR, Bortell R, Jurczyk A. Role of cilia in normal pancreas function and in diseased states. Birth Defects Res C Embryo Today. 2014;102(2):126-38. Epub 2014/05/28. doi: 10.1002/bdrc.21064. PubMed PMID: 24861006.

59. Marshall JD, Maffei P, Collin GB, Naggert JK. Alstrom syndrome: genetics and clinical overview. Curr Genomics. 2011;12(3):225-35. Epub 2011/11/02. doi: 10.2174/138920211795677912. PubMed PMID: 22043170; PMCID: PMC3137007.

60. Starks RD, Beyer AM, Guo DF, Boland L, Zhang Q, Sheffield VC, et al. Regulation of Insulin Receptor Trafficking by Bardet Biedl Syndrome Proteins. PLoS Genet. 2015;11(6):e1005311. Epub 2015/06/24. doi: 10.1371/journal.pgen.1005311. PubMed PMID: 26103456; PMCID: PMC4478011.

61. Huang-Doran I, Bicknell LS, Finucane FM, Rocha N, Porter KM, Tung YC, et al. Genetic defects in human pericentrin are associated with severe insulin resistance and diabetes. Diabetes. 2011;60(3):925-35. Epub 2011/01/29. doi: 10.2337/db10-1334. PubMed PMID: 21270239; PMCID: PMC3046854.

62. Jurczyk A, Gromley A, Redick S, San Agustin J, Witman G, Pazour GJ, et al. Pericentrin forms a complex with intraflagellar transport proteins and polycystin-2 and is required for primary cilia assembly. J Cell Biol. 2004;166(5):637-43. Epub 2004/09/01. doi: 10.1083/jcb.200405023. PubMed PMID: 15337773; PMCID: PMC2172416.

63. Monies D, Abouelhoda M, Assoum M, Moghrabi N, Rafiullah R, Almontashiri N, et al. Lessons Learned from Large-Scale, First-Tier Clinical Exome Sequencing in a Highly Consanguineous Population. Am J Hum Genet. 2019;104(6):1182-201. Epub 2019/05/28. doi: 10.1016/j.ajhg.2019.04.011. PubMed PMID: 31130284; PMCID: PMC6562004.

64. Shalev SA. Characteristics of genetic diseases in consanguineous populations in the genomic era: Lessons from Arab communities in North Israel. Clin Genet. 2019;95(1):3-9. Epub 2018/02/11. doi: 10.1111/cge.13231. PubMed PMID: 29427439.

65. Huang CY, Tan TH. DUSPs, to MAP kinases and beyond. Cell Biosci. 2012;2(1):24. Epub 2012/07/10. doi: 10.1186/2045-3701-2-24. PubMed PMID: 22769588; PMCID: PMC3406950.

66. Matsushita T, Chan YY, Kawanami A, Balmes G, Landreth GE, Murakami S. Extracellular signal-regulated kinase 1 (ERK1) and ERK2 play essential roles in osteoblast differentiation and in supporting osteoclastogenesis. Mol Cell Biol. 2009;29(21):5843-57. Epub 2009/09/10. doi: 10.1128/MCB.01549-08. PubMed PMID: 19737917; PMCID: PMC2772724.

67. Coyne ES, Bedard N, Wykes L, Stretch C, Jammoul S, Li S, et al. Knockout of USP19 Deubiquitinating Enzyme Prevents Muscle Wasting by Modulating Insulin and Glucocorticoid Signaling. Endocrinology. 2018;159(8):2966-77. Epub 2018/06/15. doi: 10.1210/en.2018-00290. PubMed PMID: 29901692.

68. Cooper TA. Use of minigene systems to dissect alternative splicing elements. Methods. 2005;37(4):331-40. Epub 2005/11/30. doi: 10.1016/j.ymeth.2005.07.015. PubMed PMID: 16314262.

69. Fraile-Bethencourt E, Diez-Gomez B, Velasquez-Zapata V, Acedo A, Sanz DJ, Velasco EA. Functional classification of DNA variants by hybrid minigenes: Identification of 30 spliceogenic variants of BRCA2 exons 17 and 18. PLoS Genet. 2017;13(3):e1006691. Epub 2017/03/25. doi: 10.1371/journal.pgen.1006691. PubMed PMID: 28339459; PMCID: PMC5384790.

70. Alvarez-Satta M, Castro-Sanchez S, Pousada G, Valverde D. Functional analysis by minigene assay of putative splicing variants found in Bardet-Biedl syndrome patients. J Cell Mol Med. 2017;21(10):2268-75. Epub 2017/05/16. doi: 10.1111/jcmm.13147. PubMed PMID: 28502102; PMCID: PMC5618670.

71. Acedo A, Hernandez-Moro C, Curiel-Garcia A, Diez-Gomez B, Velasco EA. Functional classification of BRCA2 DNA variants by splicing assays in a large minigene with 9 exons. Hum Mutat. 2015;36(2):210-21. Epub 2014/11/11. doi: 10.1002/humu.22725. PubMed PMID: 25382762; PMCID: PMC4371643.

72. Fraile-Bethencourt E, Valenzuela-Palomo A, Diez-Gomez B, Caloca MJ, Gomez-Barrero S, Velasco EA. Minigene Splicing Assays Identify 12 Spliceogenic Variants of BRCA2 Exons 14 and 15. Front Genet. 2019;10:503. Epub 2019/06/14. doi: 10.3389/fgene.2019.00503. PubMed PMID: 31191615; PMCID: PMC6546720.

73. Gaildrat P, Killian A, Martins A, Tournier I, Frébourg T, Tosi M. Use of Splicing Reporter Minigene Assay to Evaluate the Effect on Splicing of Unclassified Genetic Variants. JAMA2010. p. 249-57.

74. Yamamura T, Nozu K, Miyoshi Y, Nakanishi K, Fujimura J, Horinouchi T, et al. An in vitro splicing assay reveals the pathogenicity of a novel intronic variant in ATP6V0A4 for autosomal recessive distal renal tubular acidosis. BMC Nephrol. 2017;18(1):353. Epub 2017/12/06. doi: 10.1186/s12882-017-0774-4. PubMed PMID: 29202719; PMCID: PMC5716019.

75. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43. Epub 2020/05/29. doi: 10.1038/s41586-020-2308-7. PubMed PMID: 32461654; PMCID: PMC7334197.

CHAPTER III. Sex-specific recombination patterns predict parent of origin for recurrent genomic disorders

Trenell J. Mosley, H. Richard Johnston, David J. Cutler, Michael E. Zwick, Jennifer G. Mulle

Adapted from an original publication under review as cited:
Mosley, T. J., Johnston, H. R., Cutler D. J., Zwick, M. E., Mulle, J. G. Sex-specific recombination predicts parent of origin for recurrent genomic disorders. BMC Med. Genomics.

Introduction

Genomic disorders are caused by pathological structural variation in the human genome usually arising de novo during parental meiosis [1-4]. The most common pathogenic variety of these rearrangements are copy number variants (CNVs), *i.e.*, a deletion or duplication of > 1 kb of genetic material [3,5,6]. The clinical phenotypes of genomic disorders are varied. They include congenital dysmorphisms, neurodevelopmental, neurodegenerative and neuropsychiatric manifestations, and even more common complex phenotypes such as obesity and hypertension [7-12]. CNVs have been observed in 10% of sporadic cases of autism [13,14], 15% of schizophrenia cases [15,16], and 16% of cases of intellectual disability [17]. These and other associations highlight the importance of structural variation to human health and the need to understand the factors influencing how they arise.

There is an intense interest in understanding the mechanisms by which CNVs form [18,19]. In several regions of the genome, de novo CNVs with approximately the same breakpoints recur in independent meioses (recurrent CNVs) [1,20]. The presence of segmental duplications flanking these intervals is a hallmark feature of recurrent CNVs. It is hypothesized that misalignment and subsequent recombination between non-allelic low copy repeat (LCR) segments within the segmental duplication regions is the formative event giving rise to the CNV[21,22], so called non-allelic homologous recombination (NAHR). Risk factors that may favor NAHR have been investigated and include sequence composition and orientation of the LCRs themselves [21,23] as well as the presence of inversions at the locus [24,25].

Parental sex bias for the origin of recurrent de novo CNVs remains unexplained. De novo deletions at the 16p11.2 and 17q11.2 loci are more likely to arise on maternally inherited chromosomes [26-29]. Deletions at the 22q11.2 locus show a slight maternal bias as well [30]. In contrast, deletions at the 5q35.3 locus (Sotos syndrome [MIM: 117550]) display a paternal origin bias [31,32]. Deletions at the 7q11.23 locus (Williams syndrome [MIM: 194050]) do not show a bias in parental origin [24]. While it has been suggested that sex-specific recombination rates might influence sex biases in NAHR [26], this hypothesis has not been formally tested.

The majority of recurrent CNVs are thought to form during meiosis, when homologous chromosomes align and synapse during prophase I [33]. It is well established that meiosis differs significantly between males and females. In males, spermatagonia continuously divide and complete meiosis throughout postpubescent life with all four products of meiosis resulting in gametes. In contrast, in human females oogonia are established in fetal life and enter into an extended period of prolonged stasis in prophase I of meiosis until they complete meiosis upon ovulation and fertilization [34]. Additionally, in female meiosis, only one of four products of meiosis result in a gamete. Sexual dymorphism in meiosis extends to the patterns and processes of recombination during meiosis [33]. Here we seek to ask whether local sex-specific rates in meiotic recombination can predict the parental bias for the origin of recurrent de novo CNVs.

Methods

Parent of Origin Determination

Literature Search and Parental Origin Data Curation

For this analysis, we considered the 55 known genomic disorder CNV loci described in Coe et al., 2014 [7]. A locus was eligible for inclusion in the current analysis if it is flanked by LCRs, i.e., mediated by NAHR, and not imprinted (n = 38 eligible loci). For each of these 38 loci, we performed a systematic PubMed search to identify published data on parental origin. Studies were admitted to this paper’s analysis when the following criteria were met: (1) the study detailed parent of origin data for one of the 38 eligible NAHR-mediated loci as designated by Coe et al., 2014[7], (2) the authors of the study interrogated the entire canonical CNV interval to confirm presence of a deletion or duplication in the patients, (3) the authors determined the investigated CNVs were de novo, and (4) the study clearly treated monozygotic twins as one meiotic event and not two (Supplemental Methods; Table S3-1; Table S3-2). The literature search led to a manual review of 1,268 papers, out of which we identified 77 manuscripts across 24 loci with suitable data for analysis: 1q21.1 [35-39], 1q21.1 TAR [40], 2q13 [37], 3q29 [37-42], 5q35 [31,32], 7q11.23 [24,40,43-54], 8p23.1 [55,56], 11q13.2q13.4 [57], 15q13.3 [38,40,58], 15q24 (AC, AD, BD, and BE intervals) [59-64], 15q25.2 [65-67], 16p11.2 [26,37,40,68-70], distal 16p11.2 [37,38,70], 16p11.2p12.1 [71], 16p31.11 [37], 17p11.2 [72-76], 17q11.2 [28,29,77], 17q12 [37,38,78], 17q21.31 [19,25,67,79-84], 17q23.1q23.2 [69,85] and 22q11.2 [30,43,53,86-102] (Table 3-1). For the remaining 14 loci, no published parent of origin data could be identified. At the 3q29 locus we generated new data to determine the parent of origin for de novo events (http://genome.emory.edu/3q29/).

Determination of Parental Origin for 3q29 Deletion

Study Subject Recruitment: This study was approved by Emory University's Institutional Review Board (IRB00064133). Individuals with a clinically confirmed diagnosis of 3q29 deletion were ascertained through the internet-based 3q29 registry (https://3q29deletion.patientcrossroads.org/) as previously described [103]. Blood samples were obtained from 14 families. SNP genotyping was performed on 12 of the 14 families (10 full trios, 2 mother-child pairs) using the Illumina GSA-24 v 3.0 array. For 2 full trios (6 samples), parent of origin was determined from whole genome sequence data on Illumina's NovaSeq 6000 platform. Quality control was performed with PLINK 1.9 [104] and our custom pipeline (Supplemental Methods).

Parental Origin Analysis: Parental origin of the 3q29 deletion was determined for all 14 families using PLINK 1.9 [104]. SNPs located within the 3q29 deletion region (chr3:196029182-197617792; hg38) were isolated for analysis and the pattern of Mendelian errors (MEs) were analyzed. The parent with the most MEs was considered the parent of origin for the 3q29 deletion (Supplemental Methods). The mean age of fathers in our 3q29 cohort was collected from self-reported data in conjunction with the Emory University 3q29 project (http://genome.emory.edu/3q29/) and compared to the U.S. average (NCHS; https://www.cdc.gov/nchs/index.htm) via a two-tailed two-sample t-test using R [105].

Calculation of Recombination Rates and Ratios

Chromosome male and female recombination rates (cM/Mb) were obtained from the deCODE sex-specific maps, which are based on over 4.5 million crossover recombination events from 126,427 meioses, with an average resolution of 682 base pairs [106]. The recombination rate (cM/Mb) data from deCODE is publicly available as recombination rates calculated for a physical genomic interval bounded by two SNP markers (Supplemental Methods). Therefore, for our calculation of the average male and female recombination rates, each bounded recombination rate was weighted by the total number of base pairs contained within the respective SNP marker interval. Weighted rates were then averaged across the CNV interval for males and females, separately. The ratio of the weighted average male and female recombination rates was then calculated for each CNV interval by dividing the weighted average male recombination rate by the weighted average female recombination rate (Figure S3-1). To account for slight differences in the recombination rate ratios calculated for the different LCR22 intervals at the 22q11.2 locus we used an adjusted recombination rate ratio composed of the weighted recombination rate ratios calculated for each LCR22 interval. Weights were based on the estimated population prevalence of the different 22q11.2 deletion intervals (Table S3-3) [107].

Logistic Regression Analysis

Parental origin data was curated for CNVs at the 24 CNV loci from 77 independent studies; only independent samples were included in the analysis (duplicate or overlapping samples were removed). For each CNV locus the male to female recombination rate ratio was calculated as described above. A logistic regression model was fitted to the data with the loge-transformed male to female recombination rate ratio as the predictor and parental origin (paternal vs. maternal) as the response variable. We performed a secondary analysis stratified by deletions and duplications. See Table 3-2 and Table S3-4 for the data calculated and used in the logistic regression analyses.

Linear Regression Analysis

For linear regression, locus-specific estimates for parental origin were derived by combining the data from all published studies for a given locus. To alleviate the uncertainty introduced by small sample sizes, only those loci with more than 10 observations were included. The loge-transformed combined male to female parental origin count ratio for each locus was regressed on the calculated average loge-transformed average male to female recombination rate ratio for that locus' CNV interval. Each locus was weighted based on its sample size.

Results

Recurrent Genomic Disorder Loci Literature Search

We conducted a systematic literature search for the 38 non-imprinted and NAHR-mediated CNV loci in Coe et al., 2014 [7] (Table 3-1; Table S3-1). We identified parent-of-origin studies that met inclusion criteria as stated in Methods. 77 studies met inclusion criteria; from these 77 studies, data were curated for 24 loci, including copy number variants at 1q21.1 [35-39], 1q21.1 TAR [40], 2q13 [37], 3q29 [37-42], 5q35 [31,32], 7q11.23 [24,40,43-54], 8p23.1 [55,56], 11q13.2q13.4 [57], 15q13.3 [38,40,58], 15q24 (AC, AD, BD, and BE intervals) [59-64], 15q25.2 [65-67], 16p11.2 [26,37,40,68-70], distal 16p11.2 [37,38,70], 16p11.2p12.1 [71], 16p31.11 [37], 17p11.2 [72-76], 17q11.2 [28,29,77], 17q12 [37,38,78], 17q21.31 [19,25,67,79-84], 17q23.1q23.2 [69,85] and 22q11.2 [30,43,53,86-102] (Table 3-2). Each locus has between one and twenty independent studies representing in total 1,977 de novo deletion (N = 1,913) and duplication (N = 64) events (Table 3-2).

Parent of Origin of 3q29 Deletion

We determined parent of origin in 12 full trios where a proband had a de novo 3q29 deletion; in 2 additional trios where only proband and maternal DNA samples were available, parent of origin was inferred. For the 12 trios evaluated by SNP arrays, in all cases, the number of Mendelian errors between the presumed inherited (intact) parental allele was zero, and the mean Mendelian errors for the presumed de novo parent of origin allele were 41, with a range of 27-66. For the two trios evaluated with sequence data, Mendelian errors were 20-33-fold elevated when comparing the inherited vs de novo parent. In these 14 trios, 13 deletions (92.9%) arose on the paternal genome indicating a significant departure from the null expectation of 50%

(p = 0.002, binomial exact). When accounting for only full trios, 11 of 12 (91.7%) deletions arose on paternal haplotypes (p = 0.006, binomial exact), altogether indicating there is a paternal bias for origin of the 3q29 deletion (Table S3-4). We examined the age distribution of male parents in our cohort; the mean age is 34 years (median = 34 years) and is not significantly different from the 2018 U.S national average, (31.8 years) (p = 0.08, Two-tailed two sample t-test), These data indicate the bias in the 3q29 sample is unlikely to be due to oversampling of older fathers (Table S3-5).

Meiotic Recombination and Parental Origin

We tested the hypothesis that sex-dependent differences in meiotic recombination could explain the parental biases observed for recurrent genomic disorder loci mediated by NAHR. We determined the male and female origin counts of the CNVs curated from the literature search. Of the 1,977 CNVs, 870 were paternal in origin and 1,107 were of maternal origin. We calculated the average male and female recombination rates (cM/Mb) across the CNV intervals at all 24 loci using recombination rates published by the deCODE genetics group [106] (Figure S3-2 to S3-12). We fit a simple logistic model to the data, with the male-to-female recombination rate ratio as the predictor and parental origin as the response variable (Table 3-2; Table S3-4). Our data reveal that the sex-dependent recombination rate ratio significantly predicts parental de novo origin of a given CNV ($p = 1.07 \times 10^{-14}$, $\beta = 0.6606$, $CI_{95\%} = (0.4980,0.8333)$, OR = 1.936) (Figure 3-1). In other words: for a given region, the higher the male recombination rate is relative to the female rate, the more likely a CNV formed in that region will be paternal in origin. Stratified analyses on deletions and duplications separately lead to a nearly identical model. (Deletions: $p = 8.88 \times 10^{-14}$, $\beta = 0.6721$, $CI_{95\%} = (0.5009, 0.8546)$, OR = 1.9584; Duplications: p

$= 0.02$, $\beta = 0.8304$, $CI_{95\%} = (0.1508, 1.6017)$, OR = 2.2942) (Figure S3-13 to S3-14; Table S3-6 to S3-7). Simple linear regression on the subset of CNV loci with more than 10 samples, shows the striking correlation between relative recombination rates and parental origin, where relative recombination rates explain 85% of the variance in parental bias (Figure S3-15; Table S3-8). Our logistic model can be used to predict paternal origin rates for any locus with estimable recombination in males and females, and we have done so (Table S3-9). CNVs at the 15q13.3 and 17q23 both are predicted to have a paternal origin approximately 60% of the time, while at the 16p11.2 distal locus CNVs are predicted to have a maternal origin 76% of the time (Table S3-9). If correct, our model would predict these loci exhibit a bias in parental origin.

Discussion

Parent of origin bias for de novo events at recurrent CNV loci has been well-documented but has lacked a compelling explanation. Our analysis of data gathered on 1,977 CNVs from 77 published reports demonstrate that sex-specific variation in local meiotic recombination rates predicts parent of origin at recurrent CNV loci. Human male and female meiotic recombination rates and patterns differ greatly across the broad scale of human chromosomes. Recombination events are nearly uniformly distributed across the chromosome arms in females but tend to be clustered closer to the telomeres in males [108]. We note that this pattern has been previously recognized [26]. Here we have formally tested the hypothesis that recombination variation drives parent of origin variation using a rigorous, statistical framework (Figure 3-1) and provided an estimate for the variance in parent of origin bias that is due to sex-specific recombination rates (Figure S3-15).

Investigations into the mechanism by which recurrent CNVs arise have focused on LCRs and their makeup [1,109]. These regions are composed of units of sequence repeats that vary in orientation, percent homology, length, and copy number. Consequently, LCRs are mosaics of varying units, imparting complexity to LCR architecture [23]. The frequency of NAHR events mediated by LCRs is a function of these characteristics, and other features of the genomic architecture [21]. Specifically, the rate of NAHR is known to correlate positively with LCR length and percent homology and decrease as the distance between LCRs increases [18,21]. However, because LCRs are challenging to study with short-read sequencing technology, the population-level variability of these regions is not well described [110]. Recent breakthroughs with long-read sequencing and optical mapping have revealed remarkable variation in LCRs [111-113], and haplotypes with higher risks for CNV formation have now been identified [114]. LCRs are substrates for NAHR [1], and thus are subject to the recombination process. Local recombination rates may influence how likely an NAHR event will happen between two LCRs. Therefore, when analyzing LCR haplotypes and their susceptibility to NAHR, one would need to take into account sex-differences in recombination. For example, at loci with maternal biases, specific risk haplotypes may be required for males to form CNVs and vice versa. Greater enrichment of GC content, homologous core duplicons or the PRDM9 motifs, or other recombination-favoring factors may also be required [1,18]

Variation in recombination rates between sexes is well established [108,115-118]. Prediction of individual risk may also need to consider individual variation in meiotic recombination, particularly due to heritable variation and presence or absence of inversion polymorphisms [117,119]. Variants in several genes, including PRDM9, have been shown to affect recombination rates and the distribution of double-stranded breaks in mammals [120,121].

Common alleles in PRDM9 are evidenced to affect the percentage of recombination events within individuals that take place at hotspots [120], and variants in RNF212 are associated with opposite effects on recombination rate between males and females [116,121]. The unexplained variance in our study may be due to these additional factors, which are rich substrates for future study.

Many human genetic studies have observed correlations between inversion polymorphisms and genomic disorder loci [25,122]. Because these inversions are copy-number neutral and often located in complex repeat regions, [123] they can be difficult to assay with current high-throughput strategies and their true impact remains to be explored. One model proposes that during meiosis these regions may fail to synapse properly and increase the probability of NAHR [124,125]. Another theory suggests formation of inversions increase directly oriented content in LCRs leading to a NAHR-favorable haplotype [126]. Supporting these theories, inversion polymorphisms have been identified at the majority of recurrent CNV loci [24,25,30,122,124,126,127]. At the 7q11.23, 17q21.31, and 5q35 loci [24,25,127], compelling data indicts inversions as a highly associated marker of CNV formation. However, heterozygous inversions are known to suppress recombination perturbing the local pattern of recombination and altering the fate of chiasmata [119]. The analysis presented here strongly suggests that recombination is the driving force for CNV formation giving rise to an alternate explanation for the association between inversions and CNVs; they are both the consequence (and neither one the cause) of recombination between non-allelic homologous LCRs. Inversions and CNVs appear to be associated because both are being initiated by aberrant recombination. Viewing the system in this manner also explains the frequency of individual inversions at CNV loci. Inversions are arising via rare aberrant recombination, like CNVs, but subsequently being driven to higher frequency by natural selection, because they act to suppress recombination and "save offspring" from deleterious genomic disorders. Of course, frequent mutations leading to inversions and the details of LCR structure such as relative orientation and homology within a genomic region may promote or impede CNV formation in a locus-specific manner [128-130]. Further exploration of this relationship with improved genomic mapping can test these alternative models [131]. One testable prediction of the model described here is that inversions should be at higher frequency at loci giving rise to highly deleterious CNVs, as opposed to loci harboring recurrent benign CNVs.

