UNIVERSITY OF CALIFORNIA

Los Angeles

Analyses of Next-Generation Sequencing Data to Identify Genes Associated with Complex Neuropsychiatric Disorders

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Bioinformatics

by

Aliz Raksi Rao

2017

ABSTRACT OF THE DISSERTATION

Analyses of Next-Generation Sequencing Data to Identify Genes Associated with Complex Neuropsychiatric Disorders

by

Aliz Raksi Rao

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2017

Professor Stanley F. Nelson, Chair

Over the past decade, decreases in the cost of DNA sequencing has allowed for a surge in the amount of data being generated. This has led to the discovery of genes causal for hundreds of Mendelian disorders and genes associated with many complex disorders. Given the opportunity to use sequencing data to tackle neuropsychiatric diseases with a complex genetic architecture, I take a data-first approach to study two diseases, bipolar disorder and autism spectrum disorder (ASD). Combining information gleaned from next-gen sequencing (NGS) data with the latest analytic methods shines new light on the biology of these diseases.

In the first part of this dissertation, I present a whole-exome analysis of nine affected individuals from four families in which bipolar disorder was transmitted over several generations, and six unrelated, affected individuals. Our results demonstrate the genetic heterogeneity of bipolar disorder and provide support for rare-variant oligogenic disease model.

In the second part, I present an approach to identify rare variants associated with autism spectrum disorder (ASD) in a whole-genome sequencing study of 71 individuals diagnosed with ASD and their family members. I demonstrate that by incorporating knowledge of population-wide variant frequencies to analyses of NGS data and taking an approach sensitive to complex family structures, as opposed to utilizing only case-control or trio data, one can identify patterns that would otherwise have been missed and thus gain novel insights into disease etiology. Finally, I present a mutational burden dataset called SORVA (Significance Of Rare VAriants), which is useful in vetting candidate variants and genes from NGS studies. In effect, my studies of complex disorders using next-gen sequencing show the field is constantly evolving with new computational approaches allowing for many advances being made in the areas of psychiatric disorders and ASD, in particular.

The dissertation of Aliz Raksi Rao is approved.

Eleazar Eskin

Janet S. Sinsheimer

Rita M. Cantor

Stanley F. Nelson, Committee Chair

University of California, Los Angeles

2017

DEDICATION

To my father, whose curiosity and enthusiasm for science I wholeheartedly share. Thank you for showing me the path and supporting me throughout this journey of discovery.

To my spouse and partner, who has demonstrated incredible patience by supporting me along the way.

"Success is the sum of small efforts—
repeated day in and day out."

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ASD	Autism spectrum disorder
BD	Bipolar disorder
CADD	Combined Annotation–Dependent Depletion
CES	Clinical exome sequencing
DAVID	Database for Annotation, Visualization, and Integrated Discovery
ExAC	Exome Aggregation Consortium
GATK	Genome Analysis Toolkit
GWAS	Genome-wide association study
HPE	Holoprosencephaly
IAN	Interactive Autism Network
LOF	Loss-of-function
MAF	Minor allele frequency
NIMH	National Institute of Mental Health
NGS	Next-generation sequencing
OMIM	Online Mendelian Inheritance in Man
PDD-NOS	Pervasive developmental disorder not otherwise specified
SNV	Single nucleotide variant
SORVA	Significance of Rare Variants
SVS	SNP & Variation Suite
SZ	Schizophrenia
TADA	Transmission And De novo Association test
VUS	Variant of uncertain significance
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

ACKNOWLEDGEMENTS

No research effort is the work of a single individual, and this thesis is no different. First and foremost, I would like to acknowledge Professor Stanley F. Nelson for his support as the chair of my committee. In addition to providing insight and guidance during my research, he has demonstrated how to present my findings in a clear and succinct manner, and encouraged me in working independently, which taught me valuable lessons in management of both people and time.

During my tenure at UCLA, I've had the pleasure of working along side numerous bright researchers. I would like to thank fellow scientists Dr. Michael Yourshaw, Kevin Squire, Bret Harry, Ascia Eskin, Dr. Ferenc Raksi, Dr. Valerie Arboleda and Dr. Richard Wang for their support and helpful discussions.

Chapter Two, in full, is a reprint of a published manuscript: Rao, A.R., Yourshaw, M., Christensen, B., Nelson, S.F., Kerner, B. Rare deleterious mutations are associated with disease in bipolar disorder families. Molecular Psychiatry. 2017; 22:1009–1014. doi:10.1038/mp.2016.181. The dissertation author was the primary author of this paper, and was responsible for the computational analysis and writing the manuscript. The other authors were Michael Yourshaw, Bryce Christensen, Stanley F. Nelson, and Berit Kerner.

Chapter Three, in full, is from material in preparation for publication: Rao, A.R., Lee, H., Marvin, A.R., Lipkin, P.H., Nelson, S.F. Whole-genome sequencing of web-based recruited individuals with autism spectrum disorders reveals novel candidate genes. The dissertation author was the primary author of this paper, and was responsible for the research. The other authors were Hane Lee, Alison R. Marvin, Paul H. Lipkin and Stanley F. Nelson.

Chapter Four, in part, is from material in preparation for publication in BMC Medical Genomics, 2017: Rao, A.R., Nelson, S.F. Calculating the statistical significance of rare variants causal for Mendelian and complex disorders. The dissertation author was the primary author of this paper, and was responsible for the research.

Funding to the author was provided by National Institutes of Health (NIH) Training Grant No. T32HG002536. I thank the NIH/National Center for Advancing Translational Science (NCATS) (UCLA CTSI Grant Number UL1TR000124) for their support.

VITA

2010	B.S., Computational and Systems Biology, University of California, Los Angeles
2010	B.S., Ecology, Behavior and Evolution, University of California, Los Angeles
2011	M.S., Bioinformatics, University of California, Los Angeles
2011-2013	Graduate Student Trainee, Genomic Analysis Training Program, University of California, Los Angeles, California

PUBLICATIONS

Rao AR, Lee H, Marvin AR, Lipkin PH, Nelson SF. Whole-genome sequencing of web-based recruited individuals with autism spectrum disorders reveals novel candidate genes. *In revision.*

Rao AR, Nelson SF. (2017) Calculating the statistical significance of rare variants causal for Mendelian and complex disorders. *BMC Medical Genomics. In peer review.*

Rao AR, Yourshaw M, Christensen B, Nelson SF, Kerner B. (2017) Rare deleterious mutations are associated with disease in bipolar disorder families. *Molecular Psychiatry.* 22:1009–1014.

Xue Y, Schoser B, Rao AR, Quadrelli R, Vaglio A, Rupp V, Beichler C, Nelson SF, Windpassinger C, Wilcox WR. (2015) Exome Sequencing Identified A Splice Site Mutation in FHL1 that Causes Uruguay Syndrome, an X-linked Disorder with Skeletal Muscle Hypertrophy and Premature Cardiac Death. *Circulation: Cardiovascular Genetics.* 9(2):130.

Yourshaw M, Taylor SP, Rao AR, Martin MG, Nelson SF. (2014) Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. *Briefings in Bioinformatics*. 16(2):255–64.

Kerner B, Rao AR, Christensen B, Dandekar S, Yourshaw M, Nelson SF. (2013) Rare genomic variants link bipolar disorder with anxiety disorders to CREB-regulated intracellular signaling pathways. *Frontiers in Psychiatry*. 4(6):154.

Raksi A, Pellegrini M. (2011) Regulation of the yeast metabolic cycle by transcription factors with periodic activities. *BMC Systems Biology*. 5:160.

Campos DP, Bander LA, Raksi A, Blumstein DT. (2009). Perch exposure and predation risk: a comparative study in passerines. *Acta Ethologica*, 12(2):93–98.

Chapter 1

Introduction

1.1 Challenges of studying complex disorders via next-gen sequencing

In the current era of rapidly decreasing sequencing costs, ever larger next-generation sequencing (NGS) datasets are enabling the analysis of complex genetic diseases that were previously intractable. By sequencing only a couple of affected individuals, the first NGS studies revealed the genetic cause of rare Mendelian diseases through whole-exome or whole-genome sequencing (Lupski *et al.*, 2010; Murdock *et al.*, 2011; S. B. Ng, Bigham, *et al.*, 2010; S. B. Ng, Buckingham, *et al.*, 2010; Sobreira *et al.*, 2010). More recent studies seek to identify the genetic underpinnings of common disorders with both genetic and environmental risk factors, such as autism spectrum disorder (ASD), coronary heart disease, and late-onset Alzheimer's disease (M. Chahrour *et al.*, 2016; Myocardial Infarction Genetics Consortium Investigators *et al.*, 2014; Sirkis *et al.*, 2016; TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute *et al.*, 2014). These recent studies involve sequencing thousands of individuals, as researchers attempt to discover rare variants that contribute to common diseases, and as the overall scientific returns of hunting for genetic variants are diminishing relative to the increasing size of the projects (Farfel *et al.*, 2016). Through genome-wide association studies (GWAS) and NGS studies, hundreds of genes and gene variants have been associated with disorders such as ASD, deafness and cardiometabolic risk factors for coronary heart disease (Banerjee-Basu and Packer, 2010; Orho-Melander, 2015; Wright and Hastie, 2007).

However, high-throughput methods that are not hypothesis-driven may associate genetic findings with disease by chance, and functional validation is often lacking to support these novel disease associations. As a result, genes that more frequently exhibit rare protein-altering variants are more frequently associated with disease phenotypes (Shyr *et al.*, 2014), and it is important to validate findings through functional studies or robust statistical methods. The fact that many disease associations may be a result of chance is especially problematic for psychiatric or neurological disease, since many kinds of experiments to functionally validate findings, e.g. methods that would require invasive techniques on the human brain, are technically or ethically impossible (Markram, 2013); mouse models don't exist for many psychiatric diseases; and higher brain functions cannot be modeled *in vitro*. To summarize, abnormalities of the brain and neuronal wiring cannot easily be studied, and instead, *in silico* experiments and robust statistical tests must be used to provide evidence for disease association findings.

1.2 Current methods to analyze large NGS datasets

Following the age of GWAS and linkage analyses, the increase in next-gen sequencing studies required the development of new statistical tools to analyze the deluge of data. There have been developments on many fronts, including improvements in the prediction of disease pathogenicity of a variant, the development of various collapsing methods and tools to estimate mutational burden to evaluate whether a gene is intolerant of variation in the population.

Early methods to predict variant pathogenicity such as SIFT and PolyPhen2 use conservation and other sequence-based features to identify damaging variants (Adzhubei *et al.*, 2010; P. C. Ng and Henikoff, 2003). Newer methods such as CADD, GWAVA, Eigen and GAVIN are metaannotations, which integrate data from multiple existing tools using supervised or unsupervised learning to achieve even greater sensitivity and specificity in classifying disease causing variants (Kircher *et al.*, 2014; Ritchie *et al.*, 2014; van der Velde *et al.*, 2017). Annotating variants for deleteriousness is valuable for then prioritizing variants for further assessment using population allele frequencies, cosegregation analyses, disease association studies, or a second-tier test that complements the primary variant annotation tool (van der Velde *et al.*, 2015). An example of such a test would be one that ranks genes, as opposed to individual variants, based on a prior likelihood of it being associated with disease. The idea is that the genes most likely to contribute to disease are the genes in the human genome that are sensitive to mutational changes. Samocha et al. (2014) identified such genes that are under selective constraint and the resulting missense Z scores and pLI scores are a measure of deficit in missense or loss-of-function (LOF) variants compared to the expectation generated from predicted mutation rates. LOF variants, which include frameshift variants, stop gain variants, stop loss variants, i.e. nonsense variants, and variants affecting splice sites, are less likely tolerated in genes essential for cellular functions, and identifying LOF variants in a gene with a high pLI score can provide evidence towards such a variant being pathogenic. The pLI score has become a widely used measure for vetting genes and variants; however, methods continue to improve as reference datasets increase in size and methods are refined to provide LOF intolerance measures at a finer scale, e.g. across regions spanning known protein domains.

In the NGS era, improvements in all of these areas have contributed to the greater understanding of Mendelian and complex disease genetics. Next, I will review the genetic underpinnings of two such complex genetic disorders, namely bipolar disorder and ASD.

1.3 The genetic basis of bipolar disorder

Bipolar disorder (BD) is a severe mental disorder characterized by recurrent manic and depressive episodes with a prevalence ~1% (Kawakami, 2014; Merikangas *et al.*, 2007). Family, twin, and adoption studies have provided strong evidence for the importance of genetic factors in the etiology of bipolar disorder. Despite its estimated 0.7 to 0.8 heritability (Sullivan *et al.*, 2012), identifying the specific genetic causes of BD has proved challenging.

Initially, linkage studies identified regions of interest including 4p16, 12q23-q24, 16p13, 21q22, and Xq24-q26, and several regions on chromosome 18 (N Craddock and Jones, 1999). In the past decade, genotyping of large collections of cases and controls in genome-wide association studies have revealed individual loci associated with bipolar disorder. The SNP rs1006737 in the gene *CACNA1C* is the most replicated and most studied common genomic variant associated with bipolar disorder to date (Ferreira *et al.*, 2008; Kerner, 2014; Moskvina *et al.*, 2009; Sklar *et al.*, 2008). This gene encodes a calcium channel in ventricular cardiac muscle and is also present in smooth muscle, many secretory cells, and throughout the brain, and the protein plays a role in dendritic signaling (Striessnig *et al.*, 2014; Wheeler *et al.*, 2012). However, other studies could not replicate this association (Kloiber *et al.*, 2012; Zhang *et al.*, 2013), and research suggests that the majority of bipolar disorder may involve the interaction of multiple genes (epistasis) or other complex genetic mechanisms (N Craddock and Jones, 1999).

An important role for rare single-nucleotide variants (SNVs) in complex diseases has been proposed based on theoretical grounds (Keinan and Clark, 2012; Pritchard, 2001), and, more recently, high-throughput whole-exome sequencing (WES) and whole-genome sequencing (WGS) has enabled the identification of such variants associated with bipolar disorder. Sample sizes of WES and WGS studies of bipolar disorder have been small compared to those of schizophrenia and autism; the largest autism and schizophrenia studies analyzed the exomes of thousands of cases and controls (De Rubeis *et al.*, 2014; Iossifov *et al.*, 2014a; Purcell *et al.*, 2014), whereas the largest bipolar studies to date consisted of whole-exome sequencing of 237 trios (Kataoka *et al.*, 2016), and whole-genome sequencing of 200 individuals from 41 families with BD. This was followed by targeted sequencing of 26 candidate genes in an additional 3,014 cases and 1,717 controls (Ament *et al.*, 2015). Nevertheless, several novel candidate genes have emerged from the limited number of case-control and family-based sequencing studies of BD that have been published (Ament *et al.*, 2015; Collins *et al.*, 2013; Cruceanu *et al.*, 2013; Georgi *et al.*, 2014; Goes *et al.*, 2016; Kataoka *et al.*, 2016; Lescai *et al.*, 2017; Strauss *et al.*, 2014). Ament *et al.* (2015) found evidence for an excess of rare variants in pathways associated with γ-aminobutyric acid and calcium channel signaling, highlighting rare variant associations in *ANK3*, a synaptic scaffolding gene; voltage-gated calcium channel genes *CACNA1B*, *CACNA1C*, *CACNA1D*, *CACNG2; CAMK2A,* a prominent kinase in the central nervous system that may function in long-term potentiation and neurotransmitter release; and *NGF*, which is involved in the regulation of growth and the differentiation of sympathetic and certain sensory neurons. Another study (Goes *et al.*, 2016), although underpowered to implicate rare variants in individual genes, highlights the gene *KDM5B* (also known as *JARIDB1*) that encodes a histone H3 lysine 4 (H3K4) demethylase that has been linked to neural differentiation in embryonic stem cells (Schmitz *et al.*, 2011). The histone H3K4 methylation pathway has already been associated as a pathway strongly associated with BD based on GWAS (The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium, 2015). Other BD-associated pathways derived from a GWAS meta-analysis include corticotropin-releasing hormone signaling, cardiac β-adrenergic signaling, phospholipase C signaling, glutamate receptor signaling, endothelin 1 signaling, and cardiac hypertrophy signaling (Nurnberger *et al.*, 2014). There appears to be a notable overlap in susceptibility between bipolar disorder and schizophrenia (Nick Craddock and Sklar, 2013). Disease associated pathways shared between these two disorders include calcium- and glutamate signaling, neuropathic pain signaling in dorsal horn neurons, and calmodulin binding (Forstner *et al.*, 2017). The common theme between most studies of BD, however, is that many genes are involved in BD, the groups of risk variants may be different in different families, and exonic variants of major effect are unlikely to exist in this disorder (Kember and Bućan, 2016). Recent studies also point to a role for *de novo* LOF and protein-altering mutations in the etiology of bipolar disorder (Kataoka *et al.*, 2016), and the trend points towards sequencing larger cohorts, which will undoubtedly reveal further insights about the genetics of bipolar disorder in the future.

1.4 The genetic basis of autism spectrum disorders

Similarly to bipolar disorder, autism spectrum disorder (ASD) is a complex disease with high heritability and genetic heterogeneity. Additionally, ASD also has high phenotypic heterogeneity and consists of a constellation of neurodevelopmental presentations including autistic disorder, Asperger syndrome, childhood disintegrative disorder, and pervasive developmental disorder not otherwise specified (PDD-NOS) (American Psychiatric Association, 2013). To date, mutations in hundreds of genes have been associated to varying degrees with increased ASD risk. Most genes contribute to ASD risk by a small amount, with the notable exception of genes causal for various syndromes that also met criteria for ASD diagnosis, including *FMR1* for Fragile X syndrome (FXS), *TSC1/2* for Tuberous Sclerosis Complex (TSC), and *MECP2* for Rett syndrome (RTT) (Schaefer and Mendelsohn, 2008). Dominant, recessive, oligogenic/polygenic, and gene × environment mechanisms all clearly play a role in ASD; however, their individual contributions in different ASD subpopulations are still to be fully elucidated (M. Chahrour *et al.*, 2016).

As focus has shifted from GWAS to large-scale WES studies in the past decade, the largest studies have taken the approach of sequencing trios to detect *de novo* mutations in affected individuals. These include several studies published in 2012 and 2014, which analyzed WES data from over 4000 affected children combined (De Rubeis *et al.*, 2014; Iossifov *et al.*, 2014b, 2012; Neale *et al.*, 2012; O'Roak, Vives, Girirajan, *et al.*, 2012; Sanders *et al.*, 2012). By now, hundreds of candidate genes have emerged, and dozens have been confirmed as high-confidence ASD genes based on their recurrent disruption by *de novo* mutations in unrelated probands (De Rubeis *et al.*, 2014; Iossifov *et al.*, 2014a; O'Roak *et al.*, 2014; O'Roak, Vives, Fu, *et al.*, 2012). These include fragile X mental retardation protein targets, chromatin modifiers (e.g., *CHD8*, *CHD2*, *ARID1B*), embryonically expressed genes (e.g., *TBR1*, *DYRK1A, PTEN*), and nominal enrichment for postsynaptic density proteins (e.g., *GRIN2B*, *GABRB3*, *SHANK3*) (M. Chahrour *et al.*, 2016). Networks constructed using these high-confidence ASD risk genes identify several key pathways as being disrupted in ASD, including translational control and chromatin regulation (Hormozdiari *et al.*, 2015; O'Roak, Vives, Girirajan, *et al.*, 2012; Parikshak *et al.*, 2013; Willsey *et al.*, 2013).

Other studies have focused on identifying recessive and hemizygous variants conferring ASD risk; examples of such genes include *CNTNAP2* (Strauss *et al.*, 2006), *SLC9A9/NHE9* (Morrow *et al.*, 2008), *BCKDK* (Novarino *et al.*, 2012), and *CC2D1A* (Manzini *et al.*, 2014).

Some genes that have been found to contain recessive mutations are genes that, had they been completely inactivated, would cause severe neurological syndromes. For example, Chahrour *et al.* (2012) identified *UBE3B* as a candidate gene, and this gene is also associated with a syndrome of intellectual disability and microcephaly (Basel-Vanagaite *et al.*, 2012). Similarly, a complete loss of ASD candidate genes *AMT*, *PEX7*, and *VPS13B* will lead, respectively, to nonketotic hyperglycinemia, rhizomelic chondrodysplasia punctata, and Cohen syndrome (Yu *et al.*, 2013).

To further complicate the architecture of ASD genetics, there are sex differences: ASD is four times more common in males than in females (Christensen *et al.*, 2016). Multiple theories have been proposed, including genetic, epigenetic, and hormonal explanations (Baron-Cohen *et al.*, 2011; Robinson *et al.*, 2013; Werling and Geschwind, 2013). Multiple studies have identified an excess of maternally inherited protein-damaging variants and copy-number variants (CNVs) in cases (Bonora *et al.*, 2014; Griswold *et al.*, 2015; Krumm *et al.*, 2015; Yuen *et al.*, 2016), and many of these variants and CNVs overlap genes previously identified to contain *de novo* variants in other ASD studies, although statistical evidence for individual gene findings remains insufficient to establish the genes as high-confidence ASD genes (Abrahams *et al.*, 2013).

To further our understanding of the genetics of both ASD and bipolar disorder, larger cohorts will need to be studied, as hundreds of genes may be mutated in only a handful of cases. One approach to recruit the large number of families is to utilize web-based recruiting methods, such as the system implemented by IAN Genetics, which allows families to enroll in ASD studies regardless of their proximity to study sites. We utilized this cohort to select individuals for WGS in an attempt to identify rare variants of large effect contributing to ASD.

In the following chapters, I present findings from WES and WGS studies involving two complex disorders, bipolar disorder and autism spectrum disorder, and present a mutational burden dataset based on a reference population that aids the identification of candidate genes and variants for follow-up studies from NGS datasets.

1.5 Bibliography

Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., Menashe, I., Wadkins, T., Banerjee-Basu, S., and Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Molecular autism*, **4**(1), 36. doi:10.1186/2040-2392-4-36

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, **7**(4), 248–249. doi:10.1038/nmeth0410-248

Ament, S. A., Szelinger, S., Glusman, G., Ashworth, J., Hou, L., Akula, N., Shekhtman, T., Badner, J. A., Brunkow, M. E., Mauldin, D. E., Stittrich, A.-B., Rouleau, K., Detera-Wadleigh, S. D., Nurnberger, J. I., Edenberg, H. J., Gershon, E. S., Schork, N., Price, N. D., Gelinas, R., et al. (2015). Rare variants in neuronal excitability genes influence risk for bipolar disorder. *Proceedings of the National Academy of Sciences*, **112**(11), 3576–3581. doi:10.1073/pnas.1424958112

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association.

Banerjee-Basu, S., and Packer, A. (2010). SFARI Gene: an evolving database for the autism research community. *Disease models & mechanisms*, **3**(3–4), 133–5. doi:10.1242/dmm.005439

Baron-Cohen, S., Lombardo, M. V., Auyeung, B., Ashwin, E., Chakrabarti, B., and Knickmeyer, R. (2011). Why Are Autism Spectrum Conditions More Prevalent in Males? *PLoS Biology*, **9**(6), e1001081. doi:10.1371/journal.pbio.1001081

Basel-Vanagaite, L., Dallapiccola, B., Ramirez-Solis, R., Segref, A., Thiele, H., Edwards, A., Arends, M. J., Miró, X., White, J. K., Désir, J., Abramowicz, M., Dentici, M. L., Lepri, F., Hofmann, K., Har-Zahav, A., Ryder, E., Karp, N. A., Estabel, J., Gerdin, A.-K. B., et al. (2012). Deficiency for the Ubiquitin Ligase UBE3B in a Blepharophimosis-Ptosis-Intellectual-Disability Syndrome. *The American Journal of Human Genetics*, **91**(6), 998–1010. doi:10.1016/j.ajhg.2012.10.011

Bonora, E., Graziano, C., Minopoli, F., Bacchelli, E., Magini, P., Diquigiovanni, C., Lomartire, S., Bianco, F., Vargiolu, M., Parchi, P., Marasco, E., Mantovani, V., Rampoldi, L., Trudu, M., Parmeggiani, A., Battaglia, A., Mazzone, L., Tortora, G., IMGSAC, E., et al. (2014). Maternally inherited genetic variants of CADPS2 are present in autism spectrum disorders and intellectual disability patients. *EMBO molecular medicine*, **6**(6), 795–809. doi:10.1002/emmm.201303235

Chahrour, M. H., Yu, T. W., Lim, E. T., Ataman, B., Coulter, M. E., Hill, R. S., Stevens, C. R., Schubert, C. R., Greenberg, M. E., Gabriel, S. B., Walsh, C. A., and ARRA Autism Sequencing Collaboration. (2012). Whole-Exome Sequencing and Homozygosity Analysis Implicate Depolarization-Regulated Neuronal Genes in Autism. *PLoS Genet*, **8**(4), e1002635. doi:10.1371/journal.pgen.1002635

Chahrour, M., O'Roak, B. J., Santini, E., Samaco, R. C., Kleiman, R. J., and Manzini, M. C. (2016). Current Perspectives in Autism Spectrum Disorder: From Genes to Therapy. *The Journal of Neuroscience*, **36**(45), 11402–11410. doi:10.1523/JNEUROSCI.2335-16.2016

Christensen, D. L., Baio, J., Braun, K. V. N., Bilder, D., Charles, J., Constantino, J. N., Daniels, J., Durkin, M. S., Fitzgerald, R. T., Kurzius-Spencer, M., Lee, L.-C., Pettygrove, S., Robinson, C., Schulz, E., Wells, C., Wingate, M. S., Zahorodny, W., Yeargin-Allsopp, M., and Centers for Disease Control and Prevention (CDC). (2016). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. *MMWR. Surveillance Summaries*, **65**(3), 1–23. doi:10.15585/mmwr.ss6503a1

Collins, A. L., Kim, Y., Szatkiewicz, J. P., Bloom, R. J., Hilliard, C. E., Quackenbush, C. R., Meier, S., Rivas, F., Mayoral, F., Cichon, S., Nöthen, M. M., Rietschel, M., and Sullivan, P. F. (2013). Identifying bipolar disorder susceptibility loci in a densely affected pedigree. *Molecular Psychiatry*, **18**(12), 1245–1246. doi:10.1038/mp.2012.176

Craddock, N., and Jones, I. (1999). Genetics of bipolar disorder. *Journal of medical genetics*, **36**(8), 585–94.

Craddock, N., and Sklar, P. (2013). Genetics of bipolar disorder. *The Lancet*, **381**(9878), 1654–1662. doi:10.1016/S0140-6736(13)60855-7

Cruceanu, C., Ambalavanan, A., Spiegelman, D., Gauthier, J., Lafrenière, R. G., Dion, P. A., Alda, M., Turecki, G., and Rouleau, G. A. (2013). Family-based exome-sequencing approach identifies rare susceptibility variants for lithium-responsive bipolar disorder. *Genome*, **56**(10), 634–640. doi:10.1139/gen-2013-0081

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Ercument Cicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., Singh, T., Klei, L., Kosmicki, J., Fu, S.-C., Aleksic, B., Biscaldi, M., Bolton, P. F., Brownfeld, J. M., Cai, J., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, **515**(7526), 209–215. doi:10.1038/nature13772

Farfel, J. M., Yu, L., Buchman, A. S., Schneider, J. A., De Jager, P. L., and Bennett, D. A. (2016). Relation of genomic variants for Alzheimer disease dementia to common neuropathologies. *Neurology*, **87**(5), 489–496. doi:10.1212/WNL.0000000000002909

Ferreira, M. A. R., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L., Fan, J., Kirov, G., Perlis, R. H., Green, E. K., Smoller, J. W., Grozeva, D., Stone, J., Nikolov, I., Chambert, K., Hamshere, M. L., Nimgaonkar, V. L., Moskvina, V., Thase, M. E., et al. (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics*, **40**(9), 1056–1058. doi:10.1038/ng.209

Forstner, A. J., Hecker, J., Hofmann, A., Maaser, A., Reinbold, C. S., Mühleisen, T. W., Leber, M., Strohmaier, J., Degenhardt, F., Treutlein, J., Mattheisen, M., Schumacher, J., Streit, F., Meier, S., Herms, S., Hoffmann, P., Lacour, A., Witt, S. H., Reif, A., et al. (2017). Identification of shared risk loci and pathways for bipolar disorder and schizophrenia. *PLOS ONE*, **12**(2), e0171595. doi:10.1371/journal.pone.0171595

Georgi, B., Craig, D., Kember, R. L., Liu, W., Lindquist, I., Nasser, S., Brown, C., Egeland, J. A., Paul, S. M., and Bućan, M. (2014). Genomic View of Bipolar Disorder Revealed by Whole Genome Sequencing in a Genetic Isolate. *PLoS Genetics*, **10**(3), e1004229. doi:10.1371/journal.pgen.1004229

Goes, F. S., Pirooznia, M., Parla, J. S., Kramer, M., Ghiban, E., Mavruk, S., Chen, Y.-C., Monson, E. T., Willour, V. L., Karchin, R., Flickinger, M., Locke, A. E., Levy, S. E., Scott, L. J., Boehnke, M., Stahl, E., Moran, J. L., Hultman, C. M., Landén, M., et al. (2016). Exome Sequencing of Familial Bipolar Disorder. *JAMA Psychiatry*, **73**(6), 590. doi:10.1001/jamapsychiatry.2016.0251

Griswold, A. J., Dueker, N. D., Van Booven, D., Rantus, J. A., Jaworski, J. M., Slifer, S. H., Schmidt, M. A., Hulme, W., Konidari, I., Whitehead, P. L., Cuccaro, M. L., Martin, E. R., Haines, J. L., Gilbert, J. R., Hussman, J. P., and Pericak-Vance, M. A. (2015). Targeted massively parallel sequencing of autism spectrum disorder-associated genes in a case control cohort reveals rare loss-of-function risk variants. *Molecular Autism*, **6**(1), 43. doi:10.1186/s13229-015-0034-z

Hormozdiari, F., Penn, O., Borenstein, E., and Eichler, E. E. (2015). The discovery of integrated gene networks for autism and related disorders. *Genome Research*, **25**(1), 142–154. doi:10.1101/gr.178855.114

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., et al. (2014a). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, **515**(7526), 216–221. doi:10.1038/nature13908

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., et al. (2014b). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, **515**(7526), 216–221. doi:10.1038/nature13908

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y., Narzisi, G., Leotta, A., Kendall, J., Grabowska, E., Ma, B., Marks, S., Rodgers, L., Stepansky, A., Troge, J., Andrews, P., Bekritsky, M., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron*, **74**(2), 285–299. doi:10.1016/j.neuron.2012.04.009

Kataoka, M., Matoba, N., Sawada, T., Kazuno, A.-A., Ishiwata, M., Fujii, K., Matsuo, K., Takata, A., and Kato, T. (2016). Exome sequencing for bipolar disorder points to roles of de novo loss-of-function and protein-altering mutations. *Molecular psychiatry*, **21**(7), 885–93. doi:10.1038/mp.2016.69

Kawakami, N. (2014). *Large scale epidemiology study of the prevalence of mental disorders: World Mental Health Japan Survey Second. A report of the Health Labour Sciences Research Grant from The Ministry of Health Labour and Welfare (H25-Seishin-Ippan-006)*. Tokyo.