To our knowledge, this study is the first comprehensive investigation of parental origin of recurrent, NAHR-mediated CNV loci. Investigations of predominantly nonrecurrent CNVs show paternal bias [132-134]. Unlike recurrent CNVs, nonrecurrent CNVs are mostly formed via non-homologous end joining (NHEJ) and replicative mechanisms [1,135,136]. The standing hypothesis is that replication-based mechanisms of nonrecurrent CNV formation, which are known to accumulate errors in male germlines, contribute to this bias [132]. Our study reinforces the idea that the factors influencing recurrent CNVs differ from those impacting nonrecurrent CNVs. Future genome-wide analyses with larger sample sizes can further help refine our understanding of the divergent forces at play affecting recurrent and nonrecurrent CNV formation.

We conducted a comprehensive literature search at 38 loci and ultimately identified 1,977 samples for analysis. We note that the majority of the data come from 7 well-studied loci (Table 3-1). While we thoroughly curated the data in a systematic way, it is possible that our data is subject to publication bias, where loci that exhibit parent of origin biases are more likely to have parental origin reported. Further exacerbating potential publication bias, genetic testing for the affected patient (and even more so for the parents) can be difficult to obtain due to concerns such as insurance coverage, potential future discrimination, and privacy concerns [137-140]. However, we note individuals with CNVs are generally not ascertained or recruited under the expectation that recombination affects parent of origin, and therefore, any potential publication or ascertainment bias is unlikely to confound the results of our analysis. Analysis of a larger cohort of CNV loci including benign CNVs will give greater insight into the role of recombination, and sex differences in recombination influencing parent of origin in CNVs.

Our estimates of recombination rates summarize CNV-scale (broad-scale) patterns of recombination, rather than fine-scale patterns near the sites of relevant recombination events that form these CNVs—LCRs. For example, local sex-specific hotspots within LCRs could be the underlying drivers behind the correlation between recombination rates and parental origin. Given the nature of repetitive regions like LCRs and our inability to adequately interrogate them with current sequencing technologies, accurate recombination data across and within the LCR regions is not available. In other words, the data is currently insufficient to conclude whether or not these broad scale patterns are tightly correlated with fine-scale recombination rates in and around the LCRs. The best available data in the field allows us to infer the following: broad scale patterns of recombination tightly predict patterns of parental origin.

Conclusions

In this study we determined male and female differences in meiotic recombination rates significantly predict parent of origin for recurrent CNV loci. Combining the sex-specific recombination landscape and the mechanistic factors underlying it with a more detailed understanding of existing structural factors at genomic disorder loci can be expected to help guide standards used to identify and perform genetic counseling for individuals at risk of genomic rearrangement.

Tables

Table 3-1: Summary of CNV loci included in literature search and curated studies

Locus	MIM Number	[a]# Studies Included	Study References
1q21.1 TAR	274000	1	[40]
1q21.1	612474/612475	5	[35-39]
2q11.2	—	0	—
2q11.2q13	—	0	—
2q13	—	1	[37]
3q29	609425/611936	7	[37-42], This Study
5q35	117550	2	[31,32]
7q11.23	194050/609757	14	[24,40,43-54]
7q11.23 distal	613729	0	—
7q11.23 proximal	—	0	—
8p23.1	—	2	[55,56]
10q23	612242	0	—
11q13.2q13.4	—	1	[57]
15q11.2	615656	0[b]	—
15q13.3	612001	3	[38,40,58]
[c]15q24	—	6	[59-64]
15q25.2	614294	3	[65-67]
15q25.2 (Cooper)	—	0	—
16p11.2	611913/614671	6	[26,37,40,68-70]
16p11.2 distal	613444	3	[37,38,70]
16p11.2p12.2	—	0	—
16p11.2p12.1	—	1	[71]
16p12.1	136570	0	—
16p13.11	—	1	[37]
17p11.2	182290/610883	5	[72-76]
[d]17p11.2p12	118220/162500	0	—
17q11.2	613675/618874	3	[28,29,77]
17q12	614526/614527	3	[37,38,78]
17q21.31	610443/613533	9	[19,25,67,79-84]]
17q23	—	0	—
17q23.1q23.2	613355/613618	2	[69,85]
22q11.2	188400/192430	20	[30,43,53,86-102]
22q11.2 distal	611867	0	—

[a]Independent studies from which the parent of origin data for the current analysis were obtained. Studies may be repeated between loci. [b]Recombination rates could not be calculated for 15q11.2 as the breakpoints were outside range of recombination maps. [c]15q24 locus is represented as 6 different

intervals in Coe et al., 2014 [8]. [d]17p11.2p12 excluded due to inconsistencies in mechanism of formation. See Supplemental Methods.

Table 3-2: Summary of genomic disorder loci CNVs recombination calculations

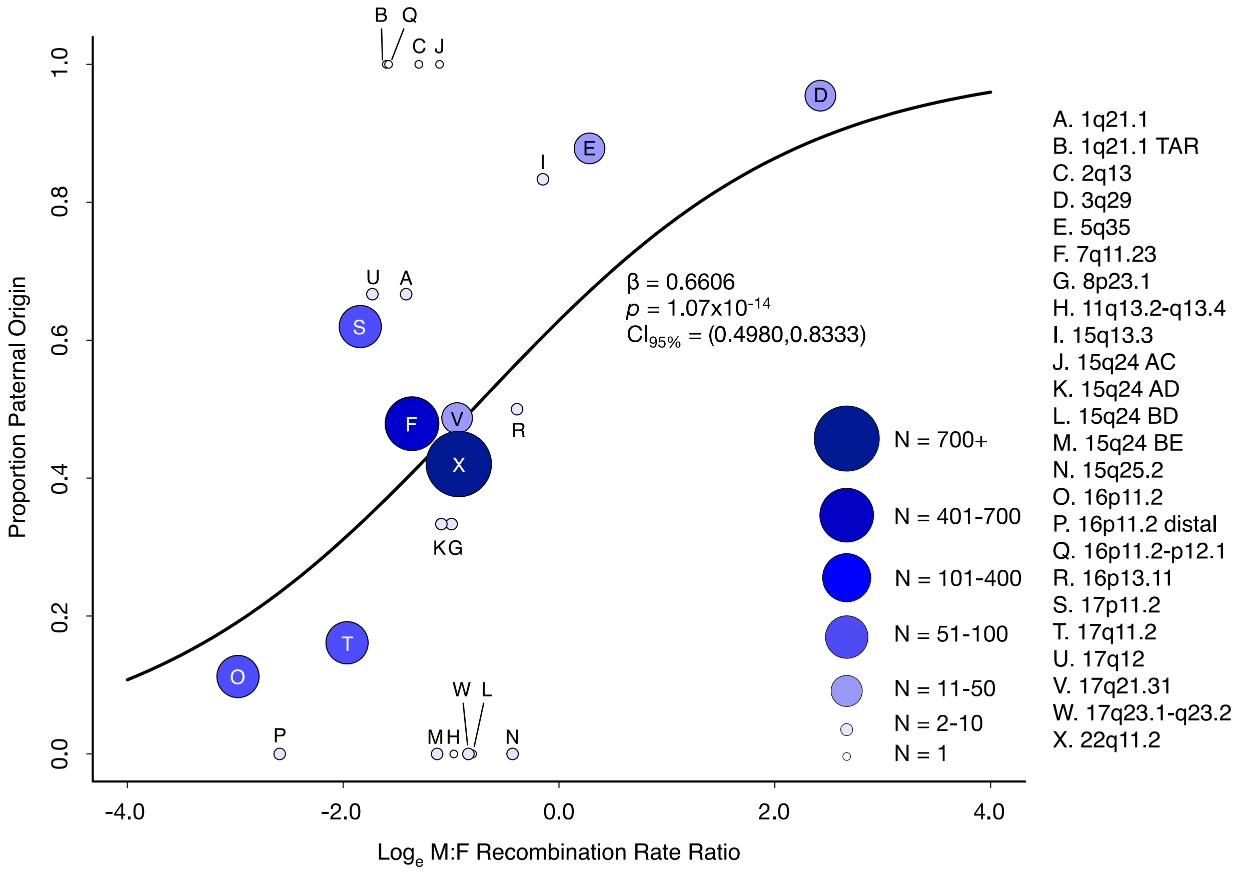
Locus	CNV Type	BED Coordinates[7]	# Samples (%)	[a]M:F Origin Counts	Del/Dup Counts	[b]Avg. Male Recomb. Rate [106]	[b]Avg. Female Recomb. Rate [106]	[c]Log$_e$ M:F Recomb. Ratio [106]
1q21.1 TAR	Del	chr1:145686999-146048495	1 (0.05%)	1:0	1/0	0.15712388	0.77814863	-1.599883
1q21.1	Del/Dup	chr1:147101794-147921262	9 (0.46%)	6:3	7/2	0.12331689	0.50839541	-1.416626
2q13	Dup	chr2:110625954-112335952	1 (0.05%)	1:0	0/1	0.44854539	1.64377881	-1.298743
3q29	Del	chr3:195988732-197628732	22 (1.11%)	21:1	22/0	3.1305211	0.27775988	2.422197
5q35	Del	chr5:176290391-177630393	41 (2.07%)	36:5	41/0	1.29955355	0.97941355	0.282822
7q11.23	Del/Dup	chr7:73328061-74727726	618 (31.26%)	296:322	598/20	0.49353554	1.92657023	-1.361890
8p23.1	Del/Dup	chr8:8235068-12035082	3 (0.15%)	1:2	1/2	0.67201752	1.81857951	-0.995527
11q13.2q13.4	Del	chr11:67985953-71571306	1 (0.05%)	0:1	1/0	0.8431765	2.23501635	-0.974828
15q13.3	Del/Dup	chr15:30840505-32190507	6 (0.30%)	5:1	5/1	1.63640726	1.89901039	-0.148828
15q24 AC	Del	chr15:72670606-75240606	1 (0.05%)	1:0	1/0	0.28479919	0.86129537	-1.106653
15q24 AD	Del	chr15:72670606-75720604	3 (0.15%)	1:2	3/0	0.27613544	0.82152961	-1.090277
15q24 BD	Del	chr15:73720606-75720604	1 (0.05%)	0:1	1/0	0.30739967	0.68432207	-0.800280
15q24 BE	Del	chr15:73720606-77840603	2 (0.10%)	0:2	2/0	0.23485125	0.72623183	-1.128917
15q25.2	Del	chr15:82513967-84070244	5 (0.25%)	0:5	5/0	0.21225081	0.32633295	-0.430177
16p11.2	Del/Dup	chr16:29641178-30191178	98 (4.96%)	11:87	79/19	0.06570935	1.28740565	-2.974798
16p11.2 distal	Del/Dup	chr16:28761178-29101178	4 (0.20%)	0:4	3/1	0.12150949	1.61662624	-2.600350
16p11.2p12.1	Dup	chr16:21341178-29431178	1 (0.05%)	1:0	0/1	0.5534655	2.68382469	-1.578799
16p13.11	Del/Dup	chr16:15408642-16198642	2 (0.10%)	1:1	1/1	1.67072378	2.46524529	-0.389038
17p11.2	Del/Dup	chr17:16805961-20576095	71 (3.59%)	44:27	59/12	0.1888066	1.19115966	-1.841959
17q11.2	Del	chr17:30838856-31888868	62 (3.14%)	10:52	62/0	0.26024285	1.85442774	-1.963716
17q12	Del	chr17:36460073-37846263	6 (0.30%	4:2	6/0	0.64750654	3.64754421	-1.728720
17q21.31	Del/Dup	chr17:45626851-46106851	39 (1.97%)	19:20	35/4	0.38234179	0.98304273	-0.944338
17q23.1q23.2	Del	chr17:59987857-62227857	2 (0.10%)	0:2	2/0	0.56466054	1.30765625	-0.839748
22q11.2	Del	[d]chr22:18924718-21111383	978 (49.47%)	411:567	978/0	1.45946494	3.69205976	-0.920692

All	**Del/Dup**	**—**	**1977 (100%)**	**870:1107**	**1913/64**	**—**	**—**	**—**

Summarized CNV data. Data are consolidated by locus. BED coordinates correspond to hg38 (LiftOver from hg18 coordinates in Coe et al., 2014). [a]Male to female CNV parent of origin counts. [b]Average male and female recombination rates are the average of the recombination rates calculated for each sample observed for the locus, e.g., 0.123331689 is the average male recombination rate calculated from the male recombination rates of the 9 1q21.1 CNVs. [c]Natural log-transformed average male to female recombination rate ratio for the locus. [d]Breakpoints cited by ClinGen for ~3.0 Mb LCR22A-LCR22D interval.

Figures

Figure 3-1. Recombination rates associate with parental origin. Predicted (curve) and observed paternal origin proportions for 1,977 CNVs from 24 loci. Curated parent of origin data from 77 published studies are collapsed by loci into single data points; recombination rate ratios are the average of the metric for all CNVs within the data point. Data points size and color correspond to the number of CNV data collapsed into the data point. Recombination rate ratios predict parental origin for CNV mediated by NAHR (p=1.07 x10-14, β = 0.6606, CI95% =(0.4980,0.8333), OR = 1.936).



Supplemental Methods

Literature Search and Data Curation

CNV loci were curated from Coe et al., 2014 [7]. This paper is an expansion of Cooper et al., 2011[141], and includes 55 known CNV loci associated with genomic disorders. We applied a set of exclusion criteria to the 55 loci in order to filter for loci in which CNVs are flanked by LCRs, *i.e.*, mediated by NAHR. As individuals with CNV at imprinted loci are generally ascertained by phenotype, and the phenotype is determined by the parent of origin, imprinted loci would introduce an ascertainment bias to the analysis and were therefore excluded. Loci were determined to be imprinted if the canonical interval overlapped imprinted genes as indicated by the Geneimprint database (http://www.geneimprint.com/). In total, 17 loci were excluded from further analysis because the loci were not flanked by LCRs and/or the loci were imprinted (Table S3-1). For the remaining 38 loci we conducted a systematic PubMed literature search for studies that reported parental origin of CNVs at these loci. On PubMed, loci were searched using a phrase with the format: *cytogenic locus OR syndrome.* The number of results/hits was recorded and if the initial search produced more than 100 hits, a sub-search was performed. This sub-search used the following format: *(cytogenic locus OR syndrome) + parental bias OR parental origin OR transmission bias OR parent-of-origin OR parent of origin OR maternal bias OR paternal bias OR paternal origin OR maternal origin).*

Studies were included in analysis:

- If the study reported parental origin
- If the authors of the study adequately interrogated the patients for presence of the CNV(s). This included stating that the CNV(s) was previously or currently confirmed,

and/or a confirmation via FISH, aCGH, marker PCR, SNP array, whole-genome sequencing, etc.

- If the authors of the study stated the CNVs were *de novo*.
- CNVs in monozygotic twins were clearly treated as the result of a single meiotic event. MZ twins are the product of one egg fertilized with one sperm, that then splits into two zygotes, thus the CNV present originated in one of the *single* gametes.

We note that 17q11.2q12, a known genomic disorder locus associated with Charcot-Marie-Tooth disease type 1A (CMT1A; duplication) and hereditary neuropathy with liability to pressure palsies (HNPP; deletion), is mediated by NAHR, and thus applicable for inclusion in our analysis. However, subsequent research on the locus produced reports of a sex-dependent bias in both the mechanism for formation of the associated CNVs, and the resulting phenotype [142,143]. CNVs of paternal origin are generated via NAHR between homologous chromosomes during meiosis and are largely duplications (resulting in CMT1A), whereas CNVs of maternal origin are produced via intrachromosomal rearrangement between sister chromatids and result in equal numbers of deletions and duplications (resulting in CMT1A and HNPP). This is likely to cause a complex ascertainment bias and introduce a confounder associated with this locus. For this reason, we excluded the 17q11.2q12 locus from this study.

77 studies in total satisfied inclusion criteria and included parental origin data for 24 loci encompassing 1,977 deletion and duplication events (Table 3-1 and Table 3-2). All search phrases and studies curated as part of the current analysis and the loci used in the logistic regression analysis are listed in Table S3-1 to S3-2 and Table S3-3, respectively.

Study Subject Recruitment

Individuals with clinically confirmed diagnosis of 3q29 deletion were ascertained through the internet-based 3q29 registry (https://3q29deletion.patientcrossroads.org/) as previously described [103,144]. We obtained blood samples and determined parental origin of the 3q29 deletion in 14 families. Of the 14 families, 12 were full trios. The remaining two families were both mother and child pairs.

Sample Collection and Banking

Whole blood was collected from proband, mother and father as previously reported [103] and banked at the NIMH Repository and Genomics Resource (NRGR; Piscataway, New Jersey, USA).

DNA Isolation

DNA samples were isolated from/obtained from biobanked samples at the NIMH Repository and Genomics Resource (NRGR; Piscataway, New Jersey, USA). The source of DNA was either whole blood or LCLs derived from biobanked blood samples.

SNP Genotyping and QC

SNP genotyping was performed on 12 of the 14 families (10 full trios, 2 mother-child pairs) by AKESOgen (Peachtree Corners, Georgia, USA) on the Illumina GSA-24 v 3.0 array, which contains 654,027 SNPs genome wide. DNA from participants was normalized and genotyped according to AKESOgen/Illumina protocols. Data was returned as separate final reports that were combined into one deduplicated final report, and converted into PLINK format for quality control. QC was performed with PLINK 1.9 [104]. Briefly, indel calls, unmappable SNPs, and SNPs with call rates less that 97% were dropped from the SNP call set (n = 35,187), leaving 611,986 SNPs genome wide. Samples reported sex and family relationships were verified using this set of quality SNPs and PLINK 1.9. F coefficient estimates for the X chromosome were calculated and sex assignment was inferred for each sample in the batch. Before sex was inferred, the --*split-x* flag with the *hg38* modifier was used to identify pseudoautosomal regions of the X chromosomes for subsequent removal during the sex check. The default parameters for the --*check-sex* flag were used to infer sex. A sample with an F coefficient less than or equal to 0.2 was assigned as male, and a sample with an F coefficient greater than or equal to 0.9 was assigned as female. Any samples with an opposite sex assignment than indicated by the given pedigree were flagged and investigated for possible sample swapping or sample mixture. Expected relationships between related samples of the batch were verified with PLINK 1.9. Variants were LD-pruned using the --*indep-pairwise* flag using 50, 5, and 0.2 for the variant count window size, variant count step size and r^2, respectively. The --*genome* flag was used to infer relationships (coefficient of relatedness; *r*) on this set of pruned SNPs. All samples sex information and relationship information were concordant with our expectation based on information provided by the families.

Whole-genome Sequencing

For 2 full trios (families 3206 and 3147; 6 samples), parent of origin was determined from whole genome sequence data. All samples were sequenced at the Hudson–Alpha Institute of Biotechnology (Birmingham, Alabama, USA) using their published protocols. Sequencing was performed to approximately 30X coverage per genome on the Illumina NovaSeq 6000 platform.

Following sequencing all base-calling was performed using standard Illumina software to generate the final FASTQ files for each sample.

Sequence Alignment: PEMapper

FASTQ files were aligned on a per sample basis with PEMapper [145] using default parameters and a Smith-Waterman alignment threshold of 95%, as recommended for 150-bp paired-end reads. Alignment was performed relative to the human Hg38 reference as reported by the University of California at Santa Cruz (UCSC) Genome Browser on July 1, 2015. The output from PEMapper, pileup and indel files, were used as input for variant calling with PECaller [145]. Pileup files contained the number of reads where an A, C, G, or T nucleotide was seen together with the number of times that base appeared deleted or there was an insertion immediately after the base. Indel files contained the nucleotide sequence of the deletions and insertions indicated in the pileup files. Alignment performance was checked before moving to variant calling. No samples were removed on the basis of failed sequence alignment.

Variant Calling: PECaller

Variant calling was performed in a single batch using PECaller [145], which assumes multiple samples all done on the same technology will be available. Optimal PECaller performance is achieved when at least 50 genomes are called in batch; 57 control genomes were included with the genomes from families 3206 and 3147 (63 genomes total). PECaller was run with the default theta value of 0.001 and a 95% posterior probability for a genotype to be considered called. A posterior probability of less than 95% was considered a missing call. Calls were produced for the repeat-masked (unique) subset of the human Hg38 reference as reported by the University of

California at Santa Cruz (UCSC) Genome Browser on July 1, 2015. The initial .snp file output from PECaller was used in a subsequent step to merge SNP variant calls with INDEL variant calls, producing a final “merged” .snp file. This raw file was used for site and sample quality control.

Whole-genome Sequence Quality Control

Quality control was performed on a per-site and per-sample basis. The following metrics were used to flag and/or exclude samples and variant sites from QC and analysis, and were calculated using a custom QC pipeline consisting of multiple in-house-developed scripts, PLINK 1.9 [104], R [105], and Bystro [146]:

1. *Per-site QC: Missing call rate:* The missing call rates for variant sites were calculated as described above. Variants with a missing call rate greater than or equal to 10% were removed from subsequent QC *and* variant analysis.
2. *Sample Mixture Check:* Possible sample mixture was checked by calculate the ratios of minor allele homozygous calls to heterozygous calls. This number varies between call batches, and thus cannot be compared across different calling experiments. However, non-mixed samples within the same calling batch should exhibit similar ratios. The ratios were calculated using Bystro. No samples were removed on the basis of possible sample mixture.
3. *Per-sample QC: Transition:Transversion Ratio:* Transition:transversion (Ti:Tv) ratios were calculated for each genome in the variant calling batch using a script developed in-house Bystro. Based on population expectations, the Ti:Tv ratio for an individual genome

is expected to be approximately 2.00, with a ratio of 2.04 representing a quality genome. The batch mean Ti:Tv ratio were calculated using Bystro. The control genomes used in batch calling were previously validated for calling performance, therefore a *mean* Ti:Tv ratio less than 2.00 suggests a failed variant-calling experiment. As such, the entire sample batch is resubmitted for variant calling. No samples were removed from analysis on the basis of Ti:Tv ratio.

4. *Per-sample QC: Silent:Replacement Ratio:* Silent:replacement (sil:rep) ratios were calculated for each genome in the variant calling batch using Bystro. The expected sil:rep ratio for a single genome is expected for fall between 1.05 and 1.15, with 1.15 indicating a quality genome. The batch mean sil:rep ratio and standard deviation were calculated using Bystro. A mean sil:rep less than 1.05 suggested a failed variant-calling experiment and the sample batch was resubmitted for variant calling. Any samples with a sil:rep ratio less than 1.05 were flagged and removed from subsequent QC. No samples were removed from analysis on the basis of sil:rep ratio.
5. *Per-sample QC: Missing call rate:* The missing call rates for samples were calculated using PLINK. The merged .snp file generated after the indel merging process was converted to a VCF [v4.0] format (snp_to_vcf2), the appropriate VCF headers were appended to file, and multiallelic variants were split using BCFtools 1.3 [147], before the final BCF was loaded into PLINK. The following flags were used during loading: --*bcf*, and --*keep-allele-order*. Per sample missing call rates were calculated using the –*missing* flag in PLINK, and the batch mean missing call rate and standard deviation was calculated using R. A mean missing call rate greater than or equal to 3% indicated a failed variant-calling experiment and the sample batch was resubmitted for variant

calling. No samples in the current analysis were removed on the basis of low call rate.

6. *Sex Check:* PLINK was used to calculate the F coefficient estimates for the X chromosome and impute sex assignment for each sample in the batch. Before sex was inferred, the --*split-x* flag with the *hg38* modifier was used to identify pseudoautosomal regions of the X chromosomes for subsequent removal during the sex check. The default parameters for the --*check-sex* flag were used to infer sex. A sample with an F coefficient less than or equal to 0.2 was assigned as female, and a sample with an F coefficient greater than or equal to 0.9 was assigned as male. All samples' inferred sex matched our expectation based on provided information.
7. *Relationship Inference:* Expected relationships between related samples of the batch were verified with PLINK. Variants were LD-pruned using the --*indep-pairwise* flag using 50, 5, and 0.2 for the variant count window size, variant count step size and r^2, respectively. The --*genome* flag was used to infer relationships (coefficient of relatedness; *r*) on this set of pruned SNPs. Among the control genomes there was a known parent-offspring relationship, which was used as a positive control, while the remaining control genomes were known to be unrelated. All samples' inferred relationships matched our expectations based on information provided by the families.