Keinan, A., and Clark, A. G. (2012). Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science*, **336**(6082), 740–743. doi:10.1126/science.1217283

Kember, R. L., and Bućan, M. (2016). Promising 2-Pronged Approach to Genetic Basis of Bipolar Disorder. *JAMA Psychiatry*, **73**(6), 553. doi:10.1001/jamapsychiatry.2016.0298

Kerner, B. (2014). Genetics of bipolar disorder. *The application of clinical genetics*, **7**, 33–42. doi:10.2147/TACG.S39297

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, **advance on**. doi:10.1038/ng.2892

Kloiber, S., Czamara, D., Karbalai, N., Müller-Myhsok, B., Hennings, J., Holsboer, F., and Lucae, S. (2012). ANK3 and CACNA1C – Missing genetic link for bipolar disorder and major depressive disorder in two German case-control samples. *Journal of Psychiatric Research*, **46**(8), 973–979. doi:10.1016/j.jpsychires.2012.04.017

Krumm, N., Turner, T. N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B. P., Stessman, H. A., He, Z.-X., Leal, S. M., Bernier, R., and Eichler, E. E. (2015). Excess of rare, inherited truncating mutations in autism. *Nature genetics*, **47**(6), 582–8. doi:10.1038/ng.3303

Lee, H., Marvin, A. R., Watson, T., Piggot, J., Law, J. K., Law, P. A., et al. (2010). Accuracy of phenotyping of autistic children based on internet implemented parent report. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *153B*(6), 1119–1126. doi:10.1002/ajmg.b.31103

Lescai, F., Als, T. D., Li, Q., Nyegaard, M., Andorsdottir, G., Biskopstø, M., Hedemand, A., Fiorentino, A., O'Brien, N., Jarram, A., Liang, J., Grove, J., Pallesen, J., Eickhardt, E., Mattheisen, M., Bolund, L., Demontis, D., Wang, A. G., McQuillin, A., et al. (2017). Whole-exome sequencing of individuals from an isolated population implicates rare risk variants in bipolar disorder. *Translational Psychiatry*, **7**(2), e1034. doi:10.1038/tp.2017.3

Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C. Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D. A., McGuire, A. L., Zhang, F., Stankiewicz, P., Halperin, J. J., Yang, C., Gehman, C., Guo, D., Irikat, R. K., Tom, W., et al. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *The New England journal of medicine*, **362**(13), 1181–91. doi:10.1056/NEJMoa0908094

Manzini, M. C., Xiong, L., Shaheen, R., Tambunan, D. E., Di Costanzo, S., Mitisalis, V., Tischfield, D. J., Cinquino, A., Ghaziuddin, M., Christian, M., Jiang, Q., Laurent, S., Nanjiani, Z. A., Rasheed, S., Hill, R. S., Lizarraga, S. B., Gleason, D., Sabbagh, D., Salih, M. A., et al. (2014). CC2D1A Regulates Human Intellectual and Social Function as well as NF-κB Signaling Homeostasis. *Cell Reports*, **8**(3), 647–655. doi:10.1016/j.celrep.2014.06.039

Markram, H. (2013). Seven challenges for neuroscience. *Functional Neurology*, **28**(3), 145–151.

Merikangas, K. R., Akiskal, H. S., Angst, J., Greenberg, P. E., Hirschfeld, R. M. A., Petukhova, M., and Kessler, R. C. (2007). Lifetime and 12-Month Prevalence of Bipolar Spectrum Disorder in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **64**(5), 543. doi:10.1001/archpsyc.64.5.543

Morrow, E. M., Yoo, S.-Y., Flavell, S. W., Kim, T.-K., Lin, Y., Hill, R. S., Mukaddes, N. M., Balkhy, S., Gascon, G., Hashmi, A., Al-Saad, S., Ware, J., Joseph, R. M., Greenblatt, R., Gleason, D., Ertelt, J. A., Apse, K. A., Bodell, A., Partlow, J. N., et al. (2008). Identifying Autism Loci and Genes by Tracing Recent Shared Ancestry. *Science*, **321**(5886), 218–223. doi:10.1126/science.1157657

Moskvina, V., Craddock, N., Holmans, P., Nikolov, I., Pahwa, J. S., Green, E., Owen, M. J., O'Donovan, M. C., and O'Donovan, M. C. (2009). Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Molecular Psychiatry*, **14**(3), 252–260. doi:10.1038/mp.2008.133

Murdock, D. R., Clark, G. D., Bainbridge, M. N., Newsham, I., Wu, Y.-Q., Muzny, D. M., Cheung, S. W., Gibbs, R. A., and Ramocki, M. B. (2011). Whole-Exome Sequencing Identifies Compound Heterozygous Mutations in WDR62 in Siblings With Recurrent Polymicrogyria. *American journal of medical genetics. Part A*, **0**(9), 2071–2077. doi:10.1002/ajmg.a.34165

Myocardial Infarction Genetics Consortium Investigators, Stitziel, N. O., Won, H.-H., Morrison, A. C., Peloso, G. M., Do, R., Lange, L. A., Fontanillas, P., Gupta, N., Duga, S., Goel, A., Farrall, M., Saleheen, D., Ferrario, P., König, I., Asselta, R., Merlini, P. A., Marziliano, N., Notarangelo, M. F., et al. (2014). Inactivating mutations in NPC1L1 and protection from coronary heart disease. *The New England Journal of Medicine*, **371**(22), 2072–2082. doi:10.1056/NEJMoa1405386

Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., Polak, P., Yoon, S., Maguire, J., Crawford, E. L., Campbell, N. G., Geller, E. T., Valladares, O., Schafer, C., Liu, H., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, **485**(7397), 242–245. doi:10.1038/nature11011

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, **31**(13), 3812–4.

Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., Beck, A. E., Tabor, H. K., Cooper, G. M., Mefford, H. C., Lee, C., Turner, E. H., Smith, J. D., Rieder, M. J., Yoshiura, K.-I., Matsumoto, N., Ohta, T., Niikawa, N., Nickerson, D. A., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics*, **42**(9), 790–3. doi:10.1038/ng.646

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., and Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, **42**(1), 30–5. doi:10.1038/ng.499

Novarino, G., El-Fishawy, P., Kayserili, H., Meguid, N. A., Scott, E. M., Schroth, J., Silhavy, J. L., Kara, M., Khalil, R. O., Ben-Omran, T., Ercan-Sencicek, A. G., Hashish, A. F., Sanders, S. J., Gupta, A. R., Hashem, H. S., Matern, D., Gabriel, S., Sweetman, L., Rahimi, Y., et al. (2012). Mutations in BCKD-kinase lead to a potentially treatable form of autism with epilepsy. *Science*, **338**(6105), 394–397. doi:10.1126/science.1224631

Nurnberger, J. I., Koller, D. L., Jung, J., Edenberg, H. J., Foroud, T., Guella, I., Vawter, M. P., and Kelsoe, J. R. (2014). Identification of Pathways for Bipolar Disorder. *JAMA Psychiatry*, **71**(6), 657. doi:10.1001/jamapsychiatry.2014.176

O'Roak, B. J., Stessman, H. A., Boyle, E. A., Witherspoon, K. T., Martin, B., Lee, C., Vives, L., Baker, C., Hiatt, J. B., Nickerson, D. A., Bernier, R., Shendure, J., and Eichler, E. E. (2014). Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nature communications*, **5**, 5595. doi:10.1038/ncomms6595

O'Roak, B. J., Vives, L., Fu, W., Egertson, J. D., Stanaway, I. B., Phelps, I. G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J. B., Turner, E. H., Levy, R., O'Day, D. R., Krumm, N., Coe, B. P., Martin, B. K., Borenstein, E., et al. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*, **338**(6114), 1619–1622. doi:10.1126/science.1227764

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., Levy, R., Ko, A., Lee, C., Smith, J. D., Turner, E. H., Stanaway, I. B., Vernot, B., Malig, M., Baker, C., Reilly, B., Akey, J. M., Borenstein, E., Rieder, M. J., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, **485**(7397), 246–250. doi:10.1038/nature10989

Orho-Melander, M. (2015). Genetics of coronary heart disease: towards causal mechanisms, novel drug targets and more personalized prevention. *Journal of Internal Medicine*, **278**(5), 433–446. doi:10.1111/joim.12407

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., Horvath, S., and Geschwind, D. H. (2013). Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell*, **155**(5), 1008–1021. doi:10.1016/j.cell.2013.10.031

Pritchard, J. K. (2001). Are Rare Variants Responsible for Susceptibility to Complex Diseases? *The American Journal of Human Genetics*, **69**(1), 124–137. doi:10.1086/321272

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S. E., Kähler, A., Duncan, L., Stahl, E., Genovese, G., Fernández, E., Collins, M. O., Komiyama, N. H., Choudhary, J. S., Magnusson, P. K. E., Banks, E., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, **506**(7487), 185–190. doi:10.1038/nature12975

Ritchie, G. R. S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nature Methods*, **11**(3), 294–296. doi:10.1038/nmeth.2832

Robinson, E. B., Lichtenstein, P., Anckarsater, H., Happe, F., and Ronald, A. (2013). Examining and interpreting the female protective effect against autistic behavior. *Proceedings of the National Academy of Sciences*, **110**(13), 5258–5262. doi:10.1073/pnas.1211070110

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook Jr, E. H., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, **46**(9), 944–950. doi:10.1038/ng.3050

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., Ercan-Sencicek, A. G., DiLullo, N. M., Parikshak, N. N., Stein, J. L., Walker, M. F., Ober, G. T., Teran, N. A., Song, Y., El-Fishawy, P., Murtha, R. C., Choi, M., Overton, J. D., Bjornson, R. D., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, **485**(7397), 237–241. doi:10.1038/nature10945

Schaefer, G. B., and Mendelsohn, N. J. (2008). Genetics evaluation for the etiologic diagnosis of autism spectrum disorders. *Genetics in Medicine*, **10**(1), 4–12. doi:10.1097/GIM.0b013e31815efdd7

Schmitz, S. U., Albert, M., Malatesta, M., Morey, L., Johansen, J. V, Bak, M., Tommerup, N., Abarrategui, I., and Helin, K. (2011). Jarid1b targets genes regulating development and is involved in neural differentiation. *The EMBO Journal*, **30**(22), 4586–4600. doi:10.1038/emboj.2011.383

Shyr, C., Tarailo-Graovac, M., Gottlieb, M., Lee, J. J., Karnebeek, C. van, and Wasserman, W. W. (2014). FLAGS, frequently mutated genes in public exomes. *BMC Medical Genomics*, **7**(1), 64. doi:10.1186/s12920-014-0064-y

Sirkis, D. W., Bonham, L. W., Aparicio, R. E., Geier, E. G., Ramos, E. M., Wang, Q., Karydas, A., Miller, Z. A., Miller, B. L., Coppola, G., and Yokoyama, J. S. (2016). Rare TREM2 variants associated with Alzheimer's disease display reduced cell surface expression. *Acta Neuropathologica Communications*, **4**, 98. doi:10.1186/s40478-016-0367-7

Sklar, P., Smoller, J. W., Fan, J., Ferreira, M. A. R., Perlis, R. H., Chambert, K., Nimgaonkar, V. L., McQueen, M. B., Faraone, S. V, Kirby, A., de Bakker, P. I. W., Ogdie, M. N., Thase, M. E., Sachs, G. S., Todd-Brown, K., Gabriel, S. B., Sougnez, C., Gates, C., Blumenstiel, B., et al. (2008). Whole-genome association study of bipolar disorder. *Molecular Psychiatry*, **13**(6), 558–569. doi:10.1038/sj.mp.4002151

Sobreira, N. L. M., Cirulli, E. T., Avramopoulos, D., Wohler, E., Oswald, G. L., Stevens, E. L., Ge, D., Shianna, K. V., Smith, J. P., Maia, J. M., Gumbs, C. E., Pevsner, J., Thomas, G., Valle, D., Hoover-Fong, J. E., and Goldstein, D. B. (2010). Whole-Genome Sequencing of a Single Proband Together with Linkage Analysis Identifies a Mendelian Disease Gene. *PLoS Genetics*, **6**(6). doi:10.1371/journal.pgen.1000991

Strauss, K. A., Markx, S., Georgi, B., Paul, S. M., Jinks, R. N., Hoshi, T., McDonald, A., First, M. B., Liu, W., Benkert, A. R., Heaps, A. D., Tian, Y., Chakravarti, A., Bucan, M., and Puffenberger, E. G. (2014). A population-based study of KCNH7 p.Arg394His and bipolar spectrum disorder. *Human Molecular Genetics*, **23**(23), 6395–6406. doi:10.1093/hmg/ddu335

Strauss, K. A., Puffenberger, E. G., Huentelman, M. J., Gottlieb, S., Dobrin, S. E., Parod, J. M., Stephan, D. A., and Morton, D. H. (2006). Recessive Symptomatic Focal Epilepsy and Mutant Contactin-Associated Protein-like 2. *New England Journal of Medicine*, **354**(13), 1370–1377. doi:10.1056/NEJMoa052773

Striessnig, J., Pinggera, A., Kaur, G., Bock, G., and Tuluc, P. (2014). L-type Ca $^{2+}$ channels in heart and brain. *Wiley Interdisciplinary Reviews: Membrane Transport and Signaling*, **3**(2), 15–38. doi:10.1002/wmts.102

Sullivan, P. F., Daly, M. J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics*, **13**(8), 537–551. doi:10.1038/nrg3240

TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute, Crosby, J., Peloso, G. M., Auer, P. L., Crosslin, D. R., Stitziel, N. O., Lange, L. A., Lu, Y., Tang, Z., Zhang, H., Hindy, G., Masca, N., Stirrups, K., Kanoni, S., Do, R., Jun, G., Hu, Y., Kang, H. M., Xue, C., et al. (2014). Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *The New England Journal of Medicine*, **371**(1), 22–31. doi:10.1056/NEJMoa1307095