Parental Origin Analysis

Parental origin of the 3q29 deletion was determined for 12 trios --10 full trios and 2 trios for which only the child and mother's info was available -- using SNP array data. Briefly, using PLINK 1.9 [104], 404 SNPs located within the 3q29 deletion interval (chr3:196029182-197617792; hg38) were isolated for analysis. Mendelian errors (MEs) were called for these SNPs using PLINK’s --*mendel* function with the -*duos* modifier to also call MEs for the mother-daughter pairs. The parent with the most mendelian errors was considered the parent of origin for the 3q29 deletion. Parental origin was determined using WGS data for two trios (3147 and 3206). Briefly, variants in the 3q29 critical region were called using PECaller [145]. The variants with a sample minor allele frequency (MAF) less that 10% were filtered from this set of SNPs, and MEs were called. As in the SNP array analysis, the parent with the most MEs was considered the parent of origin for the 3q29 deletion

Paternal Age Analysis

Age of fathers at birth data for ~3 million U.S. births in 2018 (latest data available) were obtained from the National Center for Health Statistics (NCHS) (https://www.cdc.gov/nchs/index.htm). The mean age of parents in our 3q29 cohort was collected from self-reported data in conjunction with the Emory University 3q29 project (http://genome.emory.edu/3q29/) and compared to the U.S. average via a two-tailed two-sample t-test using R [105].

Breakpoint Usage Determination

We calculated average male and recombination rates over CNV intervals as determined by reported or canonical breakpoints [7]. When possible, we used breakpoints reported by the authors of the studies to calculate average male and female recombination rates and used the UCSC LiftOver tool to convert the breakpoints to hg38. To reduce the possibility of the failure of breakpoints reported on older human genome builds to successfully liftover to hg38, we used breakpoints reported by the authors only if they were reported on human genome build 19 (hg19) or higher. Otherwise, the breakpoints cited by Coe et al., 2014 were used. We note that each interval from Coe et al., 2014, except the 1q21.1 TAR, locus successfully translated to hg38 coordinates with UCSC LiftOver. For 1q21.1 TAR, we used canonical breakpoints cited in the Clinical Genome Resource (ClinGen) [148]. We also note the breakpoints cited in Coe et al., 2014 for the 22q11.2 region correspond to the ~1.5 Mb LCR22A-LCR22B CNV interval and not the more common ~3 Mb LCR22A-LCR22D CNV interval. Where possible to distinguish between the 1.5 Mb and 3.0 Mb 22q11.2 CNVs, we used the appropriate ~1.5 Mb or ~3.0 Mb breakpoints cited in ClinGen.

Calculation of Recombination Rates and Ratios

Chromosome male and female recombination rates (cM/Mb) were obtained from the deCODE sex-specific maps [106]. The recombination rate (cM/Mb) data from deCODE is publicly available as recombination rates calculated for variably-sized physical genomic intervals bounded by two SNP markers. Therefore, for our calculation of the average male and female recombination rates, each bounded recombination rate was weighted by the total number base pairs contained within the respective SNP marker interval. Weighted rates were then averaged across the CNV interval (See Breakpoint Usage below) for males and females, separately. The ratio of weighted average male and female recombination rates was then calculated for each CNV interval by dividing the weighted average male recombination rate by the weighted average female recombination rate. To account for slight differences in the recombination ratios calculated for the different LCR22 intervals at the 22q11.2 locus we used an adjusted recombination ratio composed of the weighted recombination rate ratios calculated for each LCR22 interval. Weights were based on the estimated population prevalence of the different

22q11.2 deletion intervals (Table S3-3) [107].The data from deCODE is presented as binned rates across separate chromosomes. As such, each binned recombination rate was weighted by the total basepairs of CNV contained within the respective bin (breakpoints cited in Coe et al, 2014 [7]). Weighted binned rates were then averaged across the CNV interval.

Logistic Regression Analysis

Parental origin data was curated for CNVs at the 24 CNV loci from 77 independent studies; only independent samples were included in the analysis (duplicate or overlapping samples were removed). For each CNV locus the male to female recombination rate ratio was calculated as described above. A logistic regression model was fitted to the data with the $\log_e$-transformed male to female recombination rate ratio as the predictor and parental origin (paternal vs. maternal) as the response variable using R [105]. We performed a secondary analysis stratified by deletions/duplications. See Table 3-2 and Table S3-4 for the data calculated and used in the logistic regression.

Linear Regression Analysis

Locus-specific estimates for parental origin were derived by combining the data from all published studies for a given locus. To alleviate the uncertainty introduced by small sample sizes, only those loci with more than 10 observations were included. The $\log_e$-transformed combined male to female parental origin count ratios for each locus was regressed on the calculated averaged $\log_e$-transformed average male to female recombination rate ratio for that locus' CNV interval using R [105]. Each locus was weighted based on its sample size. A combined analysis (deletions and duplications) was performed under the assumption that an

NAHR event produces reciprocal deletion and duplication products, formation of both types of CNVs would be subject to the same biological forces. Thus, for each locus, duplications and deletions were treated equally and grouped under one locus.

Sensitivity Analysis

A sensitivity analysis was conducted for the combined linear regression by iteratively running the linear model in R [105]. On each iteration one data point was removed from the model in order to identify potential influencing points. Results from the analysis are listed in Table S3-8.

Supplemental Tables

Table S3-1. Exclusion/Inclusion statuses and literature search results of genomic disorder loci conducted January 2021.

Can be accessed online at: https://emory-my.sharepoint.com/:x:/r/personal/tmosle3_emory_edu/Documents/TMosley_Dissertation%20Links/TableS3-1_Additional%20File%202_SupplementalTable1_Litsearch_FINAL.csv?d=w18f6c4f160dc49dda6ad6c4238f11eb0&csf=1&web=1&e=BzJJsm

Table S3-2. List of 1,268 search results curated from literature search

Can be accessed online at: https://emory-my.sharepoint.com/:x:/r/personal/tmosle3_emory_edu/Documents/TMosley_Dissertation%20Links/TableS3-2_Additional%20File%203_SupplementalTable2_search_details_FINAL.csv?d=wbbf60e49a70a4c84b84320b195315d68&csf=1&web=1&e=gzK3Zx

Table S3-3. LCR22 Recombination rate data

LCR Interval	Begin	End	Pop. Frequency	[b]Avg. Male Recomb. Rate [106]	[b]Avg. Female Recomb. Rate [106]	[c]M:F Recomb. [106]
22q11.2 AB	18872532	20326091	0.050	1.28300587	3.42708495	0.37437236
22q11.2 AC	18872532	20702815	0.025	1.52325126	3.61442402	0.42143679
22q11.2 AD	18872532	21337106	0.850	1.36673167	3.44873357	0.39629958
22q11.2 BC	20326091	20702815	0.025	2.45023972	4.33727307	0.56492632
22q11.2 BD	20326091	21337106	0.025	1.48712659	3.47986356	0.42735198
22q11.2 CD	20702815	21337106	0.025	0.91497120	2.97050338	0.30801891

Recombination rate calculations for chr 22q11.2 LCR intervals. Begin and End coordinates curated from UCSC Genome Browser, hg38. [a]Male to female CNV parent of origin counts. [b]Average male and female recombination rates are as described in Table 3-2. [c]22q11.2 weighted average was calculated by weighting the M:F recombination rate ratio for each interval by the interval population frequency.

Table S3-4. Logistic regression data for 1,977 CNVs

Can be accessed online at: https://emory-my.sharepoint.com/:x:/r/personal/tmosle3_emory_edu/Documents/TMosley_Dissertation%20Links/TableS3-4_Additional%20File%204_SupplementalTable4_logistic_reg_data_FINAL.csv?d=w493d45c574af436c9f7dc991d08e15b1&csf=1&web=1&e=p7A5CA

Table S3-5. Demographic data for 3q29 cohort and parental origin of the 3q29 deletion

Family ID	Subject ID	[a]Father ID	[a]Mother ID	Sex	[b]Pat:Mat MEs	Deletion Parental Origin	[c]Parent of Origin Age
93315602	834-3156-1031	N/A	834-3156-2096	F	--:0	Pat	24y
93315602	834-3156-2096			F			
93316202	834-3162-1031	N/A	834-3162-2096	F	--:0	Pat	40y
93316202	834-3162-2096			F			
93316402	834-3164-1001	834-3164-2046	834-3164-2096	M	37:0	Pat	32y
93316402	834-3164-2046			M			
93316402	834-3164-2096			F			
93316802	834-3168-1001	834-3168-2046	834-3168-2096	M	36:0	Pat	43y
93316802	834-3168-2046			M			
93316802	834-3168-2096			F			
93318102	834-3181-1001	834-3181-2046	834-3181-2096	M	44:0	Pat	34y
93318102	834-3181-2046			M			
93318102	834-3181-2096			F			
93318902	834-3189-1001	834-3189-2046	834-3189-2096	M	67:0	Pat	29y
93318902	834-3189-2046			M			
93318902	834-3189-2096			F			
93322602	834-3226-1031	834-3226-2046	834-3226-2096	F	0:40	Mat	41y
93322602	834-3226-2046			M			
93322602	834-3226-2096			F			
93324602	834-3246-1001	834-3246-2046	834-3246-2096	M	44:0	Pat	38y
93324602	834-3246-2046			M			
93324602	834-3246-2096			F			
93325702	834-3257-1031	834-3257-2046	834-3257-2096	F	36:0	Pat	38y
93325702	834-3257-2046			M			
93325702	834-3257-2096			F			
93327702	834-3277-1001	834-3277-2046	834-3277-2096	M	54:0	Pat	28y
93327702	834-3277-2046			M			
93327702	834-3277-2096			F			
93339602	834-3396-1031	834-3396-2046	834-3396-2096	F	33:0	Pat	43y

93339602	834-3396-2046			M			
93339602	834-3396-2096			F			
93342002	834-3420-1001	834-3420-2046	834-3420-2096	M	27:0	Pat	36y
93342002	834-3420-2046			M			
93342002	834-3420-2096			F			
93320602	834-3206-1031	834-3206-2096	834-3206-2046	F	[d]401:20	Pat	32y
93320602	834-3206-2046			M			
93320602	834-3206-2096			F			
93314702	834-3147-1001	834-3147-2046	834-3147-2096	M	[d]743:22	Pat	29y
93314702	834-3147-2046			M			
93314702	834-3147-2096			F			

Demographic data is self-reported. [a]N/A indicates information is not available. Grandparental samples were not collected. [b]Number of informative SNPs per parent supporting parental origin given in the order father:mother; -- indicates parent unavailable. [c]Age of parent of origin corresponds to parent's age at birth of affected child. [d]Family 3206 and 3147 parental origin determined with whole-genome sequencing data.

Table S3-6. Summary of deletions grouped by locus, parental origin and recombination data

Locus	BED Coordinates [7]	# Samples (%)	[a]M:F Origin Counts	[b]Avg. Male Recombination Rate [106]	[b]Avg. Female Recombination Rate [106]	[c]Log$_e$ M:F Recombination Ratio [106]
1q21.1	chr1:147101794-147921262	7 (0.37%)	4:3	0.12887809	0.53120997	-1.4162905
1q21.1 TAR	dchr1:145686999-146048495	1 (0.05%)	1:0	0.15712388	0.77814863	-1.599883
3q29	chr3:195988732-197628732	22 (1.15%)	21:1	3.1305211	0.27775988	2.42219776
5q35	chr5:176290391-177630393	41 (2.14%)	36:5	1.29955355	0.97941355	0.28282209
7q11.23	chr7:73328061-74727726	598 (31.26%)	287:311	0.49353298	1.92655808	-1.3619006
8p23.1	chr8:8235068-12035082	1 (0.05%)	1:0	0.67201752	1.81857951	-0.9955266
11q13.2q13.4	chr11:67985953-71571306	1 (0.05%)	0:1	0.8431765	2.23501635	-0.9748275
15q13.3	chr15:30840505-32190507	5 (0.26%)	4:1	1.63838993	1.90155369	-0.1489573
15q24 AC	chr15:72670606-75240606	1 (0.05%)	1:0	0.28479919	0.86129537	-1.1066532
15q24 AD	chr15:72670606-75720604	3 (0.16%)	1:2	0.27613544	0.82152961	-1.0902765
15q24 BD	chr15:73720606-75720604	1 (0.05%)	0:1	0.30739967	0.68432207	-0.8002799
15q24 BE	chr15:73720606-77840603	2 (0.10%)	0:2	0.23485125	0.72623183	-1.128917
15q25.2	chr15:82513967-84070244	5 (0.26%)	0:5	0.21225081	0.32633295	-0.4301495
16p11.2	chr16:29641178-30191178	79 (4.13%)	9:70	0.06565904	1.28716751	-2.9757241
16p11.2 distal	chr16:28761178-29101178	3 (0.16%)	0:3	0.11421814	1.52707077	-2.5929965
16p13.11	chr16:15408642-16198642	1 (0.05%)	1:0	1.66089552	2.45240302	-0.3897114
17p11.2	chr17:16805961-20576095	59 (3.08%)	35:24	0.1888066	1.19115966	-1.8419594
17q11.2	chr17:30838856-31888868	62 (3.24%)	10:52	0.26024285	1.85442774	-1.9637162
17q12	chr17:36460073-37846263	6 (0/31%)	4:2	0.64750654	3.64754421	-1.7286805
17q21.31	chr17:45626851-46106851	35 (1.83%)	18:17	0.38234179	0.98304273	-0.9443376
17q23.1q23.2	chr17:59987857-62227857	2 (0.10%)	0:2	0.56466054	1.30765625	-0.839767
22q11.2	chr22:18924718-21111383	978 (51.12%)	411:567	1.45946494	3.69205976	-0.9281146
All	—	**1913 (100%)**	**844:1069**	—	—	—

Summarized duplication data. Data are consolidated by locus. BED coordinates correspond to hg38 (LiftOver from hg18 coordinates in Coe et al., 2014). [a]Male to female CNV parent of origin counts. [b]Average male and female recombination rates are as described in Table 3-2. [c]Natural log-transformed average male to female recombination rate ratio for the locus

Table S3-7. Summary of duplications grouped by locus, parental origin and recombination data

Locus	BED Coordinates [7]	# Samples (%)	[a]M:F Origin Counts	[b]Avg. Male Recombination Rate [106]	[b]Avg. Female Recombination Rate [106]	[c]Log_e M:F Recombination Ratio [106]
1q21.1	chr1:147101794-147921262	2 (3.13%)	2:0	0.1038527	0.42854444	-1.4174209
2q13	chr2:110625954-112335952	1 (1.56%)	1:0	0.44854539	1.64377881	-1.2987431
7q11.23	chr7:73328061-74727726	20 (31.25%)	9:11	0.49361225	1.92693374	-1.361935
8p23.1	chr8:8235068-12035082	2 (3.13%)	0:2	0.67201752	1.81857951	-0.9955266
15q13.3	chr15:30840505-32190507	1 (1.56%)	1:0	1.62649391	1.88629391	-0.1481873
16p11.2	chr16:29641178-30191178	19 (29.69%)	2:17	0.06591856	1.28839583	-2.9727332
16p11.2 distal	chr16:28761178-29101178	1 (1.56%)	0:1	0.14338352	1.88529263	-2.5763154
16p11.2p12.1	chr16:21341178-29431178	1 (1.56%)	1:0	0.5534655	2.68382469	-1.5787988
16p13.11	chr16:15408642-16198642	1 (1.56%)	0:1	1.68055204	2.47808756	-0.3883648
17p11.2	chr17:16805961-20576095	12 (18.75%)	9:3	0.1888066	1.19115966	-1.8419594
17q21.31	chr17:45626851-46106851	4 (6.25%)	1:3	0.38234179	0.98304273	-0.9443376
All	—	**64 (100%)**	**26:38**	—	—	—

Summarized duplication data. Data are consolidated by locus. BED coordinates correspond to hg38 (LiftOver from hg18 coordinates in Coe et al., 2014). [a]Male to female CNV parent of origin counts. [b]Average male and female recombination rates are as described in Table 3-2. [c]Natural log-transformed average male to female recombination rate ratio for the locus

Table S3-8. Sensitivity analysis results for linear regression analysis with deletions and duplications combined

Locus Removed	CNV Type	[a]r^2	*p*-value	Beta
None	Deletion/Duplication	0.8512	0.0011	0.9540
3q29	Deletion	0.7221	0.0155	0.9823
5q35	Deletion	0.8585	0.0027	0.8991
7q11.23	Deletion/Duplication	0.8906	0.0014	0.9719
16p11.2	Deletion/Duplication	0.7862	0.0078	0.9154
17p11.2	Deletion/Duplication	0.9042	0.0010	0.9729
17q11.2	Deletion	0.8597	0.0026	0.9204
17q21.31	Deletion/Duplication	0.8543	0.0029	0.9446
22q11.2	Deletion	0.9061	0.0009	0.9785

[a] Multiple r^2 value as reported by R [105].

Table S3-9. Predicted probability of paternal origin for loci with N < 10

Locus	BED Coordinates	Sample (N)	Pred. Paternal Probability
1q21.1	chr1:147101794-147921262	9	0.39899
1q21.1 TAR	chr1:145686999-146048495	1	0.37028
2q11.2	chr2:96060525-97010536	0	0.37428
2q11.2q13	chr2:100077106-107827112	0	0.53549
2q13	chr2:110625954-112335952	1	0.41771
7q11.23 distal	chr7:75332889-77032747	0	0.36502
7q11.23 proximal	chr7:67017578-72805248	0	0.47948
8p23.1	chr8:8235068-12035082	3	0.46711
10q23	chr10:80200264-87040263	0	0.53914
11q13.2q13.4	chr11:67985953-71571306	1	0.47051
15q13.3	chr15:30840505-32190507	6	0.60525
15q24 AC	chr15:72670606-75240606	1	0.44889
15q24 AD	chr15:72670606-75720604	3	0.45157
15q24 BD	chr15:73720606-75720604	1	0.44054
15q24 BE	chr15:73720606-77840603	2	0.44525
15q25.2	chr15:82513967-84070244	5	0.56021
15q25.2 Cooper	chr15:84595765-85155765	0	0.44832
16p11.2 distal	chr16:28761178-29101178	4	0.23931
16p11.2p12.1	chr16:21341178-29431178	0	0.37354
16p11.2p12.2	chr16:21601178-29031178	1	0.37478
16p12.1	chr16:21931178-22451178	0	0.51951
16p13.11	chr16:15408642-16198642	2	0.56638
17q12	chr17:36460073-37846263	6	0.35069
17q23	chr17:59577857-59997857	0	0.60003
17q23.1q23.2	chr17:59987857-62227857	2	0.49424
22q11.2 distal	chr22:21555711-23307813	0	0.58724

Probability of paternal origin for loci predicted from combined logistic regression: parental origin ~ $\log_e$(M:F recombination rate ratio).

Supplemental Figures

Figure S3-1. Schematic of recombination rate calculations. The recombination rate (cM/Mb) data from deCODE is publicly available as recombination rates estimated for variably-sized physical intervals bounded by two SNP markers (red and blue rulers). A representative image of the 3q29 locus is shown. Raw male (blue) and female (red) recombination rates are summarized and binned to demonstrate differences in male and female rates. Calculations were completed with raw recombination rate data. First, weighted average male and female recombination rates were calculated by weighting the estimated recombination rate within a respective SNP interval by the total number base pairs contained within that SNP interval. Weighted recombination rates were then averaged across the CNV interval for males and females, separately. The ratio of weighted average male and female recombination rates was then calculated for each CNV interval by dividing the weighted average male recombination rate by the weighted average female recombination rate. See Figures S3-2 to S3-12 for plotted raw male and female recombination rates for all loci included in the current analysis.

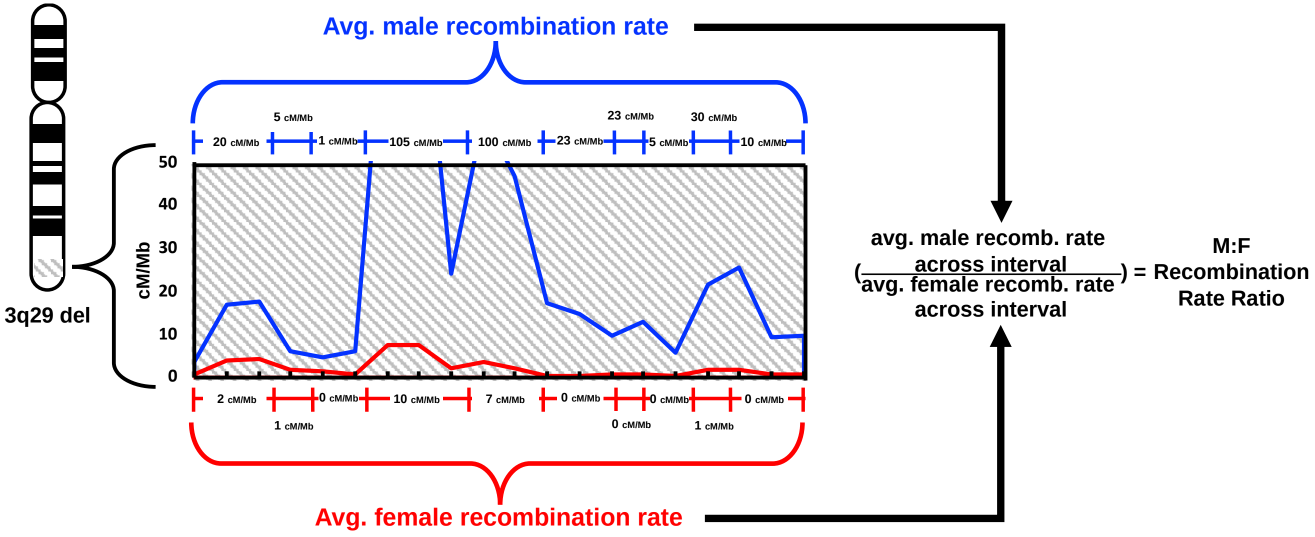


Figure S3-2. Raw deCODE recombination across 1q21.1 TAR and 1q21.1 regions. Male (blue) and female (red) recombination across the canonical 1q21.1 TAR and 1q21.1 regions (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio.