The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium. (2015). Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience*, **18**(2), 199–209. doi:10.1038/nn.3922

van der Velde, K. J., de Boer, E. N., van Diemen, C. C., Sikkema-Raddatz, B., Abbott, K. M., Knopperts, A., Franke, L., Sijmons, R. H., de Koning, T. J., Wijmenga, C., Sinke, R. J., and Swertz, M. A. (2017). GAVIN: Gene-Aware Variant INterpretation for medical sequencing. *Genome Biology*, **18**(1), 6. doi:10.1186/s13059-016-1141-7

van der Velde, K. J., Kuiper, J., Thompson, B. A., Plazzer, J.-P., van Valkenhoef, G., de Haan, M., Jongbloed, J. D. H., Wijmenga, C., de Koning, T. J., Abbott, K. M., Sinke, R., Spurdle, A. B., Macrae, F., Genuardi, M., Sijmons, R. H., and Swertz, M. A. (2015). Evaluation of CADD Scores in Curated Mismatch Repair Gene Variants Yields a Model for Clinical Validation and Prioritization. *Human Mutation*, **36**(7), 712–719. doi:10.1002/humu.22798

Werling, D. M., and Geschwind, D. H. (2013). Understanding sex bias in autism spectrum disorder. *Proceedings of the National Academy of Sciences*, **110**(13), 4868–4869. doi:10.1073/pnas.1301602110

Wheeler, D. G., Groth, R. D., Ma, H., Barrett, C. F., Owen, S. F., Safa, P., and Tsien, R. W. (2012). CaV1 and CaV2 Channels Engage Distinct Modes of Ca2+ Signaling to Control CREB-Dependent Gene Expression. *Cell*, **149**(5), 1112–1124. doi:10.1016/j.cell.2012.03.041

Willsey, A. J., Sanders, S. J., Li, M., Dong, S., Tebbenkamp, A. T., Muhle, R. A., Reilly, S. K., Lin, L., Fertuzinhos, S., Miller, J. A., Murtha, M. T., Bichsel, C., Niu, W., Cotney, J., Ercan-Sencicek, A. G., Gockley, J., Gupta, A. R., Han, W., He, X., et al. (2013). Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. *Cell*, **155**(5), 997–1007. doi:10.1016/j.cell.2013.10.020

Wright, A., and Hastie, N. D. (2007). Genes and common diseases: genetics in modern medicine. *Genes and common diseases: genetics in modern medicine.*

Yu, T. W., Chahrour, M. H., Coulter, M. E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., Schmitz-Abe, K., Harmin, D. A., Adli, M., Malik, A. N., D'Gama, A. M., Lim, E. T., Sanders, S. J., Mochida, G. H., Partlow, J. N., Sunu, C. M., Felie, J. M., Rodriguez, J., Nasir, R. H., et al. (2013). Using whole-exome sequencing to identify inherited causes of autism. *Neuron*, **77**(2), 259–273. doi:10.1016/j.neuron.2012.11.002

Yuen, R. K. C., Merico, D., Cao, H., Pellecchia, G., Alipanahi, B., Thiruvahindrapuram, B., Tong, X., Sun, Y., Cao, D., Zhang, T., Wu, X., Jin, X., Zhou, Z., Liu, X., Nalpathamkalam, T., Walker, S., Howe, J. L., Wang, Z., MacDonald, J. R., et al. (2016). Genome-wide characteristics of de novo mutations in autism. *NPJ genomic medicine*, **1**, 16027-1-16027–10. doi:10.1038/npjgenmed.2016.27

Zhang, J., Cai, J., Zhang, X., Ni, J., Guo, Z., Zhang, Y., Lu, W., and Zhang, C. (2013). Does the Bipolar Disorder-Associated CACNA1C Gene Confer Susceptibility to Schizophrenia in Han Chinese? *Journal of Molecular Neuroscience*, **51**(2), 474–477. doi:10.1007/s12031-013-0079-4

Chapter 2

Rare deleterious mutations are associated with disease in bipolar disorder families

2.1 Abstract

Bipolar disorder (BD) is a common, complex and heritable psychiatric disorder characterized by episodes of severe mood swings. The identification of rare, damaging genomic mutations in families with BD could inform about disease mechanisms and lead to new therapeutic interventions. To determine whether rare, damaging mutations shared identity-by-descent in families with BD could be associated with disease, exome sequencing was performed in multigenerational families of the NIMH BD Family Study followed by *in silico* functional prediction. Disease association and disease specificity was determined using 5090 exomes from the Sweden-Schizophrenia Population-Based Case-Control exome sequencing study. We identified 14 rare and likely deleterious mutations in 14 genes that were shared identity-by-descent among affected family members. The variants were associated with BD (P < 0.05 after Bonferroni's correction) and disease specificity was supported by the absence of the mutations in patients with schizophrenia (SZ). In addition, we found rare, functional mutations in known causal genes for neuropsychiatric disorders including holoprosencephaly and epilepsy. Our results demonstrate that exome sequencing in multigenerational families with BD is effective in identifying rare genomic variants of potential clinical relevance and also disease modifiers

related to coexisting medical conditions. Replication of our results and experimental validation are required before disease causation could be assumed.

2.2 Introduction

Bipolar disorder (BD) is a severe psychiatric disorder characterized by episodes of extremely elevated, expansive or irritable mood, grandiosity, flight of ideas, distractibility or agitation, which could lead to marked impairment in social and occupational functioning (American Psychiatric Association, 2013). Episodes of mania are often followed by severe and disabling depression. In general, BD is conceptualized as a complex disease with genetic and environmental risk factors (Craddock and Jones, 1999). Heritability estimates range from 58% to 93% with a monozygotic twin concordance rate of about 0.43 (Kieseppä *et al*., 2004; Song *et al.*, 2015). Nevertheless, the etiology of the disease remains unknown. Linkage studies and genome-wide association studies (GWAS) have suggested chromosomal and genomic regions potentially related to BD, but the identification of disease causing variants remains largely elusive (Craddock and Sklar, 2013; Kerner, 2014). Exome-wide sequencing offers now a new opportunity to lead these investigations to a new level.

BD is a common psychiatric disorder with a population prevalence of 2-3% (Kessler *et al.*, 2005a; Kessler *et al*., 2005b). However, families in which the disorder is transmitted over several generations are very rare. In the hope of finding genetic risk factors for BD with strong effect, the National Institute of Mental Health (NIMH) ascertained a number of these families in which a Mendelian mode of transmission was suggested by the pattern of disease segregation (Nurnberger *et al.*, 1997). However, after initial enthusiasm it was quickly realized that a single genetic risk factor with strong effect would most likely not explain the susceptibility to BD even in individual families (O'Rourke *et al*., 1983; Crow and DeLisi, 1998; DeLisi and Crow, 1998). Instead, mathematical model fitting suggested an oligogenic risk profile as the most likely cause of the disease, but indicated also substantial interfamilial heterogeneity (Craddock *et al.*, 1995). Early linkage studies were not equipped to perform well under this scenario and knowledge about the human genome was still in its infancy.

Although these early attempts had been unsuccessful in finding rare disease-causing genes in BD, the search for common genomic polymorphisms as disease modifiers of BD dominated the literature. Many reviews on this subject have been published and it is beyond the scope of this paper to cover this extensive literature. Instead, it is the intent of this paper to collect and present supporting evidence for the hypothesis that rare mutations might contribute to the risk of developing BD under an oligogenic mode of inheritance.

With human genome data available and falling sequencing costs, the time seems to be right to revisit the original models of disease transmission in the families of the NIMH BD genetics initiative. We conducted a family-based exome sequencing study in multigenerational families of the NIMH to test the hypothesis that several rare functional mutations in gene-coding regions are co-transmitted over several generations and shared identity-by-descent among the affected family members. We expected that the mutated genes would cluster into functional pathways suggesting potential disease path mechanisms. Large, population-based samples of patients with schizophrenia and healthy controls were also available to test disease association and disease specificity.

2.3 Materials and Methods

2.3.1 Sample selection

The analysis presented in this article was based on publicly available data and biomaterial from families of the NIMH-Bipolar Genetics Initiative (NIMH Genetics—Bipolar Disorder, 2015). We selected nine affected individuals from four Caucasian families in which BD was transmitted over several generations following an apparently Mendelian mode of inheritance. In three families, we selected the two most distantly related affected family members for exome-wide sequencing. In one family, we selected three affected individuals, as the disease appeared to be transmitted through the paternal and the maternal lineage. The ethnicity of the individuals was determined based on self-report. All affected and unaffected family members, and also the independent patients had been interviewed with the Diagnostic Interview for Genetic Studies by trained health care professionals blinded to the clinical diagnosis. The Diagnostic Interview for Genetic Studies is an extensively validated, structured clinical instrument developed by principal investigators at the NIMH for the assessment and differential diagnosis of major mood and psychotic disorders. Medical and psychiatric comorbidities were also recorded (Nurnberger *et al.*, 1994). Non-hierarchical Best Estimate consensus diagnoses were reached by at least three independent raters according to DSM-IV criteria (Leckman *et al.*, 1982; American Psychiatric Association, 2000). In addition, we randomly selected six unrelated individuals with BD for exome-wide sequencing, who had been evaluated under the same procedures (Fromer *et al.*, 2012).

2.3.2 Exome sequencing and bioinformatics analysis

DNA was isolated from immortalized lymphoblastoid cell lines. Genomic DNA extraction, library preparation, sequencing, and data analysis were performed using established procedures. Exome capture was carried out using the Illumina TruSeq Exome Enrichment Kit (Illumina, San Diego, CA, USA) and the DNA was sequenced using the HiSeq 2000 for a 100-bp paired-end run (Illumina). An average of 50 million independent paired reads were generated per sample to provide a mean 10-fold coverage across the RefSeq protein-coding exons and flanking intronic sequence (±2 bp) of 487.5% of these bases and a mean 20-fold coverage of 78.9% of the targeted sequences (Supplementary Methods). As technical controls during the sequencing process and to guard against technical artifacts, we used the DNA of 200 unrelated individuals who were sequenced in our laboratory under the same exon capture and sequencing conditions.

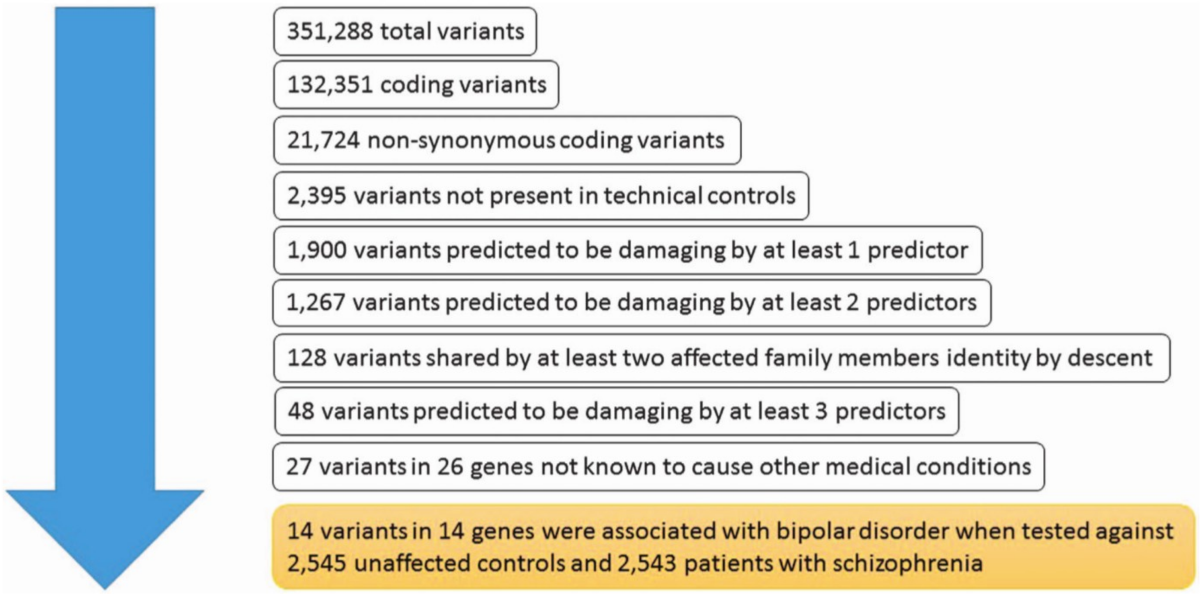

351,288 total variants
132,351 coding variants
21,724 non-synonymous coding variants
2,395 variants not present in technical controls
1,900 variants predicted to be damaging by at least 1 predictor
1,267 variants predicted to be damaging by at least 2 predictors
128 variants shared by at least two affected family members identity by descent
48 variants predicted to be damaging by at least 3 predictors
27 variants in 26 genes not known to cause other medical conditions
14 variants in 14 genes were associated with bipolar disorder when tested against 2,545 unaffected controls and 2,543 patients with schizophrenia

Figure 2.1. Selection algorithm for rare variants in families with bipolar disorder. The figure delineates the algorithm that was used to select potentially disease-causing mutations in four families with bipolar disorder.

2.3.3 Variant annotation, filtering and interpretation

Single-nucleotide variants and small structural variants including insertions and deletions were annotated using Golden Helix SNP & Variation Suite (SVS) v8.1 (Bozeman, MT, USA). Variants were filtered based on evidence for identity-by-descent sharing among affected family members, minor allele frequency $\leqslant$ 0.01%, and predictions regarding consequence on protein function by the following in silico prediction tools: SIFT, PolyPhen 2, LRT, MutationTaster, Mutation Assessor and FATHMM (Ng and Henikoff, 2001; Ramensky *et al.*, 2002; Chun and Fay, 2009; Adzhubei *et al.*, 2010; Pollard *et al.*, 2010; Schwarz *et al*., 2010; Reva *et al*., 2011) (Figure 2.1). The filtered variants were then genotyped in additional affected family members. In addition, all selected variants were also genotyped in at least one unaffected family member per family. On the basis of these results, we selected variants that were present in the affected family members and absent in the unaffected family members. Finally, we used the exome data from the Sweden-Schizophrenia Population-Based Case-Control Exome Sequencing data set (dbGAP accession: phs000473.v1.p1) for a case–control association analysis on the selected variants. This data set contained exomic data of 2545 individuals with SZ and 2545 controls.

2.3.4 Statistical analysis

To determine the statistical significance of mutation frequency differences between cases and controls, we used the Fisher's exact test for rare variants (St. Pierre *et al*., 1976) and corrected for multiple testing using the Bonferroni procedure (Simes, 1986). In this analysis, the family was considered to be the unit of observation because only variants shared among the affected family members were included in the analysis. Pathway analysis and gene set-enrichment analysis of variants that were significant in the Fisher's exact test were performed in the

Database for Annotation, Visualization, and Integrated Discovery (DAVID) Bioinformatics Resource 6.7 (Benjamini and Hochberg, 1995; Dennis *et al*., 2003; Huang *et al*., 2009a; Huang *et al.*, 2009b).