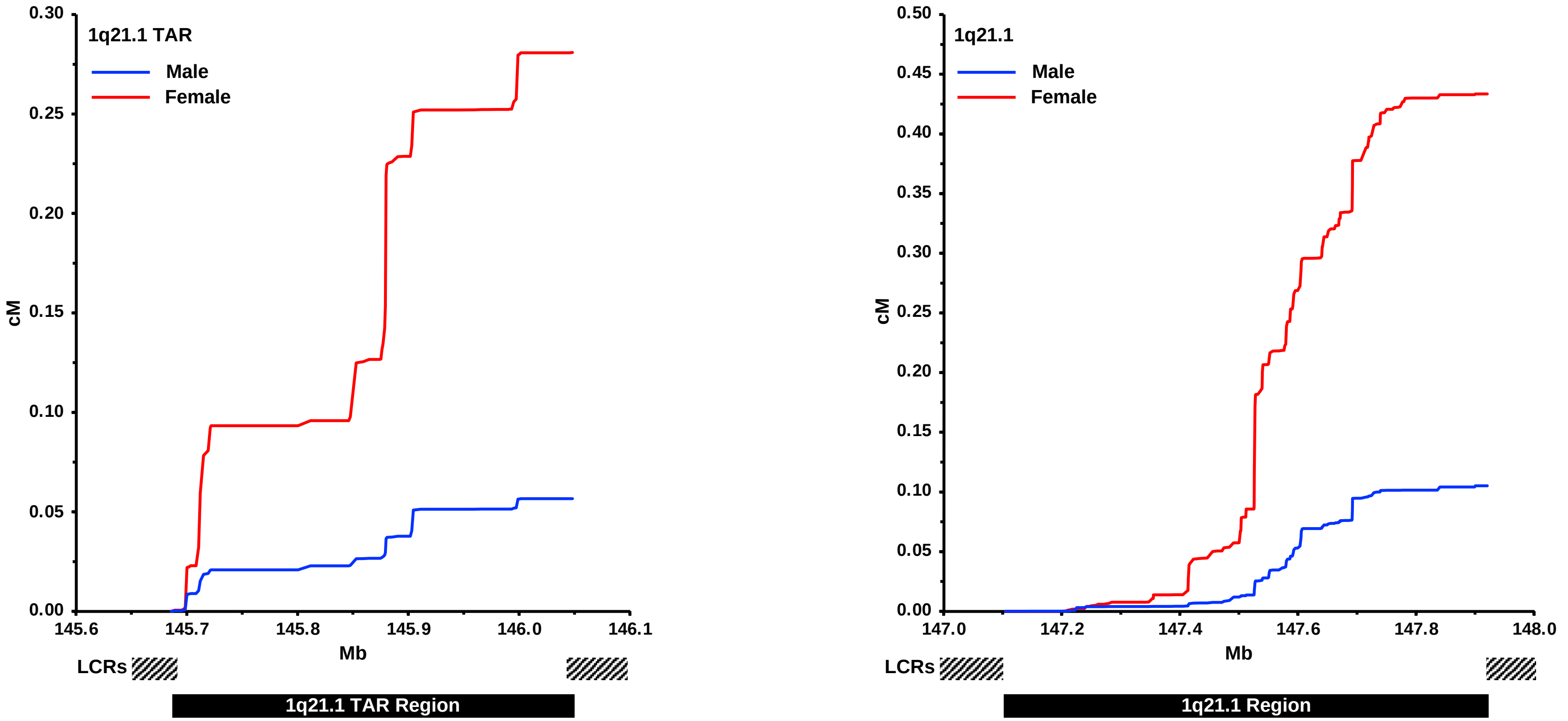


Figure S3-3. Raw deCODE recombination across 2q13 and 3q29 regions. Male (blue) and female (red) recombination across the canonical 2q13 and 3q29 regions (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio.

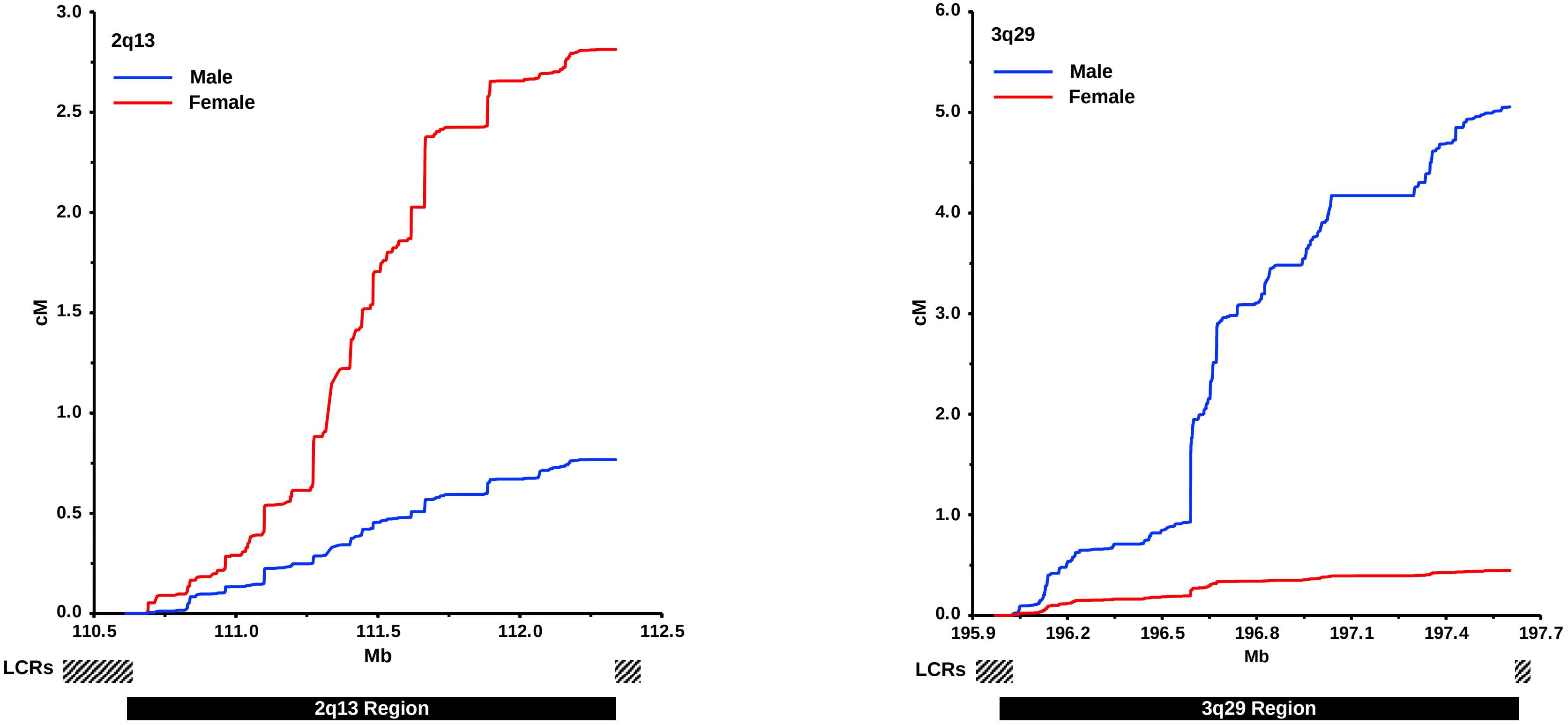


Figure S3-4. Raw deCODE recombination across 5q35 and 7q11.23 regions. Male (blue) and female (red) recombination across the canonical 5q35 and 7q11.23 regions (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio.

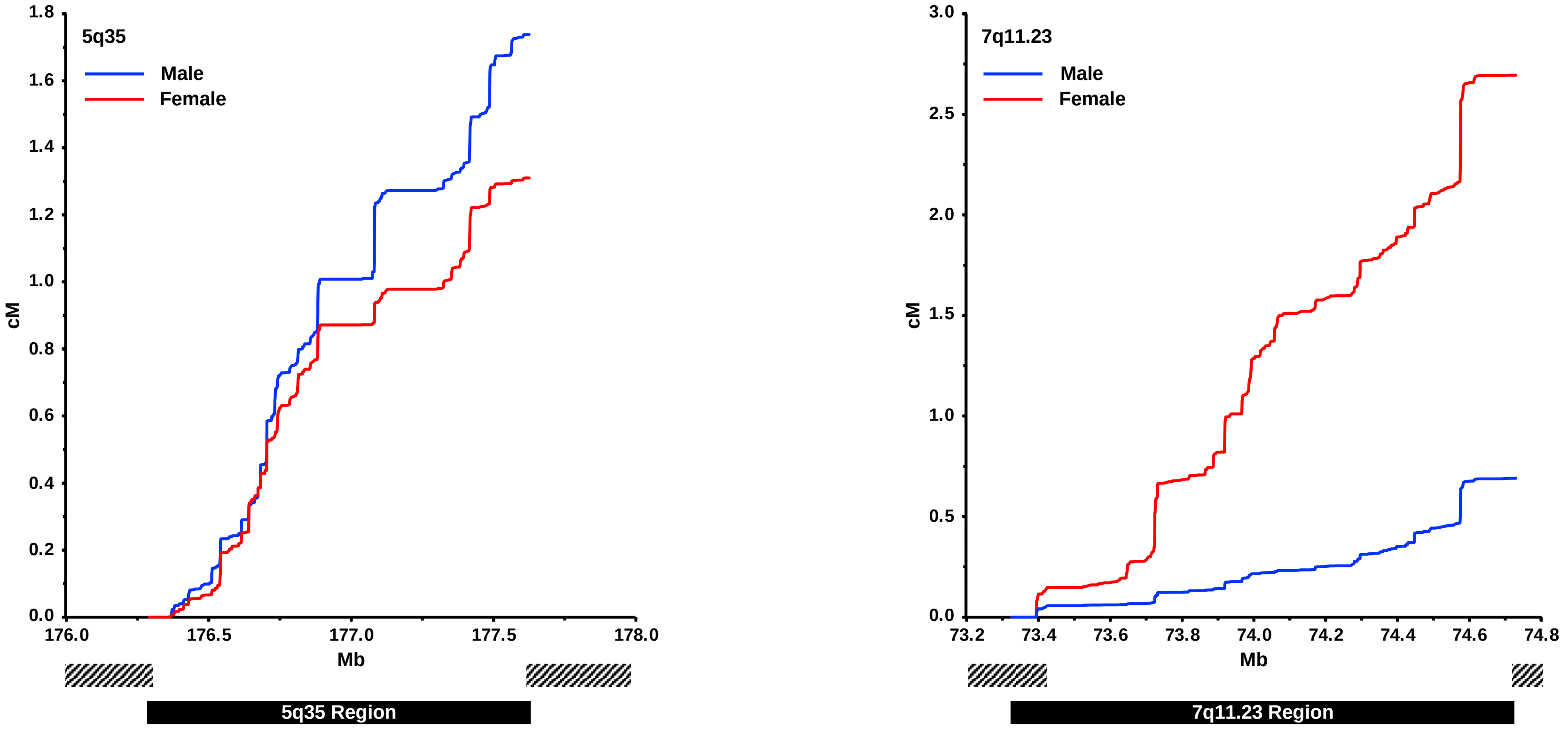


Figure S3-5. Raw deCODE recombination across 8p23.1 and 11q13.2q13.4 regions. Male (blue) and female (red) recombination across the canonical 8p23.1 and 11q13.2q13.4 regions (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio.

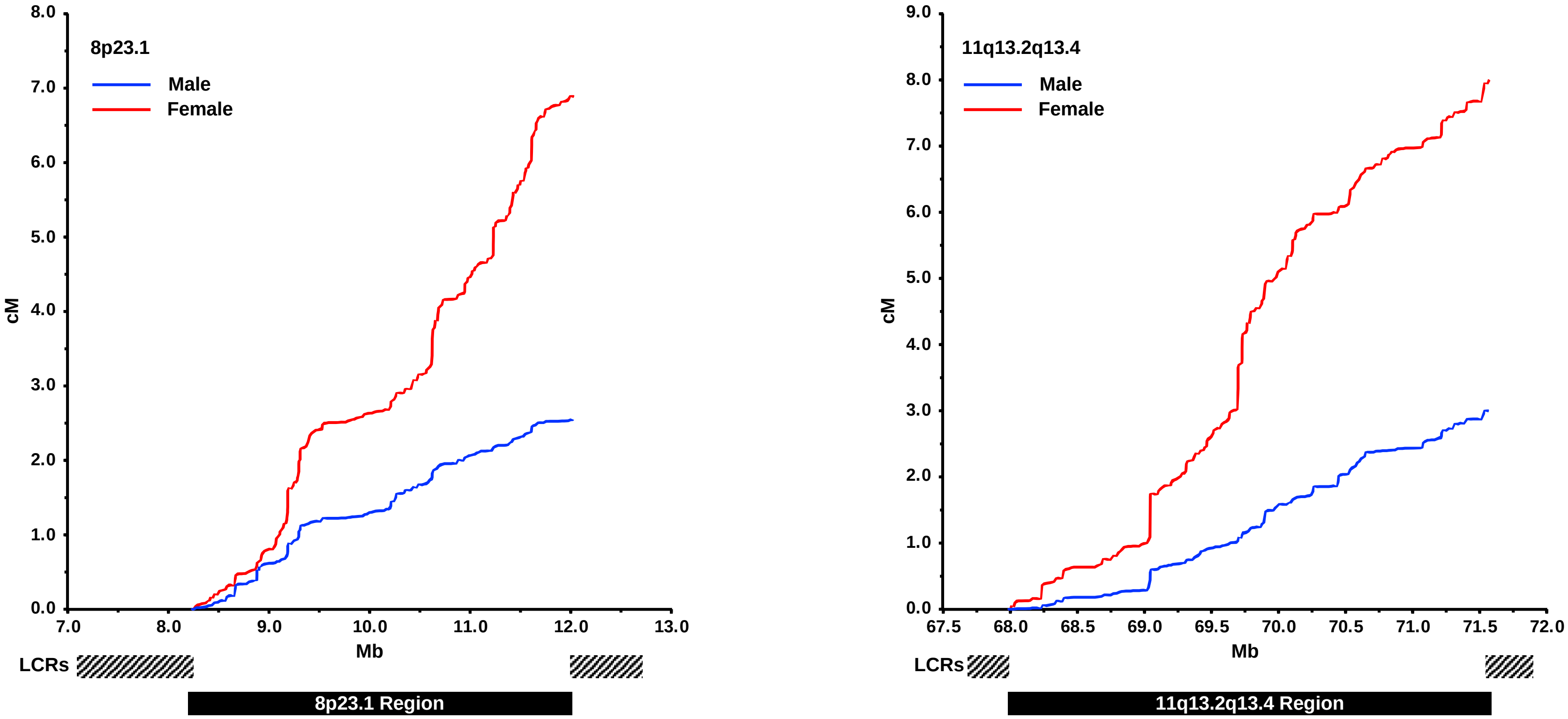


Figure S3-6. Raw deCODE recombination across 15q13.3 and 15q24. Male (blue) and female (red) recombination across the canonical 15q13.3 and 15q24 regions (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). LCRs demarking different 15q24 intervals are denoted with letters (A-E). X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio.

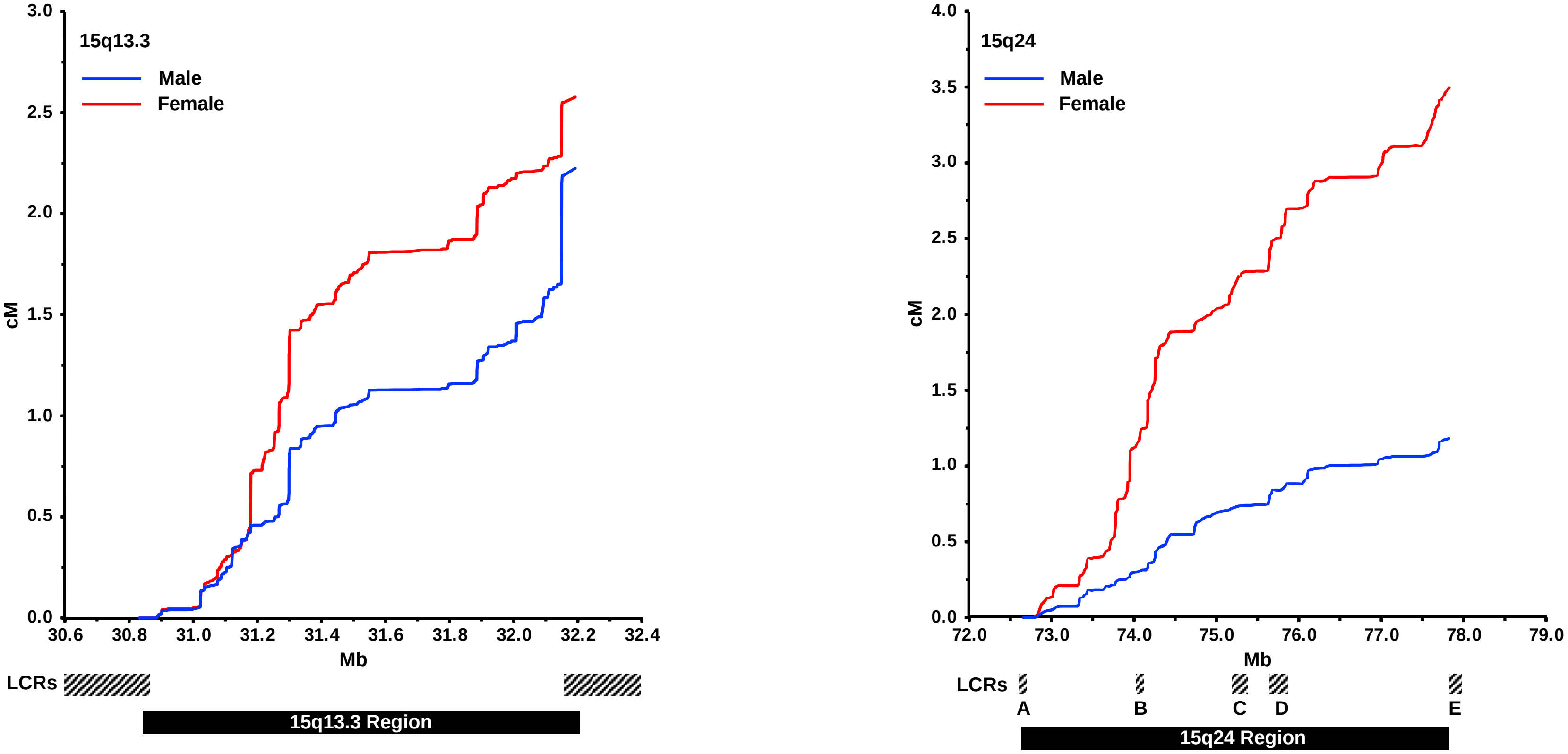


Figure S3-7. Raw deCODE recombination across 15q25.2 and 16p11.2 regions. Male (blue) and female (red) recombination across the canonical 15q25.2 and 16p11.2 regions (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio.

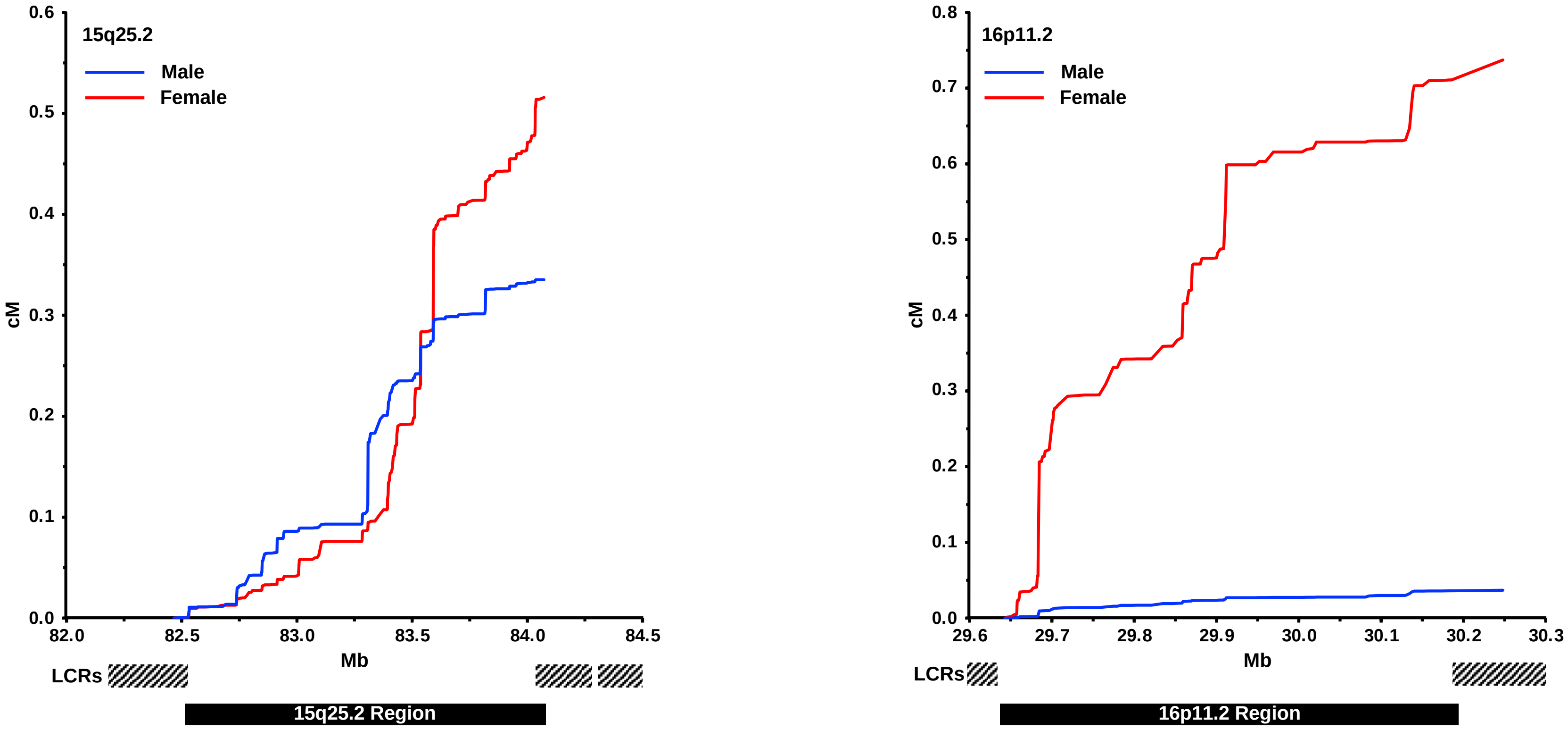
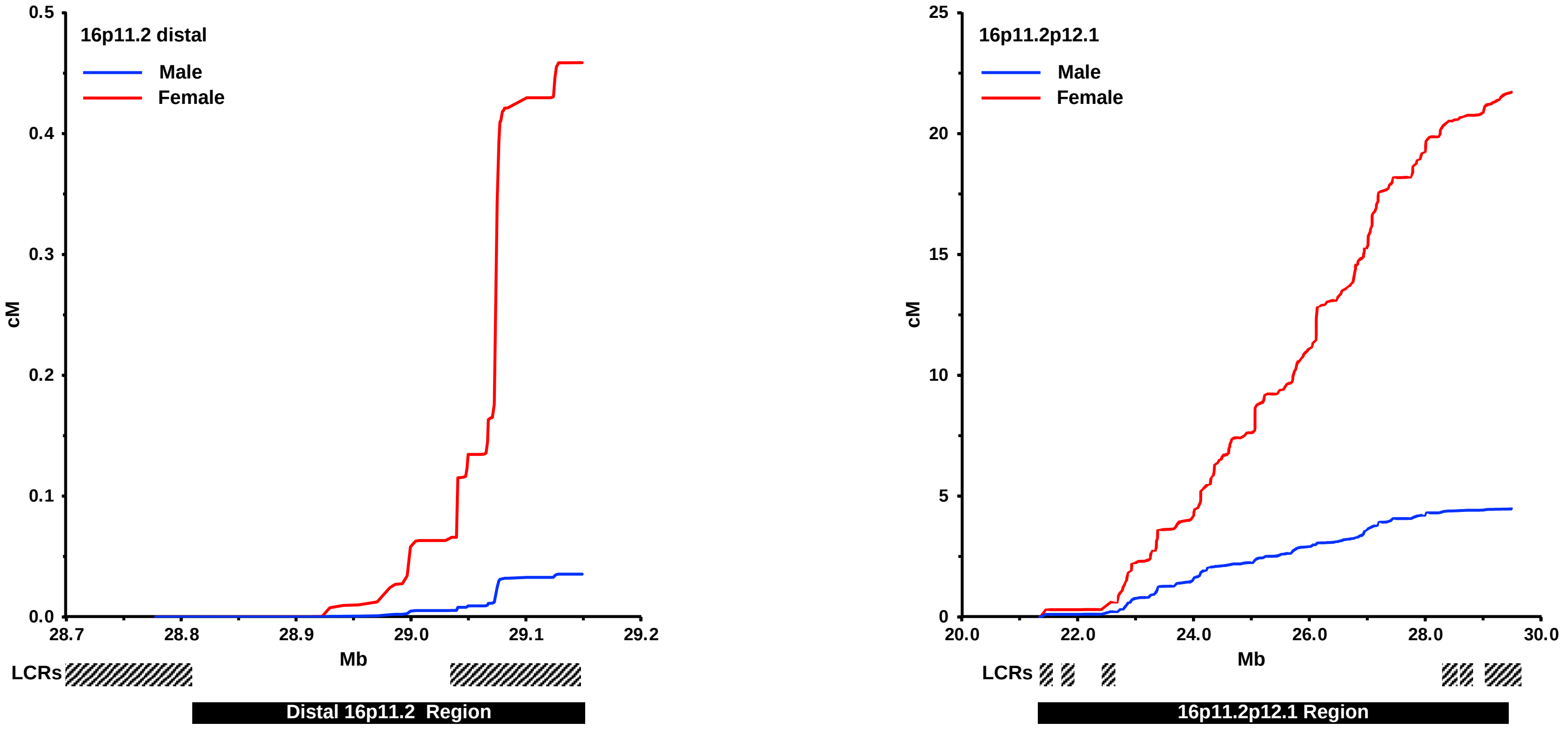


Figure S3-8. Raw deCODE recombination across distal 16p11.2 and 16p11.2p12.1 regions. Male (blue) and female (red) recombination across the canonical distal 16p11.2 and 16p11.2p12.1 regions (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio



.