2.4 Results

2.4.1 Sample characteristics

In four multigenerational families, multiple individuals were affected with a severe and complex type of BD (Table 2.1). The patients had been diagnosed with BD on average at 18 years of age (s.d. = 7.7), and at the time of interview, the majority of the patients had been ill for at least 15 years. Only one-fifth of the patients were male (20%). Almost all selected patients (93%) had been diagnosed with BD type 1 (BD1) according to DSM-IV criteria, but one independent patient carried the diagnosis of BD type 2 (BD 2). Eight patients (53%) fulfilled criteria for rapid cycling BD, a disease subtype characterized by at least four separate mood episodes over the course of one year. Ten patients (67%) had experienced symptoms of hallucinations and/or delusions, and ten patients (67%) had attempted suicide at least once during the disease course. All patients had been diagnosed with one or more psychiatric comorbidities, including anxiety disorders (73%), substance-use disorders (60%), attention-deficit hyperactivity disorder (40%),

obsessive compulsive disorder (27%), sleep disorders (27%), eating disorders (20%) and antisocial personality disorder (20%). In addition, some patients also had medical disorders that could have contributed to the phenotype variability. Among these disorders were migraine (67%), seizure disorders (33%), thyroid disorders (20%), gastrointestinal disorders (20%), metabolic disorders (13%) and cardiovascular disorders (7%). Almost half of the sample had been diagnosed with learning disability (40%).

Table 2.1. Phenotype of affected individuals in four families with bipolar disorder

	N=*15 (%)*
Age, years (SD)	38.4 (15.6)
Age of onset of bipolar disorder, years (SD)	18.2 (7.7)
Gender, male	3 (20)
Diagnosis of bipolar disorder type 1,	14 (93)
Rapid cycling	8 (53)
Suicide attempts	10 (67)
Psychosis	10 (67)
Psychiatric comorbidity	
Anxiety disorder	11 (73)
Attention deficit hyperactivity disorder	6 (40)
Substance use disorder	9 (60)
Obsessive compulsive disorder	4 (27)
Antisocial personality disorder	3 (20)
Medical comorbidity	
Seizure disorder	5 (33)
Migraine	10 (67)
Disorders of the endocrine system	3 (20)
Disorders of the metabolic system	2 (13)
Disorders of the cardiovascular system	1 (7)
Disorders of the gastrointestinal system	3 (20)
Learning disability	6 (40)
Sleep disorder	4 (27)
Eating disorder	3 (20)

2.4.2 Identification of rare, damaging mutations

Whole-exome sequencing and genotyping of the 15 affected individuals identified 14 rare and likely damaging mutations that were shared identity-by-descent. The mutations were absent in the unaffected family members and also in the technical controls (Figure 2.1, Table 2.2). Seven of these mutations were novel and seven variants had been described previously in un-phenotyped population samples at very low frequency (Table 2.3). The variants were of high

Table 2.2. List of mutated genes in bipolar disorder families

Entrez #	*Gene*	*Name*	*Function*
9743	*ARHGAP32*	Rho GTPase activating protein 32	GTPase activator activity (GO:0005096), phosphatidylinositol binding (GO:0035091)
60314	*C12orf10*	chromosome 12 open reading frame 10	Locomotory exploration behavior (GO:0035641)
84516	*DCTN5*	dynactin 5 (p25)	Centrosome (GO:0005813)
2060	*EPS15*	epidermal growth factor receptor pathway substrate 15	calcium ion binding (GO:0005509) protein binding (GO:0005515)
2568	*GABRP*	Gamma-Aminobutyric Acid (GABA) A Receptor, Pi	GABA-A receptor activity (GO:0004890)
115399	*LRRC56*	leucine rich repeat containing 56	Unknown
4649	*MYO9A*	myosin IXA	Regulation of small GTPase-mediated signal transduction (GO:0051056)
84700	*MYO18B*	myosin XVIIIB	Vasculogenesis (GO:0001570)
284434	*NWD1*	NACHT and WD repeat domain containing 1	ATP binding (GO:0005524)
5286	*PIK3C2A*	phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 alpha	1-phosphatidylinositol-3-kinase activity (GO:0016303)
4660	*PPP1R12B*	Protein Phosphatase 1, Regulatory Subunit 12B	Small GTPase-mediated signal transduction (GO:0007264)
5829	*PXN*	Paxillin	Activation of MAPK activity (GO:0000187)
374403	*TBC1D10C*	TBC1 domain family, member 10C	Regulation of Rab GTPase activity(GO:0032313)
7158	*TP53BP1*	tumor protein p53 binding protein 1	RNA polymerase II activating transcription factor binding (GO:0001102)
143630	*UBQLNL*	ubiquilin-like	Protein binding (GO:0005515)
146862	*UNC45B*	Unc-45 Homolog B (C. Elegans)	Chaperone-mediated protein folding (GO:0061077)

quality and predicted to be damaging for the protein structure or function by at least three functional predictors (Supplementary Tables 2.S1 and 2.S2). In addition, we found one novel frameshift mutation and one known, rare deletion/insertion mutation, both with unknown functional consequences (Table 2.3). The mutations were private to the individual families, in which they were discovered, and none of the mutations were present in 2545 ethnically matched controls ($P \leqslant 1.6 \times 10^{-3}$). Furthermore, none of the 2545 exomes of patients with SZ carried the same mutations, indicating disease specificity. Three of the mutated genes, myosin IXA (*MYO9A*), TBC1 domain family, member 10C (*TBC1D10C*) and Rho GTPase activating protein 32 (*ARHGAP32*) had GTPase-activating function, but *in silico* analysis in DAVID revealed no

Table 2.3. Molecular characteristics of mutations in families with bipolar disorder

Location	*Chrom*	*Gene*	*Identifier*	*Transcript*	*Exon*	*Coding*	*Protein*
11:128838929	11q24.3	ARHGAP32	Novel	NM_014715	13	c.5090G4T	p.Gly1697Val
12:53694010	12q13.13	C12orf10	Novel	NM_021640	2	c.293A4G	p.Tyr98Cys
16:23672532	16p12.2	DCTN5	Novel	NM_001199743	4	c.278T4C	p.Ile93Thr
1:51826856	1p32.3	EPS15	rs148821171	NM_001159969	12	c.1589C4T	p.Ala530Val
5:170238979	5q35.1	GABRP	Novel	NM_014211	10	c.1040A4T	p.Glu347Val
11:549982	11p15.5	LRRC56	Novel	NM_198075	7	c.407_408insT	p.Ser136fs
15:72191038	15q23	MYO9A	Novel	NM_006901	25	c.3806G4A	p.Arg1269Gln
22:26224877	22q12.1	MYO18B	rs373113816	NM_032608	15	c.2921G4A	p.Arg974His
19:16860396	19p13.11	NWD1	rs148848880	NM_001007525	6	c.943C4T	p.Arg315Cys
11:17172051	11p15.1	PIK3C2A	Novel	NM_002645	3	c.1321T4G	p.Cys441Gly
1:202398004	1q32.1	PPP1R12B	rs199816573	NM_001167857	6	c.868G4A	p.Ala290Thr
12:120650260	12q24	PXN	Novel	NM_025157	11	c.1132C4T	p.Arg378Cys
11:67172591	11q13.2	TBC1D10C	rs201081455	NM_198517	3	c.188G4A	p.Arg63Gln
15:43762077	15q15.3	TP53BP1	rs28903074	NM_001141979	11	c.1362_1367delTATCCC	p.454_456delinsPro
11:5537397	11p15.4	UBQLNL	rs7933557	NM_145053	1	c.275A4T	p.Asp92Val
17:33504148	17q12	UNC45B	rs137917897	NM_001033576	16	c.2138G4A	p.Arg713Gln

statistically significant clustering of the mutated genes in any known pathophysiological pathway.

In addition to these 16 variants, we discovered two known, rare, compound heterozygous variants in the gene solute carrier family 22 (organic cation transporter), member 1 (*SLC22A1*) in one severely affected individual. The first mutation (rs55918055) was inherited through the paternal lineage and the second mutation (rs34059508) was inherited through the maternal lineage. These non-synonymous coding mutations were predicted to be deleterious. We also identified mutations in known, disease-causing genes for several medical conditions that could have had disease-modifying effects (Supplementary Table 2.S3). For example, a patient with seizure disorder carried a mutation in the gene prickle homolog 1 (Drosophila)] (*PRICKLE1*), a known gene for progressive myoclonic epilepsy 1B (EPM1B, MIM:612437). In one family, a novel mutation in the gene dispatched homolog 1 (Drosophila) (*DISP1*) segregated with the disease phenotype. Mutations in *DISP1* are known to cause holoprosencephaly (HPE) type 2-4

(HPE2, MIM:157170; HPE3, MIM:142945; HPE4, MIM:142946; HPE5, MIM:609637), and in addition, this gene is also known as the main suspect in the Chromosome 1q41–q42 deletion syndrome (MIM:612530). Although epilepsy and HPE could present with seizures, mood symptoms, psychosis, developmental delay and learning disabilities, mutations in these two genes could explain some of the neuropsychiatric phenotypes that segregated in two of the families. The gene Ankyrin Repeat and Kinase Domain Containing 1 (*ANKK1*), which has been related to migraine and alcohol dependence, (Neville *et al*., 2004; Ridge *et al*., 2009; Ghosh *et al.*, 2013) also carried a likely damaging mutation. The gene T-box 2 (*TBX2*) has been related to cognitive and behavioral abnormalities in the chromosome 17q23.1–q23.2 deletion syndrome (MIM:613355), and toll-like receptor 5 (*TLR5*) has been associated with systemic lupus erythematosus (SLEB1, MIM:601744). None of the variants could be replicated in the independent patients with BD.

2.5 Discussion

We identified rare, deleterious and likely disease-causing mutations in gene-coding regions through unbiased, exome-wide sequencing in families with BD. Each family carried rare mutations in several genes that were shared identity-by-descent by affected family members and the variants were absent in the unaffected family members. All variants were predicted to be damaging by several in silico functional predictors. In each several rare, damaging mutations were associated with the disease. These findings are consistent with the currently favored hypothesis of oligogenic disease causation in BD (Neuman and Rics, 1992; Gershon, 2000).

Exome-sequencing is increasingly utilized to identify very rare and likely disease-causing mutations in many neuropsychiatric disorders (Binder, 2012). Our focus on rare and even private mutations is consistent with current trends in genetic epidemiologic research; however, our study is one of the first to examine the exomes of BD patients from multigenerational families in an unbiased, genome-wide approach and to evaluate the results in the context of a large number of population-based healthy controls and patients with SZ. The results of this study reveal a complex scenario of rare and private missense and loss-of-function mutations in novel candidate genes. In addition, we found mutations in known disease-causing genes for medical conditions that could have potentially had disease-modifying effects, for example on intellectual ability or immune status.

Our results could be viewed in the context of previously published linkage analyses in the families of the NIMH genetics initiative. Genome-wide significant linkage signals have been reported in the chromosomal regions 16p12.2 and 17q12 (Dick *et al.*, 2003; Cheng *et al.*, 2006). The region 16p12.2 has also been linked to the sub-phenotype psychosis and suggestive linkage has been found to the chromosomal regions 19p13 with the same phenotype (Cheng *et al.*, 2006). However, when considering linkage results it has to be kept in mind that linkage regions on average contain hundreds of genes, and therefore, conclusions about supporting evidence of linkage results should be viewed with great caution.

Our conclusion about a causal relationship between the described variants and BD is plausible, and coherent with some pathophysiological theories. Especially GTPase-activation is a pathophysiological process that is supported by animal models and cell culture experiments (Kalkman, 2012; Akula *et al.*, 2014; Farhy *et al.*, 2014) GTPases are a target of lithium, a drug frequently used to treat BD; and therefore, a role for G-proteins in disease processes of BD has long been hypothesized (Drummond *et al*., 1988; Schreiber and Avissar, 1991; Kõks *et al*., 2004; Roybal *et al.*, 2007; Must *et al.*, 2009; Kõks *et al*., 2011; Lee *et al*., 2011; Corena-McLeod *et al*., 2013; Lee *et al.*, 2013; Farhy *et al.*, 2014; Gonzalez, 2014; Naismith *et al*., 2014; Soreca 2014; Srinivasan *et al.*, 2014; Bellivier *et al*., 2015; Carpenter *et al.*, 2015).

The patients with BD had also been diagnosed with a number of medical and neurological disorders including seizure disorders and learning disability. Therefore, it is highly likely that some of the identified genes might in fact be related to these disorders rather than to BD itself. In fact, we were able to identify rare mutations in genes that have previously been linked to seizure disorders and HPE. These conditions could potentially modify the disease expression and the disease course of BD. As none of the protein-damaging mutations were present in patients with SZ, shared genetic risk factors between BD and SZ might be uncommon.

Limitations of our study are (1) the very small sample size of BD patients in this data set. This limitation could result in an underestimation or overestimation of the effect size of these rare and private mutations. Given the rare nature of the variants in the general population, replication of individual variants is highly unlikely. Another limitation of our analysis was the dependence on *in silico* functional predictions. Many examples indicate that these predictions might not always reflect the true biological, cell-specific consequences of a specific mutation on an individual's genetic background. Therefore, it is recommended to test the functional consequences of the identified mutations and experimentally validate the effect in cell culture assays and in in vivo models. Despite obvious limitations, our results are consistent with previous publications in the literature. For example, several groups have identified rare functional mutations in BD families (Song *et al.*, 2010; Green *et al.*, 2011; Goes *et al*., 2016), even though statistical significance after correction for multiple testing in larger samples still needs to be established. In addition, rare structural variants have been associated with BD, but the functional consequences of these variants remain to be determined (Mehta *et al.*, 2014; Kember *et al*., 2015).

Although individual mutations and genes still require further support before generalizable conclusion can be drawn, it has become clear that BD is by far more heterogeneous than previously anticipated. Our results support a rare-variant oligogenic disease models in families with BD and stress the importance of protein-coding regions. On the basis of these results, we recommend to fund studies that focus on multi-generational families to identify functional mutations. Furthermore, in clinical practice, it should be recognized that in some families, BD might be transmitted with higher risk than generally anticipated in the framework of a common complex disorder, and that genetic counseling might be recommended. Exome-wide sequencing could be useful in high-risk families to identify known disease-causing mutations for neuropsychiatric disorders that might resemble BD, such as HPE and seizure disorders.

2.6 Conclusions

The results of our study indicate that rare, deleterious mutations in gene-coding regions could be related to a BD phenotype in families, in which the disease is transmitted over several generations. Exome sequencing in multigenerational families with BD is effective in identifying rare genomic variants with potential clinical relevance. Our results further support the rare-variant oligogenic disease model of BD. The disease association of the identified mutations need to be replicated and the functional consequences of the mutations validated before the information could be used in clinical settings.