Figure S3-9. Raw deCODE recombination across 16p13.11 and 17p11.2 regions. Male (blue) and female (red) recombination across the canonical 16p13.11 and 17p11.2 regions (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio.

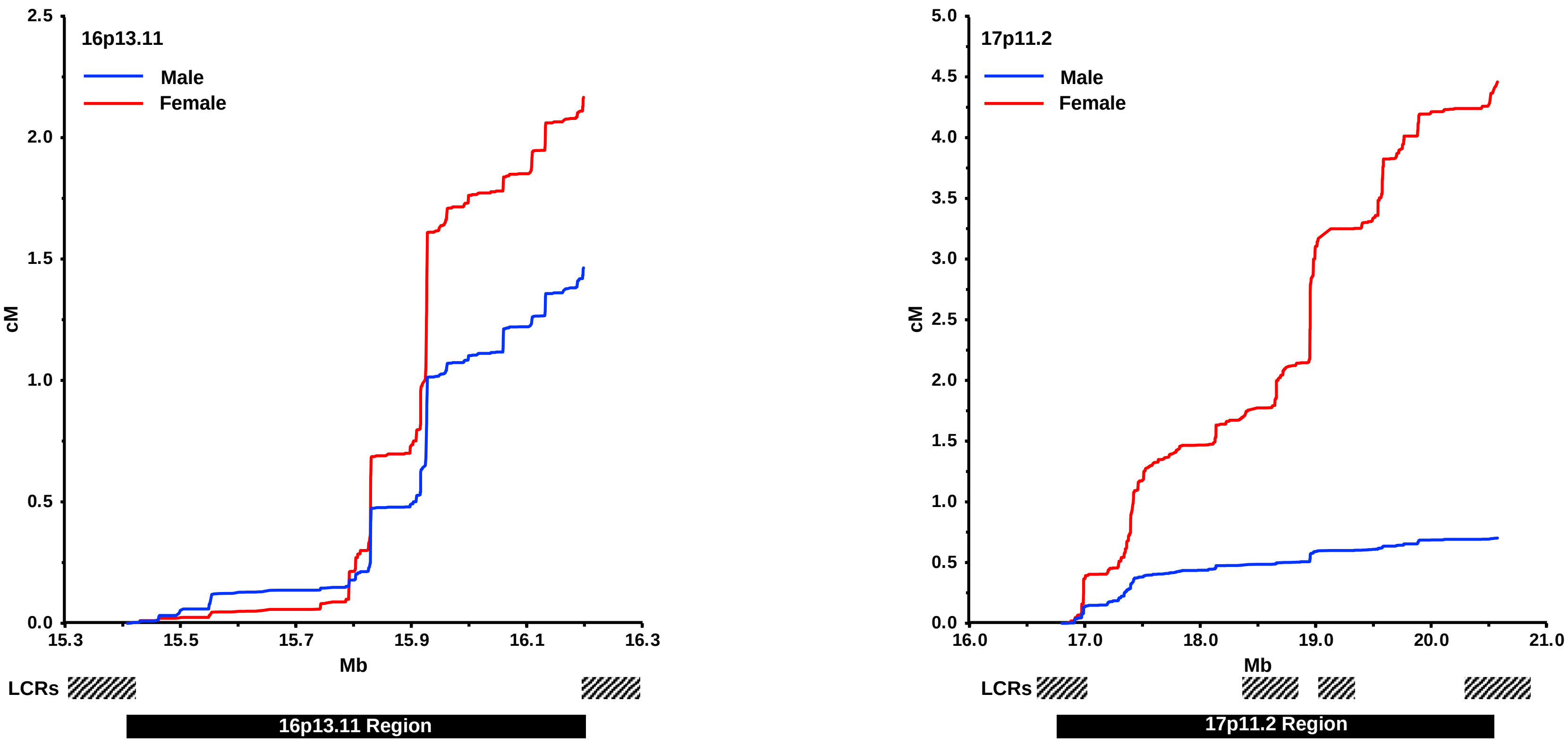


Figure S3-10. Raw deCODE recombination across 17q11.2 and 17q12 regions. Male (blue) and female (red) recombination across the canonical 17q11.2 and 17q12 regions (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio.

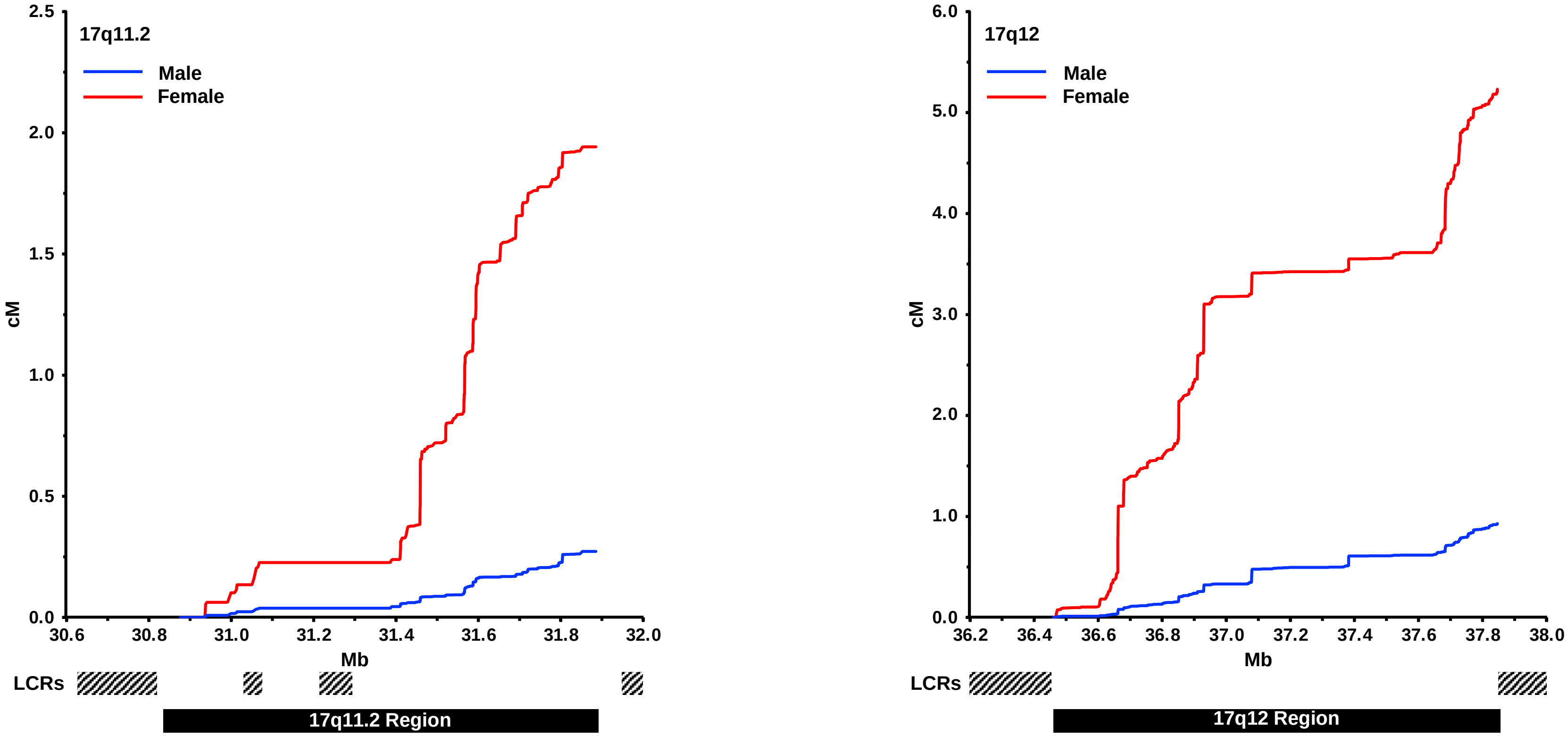


Figure S3-11. Raw deCODE recombination across 17q21.31 and 17q23.1q23.2 regions. Male (blue) and female (red) recombination across the canonical 17q21.31 and 17q23.1q23.2 regions (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio.

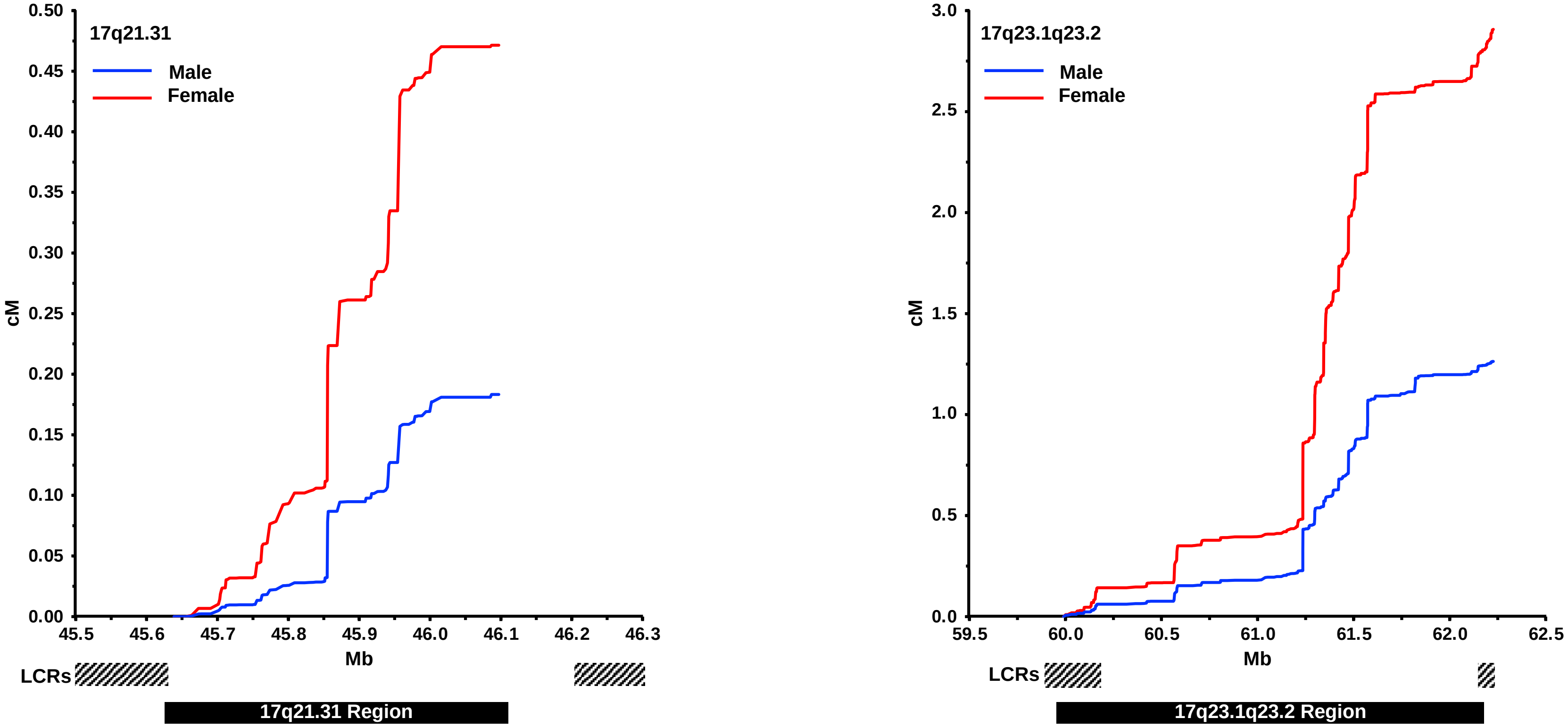


Figure S3-12. Raw deCODE recombination across 22q11.2 region. Male (blue) and female (red) recombination across the canonical 22q11.2 region (black bar). Location of flanking LCRs pulled from UCSC Genome Browser; hg38 (hatched bars). LCRs demarking different 22q11.2 intervals are denoted with letters (A-D) X-axis is position along the chromosome in Mb. Y-axis is the scaled probability of recombination (cM) across the interval. The curves summarize the rate of increase in probability of recombination as over the interval. The ratio of the right-most y-values of the male and female curves roughly equals the male-to-female recombination rate ratio.

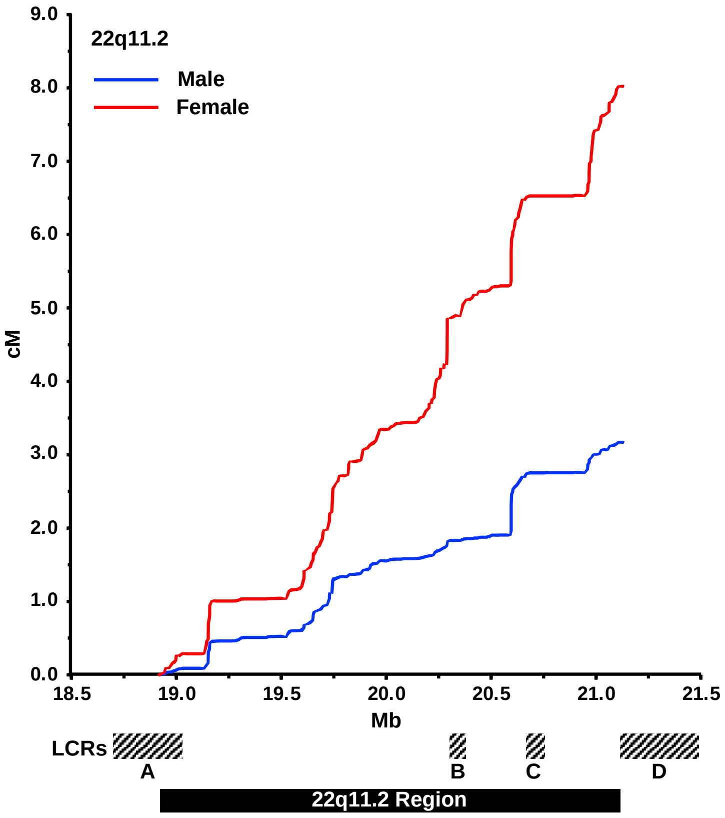


Figure S3-13. Logistic regression for deletions. Estimated (black curve) and observed paternal origin proportions for 1,913 deletions from 22 loci are shown. Curated parent of origin data from are collapsed by loci into single data points; plotted recombination rate ratios are the average of the metric for all CNVs within the data point. Data point size and color correspond to the number of CNVs collapsed into the data point. Recombination rates predict parent of origin for deletions mediated by NAHR. p=8.88x10-14, β=0.6721, CI95%=(0.5009,0.8546).

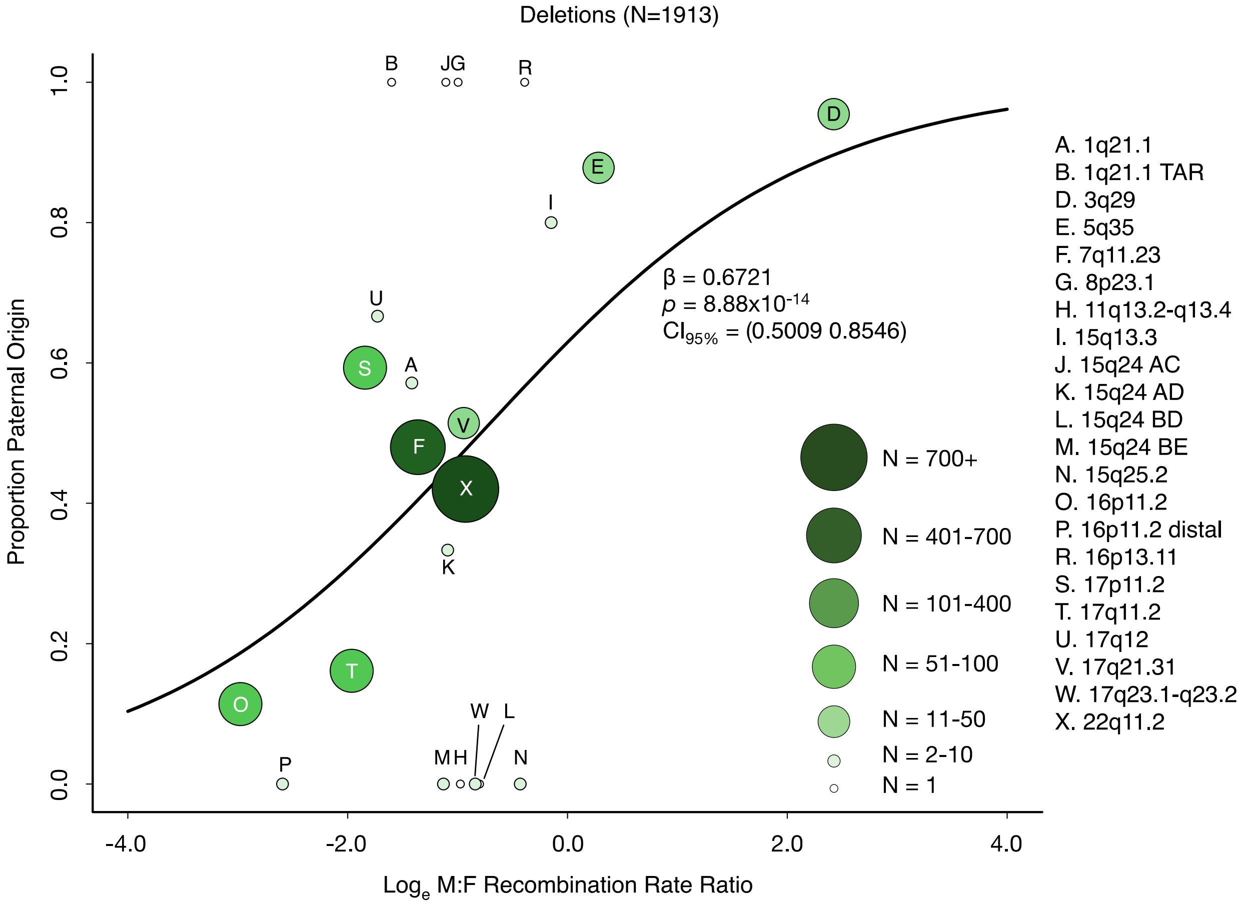


Figure S3-14. Logistic regression for duplications. Estimated (black curve) and observed paternal origin proportions for 64 duplications from 11 loci are shown. Curated parent of origin data from are collapsed by loci into single data points; plotted recombination rate ratios are the average of the metric for all CNVs within the data point. Data point size and color correspond to the number of CNVs collapsed into the data point. Recombination rates predict parent of origin for deletions mediated by NAHR. *p*=0.02, β=0.8304, $CI_{95\%}$=(0.1508,1.6017).

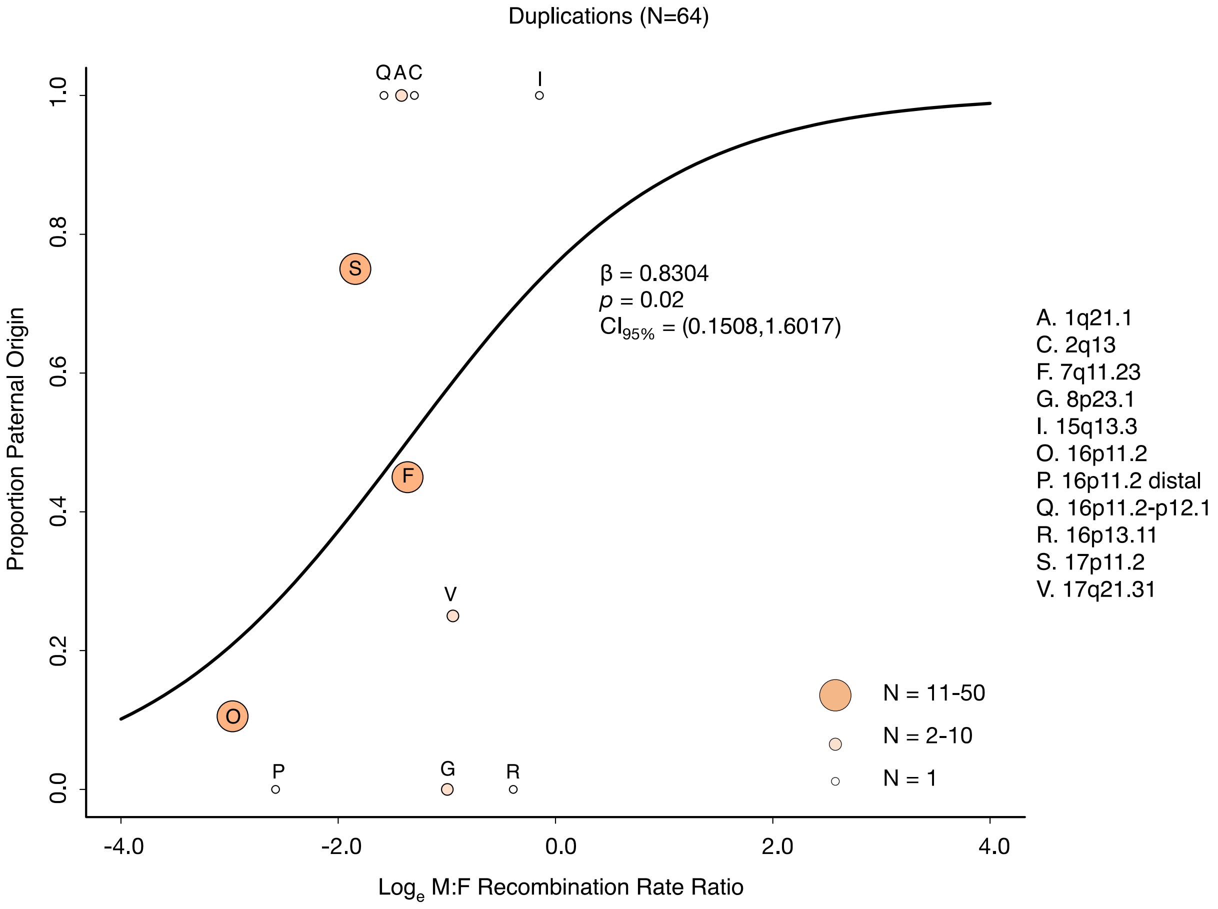
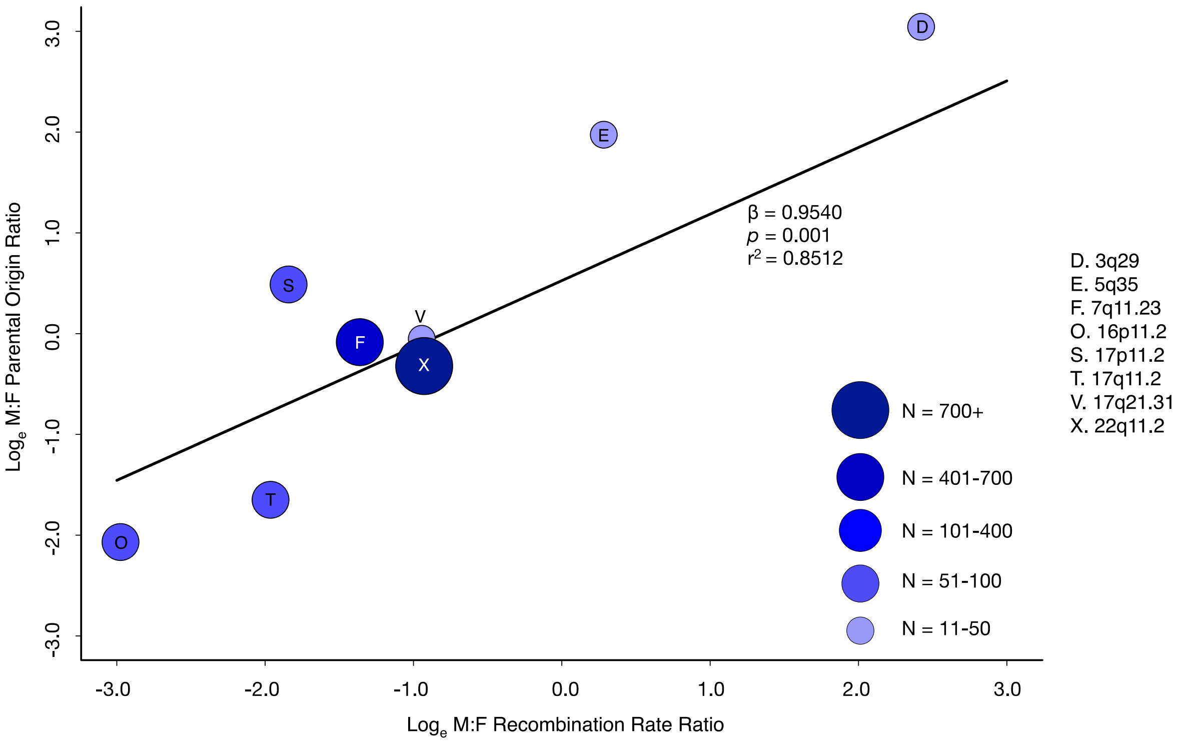


Figure S15. Linear regression with combined CNV parent of origin data. Log_e-transformed male to female parental origin ratio regressed on log_e-transformed average male to female recombination rate ratio. Curated parent of origin data from loci with ≥10 samples are collapsed by loci into single data points; plotted recombination rate ratios are the average of the metric for all CNVs within the data point. Data point size and color correspond to the number of CNVs collapsed into the data point. Recombination rates are associated with male-to-female parental origin ratios for CNVs mediated by NAHR (multiple r^2=0.8512, *p*=0.001, β=0.9540, $CI_{95\%}$=(0.5555,1.3525)). This estimate is not influenced by any particular data point as demonstrated by a sensitivity analysis (Table S3-7).



References

1. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17(4):224-38. Epub 2016/03/01. doi: 10.1038/nrg.2015.25. PubMed PMID: 26924765; PMCID: PMC4827625.

2. Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, Goldstein A, et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N Engl J Med. 2012;367(14):1321-31. Epub 2012/09/14. doi: 10.1056/NEJMoa1200395. PubMed PMID: 22970919; PMCID: PMC3494411.

3. Harel T, Lupski JR. Genomic disorders 20 years on-mechanisms for clinical manifestations. Clin Genet. 2018;93(3):439-49. Epub 2017/09/28. doi: 10.1111/cge.13146. PubMed PMID: 28950406.

4. Mefford HC. Genotype to phenotype-discovery and characterization of novel genomic disorders in a "genotype-first" era. Genet Med. 2009;11(12):836-42. Epub 2009/12/17. doi: 10.1097/GIM.0b013e3181c175d2. PubMed PMID: 20010361.

5. Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. Annu Rev Genet. 2011;45:203-26. Epub 2011/08/23. doi: 10.1146/annurev-genet-102209-163544. PubMed PMID: 21854229; PMCID: PMC6662611.

6. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85-97. Epub 2006/01/19. doi: 10.1038/nrg1767. PubMed PMID: 16418744.

7. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, Bosco P, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet. 2014;46(10):1063-71. Epub 2014/09/15. doi: 10.1038/ng.3092. PubMed PMID: 25217958; PMCID: PMC4177294.

8. Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet Med. 2011;13(9):777-84. Epub 2011/08/17. doi: 10.1097/GIM.0b013e31822c79f9. PubMed PMID: 21844811; PMCID: PMC3661946.

9. Lifton RP, Dluhy RG, Powers M, Rich GM, Cook S, Ulick S, et al. A chimaeric 11 beta-hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and

human hypertension. Nature. 1992;355(6357):262-5. Epub 1992/01/16. doi: 10.1038/355262a0. PubMed PMID: 1731223.

10. Mulle JG, Dodd AF, McGrath JA, Wolyniec PS, Mitchell AA, Shetty AC, et al. Microdeletions of 3q29 confer high risk for schizophrenia. Am J Hum Genet. 2010;87(2):229-36. Epub 2010/08/10. doi: 10.1016/j.ajhg.2010.07.013. PubMed PMID: 20691406; PMCID: PMC2917706.

11. Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, Vital A, et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat Genet. 2006;38(1):24-6. Epub 2005/12/22. doi: 10.1038/ng1718. PubMed PMID: 16369530.

12. Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature. 2010;463(7281):671-5. Epub 2010/02/05. doi: 10.1038/nature08727. PubMed PMID: 20130649; PMCID: PMC2880448.

13. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet. 2014;94(5):677-94. Epub 2014/04/29. doi: 10.1016/j.ajhg.2014.03.018. PubMed PMID: 24768552; PMCID: PMC4067558.

14. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. Science. 2007;316(5823):445-9. Epub 2007/03/17. doi: 10.1126/science.1138659. PubMed PMID: 17363630; PMCID: PMC2993504.

15. Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet. 2017;49(1):27-35. Epub 2016/11/22. doi: 10.1038/ng.3725. PubMed PMID: 27869829; PMCID: PMC5737772.

16. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science. 2008;320(5875):539-43. Epub 2008/03/29. doi: 10.1126/science.1155174. PubMed PMID: 18369103.

17. Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, Vu TH, et al. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. PLoS Genet. 2011;7(11):e1002334. Epub 2011/11/22. doi: 10.1371/journal.pgen.1002334. PubMed PMID: 22102821; PMCID: PMC3213131.

18. Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, et al. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. Genome Res. 2013;23(9):1395-409. Epub 2013/05/10. doi: 10.1101/gr.152454.112. PubMed PMID: 23657883; PMCID: PMC3759717.

19. Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, et al. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet. 2006;38(9):1038-42. Epub 2006/08/15. doi: 10.1038/ng1862. PubMed PMID: 16906162.

20. Liu P, Carvalho CM, Hastings PJ, Lupski JR. Mechanisms for recurrent and complex human genomic rearrangements. Curr Opin Genet Dev. 2012;22(3):211-20. Epub 2012/03/24. doi: 10.1016/j.gde.2012.02.012. PubMed PMID: 22440479; PMCID: PMC3378805.

21. Liu P, Lacaria M, Zhang F, Withers M, Hastings PJ, Lupski JR. Frequency of nonallelic homologous recombination is correlated with length of homology: evidence that ectopic synapsis precedes ectopic crossing-over. Am J Hum Genet. 2011;89(4):580-8. Epub 2011/10/11. doi: 10.1016/j.ajhg.2011.09.009. PubMed PMID: 21981782; PMCID: PMC3188830.

22. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437-55. Epub 2010/01/12. doi: 10.1146/annurev-med-100708-204735. PubMed PMID: 20059347.

23. Marques-Bonet T, Eichler EE. The evolution of human segmental duplications and the core duplicon hypothesis. Cold Spring Harb Symp Quant Biol. 2009;74:355-62. Epub 2009/09/01. doi: 10.1101/sqb.2009.74.011. PubMed PMID: 19717539; PMCID: PMC4114149.

24. Hobart HH, Morris CA, Mervis CB, Pani AM, Kistler DJ, Rios CM, et al. Inversion of the Williams syndrome region is a common polymorphism found more frequently in parents of children with Williams syndrome. Am J Med Genet C Semin Med Genet. 2010;154C(2):220-8. Epub 2010/04/29. doi: 10.1002/ajmg.c.30258. PubMed PMID: 20425783; PMCID: PMC2946898.

25. Koolen DA, Vissers LE, Pfundt R, de Leeuw N, Knight SJ, Regan R, et al. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. Nat Genet. 2006;38(9):999-1001. Epub 2006/08/15. doi: 10.1038/ng1853. PubMed PMID: 16906164.

26. Duyzend MH, Nuttle X, Coe BP, Baker C, Nickerson DA, Bernier R, et al. Maternal Modifiers and Parent-of-Origin Bias of the Autism-Associated 16p11.2 CNV. Am J Hum Genet. 2016;98(1):45-57. Epub 2016/01/11. doi: 10.1016/j.ajhg.2015.11.017. PubMed PMID: 26749307; PMCID: PMC4716684.

27. Lazaro C, Gaona A, Ainsworth P, Tenconi R, Vidaud D, Kruyer H, et al. Sex differences in mutational rate and mutational mechanism in the NF1 gene in neurofibromatosis type 1 patients. Hum Genet. 1996;98(6):696-9. Epub 1996/12/01. doi: 10.1007/s004390050287. PubMed PMID: 8931703.

28. Neuhausler L, Summerer A, Cooper DN, Mautner VF, Kehrer-Sawatzki H. Pronounced maternal parent-of-origin bias for type-1 NF1 microdeletions. Hum Genet. 2018;137(5):365-73. Epub 2018/05/08. doi: 10.1007/s00439-018-1888-x. PubMed PMID: 29730711.

29. Upadhyaya M, Ruggieri M, Maynard J, Osborn M, Hartog C, Mudd S, et al. Gross deletions of the neurofibromatosis type 1 (NF1) gene are predominantly of maternal origin and commonly associated with a learning disability, dysmorphic features and developmental delay. Hum Genet. 1998;102(5):591-7. Epub 1998/07/08. doi: 10.1007/s004390050746. PubMed PMID: 9654211.

30. Delio M, Guo T, McDonald-McGinn DM, Zackai E, Herman S, Kaminetzky M, et al. Enhanced maternal origin of the 22q11.2 deletion in velocardiofacial and DiGeorge syndromes. Am J Hum Genet. 2013;92(3):439-47. Epub 2013/03/05. doi: 10.1016/j.ajhg.2013.01.018. PubMed PMID: 23453669; PMCID: PMC3591861.

31. Miyake N, Kurotaki N, Sugawara H, Shimokawa O, Harada N, Kondoh T, et al. Preferential paternal origin of microdeletions caused by prezygotic chromosome or chromatid rearrangements in Sotos syndrome. Am J Hum Genet. 2003;72(5):1331-7. Epub 2003/04/11. doi: 10.1086/375166. PubMed PMID: 12687502; PMCID: PMC1180287.

32. Tatton-Brown K, Douglas J, Coleman K, Baujat G, Chandler K, Clarke A, et al. Multiple mechanisms are implicated in the generation of 5q35 microdeletions in Sotos syndrome. J Med

Genet. 2005;42(4):307-13. Epub 2005/04/05. doi: 10.1136/jmg.2004.027755. PubMed PMID: 15805156; PMCID: PMC1736029.

33. Page SL, Hawley RS. Chromosome choreography: the meiotic ballet. Science. 2003;301(5634):785-9. Epub 2003/08/09. doi: 10.1126/science.1086605. PubMed PMID: 12907787.

34. Hunt PA, Hassold TJ. Sex matters in meiosis. Science. 2002;296(5576):2181-3. Epub 2002/06/22. doi: 10.1126/science.1071907. PubMed PMID: 12077403.

35. Christiansen J, Dyck JD, Elyas BG, Lilley M, Bamforth JS, Hicks M, et al. Chromosome 1q21.1 contiguous gene deletion is associated with congenital heart disease. Circ Res. 2004;94(11):1429-35. Epub 2004/05/01. doi: 10.1161/01.RES.0000130528.72330.5c. PubMed PMID: 15117819.

36. Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K, et al. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N Engl J Med. 2008;359(16):1685-99. Epub 2008/09/12. doi: 10.1056/NEJMoa0805384. PubMed PMID: 18784092; PMCID: PMC2703742.

37. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron. 2015;87(6):1215-33. Epub 2015/09/25. doi: 10.1016/j.neuron.2015.09.016. PubMed PMID: 26402605; PMCID: PMC4624267.

38. Smajlagic D, Lavrichenko K, Berland S, Helgeland O, Knudsen GP, Vaudel M, et al. Population prevalence and inheritance pattern of recurrent CNVs associated with neurodevelopmental disorders in 12,252 newborns and their parents. Eur J Hum Genet. 2021;29(1):205-15. Epub 2020/08/12. doi: 10.1038/s41431-020-00707-7. PubMed PMID: 32778765; PMCID: PMC7852900.

39. Soemedi R, Wilson IJ, Bentham J, Darlay R, Topf A, Zelenika D, et al. Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. Am J Hum Genet. 2012;91(3):489-501. Epub 2012/09/04. doi: 10.1016/j.ajhg.2012.08.003. PubMed PMID: 22939634; PMCID: PMC3511986.

40. Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D, et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the

pathogenesis of schizophrenia. Mol Psychiatry. 2012;17(2):142-53. Epub 2011/11/16. doi: 10.1038/mp.2011.154. PubMed PMID: 22083728; PMCID: PMC3603134.

41. Malt EA, Juhasz K, Frengen A, Wangensteen T, Emilsen NM, Hansen B, et al. Neuropsychiatric phenotype in relation to gene variants in the hemizygous allele in 3q29 deletion carriers: A case series. Mol Genet Genomic Med. 2019;7(9):e889. Epub 2019/07/28. doi: 10.1002/mgg3.889. PubMed PMID: 31347308; PMCID: PMC6732294.

42. Quintero-Rivera F, Sharifi-Hannauer P, Martinez-Agosto JA. Autistic and psychiatric findings associated with the 3q29 microdeletion syndrome: case report and review. Am J Med Genet A. 2010;152A(10):2459-67. Epub 2010/09/11. doi: 10.1002/ajmg.a.33573. PubMed PMID: 20830797.

43. Baumer A, Dutly F, Balmer D, Riegel M, Tukel T, Krajewska-Walasek M, et al. High level of unequal meiotic crossovers at the origin of the 22q11. 2 and 7q11.23 deletions. Hum Mol Genet. 1998;7(5):887-94. Epub 1998/05/23. doi: 10.1093/hmg/7.5.887. PubMed PMID: 9536094.

44. Bayes M, Magano LF, Rivera N, Flores R, Perez Jurado LA. Mutational mechanisms of Williams-Beuren syndrome deletions. Am J Hum Genet. 2003;73(1):131-51. Epub 2003/06/11. doi: 10.1086/376565. PubMed PMID: 12796854; PMCID: PMC1180575.

45. Codina-Sola M, Costa-Roger M, Perez-Garcia D, Flores R, Palacios-Verdu MG, Cusco I, et al. Genetic factors contributing to autism spectrum disorder in Williams-Beuren syndrome. J Med Genet. 2019;56(12):801-8. Epub 2019/08/16. doi: 10.1136/jmedgenet-2019-106080. PubMed PMID: 31413120; PMCID: PMC6929708.

46. Depienne C, Heron D, Betancur C, Benyahia B, Trouillard O, Bouteiller D, et al. Autism, language delay and mental retardation in a patient with 7q11 duplication. J Med Genet. 2007;44(7):452-8. Epub 2007/04/03. doi: 10.1136/jmg.2006.047092. PubMed PMID: 17400790; PMCID: PMC1994965.

47. Dutra RL, Pieri Pde C, Teixeira AC, Honjo RS, Bertola DR, Kim CA. Detection of deletions at 7q11.23 in Williams-Beuren syndrome by polymorphic markers. Clinics (Sao Paulo). 2011;66(6):959-64. Epub 2011/08/03. doi: 10.1590/s1807-59322011000600007. PubMed PMID: 21808859; PMCID: PMC3129970.

48. Ghaffari M, Tahmasebi Birgani M, Kariminejad R, Saberi A. Genotype-phenotype correlation and the size of microdeletion or microduplication of 7q11.23 region in patients with

Williams-Beuren syndrome. Ann Hum Genet. 2018;82(6):469-76. Epub 2018/08/30. doi: 10.1111/ahg.12278. PubMed PMID: 30155880.

49. Masson J, Demily C, Chatron N, Labalme A, Rollat-Farnier PA, Schluth-Bolard C, et al. Molecular investigation, using chromosomal microarray and whole exome sequencing, of six patients affected by Williams Beuren syndrome and Autism Spectrum Disorder. Orphanet J Rare Dis. 2019;14(1):121. Epub 2019/06/04. doi: 10.1186/s13023-019-1094-5. PubMed PMID: 31151468; PMCID: PMC6545013.

50. Morris CA, Mervis CB, Paciorkowski AP, Abdul-Rahman O, Dugan SL, Rope AF, et al. 7q11.23 Duplication syndrome: Physical characteristics and natural history. Am J Med Genet A. 2015;167A(12):2916-35. Epub 2015/09/04. doi: 10.1002/ajmg.a.37340. PubMed PMID: 26333794; PMCID: PMC5005957.

51. Perez-Garcia D, Flores R, Brun-Gasca C, Perez-Jurado LA. Lateral preference in Williams-Beuren syndrome is associated with cognition and language. Eur Child Adolesc Psychiatry. 2015;24(9):1025-33. Epub 2014/11/29. doi: 10.1007/s00787-014-0652-6. PubMed PMID: 25431039.

52. Robinson WP, Waslynka J, Bernasconi F, Wang M, Clark S, Kotzot D, et al. Delineation of 7q11.2 deletions associated with Williams-Beuren syndrome and mapping of a repetitive sequence to within and to either side of the common deletion. Genomics. 1996;34(1):17-23. Epub 1996/05/15. doi: 10.1006/geno.1996.0237. PubMed PMID: 8661020.

53. Thomas NS, Durkie M, Potts G, Sandford R, Van Zyl B, Youings S, et al. Parental and chromosomal origins of microdeletion and duplication syndromes involving 7q11.23, 15q11-q13 and 22q11. Eur J Hum Genet. 2006;14(7):831-7. Epub 2006/04/18. doi: 10.1038/sj.ejhg.5201617. PubMed PMID: 16617304.

54. Wu YQ, Sutton VR, Nickerson E, Lupski JR, Potocki L, Korenberg JR, et al. Delineation of the common critical region in Williams syndrome and clinical correlation of growth, heart defects, ethnicity, and parental origin. Am J Med Genet. 1998;78(1):82-9. Epub 1998/06/24. doi: 10.1002/(sici)1096-8628(19980616)78:1<82::aid-ajmg17>3.0.co;2-k. PubMed PMID: 9637430.

55. Barber JC, Rosenfeld JA, Foulds N, Laird S, Bateman MS, Thomas NS, et al. 8p23.1 duplication syndrome; common, confirmed, and novel features in six further patients. Am J Med Genet A. 2013;161A(3):487-500. Epub 2013/01/25. doi: 10.1002/ajmg.a.35767. PubMed PMID: 23345203.

56. Shimokawa O, Miyake N, Yoshimura T, Sosonkina N, Harada N, Mizuguchi T, et al. Molecular characterization of del(8)(p23.1p23.1) in a case of congenital diaphragmatic hernia. Am J Med Genet A. 2005;136(1):49-51. Epub 2005/06/07. doi: 10.1002/ajmg.a.30778. PubMed PMID: 15937941.

57. Wischmeijer A, Magini P, Giorda R, Gnoli M, Ciccone R, Cecconi L, et al. Olfactory Receptor-Related Duplicons Mediate a Microdeletion at 11q13.2q13.4 Associated with a Syndromic Phenotype. Mol Syndromol. 2011;1(4):176-84. Epub 2011/03/05. doi: 10.1159/000322054. PubMed PMID: 21373257; PMCID: PMC3042121.

58. Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, et al. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet. 2008;40(3):322-8. Epub 2008/02/19. doi: 10.1038/ng.93. PubMed PMID: 18278044; PMCID: PMC2365467.

59. Chen CP, Wang LK, Chern SR, Wu PS, Chen SW, Wu FT, et al. Prenatal diagnosis and molecular cytogenetic characterization of a chromosome 15q24 microdeletion. Taiwan J Obstet Gynecol. 2020;59(3):432-6. Epub 2020/05/18. doi: 10.1016/j.tjog.2020.03.017. PubMed PMID: 32416893.

60. Gao X, Gotway G, Rathjen K, Johnston C, Sparagana S, Wise CA. Genomic Analyses of Patients With Unexplained Early-Onset Scoliosis. Spine Deform. 2014;2(5):324-32. Epub 2014/09/01. doi: 10.1016/j.jspd.2014.04.014. PubMed PMID: 27927329; PMCID: PMC4228381.

61. Huynh MT, Lambert AS, Tosca L, Petit F, Philippe C, Parisot F, et al. 15q24.1 BP4-BP1 microdeletion unmasking paternally inherited functional polymorphisms combined with distal 15q24.2q24.3 duplication in a patient with epilepsy, psychomotor delay, overweight, ventricular arrhythmia. Eur J Med Genet. 2018;61(8):459-64. Epub 2018/03/20. doi: 10.1016/j.ejmg.2018.03.005. PubMed PMID: 29549028.

62. McInnes LA, Nakamine A, Pilorge M, Brandt T, Jimenez Gonzalez P, Fallas M, et al. A large-scale survey of the novel 15q24 microdeletion syndrome in autism spectrum disorders identifies an atypical deletion that narrows the critical region. Mol Autism. 2010;1(1):5. Epub 2010/08/04. doi: 10.1186/2040-2392-1-5. PubMed PMID: 20678247; PMCID: PMC2907565.

63. Mefford HC, Rosenfeld JA, Shur N, Slavotinek AM, Cox VA, Hennekam RC, et al. Further clinical and molecular delineation of the 15q24 microdeletion syndrome. J Med Genet.

2012;49(2):110-8. Epub 2011/12/20. doi: 10.1136/jmedgenet-2011-100499. PubMed PMID: 22180641; PMCID: PMC3261729.

64. Sharp AJ, Selzer RR, Veltman JA, Gimelli S, Gimelli G, Striano P, et al. Characterization of a recurrent 15q24 microdeletion syndrome. Hum Mol Genet. 2007;16(5):567-72. Epub 2007/03/16. doi: 10.1093/hmg/ddm016. PubMed PMID: 17360722.

65. Burgess T, Brown NJ, Stark Z, Bruno DL, Oertel R, Chong B, et al. Characterization of core clinical phenotypes associated with recurrent proximal 15q25.2 microdeletions. Am J Med Genet A. 2014;164A(1):77-86. Epub 2013/12/20. doi: 10.1002/ajmg.a.36203. PubMed PMID: 24352913.

66. Palumbo O, Palumbo P, Palladino T, Stallone R, Miroballo M, Piemontese MR, et al. An emerging phenotype of interstitial 15q25.2 microdeletions: clinical report and review. Am J Med Genet A. 2012;158A(12):3182-9. Epub 2012/11/21. doi: 10.1002/ajmg.a.35631. PubMed PMID: 23166063.

67. Wagenstaller J, Spranger S, Lorenz-Depiereux B, Kazmierczak B, Nathrath M, Wahl D, et al. Copy-number variations measured by single-nucleotide-polymorphism oligonucleotide arrays in patients with mental retardation. Am J Hum Genet. 2007;81(4):768-79. Epub 2007/09/12. doi: 10.1086/521274. PubMed PMID: 17847001; PMCID: PMC2227926.

68. Egolf LE, Vaksman Z, Lopez G, Rokita JL, Modi A, Basta PV, et al. Germline 16p11.2 Microdeletion Predisposes to Neuroblastoma. Am J Hum Genet. 2019;105(3):658-68. Epub 2019/09/03. doi: 10.1016/j.ajhg.2019.07.020. PubMed PMID: 31474320; PMCID: PMC6731370.

69. Karolak JA, Gambin T, Honey EM, Slavik T, Popek E, Stankiewicz P. A de novo 2.2 Mb recurrent 17q23.1q23.2 deletion unmasks novel putative regulatory non-coding SNVs associated with lethal lung hypoplasia and pulmonary hypertension: a case report. BMC Med Genomics. 2020;13(1):34. Epub 2020/03/08. doi: 10.1186/s12920-020-0701-6. PubMed PMID: 32143628; PMCID: PMC7060516.