2.7 Supplementary Methods

2.7.1 Exome capture and re-sequencing

Genomic DNA was isolated from cell lines following standard protocols. After DNA quality control, exome capture was performed using the Illumina TruSeq™ Exome Enrichment Kit (http://www.illumina.com/documents/products/datasheets/datasheet_truseq_exome_enrichment_kit.pdf). This method offers uniform coverage across 62 Mb of exome sequence including 5'UTRs, 3'UTRs, microRNAs and other non-coding RNAs. Libraries were prepared using the TruSeq DNA sample preparation kit (Illumina Inc., San Diego, CA). The assay is designed to target 201,121 exons in 20,794 genes (based on the NCBI37/hg19 reference genome) covering about 97% of the CCDS coding exons (Pruitt *et al.*, 2009) and 96% of the RefSeq coding exons (Pruitt *et al.*, 2012). The exome capture was performed according to standard protocols and was followed by sample amplification, cluster generation, and sequencing of 100 base paired ends on the Illumina HiSeq 2000 Sequencing platform. Sequencing was done either at the *UCLA Broad Stem Cell Research Center* (*BSCRC*) or the UCLA Clinical Genomics Center. In order to maximize coverage and minimize cost we ran 4 or 6 exomes per lane, expecting average coverage of 40X per exome. The output images were analyzed using the Illumina real time analysis software.

2.7.2 Quality control (QC)

For sample quality control, artificial double stranded DNA targets were incorporated during the sample preparation step. These sequences were targeted during the sequence capture process and were used for quality control of the enzymatic steps and troubleshooting during sample preparation. A standard phiX DNA library equally distributed over all 8 sequencing lanes served as quality control during the sequencing step. Libraries loaded at a minimum of 500 000 reads per mm^2; more than 85% of the bases passed quality measures of greater than 30, based on the Illumina RTA analysis. There were a minimum of 50 million reads per sample and the percentage of reads with non-matching barcodes did not exceed four times the error rate.

2.7.3 Sequence alignment and variant calling

Sequencing reads were aligned to the reference genome using the Novoalign v3.04.01. software package (Hercus, 2009). Variants were called using the Genome Analysis Toolkit (GATK) v3.5 (McKenna *et al.*, 2010), which was run in multiple sample mode with additional exomes from our laboratory, according to published Best Practice recommendations (DePristo *et al.*, 2011; Van de Auwera *et al.*, 2002). The total number of novel and known variants was assessed and compared to the expected values based on the average number of variants called in our laboratory. We obtained additional quality control metrics using the GATK unless otherwise noted. These included alignment summary, GC bias, hybridization selection (Picard), insert size, mean quality by cycle, quality score distribution (Picard), capture specificity, efficiency of the target capture reaction at 10x coverage, percentage of duplicate reads, percentage of reads mapped to a reference, percentage of variants called in dbSNP135, and a summary of coverage statistics for the targeted regions. In addition, we confirmed sib ships by running PLINK (Purcell *et al.*, 2007) to calculate a matrix of genome-wide average IBS pairwise identities and verifying that closely related individuals had higher relatedness values and that there was no inbreeding within families.

The average coverage was 55.4X per targeted base, and 59.7% of the aligned bases mapped to the targeted regions. On average, 87.5% of the targeted bases in each individual were supported by at least 10 independent sequence reads, and 78.9% were supported by at least 20 independent reads. On average, there were approximately 19,985 dbSNP sites and 484 novel SNPs per sample. Other quality control metrics such as the ratio of heterozygous to homozygous SNPs, the ratio of transitions to transversions, and dbSNP concordance rate were within normal ranges.

2.7.4 Variant annotation, filtering and interpretation

After quality control, we performed the variant annotation using Golden Helix *SNP & Variation Suite v8.1* (Bozeman, MT: Golden Helix, Inc.; Available: http://www.goldenhelix.com) and in addition, we used custom annotation tables that were created using software that was developed in-house (Yourshaw *et al.*, 2014). We incorporated information from UniProt (The Uniprot Consortium, 2011) and GERP conservation scores from the mammalian alignment set (Cooper *et al.*, 2005). Minor allele frequencies of variants and counts for the number of individuals carrying a mutation were calculated from allele frequencies in the Phase 1 Integrated Variant Call Set of the 1000 Genomes Project and the NHLBI Exome Sequencing Project's Exome Variant Server (EVS) (Database of Single Nucleotide Polymorphisms (dbSNP); The 1000 Genomes Project Consortium, 2010; Exome Variant Server, NHLBI Exome Sequencing Project (ESP); NIEHS Environmental Genome Project).

Variants were filtered based on minor allele frequency (MAF) in both the 1000 Genomes Project and the European EVS subpopulation, functional consequences for the encoded protein (Gorlov *et al.*, 2008; Kryukov *et al.*, 2007), and evidence for identity by descent sharing among affected family members. In addition, variants were excluded from the analysis if they were present in more than one percent of our technical controls, or predicted to be benign by all three functional predictors, SIFT (Kumar *et al.*, 2009), PolyPhen2 (Adzhubei *et al.*, 2010), and MutationTaster (Schwarz *et al.*, 2010). We selected variants that were likely loss of function mutations even though functional prediction scores were not available, such as initiator codon changes, frameshift indels, stop gain/loss mutation, or splice site disrupting variants.

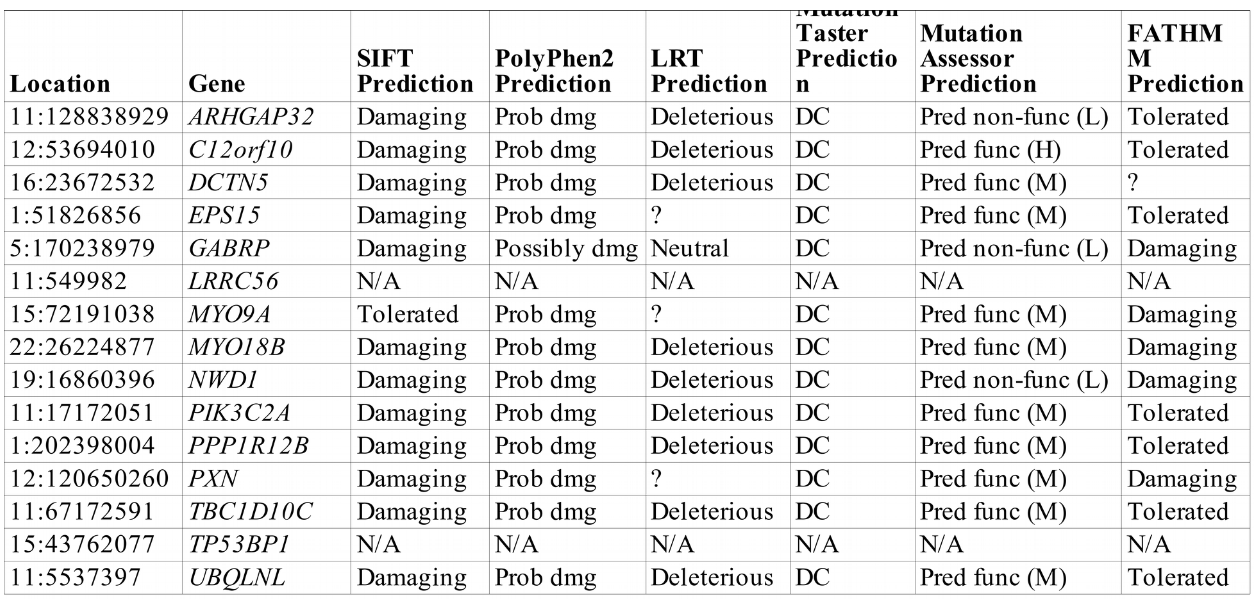
Then, we applied a stringent selection algorithm, which prioritized variants based on likely loss of function or uniform prediction of deleterious consequences by at least three predictors. A manual literature review was performed for all genes containing the selected variants. We also performed a pathway analysis on the selected genes using the Database for Annotation, Visualization and Integrated Discovery system (DAVID) (Dennis *et al.*, 2003).

2.8 Supplementary Tables

Supplementary Table 2.S1: Sequencing quality of selected mutations in families with bipolar disorder

Location	**Gene**	**Identifier**	**Quality Score**	**Filter Result**	**Avg Depth**
11:128838929	*ARHGAP32*	novel	1054.11	Pass	59.9484
12:53694010	*C12orf10*	novel	1340.32	Pass	94.4452
16:23672532	*DCTN5*	novel	847.87	Pass	130.8452
1:51826856	*EPS15*	rs148821171	2012.58	Pass	86.1419
5:170238979	*GABRP*	novel	1162.23	Pass	95.9226
11:549982	*LRRC56*	novel	858.58	Pass	70.7613
15:72191038	*MYO9A*	novel	2020.3	Pass	117.1935
22:26224877	*MYO18B*	rs373113816	639.86	Pass	59.9419
19:16860396	*NWD1*	rs148848880	2254.24	Pass	94.9226
11:17172051	*PIK3C2A*	novel	1076.19	Pass	72.4839
1:202398004	*PPP1R12B*	rs199816573	1489.24	Pass	74.4065
12:120650260	*PXN*	novel	834.52	Pass	40.9548
11:67172591	*TBC1D10C*	rs201081455	1445.25	Pass	68.1307
15:43762077	*TP53BP1*	rs28903074	4725.65	Pass	89.0452
11:5537397	*UBQLNL*	rs7933557	2693.17	Pass	110.8194
17:33504148	*UNC45B*	rs137917897	1396.38	Pass	64.9548

Supplementary Table 2.S2: *In silico* functional predictions for selected mutations in families with bipolar disorder



Location	Gene	SIFT Prediction	PolyPhen2 Prediction	LRT Prediction	Mutation Taster Prediction	Mutation Assessor Prediction	FATHMM Prediction
11:128838929	*ARHGAP32*	Damaging	Prob dmg	Deleterious	DC	Pred non-func (L)	Tolerated
12:53694010	*C12orf10*	Damaging	Prob dmg	Deleterious	DC	Pred func (H)	Tolerated
16:23672532	*DCTN5*	Damaging	Prob dmg	Deleterious	DC	Pred func (M)	?
1:51826856	*EPS15*	Damaging	Prob dmg	?	DC	Pred func (M)	Tolerated
5:170238979	*GABRP*	Damaging	Possibly dmg	Neutral	DC	Pred non-func (L)	Damaging
11:549982	*LRRC56*	N/A	N/A	N/A	N/A	N/A	N/A
15:72191038	*MYO9A*	Tolerated	Prob dmg	?	DC	Pred func (M)	Damaging
22:26224877	*MYO18B*	Damaging	Prob dmg	Deleterious	DC	Pred func (M)	Damaging
19:16860396	*NWD1*	Damaging	Prob dmg	Deleterious	DC	Pred non-func (L)	Damaging
11:17172051	*PIK3C2A*	Damaging	Prob dmg	Deleterious	DC	Pred func (M)	Tolerated
1:202398004	*PPP1R12B*	Damaging	Prob dmg	Deleterious	DC	Pred func (M)	Tolerated
12:120650260	*PXN*	Damaging	Prob dmg	?	DC	Pred func (M)	Damaging
11:67172591	*TBC1D10C*	Damaging	Prob dmg	Deleterious	DC	Pred func (M)	Tolerated
15:43762077	*TP53BP1*	N/A	N/A	N/A	N/A	N/A	N/A
11:5537397	*UBQLNL*	Damaging	Prob dmg	Deleterious	DC	Pred func (M)	Tolerated

Abbreviations: Prob dmg: Probably damaging. Possibly dmg: Possibly damaging. DC: Disease causing. Pred: Predicted. non-func: non-functional. func: functional. L: low. M: medium. H: high.

Supplementary Table 2.S3: Variants in bipolar disorder families known to cause different disorders

See material attached with the dissertation.

2.9 Bibliography

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, **7**(4), 248–249.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association. (DSM-5 ®), Fifth Edition. Arlington, VA, 2013.

Carpenter, J. S., Robillard, R., Lee, R. S. C., Hermens, D. F., Naismith, S. L., White, D., Whitwell, B., Scott, E. M., and Hickie, I. B. (2015). The relationship between sleep-wake cycle and cognitive functioning in young people with affective disorders. *PloS One*, **10**(4), e0124710.

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, **15**(7), 901–913. doi:10.1101/gr.3577405

Craddock, N, and Jones, I. (1999). Genetics of bipolar disorder. *Journal of medical genetics*, **36**(8), 585–594.

Craddock, Nick, and Sklar, P. (2013). Genetics of bipolar disorder. *The Lancet*, **381**(9878), 1654–1662.

Cross-Disorder Group of the Psychiatric Genomics Consortium, and Genetic Risk Outcome of Psychosis (GROUP) Consortium. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, **381**(9875), 1371–1379.

Crow, T. J., and DeLisi, L. E. (1998). The chromosome workshops at the 5th International Congress of Psychiatric Genetics - the weight of the evidence from genome scans. *Psychiatric Genetics*, **8**, 59–61.

Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 135). Available from: http://www.ncbi.nlm.nih.gov/SNP/.

DAVID Bioinformatics Resource 6.7; available from https://david.ncifcrf.gov/content.jspfile=citation.htm

DeLisi, L. E., and Crow, T. J. (1999). Chromosome Workshops 1998: current state of psychiatric linkage. *American Journal of Medical Genetics*, **88**(3), 215–218.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, **4**(5), P3.

DePristo, M. A., Banks, E., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**(5), 491–498.

Dunn, O. J. (1959). Estimation of the Medians for Dependent Variables. *The Annals of Mathematical Statistics*, **30**(1), 192–197.

Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) ESP5400 December 10, 2011.

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., Kirov, G., Sullivan, P. F., Hultman, C. M., Sklar, P., and Purcell, S. M. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *The American Journal of Human Genetics*, **91**(4), 597–607.

Gonzalez, R. (2014). The relationship between bipolar disorder and biological rhythms. *The Journal of Clinical Psychiatry*, **75**(4), e323-331.

Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R., and Amos, C. I. (2008). Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. *The American Journal of Human Genetics*, **82**(1), 100–112. doi:10.1016/j.ajhg.2007.09.006

Green, E. K., Grozeva, D., Sims, R., Raybould, R., Forty, L., Gordon-Smith, K., Russell, E., St. Clair, D., Young, A. H., Ferrier, I. N., Kirov, G., Jones, I., Jones, L., Owen, M. J., O'Donovan, M. C., and Craddock, N. (2011). DISC1 exon 11 rare variants found more commonly in schizoaffective spectrum cases than controls. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, **156B**(4), 490–492.

Hercus, C. (2009). *Novoalign*. www.novocraft.com. Accessed 16 January 2013.

Kember, R. L., Georgi, B., Bailey-Wilson, J. E., Stambolian, D., Paul, S. M., and Bućan, M. (2015). Copy number variants encompassing Mendelian disease genes in a large multigenerational family segregating bipolar disorder. *BMC Genetics*, **16**(1), 27.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., and Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **62**(6), 593–602.

Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R., and Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **62**(6), 617–627.

Kieseppä, T., Partonen, T., Haukka, J., Kaprio, J., and Lönnqvist, J. (2004). High concordance of bipolar I disorder in a nationwide sample of twins. *The American Journal of Psychiatry*, **161**(10), 1814–1821.