70. Redaelli S, Maitz S, Crosti F, Sala E, Villa N, Spaccini L, et al. Refining the Phenotype of Recurrent Rearrangements of Chromosome 16. Int J Mol Sci. 2019;20(5):1-17. Epub 2019/03/07. doi: 10.3390/ijms20051095. PubMed PMID: 30836598; PMCID: PMC6429492.

71. Tabet AC, Pilorge M, Delorme R, Amsellem F, Pinard JM, Leboyer M, et al. Autism multiplex family with 16p11.2p12.2 microduplication syndrome in monozygotic twins and distal

16p11.2 deletion in their brother. Eur J Hum Genet. 2012;20(5):540-6. Epub 2012/01/12. doi: 10.1038/ejhg.2011.244. PubMed PMID: 22234155; PMCID: PMC3330222.

72. Greenberg F, Guzzetta V, Montes de Oca-Luna R, Magenis RE, Smith AC, Richter SF, et al. Molecular analysis of the Smith-Magenis syndrome: a possible contiguous-gene syndrome associated with del(17)(p11.2). Am J Hum Genet. 1991;49(6):1207-18. Epub 1991/12/01. PubMed PMID: 1746552; PMCID: PMC1686451.

73. Nakamine A, Ouchanov L, Jimenez P, Manghi ER, Esquivel M, Monge S, et al. Duplication of 17(p11.2p11.2) in a male child with autism and severe language delay. Am J Med Genet A. 2008;146A(5):636-43. Epub 2007/03/06. doi: 10.1002/ajmg.a.31636. PubMed PMID: 17334992.

74. Potocki L, Shaw CJ, Stankiewicz P, Lupski JR. Variability in clinical phenotype despite common chromosomal deletion in Smith-Magenis syndrome [del(17)(p11.2p11.2)]. Genet Med. 2003;5(6):430-4. Epub 2003/11/14. doi: 10.1097/01.gim.0000095625.14160.ab. PubMed PMID: 14614393.

75. Shaw CJ, Bi W, Lupski JR. Genetic proof of unequal meiotic crossovers in reciprocal deletion and duplication of 17p11.2. Am J Hum Genet. 2002;71(5):1072-81. Epub 2002/10/11. doi: 10.1086/344346. PubMed PMID: 12375235; PMCID: PMC420000.

76. Yang SP, Bidichandani SI, Figuera LE, Juyal RC, Saxon PJ, Baldini A, et al. Molecular analysis of deletion (17)(p11.2p11.2) in a family segregating a 17p paracentric inversion: implications for carriers of paracentric inversions. Am J Hum Genet. 1997;60(5):1184-93. Epub 1997/05/01. PubMed PMID: 9150166; PMCID: PMC1712444.

77. Steinmann K, Kluwe L, Cooper DN, Brems H, De Raedt T, Legius E, et al. Copy number variations in the NF1 gene region are infrequent and do not predispose to recurrent type-1 deletions. Eur J Hum Genet. 2008;16(5):572-80. Epub 2008/01/24. doi: 10.1038/sj.ejhg.5202002. PubMed PMID: 18212816.

78. Palumbo P, Antona V, Palumbo O, Piccione M, Nardello R, Fontana A, et al. Variable phenotype in 17q12 microdeletions: clinical and molecular characterization of a new case. Gene. 2014;538(2):373-8. Epub 2014/02/04. doi: 10.1016/j.gene.2014.01.050. PubMed PMID: 24487052.

79. Digilio MC, Bernardini L, Capolino R, Digilio M, Dentici ML, Novelli A, et al. Hypopigmented skin patches in 17q21.31 microdeletion syndrome: expanding the spectrum of

cutaneous findings. Clin Dysmorphol. 2014;23(1):32-4. Epub 2013/12/05. doi: 10.1097/MCD.0000000000000019. PubMed PMID: 24300293.

80. Dubourg C, Sanlaville D, Doco-Fenzy M, Le Caignec C, Missirian C, Jaillard S, et al. Clinical and molecular characterization of 17q21.31 microdeletion syndrome in 14 French patients with mental retardation. Eur J Med Genet. 2011;54(2):144-51. Epub 2010/11/26. doi: 10.1016/j.ejmg.2010.11.003. PubMed PMID: 21094706.

81. Grisart B, Willatt L, Destree A, Fryns JP, Rack K, de Ravel T, et al. 17q21.31 microduplication patients are characterised by behavioural problems and poor social interaction. J Med Genet. 2009;46(8):524-30. Epub 2009/06/09. doi: 10.1136/jmg.2008.065367. PubMed PMID: 19502243.

82. Kirchhoff M, Bisgaard AM, Duno M, Hansen FJ, Schwartz M. A 17q21.31 microduplication, reciprocal to the newly described 17q21.31 microdeletion, in a girl with severe psychomotor developmental delay and dysmorphic craniofacial features. Eur J Med Genet. 2007;50(4):256-63. Epub 2007/06/20. doi: 10.1016/j.ejmg.2007.05.001. PubMed PMID: 17576104.

83. Koolen DA, Sharp AJ, Hurst JA, Firth HV, Knight SJ, Goldenberg A, et al. Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. J Med Genet. 2008;45(11):710-20. Epub 2008/07/17. doi: 10.1136/jmg.2008.058701. PubMed PMID: 18628315; PMCID: PMC3071570.

84. Vlckova M, Hancarova M, Drabova J, Slamova Z, Koudova M, Alanova R, et al. Monozygotic twins with 17q21.31 microdeletion syndrome. Twin Res Hum Genet. 2014;17(5):405-10. Epub 2014/06/10. doi: 10.1017/thg.2014.29. PubMed PMID: 24909117.

85. Karolak JA, Vincent M, Deutsch G, Gambin T, Cogne B, Pichon O, et al. Complex Compound Inheritance of Lethal Lung Developmental Disorders Due to Disruption of the TBX-FGF Pathway. Am J Hum Genet. 2019;104(2):213-28. Epub 2019/01/15. doi: 10.1016/j.ajhg.2018.12.010. PubMed PMID: 30639323; PMCID: PMC6369446.

86. Bassett AS, Marshall CR, Lionel AC, Chow EW, Scherer SW. Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. Hum Mol Genet. 2008;17(24):4045-53. Epub 2008/09/23. doi: 10.1093/hmg/ddn307. PubMed PMID: 18806272; PMCID: PMC2638574.

87. Baumer A, Riegel M, Schinzel A. Non-random asynchronous replication at 22q11.2 favours unequal meiotic crossovers leading to the human 22q11.2 deletion. J Med Genet. 2004;41(6):413-20. Epub 2004/06/03. doi: 10.1136/jmg.2003.016352. PubMed PMID: 15173225; PMCID: PMC1735820.

88. Brondum-Nielsen K, Christensen K. Chromosome 22q11 deletion and other chromosome aberrations in cases with cleft palate, congenital heart defects and/or mental disability. A survey based on the Danish Facial Cleft Register. Clin Genet. 1996;50(3):116-20. Epub 1996/09/01. doi: 10.1111/j.1399-0004.1996.tb02364.x. PubMed PMID: 8946108.

89. Chakraborty D, Bernal AJ, Schoch K, Howard TD, Ip EH, Hooper SR, et al. Dysregulation of DGCR6 and DGCR6L: psychopathological outcomes in chromosome 22q11.2 deletion syndrome. Transl Psychiatry. 2012;2:e105. Epub 2012/07/27. doi: 10.1038/tp.2012.31. PubMed PMID: 22832905; PMCID: PMC3337078.

90. Demczuk S, Levy A, Aubry M, Croquette MF, Philip N, Prieur M, et al. Excess of deletions of maternal origin in the DiGeorge/velo-cardio-facial syndromes. A study of 22 new patients and review of the literature. Hum Genet. 1995;96(1):9-13. Epub 1995/07/01. doi: 10.1007/BF00214179. PubMed PMID: 7607662.

91. Fokstuen S, Arbenz U, Artan S, Dutly F, Bauersfeld U, Brecevic L, et al. 22q11.2 deletions in a series of patients with non-selective congenital heart defects: incidence, type of defects and parental origin. Clin Genet. 1998;53(1):63-9. Epub 1998/04/29. doi: 10.1034/j.1399-0004.1998.531530113.x. PubMed PMID: 9550365.

92. Glaser B, Mumme DL, Blasey C, Morris MA, Dahoun SP, Antonarakis SE, et al. Language skills in children with velocardiofacial syndrome (deletion 22q11.2). J Pediatr. 2002;140(6):753-8. Epub 2002/06/20. doi: 10.1067/mpd.2002.124774. PubMed PMID: 12072882.

93. Guo T, Diacou A, Nomaru H, McDonald-McGinn DM, Hestand M, Demaerel W, et al. Deletion size analysis of 1680 22q11.2DS subjects identifies a new recombination hotspot on chromosome 22q11.2. Hum Mol Genet. 2018;27(7):1150-63. Epub 2018/01/24. doi: 10.1093/hmg/ddy028. PubMed PMID: 29361080; PMCID: PMC6059186.

94. Michaelovsky E, Gothelf D, Korostishevsky M, Frisch A, Burg M, Carmel M, et al. Association between a common haplotype in the COMT gene region and psychiatric disorders in

individuals with 22q11.2DS. Int J Neuropsychopharmacol. 2008;11(3):351-63. Epub 2007/10/24. doi: 10.1017/S1461145707008085. PubMed PMID: 17949513.

95. Molck MC, Vieira TP, Simioni M, Sgardioli IC, dos Santos AP, Xavier AC, et al. Distal 22q11.2 microduplication combined with typical 22q11.2 proximal deletion: a case report. Am J Med Genet A. 2015;167A(1):215-20. Epub 2014/11/02. doi: 10.1002/ajmg.a.36809. PubMed PMID: 25358462.

96. Morrow B, Goldberg R, Carlson C, Das Gupta R, Sirotkin H, Collins J, et al. Molecular definition of the 22q11 deletions in velo-cardio-facial syndrome. Am J Hum Genet. 1995;56(6):1391-403. Epub 1995/06/01. PubMed PMID: 7762562; PMCID: PMC1801093.

97. Saitta SC, Harris SE, Gaeth AP, Driscoll DA, McDonald-McGinn DM, Maisenbacher MK, et al. Aberrant interchromosomal exchanges are the predominant cause of the 22q11.2 deletion. Hum Mol Genet. 2004;13(4):417-28. Epub 2003/12/19. doi: 10.1093/hmg/ddh041. PubMed PMID: 14681306; PMCID: PMC2836129.

98. Saitta SC, Harris SE, McDonald-McGinn DM, Emanuel BS, Tonnesen MK, Zackai EH, et al. Independent de novo 22q11.2 deletions in first cousins with DiGeorge/velocardiofacial syndrome. Am J Med Genet A. 2004;124A(3):313-7. Epub 2004/01/07. doi: 10.1002/ajmg.a.20421. PubMed PMID: 14708107; PMCID: PMC2811370.

99. Sandrin-Garcia P, Abramides DV, Martelli LR, Ramos ES, Richieri-Costa A, Passos GA. Typical phenotypic spectrum of velocardiofacial syndrome occurs independently of deletion size in chromosome 22q11.2. Mol Cell Biochem. 2007;303(1-2):9-17. Epub 2007/04/12. doi: 10.1007/s11010-007-9450-5. PubMed PMID: 17426930.

100. Sandrin-Garcia P, Macedo C, Martelli LR, Ramos ES, Guion-Almeida ML, Richieri-Costa A, et al. Recurrent 22q11.2 deletion in a sibship suggestive of parental germline mosaicism in velocardiofacial syndrome. Clin Genet. 2002;61(5):380-3. Epub 2002/06/26. doi: 10.1034/j.1399-0004.2002.610511.x. PubMed PMID: 12081724.

101. Vervoort L, Demaerel W, Rengifo LY, Odrzywolski A, Vergaelen E, Hestand MS, et al. Atypical chromosome 22q11.2 deletions are complex rearrangements and have different mechanistic origins. Hum Mol Genet. 2019;28(22):3724-33. Epub 2019/12/31. doi: 10.1093/hmg/ddz166. PubMed PMID: 31884517; PMCID: PMC6935389.

102. Vittorini S, Sacchelli M, Iascone MR, Collavoli A, Storti S, Giusti A, et al. Molecular characterization of chromosome 22 deletions by short tandem repeat polymorphism (STRP) in

patients with conotruncal heart defects. Clin Chem Lab Med. 2001;39(12):1249-58. Epub 2002/01/19. doi: 10.1515/CCLM.2001.201. PubMed PMID: 11798086.

103. Murphy MM, Lindsey Burrell T, Cubells JF, Espana RA, Gambello MJ, Goines KCB, et al. Study protocol for The Emory 3q29 Project: evaluation of neurodevelopmental, psychiatric, and medical symptoms in 3q29 deletion syndrome. BMC Psychiatry. 2018;18(1):183. Epub 2018/06/10. doi: 10.1186/s12888-018-1760-5. PubMed PMID: 29884173; PMCID: PMC5994080.

104. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. Epub 2015/02/28. doi: 10.1186/s13742-015-0047-8. PubMed PMID: 25722852; PMCID: PMC4342193.

105. Team RC. R: A Language and Environment for Statistical Computing. Vienna, Austria2014.

106. Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science. 2019;363(6425):eaau1043. Epub 2019/01/27. doi: 10.1126/science.aau1043. PubMed PMID: 30679340.

107. McDonald-McGinn DM, Hain HS, Emanuel BS, Zackai EH. 22q11.2 Deletion Syndrome. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, Amemiya A, editors. GeneReviews((R)). Seattle (WA): University of Washington; 1993.

108. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet. 1998;63(3):861-9. Epub 1998/08/27. doi: 10.1086/302011. PubMed PMID: 9718341; PMCID: PMC1377399.

109. Sharp AJ, Cheng Z, Eichler EE. Structural variation of the human genome. Nature Reviews Genetics2006. p. 85-97.

110. Eichler EE. Masquerading repeats: paralogous pitfalls of the human genome. Genome Res. 1998;8(8):758-62. Epub 1998/09/02. doi: 10.1101/gr.8.8.758. PubMed PMID: 9724321.

111. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol.

2012;30(8):771-6. Epub 2012/07/17. doi: 10.1038/nbt.2303. PubMed PMID: 22797562; PMCID: PMC3817024.

112. Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. Nat Commun. 2019;10(1):1025. Epub 2019/03/06. doi: 10.1038/s41467-019-08992-7. PubMed PMID: 30833565; PMCID: PMC6399254.

113. Mak AC, Lai YY, Lam ET, Kwok TP, Leung AK, Poon A, et al. Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays. Genetics. 2016;202(1):351-62. Epub 2015/10/30. doi: 10.1534/genetics.115.183483. PubMed PMID: 26510793; PMCID: PMC4701098.

114. Demaerel W, Mostovoy Y, Yilmaz F, Vervoort L, Pastor S, Hestand MS, et al. The 22q11 low copy repeats are characterized by unprecedented size and structural variability. Genome Res. 2019;29(9):1389-401. Epub 2019/09/05. doi: 10.1101/gr.248682.119. PubMed PMID: 31481461; PMCID: PMC6724673.

115. Agarwal I, Przeworski M. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. Proc Natl Acad Sci U S A. 2019;116(36):17916-24. Epub 2019/08/21. doi: 10.1073/pnas.1900714116. PubMed PMID: 31427530; PMCID: PMC6731651.

116. Chowdhury R, Bois PR, Feingold E, Sherman SL, Cheung VG. Genetic analysis of variation in human meiotic recombination. PLoS Genet. 2009;5(9):e1000648. Epub 2009/09/19. doi: 10.1371/journal.pgen.1000648. PubMed PMID: 19763160; PMCID: PMC2730532.

117. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science. 2008;319(5868):1395-8. Epub 2008/02/02. doi: 10.1126/science.1151851. PubMed PMID: 18239090.

118. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. A high-resolution recombination map of the human genome. Nat Genet. 2002;31(3):241-7. Epub 2002/06/08. doi: 10.1038/ng917. PubMed PMID: 12053178.

119. Crown KN, Miller DE, Sekelsky J, Hawley RS. Local Inversion Heterozygosity Alters Recombination throughout the Genome. Curr Biol. 2018;28(18):2984-90 e3. Epub 2018/09/04. doi: 10.1016/j.cub.2018.07.004. PubMed PMID: 30174188; PMCID: PMC6156927.

120. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010;467(7319):1099-103. Epub 2010/10/29. doi: 10.1038/nature09525. PubMed PMID: 20981099.

121. Kong A, Thorleifsson G, Stefansson H, Masson G, Helgason A, Gudbjartsson DF, et al. Sequence variants in the RNF212 gene associate with genome-wide recombination rate. Science. 2008;319(5868):1398-401. Epub 2008/02/02. doi: 10.1126/science.1152422. PubMed PMID: 18239089.

122. Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, et al. Characterization of six human disease-associated inversion polymorphisms. Hum Mol Genet. 2009;18(14):2555-66. Epub 2009/04/23. doi: 10.1093/hmg/ddp187. PubMed PMID: 19383631; PMCID: PMC2701327.

123. Puig M, Casillas S, Villatoro S, Caceres M. Human inversions and their functional consequences. Brief Funct Genomics. 2015;14(5):369-79. Epub 2015/05/23. doi: 10.1093/bfgp/elv020. PubMed PMID: 25998059; PMCID: PMC4576756.

124. Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, et al. A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. Nat Genet. 2001;29(3):321-5. Epub 2001/10/31. doi: 10.1038/ng753. PubMed PMID: 11685205; PMCID: PMC2889916.

125. Rao PN, Li W, Vissers LE, Veltman JA, Ophoff RA. Recurrent inversion events at 17q21.31 microdeletion locus are linked to the MAPT H2 haplotype. Cytogenet Genome Res. 2010;129(4):275-9. Epub 2010/07/08. doi: 10.1159/000315901. PubMed PMID: 20606400; PMCID: PMC3202913.

126. Gimelli G, Pujana MA, Patricelli MG, Russo S, Giardino D, Larizza L, et al. Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. Hum Mol Genet. 2003;12(8):849-58. Epub 2003/04/02. doi: 10.1093/hmg/ddg101. PubMed PMID: 12668608.

127. Visser R, Shimokawa O, Harada N, Kinoshita A, Ohta T, Niikawa N, et al. Identification of a 3.0-kb major recombination hotspot in patients with Sotos syndrome who carry a common 1.9-Mb microdeletion. Am J Hum Genet. 2005;76(1):52-67. Epub 2004/12/08. doi: 10.1086/426950. PubMed PMID: 15580547; PMCID: PMC1196433.

128. Antonarakis SE, Rossiter JP, Young M, Horst J, de Moerloose P, Sommer SS, et al. Factor VIII gene inversions in severe hemophilia A: results of an international consortium study. Blood. 1995;86(6):2206-12. Epub 1995/09/15. doi: 10.1182/blood.V86.6.2206.bloodjournal8662206. PubMed PMID: 7662970.

129. Tam E, Young EJ, Morris CA, Marshall CR, Loo W, Scherer SW, et al. The common inversion of the Williams-Beuren syndrome region at 7q11.23 does not cause clinical symptoms. Am J Med Genet A. 2008;146A(14):1797-806. Epub 2008/06/17. doi: 10.1002/ajmg.a.32360. PubMed PMID: 18553513; PMCID: PMC2886033.

130. Zody MC, Jiang Z, Fung HC, Antonacci F, Hillier LW, Cardone MF, et al. Evolutionary toggling of the MAPT 17q21.31 inversion region. Nat Genet. 2008;40(9):1076-83. Epub 2009/01/24. doi: 10.1038/ng.193. PubMed PMID: 19165922; PMCID: PMC2684794.

131. Giner-Delgado C, Villatoro S, Lerga-Jaso J, Gaya-Vidal M, Oliva M, Castellano D, et al. Evolutionary and functional impact of common polymorphic inversions in the human genome. Nat Commun. 2019;10(1):4222. Epub 2019/09/19. doi: 10.1038/s41467-019-12173-x. PubMed PMID: 31530810; PMCID: PMC6748972.

132. Hehir-Kwa JY, Rodriguez-Santiago B, Vissers LE, de Leeuw N, Pfundt R, Buitelaar JK, et al. De novo copy number variants associated with intellectual disability have a paternal origin and age bias. J Med Genet. 2011;48(11):776-8. Epub 2011/10/05. doi: 10.1136/jmedgenet-2011-100147. PubMed PMID: 21969336.

133. Kloosterman WP, Francioli LC, Hormozdiari F, Marschall T, Hehir-Kwa JY, Abdellaoui A, et al. Characteristics of de novo structural changes in the human genome. Genome Res. 2015;25(6):792-801. Epub 2015/04/18. doi: 10.1101/gr.185041.114. PubMed PMID: 25883321; PMCID: PMC4448676.

134. Ma R, Deng L, Xia Y, Wei X, Cao Y, Guo R, et al. A clear bias in parental origin of de novo pathogenic CNVs related to intellectual disability, developmental delay and multiple congenital anomalies. Sci Rep. 2017;7:44446. Epub 2017/03/23. doi: 10.1038/srep44446. PubMed PMID: 28322228; PMCID: PMC5359547.

135. Lee JA, Carvalho CM, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell. 2007;131(7):1235-47. Epub 2007/12/28. doi: 10.1016/j.cell.2007.11.037. PubMed PMID: 18160035.

136. Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet. 2009;41(7):849-53. Epub 2009/06/23. doi: 10.1038/ng.399. PubMed PMID: 19543269; PMCID: PMC4461229.

137. Barton KS, Tabor HK, Starks H, Garrison NA, Laurino M, Burke W. Pathways from autism spectrum disorder diagnosis to genetic testing. Genet Med. 2018;20(7):737-44. Epub 2017/10/20. doi: 10.1038/gim.2017.166. PubMed PMID: 29048417; PMCID: PMC5908763.

138. Glassford MR, Purcell RH, Pass S, Murphy MM, Bassell GJ, Mulle JG. Caregiver perspectives on a diagnosis of 3q29 deletion. medRxiv. 2020:2020.09.21.20198770. doi: 10.1101/2020.09.21.20198770.

139. Phillips KA, Trosman JR, Deverka PA, Quinn B, Tunis S, Neumann PJ, et al. Insurance coverage for genomic tests. Science. 2018;360(6386):278-9. Epub 2018/04/21. doi: 10.1126/science.aas9268. PubMed PMID: 29674586; PMCID: PMC5991085.

140. Stiles D, Appelbaum PS. Cases in Precision Medicine: Concerns About Privacy and Discrimination After Genomic Sequencing. Ann Intern Med. 2019;170(10):717-21. Epub 2019/05/07. doi: 10.7326/M18-2666. PubMed PMID: 31060048; PMCID: PMC6715527.

141. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nat Genet. 2011;43(9):838-46. Epub 2011/08/16. doi: 10.1038/ng.909. PubMed PMID: 21841781; PMCID: PMC3171215.

142. Lopes J, Ravise N, Vandenberghe A, Palau F, Ionasescu V, Mayer M, et al. Fine mapping of de novo CMT1A and HNPP rearrangements within CMT1A-REPs evidences two distinct sex-dependent mechanisms and candidate sequences involved in recombination. Hum Mol Genet. 1998;7(1):141-8. Epub 1998/02/28. doi: 10.1093/hmg/7.1.141. PubMed PMID: 9384615.

143. Lopes J, Vandenberghe A, Tardieu S, Ionasescu V, Levy N, Wood N, et al. Sex-dependent rearrangements resulting in CMT1A and HNPP. Nat Genet. 1997;17(2):136-7. Epub 1997/11/05. doi: 10.1038/ng1097-136. PubMed PMID: 9326925.