Kryukov, G. V., Pennacchio, L. A., and Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *American Journal of Human Genetics*, **80**(4), 727–739.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073–1081.

Hercus C. Novoalign [Internet]. 2009. Available: www.novocraft.com

Leckman, J. F., Sholomskas, D., Thompson, W. D., Belanger, A., and Weissman, M. M. (1982). Best estimate of lifetime psychiatric diagnosis: a methodological study. *Archives of General Psychiatry*, **39**(8), 879–883.

Lee, H.-J., Rex, K. M., Nievergelt, C. M., Kelsoe, J. R., and Kripke, D. F. (2011). Delayed sleep phase syndrome is related to seasonal affective disorder. *Journal of affective disorders*, **133**(3), 573–579.

Lee, H.-J., Son, G.-H., and Geum, D. (2013). Circadian rhythm hypotheses of mixed features, antidepressant treatment resistance, and manic switching in bipolar disorder. *Psychiatry Investigation*, **10**(3), 225–232.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9), 1297–1303.

Mehta, D., Iwamoto, K., Ueda, J., Bundo, M., Adati, N., Kojima, T., and Kato, T. (2014). Comprehensive survey of CNVs influencing gene expression in the human brain and its implications for pathophysiology. *Neuroscience Research*, **79**, 22–33.

Mergy, M. A., Gowrishankar, R., Gresch, P. J., Gantz, S. C., Williams, J., Davis, G. L., Wheeler, C. A., Stanwood, G. D., Hahn, M. K., and Blakely, R. D. (2014). The rare DAT coding variant Val559 perturbs DA neuron function, changes behavior, and alters in vivo responses to psychostimulants. *Proceedings of the National Academy of Sciences*, **111**(44), E4779–E4788.

Naismith, S. L., Lagopoulos, J., Hermens, D. F., White, D., Duffy, S. L., Robillard, R., Scott, E. M., and Hickie, I. B. (2014). Delayed circadian phase is linked to glutamatergic functions in young people with affective disorders: a proton magnetic resonance spectroscopy study. *BMC psychiatry*, **14**(1), 345.

NIEHS Environmental Genome Project, Seattle, WA (URL: http://evs.gs.washington.edu/niehsExome/) version 0.0.6 September 30, 2011.

NIMH Genetics—Bipolar Disorder. Available from https://www.nimhgenetics.org/available_data/bipolar_disorder/index.php (accessed on 3 March 2015).

Nurnberger, J. I., Blehar, M. C., Kaufmann, C. A., York-Cooler, C., Simpson, S. G., Harkavy-Friedman, J., Severe, J. B., Malaspina, D., and Reich, T. (1994). Diagnostic interview for genetic studies. Rationale, unique features, and training. NIMH Genetics Initiative. *Archives of General Psychiatry*, **51**(11), 849-859; discussion 863-864.

Nurnberger, J. I., DePaulo, J. R., Gershon, E. S., Reich, T., Blehar, M. C., Edenberg, H. J., Foroud, T., Miller, M., Bowman, E., Mayeda, A., Rau, N. L., Smiley, C., Conneally, P. M., McMahon, F., Meyers, D., Simpson, S., McInnis, M., Stina, O. C., Detera-Wadleigh, S., Goldin, L., Guroff, J., Maxwell, E., Kazuba, D., Gejman, P. V., Badner, J., Sanders, A., Rice, J., Bierut, L., and Goate, A. (1997). Genomic survey of bipolar illness in the NIMH genetics initiative pedigrees: A preliminary report. *American Journal of Medical Genetics*, **74**(3), 227–237.

O'Rourke, D. H., McGuffin, P., and Reich, T. (1983). Genetic analysis of manic-depressive illness. *American Journal of Physical Anthropology*, **62**(1), 51–59.

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J., Hart, E., Suner, M.-M., Landrum, M. J., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J. L., Curwen, V., et al. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, **19**(7), 1316–1323. doi:10.1101/gr.080531.108

Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, **40**(D1), D130–D135. doi:10.1093/nar/gkr1079

Purcell, S., Neale, B., Toddbrown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., Debakker, P., and Daly, M. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, **81**(3), 559–575. doi:10.1086/519795

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*, gkr407.

Roybal, K., Theobold, D., Graham, A., DiNieri, J. A., Russo, S. J., Krishnan, V., Chakravarty, S., Peevey, J., Oehrlein, N., Birnbaum, S., Vitaterna, M. H., Orsulak, P., Takahashi, J. S., Nestler, E. J., Carlezon, W. A. Jr., and McClung, C. A. (2007). Mania-like behavior induced by disruption of CLOCK. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(15), 6406–6411.

Schwarz, J. M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, **7**(8), 575–576.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N., and Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation*, **34**(1), 57–65.

SNP & Variation Suite ™ (Version 8.1). (Available from http://www.goldenhelix.com). Bozeman, MT: Golden Helix, Inc. http://www.goldenhelix.com

Song, J., Bergen, S. E., Kuja-Halkola, R., Larsson, H., Landén, M., and Lichtenstein, P. (2015). Bipolar disorder and its relation to major psychiatric disorders: a family-based study in the Swedish population. *Bipolar Disorders*, **17**(2):184–93.

Song, W., Li, W., Noltner, K., Yan, J., Green, E., Grozeva, D., Jones, I. R., Craddock, N., Longmate, J., Feng, J., and Sommer, S. S. (2010). Identification of high risk DISC1 protein structural variants in patients with bipolar spectrum disorder. *Neuroscience Letters*, **486**(3), 136–140.

St Pierre, J., Cadieux, M., Guérault, A., and Quevillon, M. (1976). Statistical tables to detect significance between frequencies in two small samples, with particular reference to biological assays. *Revue Canadienne De Biologie / Éditée Par l'Université De Montréal*, **35**(1), 17–23.

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467: 1061–1073. doi:10.1038/nature09534

The UniProt Consortium. (2011). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, **40**(D1), D71–D75.

TruSeq(TM) Exome Enrichment Kit. (n.d.). Illumina, Inc. http://www.illumina.com/documents/products/datasheets/datasheet_truseq_exome_enrichment_kit.pdf

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2002). From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc.

Y Benjamini, Y. H. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Royal Statist. Soc., Series B*, **57**, 289–300.

Yourshaw, M., Taylor, S. P., Rao, A. R., Martín, M. G., and Nelson, S. F. (2014). Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. *Briefings in Bioinformatics*, bbu008.

Chapter 3

Whole-genome sequencing of web-based recruited individuals with autism spectrum disorders reveals novel candidate genes

3.1 Abstract

Whole-exome and whole-genome sequencing have increasingly enabled new pathogenic gene variant identification for complex disorders such as autism spectrum disorder (ASD), and have provided insights into the etiology of ASD. We report on results from whole-genome sequencing (WGS) of 188 individuals, which includes 71 subjects diagnosed with ASD and their family members, recruited using a web-based platform. Gene pathways enriched among various models include cell adhesion, cell-cell signaling and nervous system development, and molecular functions enriched include ATP-dependent microtubule motor activity and calcium ion binding. Comprehensive analysis that incorporates information about a gene's population mutational burden and the relatedness between family members reveals several novel candidate ASD genes, including *STAU2* and *PPFIA3*. Our findings also provide support for the previously suggested autism gene *KMT2C*, and genes with limited prior support, including *GBX2* and *USP54*. To conclude, we identified several candidate ASD genes that shed further light onto the etiology of this disorder.

3.2 Introduction

Autism spectrum disorder (ASD) is a highly heterogeneous group of neuro-developmental disorders, and while strong familial evidence supports a substantial genetic contribution to the etiology of ASD, specific genetic abnormalities have been identified in only a small minority of all cases. Several large-scale whole-exome sequencing (WES) studies have been carried out to-date in trios and quads to elucidate causal genes underlying autism spectrum disorders (ASD) (Iossifov *et al.,* 2012; Neale *et al.,* 2012; O'Roak, Vives, Fu, *et al.*, 2012; O'Roak, Vives, Girirajan, *et al*., 2012; Sanders *et al.,* 2012). However, genes identified as containing *de novo* mutations rarely overlap between studies, and it is difficult to interpret structural variants (SVs) and copy number variants (CNVs) from whole-exome data. Recently, whole genome sequencing studies have revealed additional ASD-relevant mutations and large structural variants, which highlight the substantial genetic heterogeneity that exists in ASD (Brandler *et al.,* 2016; Yuen *et al.,* 2015). As of June 2017, 910 genes have been associated with ASD on AutDB (Basu *et al.,* 2009). However, each gene contributes to ASD risk by a small amount and many findings are only marginally significant (Vorstman *et al.,* 2017). Much larger cohorts would need to be sequenced to identify all ASD risk genes and their relevance to the many clinical sub-phenotypes (Lee *et al.,* 2010).

As the identification of genetic components of ASD has advanced rapidly in recent years, a two-class risk genetic model has emerged for autism (Ronemus *et al.,* 2014; Zhao *et al.,* 2007). In low risk families, a highly penetrant *de novo* loss of function (LOF) variant arises in a male to cause autism. In high-risk families, a mother carries the LOF variant but is unaffected, and this variant gets passed on to the male offspring in dominant fashion, who develops autism. This model, in addition to the known heterogeneity between ASD patients, suggests that a family-centered approach to next-gen sequencing (NGS) data analysis may be useful in identifying genes causal for ASD, as opposed to the commonly used strategy of focusing on *de novo* LOF variants and case-control cohorts in datasets of ever increasing size. Furthermore, we also considered a model of autosomal recessive inheritance, since this model is suggested in a subgroup of families with autism (Betancur 2011; Lim *et al.,* 2013; Ritvo *et al.,* 1985; Yu *et al.,* 2013). When combined with newly developed statistical methods, we demonstrate that this approach is successful at identifying novel ASD candidate genes.

3.3 Methods

3.3.1 Sample recruitment

We recruited ASD-affected individuals and their families from across the United States using an online recruitment process in collaboration with IAN Genetics, a project run by the Interactive Autism Network (IAN). Phenotypic data was benchmarked to verify the accuracy of using a web-based approach to autism phenotyping, and 98% of individuals were ascertained to meet criteria for ASD (Daniels *et al.,* 2012; Lee *et al.,* 2010).

Of 1266 samples collected, we sequenced 200 individuals from larger families such as trios and quads, or individuals that had phenotypes in addition to ASD, such as seizures, mental retardation, motor delay, or psychiatric disorders. Most participants were white, and ASD screening test scores, specifically SRS T-scores and SCQ Total Scores, were representative of the population for affected and unaffected individuals (Figure 3.1). 188 samples passed quality control filters. The sequenced dataset included 11 quads (consisting of the proband, at least one sibling, and parents); 6 trios (proband and both parents); 39 probands with discordant and

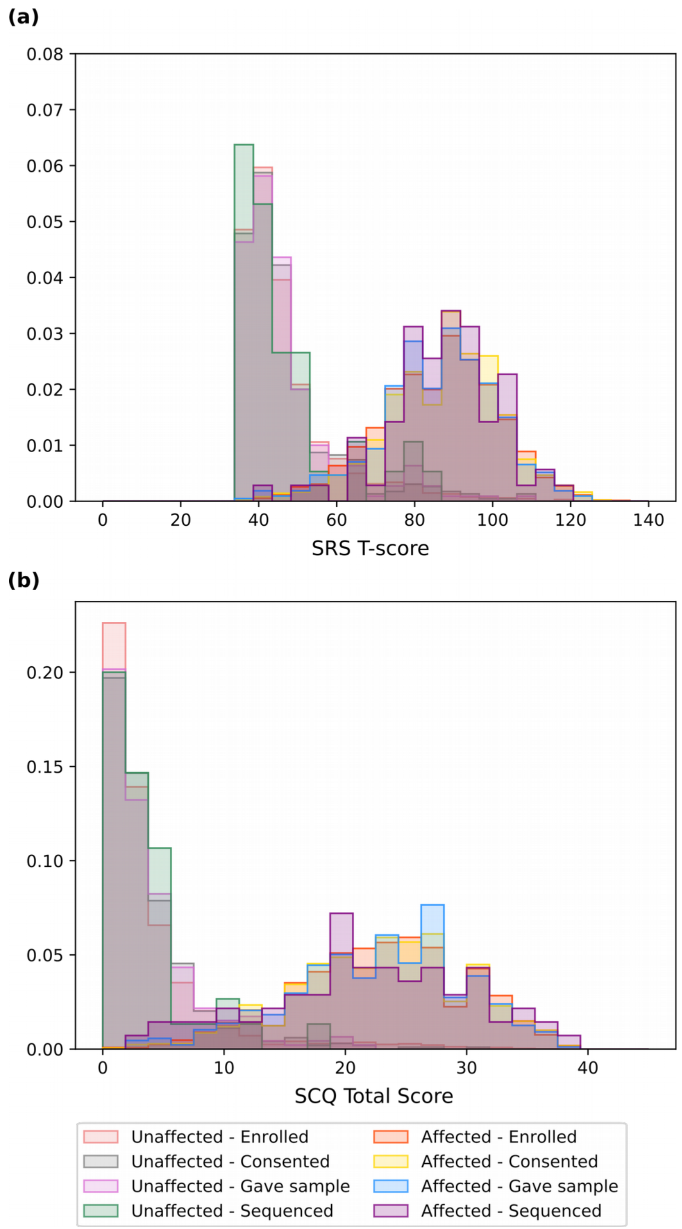


Figure 3.1. Distribution of ASD screening test scores in IAN participants

(a) Histogram of SRS T-scores of individuals enrolled in IAN, if SRS scores were available (N=5802 (affected); N=3485 (unaffected)), compared to individuals consented for participation in IAN Genetics (N=1021 (affected); N=476 (unaffected)), individuals who gave blood or saliva samples (N=442 (affected); N=228 (unaffected)), and individuals selected for sequencing (N=73 (affected); N=39 (unaffected)). SRS scores were unavailable for 7957 individuals enrolled in IAN. **(b)** Histogram of SCQ Total Scores of individuals enrolled in IAN, if SCQ scores were available (N=9483 (affected); N=6213 (unaffected)), compared to individuals consented for participation in IAN Genetics (N=1117 (affected); N=528 (unaffected)), individuals who gave blood or saliva samples (N=467 (affected); N=246 (unaffected)), and individuals selected for sequencing (N=74 (affected); N=40 (unaffected)). SRS scores were unavailable for 1565 enrolled individuals.

concordant siblings and/or one parent sequenced; and 5 unrelated, independent individuals with other phenotypes in addition to ASD. Probands had primary diagnoses that included autism; childhood disintegrative disorder (CDD); pervasive developmental disorder, not otherwise specified (PDD-NOS); and Asperger syndrome (Table 3.1).