144. Glassford MR, Rosenfeld JA, Freedman AA, Zwick ME, Mulle JG, Unique Rare Chromosome Disorder Support G. Novel features of 3q29 deletion syndrome: Results from the 3q29 registry. Am J Med Genet A. 2016;170A(4):999-1006. Epub 2016/01/08. doi: 10.1002/ajmg.a.37537. PubMed PMID: 26738761; PMCID: PMC4849199.

145. Johnston HR, Chopra P, Wingo TS, Patel V, International Consortium on B, Behavior in 22q11.2 Deletion S, et al. PEMapper and PECaller provide a simplified approach to whole-genome sequencing. Proc Natl Acad Sci U S A. 2017;114(10):E1923-E32. Epub 2017/02/23. doi: 10.1073/pnas.1618065114. PubMed PMID: 28223510; PMCID: PMC5347547.

146. Kotlar AV, Trevino CE, Zwick ME, Cutler DJ, Wingo TS. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. Genome Biol. 2018;19(1):14. Epub 2018/02/08. doi: 10.1186/s13059-018-1387-3. PubMed PMID: 29409527; PMCID: PMC5801807.

147. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93. Epub 2011/09/10. doi: 10.1093/bioinformatics/btr509. PubMed PMID: 21903627; PMCID: PMC3198575.

148. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015;372(23):2235-42. Epub 2015/05/28. doi: 10.1056/NEJMsr1406261. PubMed PMID: 26014595; PMCID: PMC4474187.

CHAPTER IV. Discussion

Trenell J. Mosley, Jennifer G. Mulle, Michael E. Zwick

Summary

Investigating rare genetic diseases (RGDs) and variation can expand our understanding of the genotype-phenotype map and provide insights into fundamental biological functions [1]. In addition, the knowledge gained from RGDs can provide information on common diseases and aid in developing interventions for patients suffering from both rare and common diseases [2,3]. Investigating RGDs requires understanding the genetic variation that can contribute to conditions, robust detection of that variation, and the means to annotate and interpret its effects in the appropriate biological context [4,5]. Advances in next-generation sequencing technologies and the public availability of informative genomics databases have helped speed the investigation of rare genetic diseases [2].

This work investigates two different classes of RGDs and the variation that contribute to them. In Chapter Two, we identified a putative pathogenic splicing single nucleotide variant for a rare primordial dwarfism disorder, variant POC1A-related (vPOC1A) syndrome. We demonstrated the molecular consequences of that variant using *in vitro* functional testing. Our data showed that the variant causes improper splicing of the message resulting in a loss of exon 9 in the *POC1A* messenger RNA; this is the first report of a noncoding splice variant associated with vPOC1A syndrome. Our finding expands the known allelic spectrum of vPOC1A syndrome and **S**hort stature, **O**nychodysplsia, **F**acial dysmorphisms, and Hyper**T**richosis (SOFT) syndrome. In addition, it adds to the growing body of work describing *POC1A* and its link to primordial dwarfism and insulin resistance.

In Chapter Three, we found that sex-specific patterns of meiotic recombination predict the parent of origin for rare CNVs at genomic disorder loci, and predicted paternal origin frequencies for several CNV including those mapped to 15q13.3, distal 16p11.2, and 17q23.

Additionally, we provided strong evidence for the presence of a significant male bias for the parent of origin at the 3q29 locus. These results reveal meiotic recombination as a sizeable factor influencing the origin of pathogenic recurrent CNVs. Furthermore, it indicates the need to consider meiotic recombination when evaluating structural risk haplotypes for CNV formation. This study also has implications for genetic counseling of individuals with CNV disorders and their families. When combined with findings from Chapter Two, this work highlights the enormous impact of genetic and genomic variation on normal biological processes, like mRNA splicing and recombination, and the advantages of studying rare disease variation to unravel the intricacies of biology.

This work also demonstrates (1) the advantages of using next-generation sequencing technologies, particularly whole-genome sequencing (WGS) and associated informatics pipelines, (2) the need for validation and functional testing, and (3) the utility of publicly available data, databases, and resources. We uncovered the intronic variant in *POC1A* in chapter two using WGS; Whole-exome sequencing (WES) would have missed the noncoding disease-causing variant [6]. Using genomics databases such as gnomAD, and predictive genetics software, *i.e.,* CADD, and Human Splice Finder, we were able to easily filter and prioritize variants, allowing us to quickly identify a small subset of nine targets to interpret and investigate [7-10]. In addition to the noncoding variant we identified, two additional coding sequence variants were classified as potentially causal, both with links to insulin and growth phenotypes. While in some cases segregation analysis would have likely eliminated at least one of the variants, in this case the consanguineous nature of the family precluded this possibility. Functional testing of the two of the three variants, enabled us to correctly identify the actual causal variant [5]. In Chapter Three, public availability of parental origin data from published literature and male and female recombination map data contributed to our discovery of recombination as a primary predictor of parent of origin in LCR-mediated CNVs. Additionally, while the majority of parental origins for the3q29 deletions were determined using SNP genotyping, a subset was determined with WGS, which can also be leveraged in future studies of structural variation.

Future Directions

Despite significant advancements in rare genetic disease research in the past 20 years, there are over 1,000 Mendelian disorders with unknown molecular causes (OMIM: https://www.omim.org/statistics/geneMap). More work is needed to improve our ability to detect and interpret genetic variation contributing to RGDs. Combining WGS with other genome-wide omics such as transcriptomics, metabolomics, and epigenomics offers a path to increased molecular diagnoses for RGDs. This multi-omics approach to RGD research enhances our ability to interpret variants by creating a complete picture of the molecular effects of a genetic variant [11]. Ultimately, this can improve time to diagnosis. In a 2017 study, the combination of Illumina short-read genome sequencing with RNA-seq increased diagnostic yield by 10% [12] and in a separate study aided in discovering SVA retrotransposon insertion in a *TAF1* intron as the cause for X-linked Dystonia-Parkinsonism [13]. The Undiagnosed Diseases Network has also used multi-omics to assist in identifying and evaluating the causal genes for RGDs, resulting in an increased diagnostic yield [1,14]. Combining WES with metabolomics also aids in the identification of potential targets for therapeutics, as demonstrated for a patient with epileptic encephalopathy and dysmorphic features and mutations in *NANS*. Metabolomic analysis identified increased levels of the substrate of NANS in a 4-year-old patient, allowing Nacetylneuraminic acid supplementation to be recognized as a potential treatment [15]. Thus, a multi-omics approach has the potential to enhance all steps of RGD research—variant identification, variant prioritization, functional validation, and therapeutics development.

Advancements in sequencing technologies can allow us to assess “unreachable” regions of the genome. Current short-read technologies, although cost-effective, are unable to accurately capture repetitive and complex regions in the human genome, which are estimated to comprise 60% of the genome [16-18]. Developments in long-read sequencing platforms such as PacBio single-molecule real-time (SMRT) sequence and Nanopore sequencing have a demonstrated utility in interrogating these regions [19]. Compared to short reads, which range between 150-1,000 base pairs in read length, long-read sequencers can generate reads ranging anywhere from 8 to 200 kilobases in size [20]. This length can span repeats or complex regions and flank unique sequences, allowing accurate mapping of reads to the genome. As such long-read sequencing enhances detection and phasing of SVs, which are enriched in repetitive regions like segmental duplications, centromeres, and telomeres and are much more likely to affect gene expression and be associated with disease than SNVs [21-23]. Long read technologies can also better detect and diagnose repeat expansion disorders. For example, the D4Z4 repeat tract in facioscapulohumeral muscular dystrophy (FSHD) can be accurately differentiated from the homologous but benign allele on chromosome 10 with long-read sequencing as well as optical mapping [24,25]. Inversions have largely been invisible to short-read sequencing, and studies estimate more than 85% of insertions are missed with short-read sequencing [26,27]. Recent efforts to capture genetic diversity have shown the increased sensitivity for inversions and insertions long-read sequence data enables, especially when combined with multiple orthogonal analysis algorithms [28]. Complex rearrangements, like chromosome shattering, *i.e.*, chromothripsis events, are signatures of cancer but are also implicated in congenital disorders [29,30]. Cretu Stancu and colleagues recently demonstrated the ability to confidently detect and characterize chromothripsis events with Nanopore sequencing in patients with congenital disease [31].

Long-read sequencing can also phase over longer lengths of DNA and when combined with other technologies such as optical mapping increase our ability to de novo assemble genomes. Two groups recently demonstrated the use of Nanopore and PacBio long-read sequencing to construct the human 8 and X chromosomes from telomere to telomere [32,33]. While there are still challenges related to accuracy and application to diploid chromosomes, the implications of this feat and others include the generation of a *complete* human reference genome, and the potential to expand our collection of reference genomes to include additional diverse genomes [27,33]. In addition, the generation of reference-free assemblies of personal genomes also enhances prospects for precision medicine [34,35]. Altogether, usage of long-read sequencing technologies will increase our understanding of the standing human genetic variation, structural and sequence-based, that could underly RGDs.

Inclusion of individuals from diverse populations in research cohorts is needed to extend our understanding of RGDs, genetic variation, and its relation to disease. These populations contain variation that otherwise is not captured by the current reference genome or databases. Numerous studies have shown we are missing a large portion of human genetic diversity and that the construction of diverse genomes and databases are needed as our understanding of the full spectrum of variation is incomplete without them [27,36,37]. Rare genetic diseases are also more prevalent in communities of color. Yet, a majority of rare disease studies are conducted on samples of predominantly white participants [38].

The potential benefits of diversity in human genome studies are well-known and demonstrated: (1) a greater understanding of variation in relatively healthy versus affected individuals, (2) greater understanding of common private variation within different populations, (3) creation of more comprehensive benign and pathogenic variation databases, and (4) increased insights into human evolution, to name a few [22,36,39,40]. Yet, the field struggles with the underrepresentation of diverse participation in research [41-43]. Structural barriers such as study design and geographic location and cultural barriers such as language and lack of trust impede participation [42]. These are just a few of the obstacles that preclude the participation of diverse individuals in research cohorts and require ethical and inclusive approaches to communication with and recruitment of study participants from underrepresented communities [41,43,44]. As an example, community-based approaches to research have a positive effect on research participation. They focus on including communities as partners in research rather than as just subjects. Actively engaging community members can reduce misconceptions about the research process and address prominent concerns around research, such as privacy, discrimination, and individual and community impacts [43,44]. The Patient-Centered Outcomes Research Institute takes a similar approach that engages a range of stakeholders, including patients, caregivers, scientists, organizations, and clinicians, to guide research towards outcomes that are important to patients and their communities [45,46]. Diverse research teams can also enhance participation, as they are more likely to include scientists with the cultural competencies needed to ethically and effectively engage with diverse communities and can decrease suspicion and mistrust from community members [47,48]. Unfortunately, the genetics field suffers from a lack of diverse scientists in the workforce due to systemic biases and barriers [49]. Moving forward, strategic efforts, such as those proposed by the 2020 NHGRI action agenda, should be made to increase exposure, recruitment, and training of diverse groups in the genetics and genomics field [48,49]. These are just two examples of approaches to increase diverse participation in research. As we continue to explore and employ well-rounded and ethical strategies, our realization of a comprehensive understanding of population variation will ultimately lead to advancements in all areas of human genetics research while also allowing the benefits of RGD research and precision medicine to be realized by patients of ethnicities and backgrounds that have otherwise been excluded [50].

Conclusions

In conclusion, this work uncovered the influence of genetic and genomic variation on rare genetic diseases. We showed a rare primordial dwarfism disorder, vPOC1A syndrome, is caused by an intronic variant in *POC1A* and further delineate the allelic spectrum that contributes to *POC1A*-related disorders. Using publicly available parental origin and recombination map data, we determined that meiotic recombination significantly influences the parental origin of CNVs associated with disorders and added the 3q29 deletion to the growing number of CNV loci with parental origin biases. Taken together, this work demonstrates the benefits of investigating rare diseases to understand human biology.

References

1. Lee CE, Singleton KS, Wallin M, Faundez V. Rare Genetic Diseases: Nature's Experiments on Human Development. iScience. 2020;23(5):101123. Epub 2020/05/19. doi: 10.1016/j.isci.2020.101123. PubMed PMID: 32422592; PMCID: PMC7229282.

2. Fernandez-Marmiesse A, Gouveia S, Couce ML. NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. Curr Med Chem. 2018;25(3):404-32. Epub 2017/07/20. doi: 10.2174/0929867324666170718101946. PubMed PMID: 28721829; PMCID: PMC5815091.

3. Stoller JK. The Challenge of Rare Diseases. Chest. 2018;153(6):1309-14. Epub 2018/01/13. doi: 10.1016/j.chest.2017.12.018. PubMed PMID: 29325986.

4. Dawkins HJS, Draghia-Akli R, Lasko P, Lau LPL, Jonker AH, Cutillo CM, et al. Progress in Rare Diseases Research 2010-2016: An IRDiRC Perspective. Clin Transl Sci. 2018;11(1):11-20. Epub 2017/08/11. doi: 10.1111/cts.12501. PubMed PMID: 28796411; PMCID: PMC5759730.

5. Rodenburg RJ. The functional genomics laboratory: functional validation of genetic variants. J Inherit Metab Dis. 2018;41(3):297-307. Epub 2018/02/16. doi: 10.1007/s10545-018-0146-7. PubMed PMID: 29445992; PMCID: PMC5959958.

6. Bick D, Jones M, Taylor SL, Taft RJ, Belmont J. Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases. J Med Genet. 2019;56(12):783-91. Epub 2019/04/27. doi: 10.1136/jmedgenet-2019-106111. PubMed PMID: 31023718; PMCID: PMC6929710.

7. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009;37(9):e67. Epub 2009/04/03. doi: 10.1093/nar/gkp215. PubMed PMID: 19339519; PMCID: PMC2685110.

8. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43. Epub 2020/05/29. doi: 10.1038/s41586-020-2308-7. PubMed PMID: 32461654; PMCID: PMC7334197.

9. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet.

2014;46(3):310-5. Epub 2014/02/04. doi: 10.1038/ng.2892. PubMed PMID: 24487276; PMCID: PMC3992975.

10. Kotlar AV, Trevino CE, Zwick ME, Cutler DJ, Wingo TS. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. Genome Biol. 2018;19(1):14. Epub 2018/02/08. doi: 10.1186/s13059-018-1387-3. PubMed PMID: 29409527; PMCID: PMC5801807.

11. Labory J, Fierville M, Ait-El-Mkadem S, Bannwarth S, Paquis-Flucklinger V, Bottini S. Multi-Omics Approaches to Improve Mitochondrial Disease Diagnosis: Challenges, Advances, and Perspectives. Frontiers in Molecular Biosciences. 2020;7(327). doi: 10.3389/fmolb.2020.590842.

12. Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun. 2017;8:15824. Epub 2017/06/13. doi: 10.1038/ncomms15824. PubMed PMID: 28604674; PMCID: PMC5499207.

13. Aneichyk T, Hendriks WT, Yadav R, Shin D, Gao D, Vaine CA, et al. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. Cell. 2018;172(5):897-909 e21. Epub 2018/02/24. doi: 10.1016/j.cell.2018.02.011. PubMed PMID: 29474918; PMCID: PMC5831509.

14. Kyle JE, Stratton KG, Zink EM, Kim Y-M, Bloodsworth KJ, Monroe ME, et al. A resource of lipidomics and metabolomics data from individuals with undiagnosed diseases. Scientific Data. 2021;8(1):114. doi: 10.1038/s41597-021-00894-y.

15. Tarailo-Graovac M, Shyr C, Ross CJ, Horvath GA, Salvarinova R, Ye XC, et al. Exome Sequencing and the Management of Neurometabolic Disorders. New England Journal of Medicine. 2016;374(23):2246-55. doi: 10.1056/NEJMoa1515792. PubMed PMID: 27276562.

16. Chiara M, Pavesi G. Evaluation of Quality Assessment Protocols for High Throughput Genome Resequencing Data. Front Genet. 2017;8:94. Epub 2017/07/25. doi: 10.3389/fgene.2017.00094. PubMed PMID: 28736571; PMCID: PMC5500642.

17. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13(5):278-89. Epub 2015/11/07. doi: 10.1016/j.gpb.2015.08.002. PubMed PMID: 26542840; PMCID: PMC4678779.

18. Mitsuhashi S, Matsumoto N. Long-read sequencing for rare human genetic diseases. J Hum Genet. 2020;65(1):11-9. Epub 2019/09/29. doi: 10.1038/s10038-019-0671-8. PubMed PMID: 31558760.

19. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015;517(7536):608-11. Epub 2014/11/11. doi: 10.1038/nature13907. PubMed PMID: 25383537; PMCID: PMC4317254.

20. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333-51. Epub 2016/05/18. doi: 10.1038/nrg.2016.49. PubMed PMID: 27184599.

21. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural variation on human gene expression. Nat Genet. 2017;49(5):692-9. Epub 2017/04/04. doi: 10.1038/ng.3834. PubMed PMID: 28369037; PMCID: PMC5406250.

22. Collins RL, Brand H, Karczewski KJ, Zhao X, Alfoldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature. 2020;581(7809):444-51. Epub 2020/05/29. doi: 10.1038/s41586-020-2287-8. PubMed PMID: 32461652; PMCID: PMC7334194.

23. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, et al. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005;77(1):78-88. Epub 2005/05/27. doi: 10.1086/431652. PubMed PMID: 15918152; PMCID: PMC1226196.

24. Dai Y, Li P, Wang Z, Liang F, Yang F, Fang L, et al. Single-molecule optical mapping enables quantitative measurement of D4Z4 repeats in facioscapulohumeral muscular dystrophy (FSHD). J Med Genet. 2020;57(2):109-20. Epub 2019/09/12. doi: 10.1136/jmedgenet-2019-106078. PubMed PMID: 31506324; PMCID: PMC7029236.

25. Mitsuhashi S, Nakagawa S, Takahashi Ueda M, Imanishi T, Frith MC, Mitsuhashi H. Nanopore-based single molecule sequencing of the D4Z4 array responsible for facioscapulohumeral muscular dystrophy. Sci Rep. 2017;7(1):14789. Epub 2017/11/03. doi: 10.1038/s41598-017-13712-6. PubMed PMID: 29093467; PMCID: PMC5665936.

26. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation.

Science. 2021;372(6537). Epub 2021/02/27. doi: 10.1126/science.abf7117. PubMed PMID: 33632895; PMCID: PMC8026704.

27. Wong KHY, Ma W, Wei CY, Yeh EC, Lin WJ, Wang EHF, et al. Towards a reference genome that captures global genetic diversity. Nat Commun. 2020;11(1):5482. Epub 2020/11/01. doi: 10.1038/s41467-020-19311-w. PubMed PMID: 33127893; PMCID: PMC7599213.

28. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10(1):1784. Epub 2019/04/18. doi: 10.1038/s41467-018-08148-z. PubMed PMID: 30992455; PMCID: PMC6467913.

29. Colnaghi R, Carpenter G, Volker M, O'Driscoll M. The consequences of structural genomic alterations in humans: genomic disorders, genomic instability and cancer. Semin Cell Dev Biol. 2011;22(8):875-85. Epub 2011/08/02. doi: 10.1016/j.semcdb.2011.07.010. PubMed PMID: 21802523.

30. Weckselblatt B, Rudd MK. Human Structural Variation: Mechanisms of Chromosome Rearrangements. Trends Genet. 2015;31(10):587-99. Epub 2015/07/26. doi: 10.1016/j.tig.2015.05.010. PubMed PMID: 26209074; PMCID: PMC4600437.

31. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nature Communications. 2017;8(1):1326. doi: 10.1038/s41467-017-01343-4.

32. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, et al. The structure, function and evolution of a complete human chromosome 8. Nature. 2021;593(7857):101-7. Epub 2021/04/09. doi: 10.1038/s41586-021-03420-7. PubMed PMID: 33828295.

33. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020;585(7823):79-84. Epub 2020/07/15. doi: 10.1038/s41586-020-2547-7. PubMed PMID: 32663838; PMCID: PMC7484160.

34. Highnam G, Mittelman D. Personal genomes and precision medicine. Genome Biol. 2012;13(12):324-. doi: 10.1186/gb-2012-13-12-324. PubMed PMID: 23253090.

35. Xiao W, Wu L, Yavas G, Simonyan V, Ning B, Hong H. Challenges, Solutions, and Quality Metrics of Personal Genome Assembly in Advancing Precision Medicine. Pharmaceutics. 2016;8(2):15. doi: 10.3390/pharmaceutics8020015. PubMed PMID: 27110816.

36. Bergstrom A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367(6484). Epub 2020/03/21. doi: 10.1126/science.aay5012. PubMed PMID: 32193295; PMCID: PMC7115999.

37. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al. Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun. 2016;7:12065. Epub 2016/07/01. doi: 10.1038/ncomms12065. PubMed PMID: 27356984; PMCID: PMC4931320.

38. Rise for Rare: The Rare Disease Diversity Coalition; 2021 [cited 2021 May 9]. Available from: https://www.rarediseasediversity.org/rise-for-rare.

39. Clyde D. Diverse human genomes. Nature Reviews Genetics. 2020;21(6):338-. doi: 10.1038/s41576-020-0235-y.

40. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290-9. doi: 10.1038/s41586-021-03205-y.

41. Claw KG, Anderson MZ, Begay RL, Tsosie KS, Fox K, Garrison NA, et al. A framework for enhancing ethical genomic research with Indigenous communities. Nature Communications. 2018;9(1):2957. doi: 10.1038/s41467-018-05188-3.

42. Mapes BM, Foster CS, Kusnoor SV, Epelbaum MI, AuYoung M, Jenkins G, et al. Diversity and inclusion for the All of Us research program: A scoping review. PLoS One. 2020;15(7):e0234962. Epub 2020/07/02. doi: 10.1371/journal.pone.0234962. PubMed PMID: 32609747; PMCID: PMC7329113.

43. Skinner HG, Calancie L, Vu MB, Garcia B, DeMarco M, Patterson C, et al. Using Community-Based Participatory Research Principles to Develop More Understandable Recruitment and Informed Consent Documents in Genomic Research. PLOS ONE. 2015;10(5):e0125466. doi: 10.1371/journal.pone.0125466.

44. Bonham VL, Citrin T, Modell SM, Franklin TH, Bleicher EW, Fleck LM. Community-based dialogue: engaging communities of color in the United states' genetics policy conversation.

J Health Polit Policy Law. 2009;34(3):325-59. Epub 2009/05/20. doi: 10.1215/03616878-2009-009. PubMed PMID: 19451407; PMCID: PMC2800818.

45. Better Research through Engagement: Patient-Centered Outcomes Research Institute; 2016 [cited 2021 May 8]. Available from: https://www.pcori.org/sites/default/files/PCORI-Better-Research-Through-Engagement.pdf.

46. Research Done Differently: Patient-Centered Outcomes Research Institute; 2019 [cited 2021 May 8]. Available from: https://www.pcori.org/sites/default/files/PCORI-Research-Done-Differently.pdf.

47. Ejiogu N, Norbeck JH, Mason MA, Cromwell BC, Zonderman AB, Evans MK. Recruitment and retention strategies for minority or poor clinical research participants: lessons from the Healthy Aging in Neighborhoods of Diversity across the Life Span study. Gerontologist. 2011;51 Suppl 1(Suppl 1):S33-S45. doi: 10.1093/geront/gnr027. PubMed PMID: 21565817.

48. Oh SS, Galanter J, Thakur N, Pino-Yanes M, Barcelo NE, White MJ, et al. Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. PLOS Medicine. 2015;12(12):e1001918. doi: 10.1371/journal.pmed.1001918.

49. Bonham VL, Green ED. The genomics workforce must become more diverse: a strategic imperative. Am J Hum Genet. 2021;108(1):3-7. Epub 2021/01/09. doi: 10.1016/j.ajhg.2020.12.013. PubMed PMID: 33417888; PMCID: PMC7820786.

50. All of Us Research Program I, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, et al. The "All of Us" Research Program. N Engl J Med. 2019;381(7):668-76. Epub 2019/08/15. doi: 10.1056/NEJMsr1809937. PubMed PMID: 31412182.