3.3.2 Sequencing and bioinformatics pipeline

We completed whole-genome sequencing of blood- and saliva-derived genomic DNA at a mean coverage of 35.8×. All samples were sequenced on the Illumina HiSeq X Ten sequencer using a 150-base paired-end single-index read format. After quality control, reads were mapped to human reference hg38 using Illumina Sequence Integration Software (ISIS) (http://support.illumina.com/sequencing/sequencing_instruments/miseq/downloads.ilmn). Samples that did not meet desired coverage metrics were re-sequenced and the data was merged

Table 3.1. Phenotypes of affected individuals in 61 families with ASD

	N = 71, %
Gender, male	54 (76)
Primary diagnosis	
Childhood disintegrative disorder	1 (1.4)
Autism	42 (58)
PDD-NOS	13 (17.8)
Asperger syndrome	7 (9.6)
Other ASD	8 (11.0)
Medical comorbidity	
Seizure disorder or epilepsy	17 (24)
Mental retardation	6 (8.5)
Motor delay	43 (61)
Cerebral Palsy	3 (4.2)
Psychiatric comorbidity	
Depression	1 (1.4)
Bipolar disorder	3 (4.2)
Attention-deficit hyperactivity disorder	22 (31)

with the first run.

After variant calling and structural variant calling, variants were annotated using GoldenHelix VarSeq v.1.4.2 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com) for protein consequence, predicted deleteriousness, and information on protein-coding genes extracted from the NCBI RefSeq gene annotation release 107. Additional annotations were obtained from the GnomAD, dbNSFP and SORVA datasets (Liu *et al.,* 2016; Rao and Nelson 2017; Samocha *et al.,* 2014). Protein coding variants were filtered according to multiple criteria. We filtered out variants that were not reliably called. Next, we filtered out synonymous variants, and we filtered out missense variants with a CADD PHRED-scaled score < 10, which would suggest that the variant is not predicted to have deleterious consequences(Kircher *et al.,* 2014). Additionally, we filtered out variants with a minor allele frequency (MAF) <= 0.1% in the NHLBI ESP dataset and those that were called in more than two independent, unaffected samples. Candidate *de novo* mutations were further filtered, and we excluded variants that were called in >= 1 technical control, or if we observed more than one read in either parent that supported the candidate *de novo* mutations. To further reduce the list of putative variants, we filtered out variants that were called in unaffected siblings or the father in each kinship. Unaffected mothers of the proband were permitted to carry the variant, under the assumption that we were seeking variants that fit the following model: autism may be caused by highly penetrant *de novo* loss of function (LOF) variants in males, or a mother may carry the LOF variant but is unaffected, and this variant gets passed on to the male offspring in dominant fashion, who develops autism. Structural variants (SVs) were called using MANTA and CNVnator, and the overlap of calls were considered to be a high-confidence call set. These methods are complimentary in that MANTA combines paired and split-read evidence to call SVs and indels from mapped paired-end sequencing reads (Chen *et al.,* 2016), while CNVnator uses read depth information to call structural variants (Abyzov *et al.,* 2011). Additional filtering was performed on *de novo* SV calls: we excluded SVs that were supported by more than one read pair or split read in either parent, and paired reads were manually inspected in IGV(Robinson *et al.,* 2011; Thorvaldsdóttir *et al.,* 2013). Validation was attempted for predicted *de novo* SNVs and SV breakpoints via Sanger sequencing of all family members.

To determine whether our dataset was comparable to previous ASD whole-exome and whole-genome sequencing studies, we calculated the fraction of individuals with *de novo* missense or LOF variants in known ASD genes, defined as non-syndromic genes with a score of 1 or 2 on SFARI. We used a test for equality of proportions to determine whether this differed significantly from the fraction observed in a previous ASD study that included whole-exome sequencing data from 2,508 trios (Iossifov *et al.,* 2012). We repeated this test with a list of 792 genes recognized to be autism-related from Butler *et al.* (2015), and curated lists of 528 known intellectual disability (ID) genes and 1156 known and candidate ID genes (Gilissen *et al.,* 2014).

To identify novel candidate ASD genes enriched for rare, protein-altering variants in our dataset, we identified genes that contained rare, deleterious variants in multiple families that segregated with ASD or and was either *de novo* or inherited from the mother. We annotated these genes with the number of individuals in the 1000 Genomes Project who have rare variants in a gene, using the Significance of Rare Variants (SORVA) standalone tool (Rao and Nelson 2017). Genes were ranked as candidate ASD genes if, in addition to observing rare variants in multiple families, the gene is also found to be intolerant of rare missense and LOF variants in the population based on SORVA. We obtained a more precise measure of the mutational burden for top candidate genes by using data from the Genome Aggregation Database (GnomAD), which aggregates sequencing data from over 120,000 individuals (Lek *et al*., 2016). We filtered out genes that had low or uneven coverage across exons in reference datasets, and genes in which independent families had variants at the same loci. These variants are likely to be technical artifacts or common variants that erroneously passed MAF thresholds due to differences in variant filtering methods in reference datasets.

Finally, we merged our data with the TADA dataset used by De Rubeis *et al*. (2014) to determine whether any previously highly ranked ASD genes receive additional support based on our findings. We used TADA v1.1 and used parameters and methods identical to those used in the previous study, with a slight difference in filtering for "probably damaging" missense variants. We consider missense variants to be probably damaging if they have a CADD score greater than 10, whereas De Rubeis *et al.* (2014) considered missense variants characterized as probably damaging by PolyPhen-2. All variants were filtered for MAF < 0.1%.

3.3.3 Validation of *de novo* events

Putative *de novo* events were validated by sequencing the carrier and both parents using Sanger sequencing methods. Reverse transcription (RT) polymerase chain reaction (PCR) was performed with MyTaq Polymerase (Bioline) in 50 mL reactions with an initial denaturation at 95°C for 1 minute, followed by 40 cycles of 95°C for 15 seconds, 59°C for 15 seconds, and and a final elongation step of 72°C for 15 seconds. Oligonucleotide primers were designed for the PCR amplification of cDNA fragments containing the *de novo* mutation or SV breakpoint of interest (Supplementary Table 3.S1). When validating a duplication breakpoint in Lysine Methyltransferase 2C (*MLL3*; *KMT2C*) in one family, we used a cDNA fragment of a ubiquitously expressed part of the gene as a positive control. PCR products were separated on a 3% agarose gel. Sanger sequencing was performed by Laragen, Inc.

3.3.4 Functional Enrichment and Network Analyses

To identify novel gene pathways that were enriched in the genes disrupted by the *de novo* mutations, we tested the mutation burden in all the gene-sets listed in the Gene Ontology Consortium's GO Enrichment Analysis tool, which connects to the PANTHER Classification System (Mi *et al.*, 2017). The gene list provided included all genes with rare (MAF < 0.1%) missense variants with CADD score > 10 or LOF variants that met either of the following models: *de novo*, possible compound heterozygous, rare homozygous, or there was a confirmed structural variant overlapping the gene (Supplementary Table 3.S2). P-values reported have been corrected for multiple testing using the Bonferroni correction. DAPPLE software was used for the genetic interaction and protein–protein interaction analysis using 1,000 permutations (Rossin *et al.*, 2011).

3.4 Results

The online sample recruitment process resulted in sample or phenotype mix-ups in 4 out of 61 families sequenced, and all 4 were resolved after analyzing the relatedness between samples and patterns of inheritance. Three mix-ups were a result of switched samples within a family, and one family reported an incorrect gender for an unaffected child.

After filtering for rare variants that segregate with affected status and are predicted to have deleterious effects on protein transcripts, we found that 25 genes contained rare variants in at least two independent families in our dataset. We identified 18 *de novo* events in 18 affected individuals whose both parents we had sequenced; 4 were stop-gain LOF mutations, 12 were missense mutations, and 2 were *de novo* SVs. All *de novo* point mutations and SV breakpoints were validated using Sanger sequencing. One gene with a *de novo* mutation, *KMT2C*, which contained a duplication at chr7:152181665-152375113 (hg38), was a known high confidence or strong candidate ASD gene, defined as genes with a score of 1 or 2 in the SFARI-Gene database (Banerjee-Basu and Packer, 2010) (N=79). Given that only 4.75% of probands are expected to carry *de novo* LOF or missense mutations in either known ASD gene, an estimate that we base on a previously published dataset of 2,508 trios (Iossifov *et al*., 2014), our observation does not differ from expectation (*P* = 0.87). Using a more permissive ASD gene list consisting of 792 genes (Butler *et al*., 2015), we observe mutations in 2 of these genes, which also does not differ from expectation (*P* = 0.86).

Since the majority of our dataset did not consist of trios, we used the transmission and *de novo* association test (TADA) to determine whether our findings provide additional support to genes ranked highly by previous TADA tests by De Rubeis *et al.* (2014); this test incorporates information from transmitted variants and case-control data in addition to *de novo* variants. The greatest increase in a gene's Bayes factor (BF), a measure of the probability that a gene is causal for autism, is seen for genes in which we observed *de novo* variants in our dataset, as expected. *De novo* variants that originate in the parental germ line are a strong source of causality for ASDs (Marshall *et al*., 2008; Ronemus *et al*., 2014; Sebat *et al*., 2007), and accordingly, TADA parameters are set to score these highly. Genes that receive added support from our data are distributed evenly across the list of genes ranked according to BF from the previous study. In other words, we were equally likely to observe *de novo* and inherited variants in genes ranked low or high by De Rubeis *et al*. (2014), highlighting the heterogeneity of ASD risk gene lists and the fact that lists have little overlap between studies (Mosca *et al*., 2017), as well as the high likelihood of novel ASD findings from our dataset.

TADA is most useful for trio and case-control data, and since our cohort also includes additional family structures, we also used a family-based approach to analyzing the WGS data. We filtered out variants present in unaffected siblings and specifically identified rare variants that meet the two-class model of ASDs. Genes in which we observed *de novo* missense or LoF variants are shown in Table 3.2.

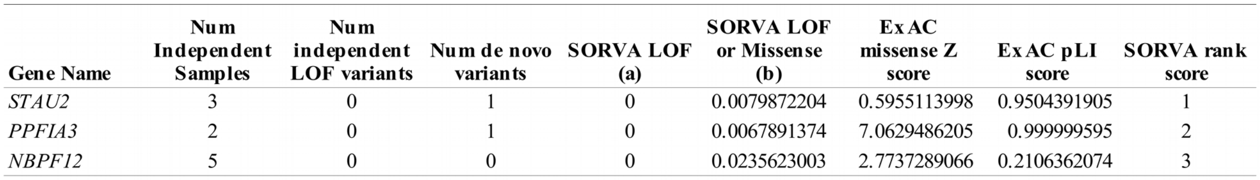
To note are genes that contained rare missense or LOF variants in multiple ASD families given that these genes are intolerant of rare missense or LOF variants in the general population (Table 3.3). Two of these genes also had high pLI scores, suggesting that they have an essential biological function. First, in the gene *STAU2* (Staufen 2), we observed one *de novo* missense

Table 3.2. De novo variants observed in ASD trios

Chr:Pos (hg38)	Gene Name	Ref/Alt	Consequence	Protein Consequence	ExAC AF	pLI score	missense Z-score	CADD PHRED scaled score	Proband IAN ID
2:230401408	*SP140L*	A/G	missense	NP_612411.4:p.Ser489Gly	0	6.07E-05	-0.1	25.3	IAN13OAR
18:76905313	*ZNF236*	G/A	missense	NP_031371.3:p.Arg730Gln	0	1	4.89	33	IAN13OAR
2:54612259	*SPTBN1*	G/T	missense	NP_003119.2:p.Met133Ile	0	1	5.7	23.1	IAN2WQQY
2:54612260	*SPTBN1*	G/C	missense	NP_003119.2:p.Gly134Arg	0	1	5.7	32	IAN2WQQY
4:153585595	*KIAA0922*	C/T	missense	NP_001124479.1:p.Thr432Ile	8.236E-06	0.99939	-0.84	24.7	IAN2WQQY
7:92532429	*RBM48*	C/T	stop_gained	NP_115496.2:p.Gln110Ter	0	5.12E-05	0.28	37	IAN2WQQY
8:73673196	*STAU2*	C/A	missense	NP_001157852.1:p.Arg107Ser	0	0.950439	0.6	26.2	IAN2WQQY
6:31951593	*CFB*	C/T	missense	NP_001701.2:p.Arg710Cys	0	0.000979	2.21	35	IAN5V6UA
11:102795207	*MMP1*	G/C	missense	NP_002412.1:p.Thr289Arg	0	2.49E-15	-2.92	25.2	IAN5V6UA
1:10326139	*KIF1B*	G/A	missense	NP_055889.2:p.Glu856Lys	1.647E-05	1	4.04	26.8	IAN8G47V
2:236167476	*GBX2*	C/T	missense	NP_001476.2:p.Ala166Thr	8.376E-06	0.567961	2.94	21.4	IAN9PGWH
2:42444447 *	*KCNG3*	G/T	stop_gained	NP_579875.1:p.Tyr266Ter	0	0.040597	3.72	40	IANK1C4U
8:42178946 *	*PLAT*	C/G	missense	NP_000921.1:p.Gly494Ala	0.0008072	0.000197	0.08	23.4	IANK1C4U
15:101705701	*TARSL2*	G/T	missense	NP_689547.2:p.Ala326Glu	0	2.86E-09	-0.23	14.83	IANRYFPL
10:73516727	*USP54*	G/T	stop_gained	NP_689799.3:p.Tyr1233Ter	0	3.94E-05	1.09	35	IR7P5E1
7:152181665-152375113	*KMT2C*	DUP	193 kb duplication of exons 2-35 of 59			1	1.53		IX8Y9J4

Variants denoted by * were in regions free of sequencing errors but could not be validated using Sanger sequencing, and are possible mosaic variants.

Table 3.3. Genes that contained rare variants in a multiple independent families and are depleted of variation in the population



Gene Name	Num Independent Samples	Num independent LOF variants	Num de novo variants	SORVA LOF (a)	SORVA LOF or Missense (b)	ExAC missense Z score	ExAC pLI score	SORVA rank score
STAU2	3	0	1	0	0.0079872204	0.5955113998	0.9504391905	1
PPFIA3	2	0	1	0	0.0067891374	7.0629486205	0.999999595	2
NBPF12	5	0	0	0	0.0235623003	2.7737289066	0.2106362074	3

Multiple independent families contained rare missense or LOF variants in these genes. Given the number of independent events, the depletion of variants in these genes in the general population, and the rarity of *de novo* events, these genes ranked highly among candidate genes for follow-up. Column (a) indicates the proportion of individuals in the 1000 Genomes Project dataset (N=2504) who are heterozygous or homozygous for a rare (MAF <= 0.001) LOF variant anywhere in the given gene, obtained from the SORVA dataset. Column (b) indicates the same, but includes missense variants, as well.

variant (R107S), as well as two families in which affected individuals had inherited rare variants from the mother (V22I and A365S). Observing rare variants in three independent families, including one *de novo* variant, is unusual given that only 0.8% of the general population carry rare LOF or missense variants in *STAU2* based on data in GnomAD, which contains data on over 120,000 unrelated individuals (Lek *et al*., 2016). The variants were in three of four double-stranded RNA-bindings domains in *STAU2*, and this gene is known to be involved in synaptic plasticity, translation, localisation, and ribonucleoprotein formation (Heraud-Farlow and Kiebler, 2014). In the second, *PPFIA3* (PTPRF Interacting Protein Alpha 3) we observed a missense *de novo* variant (S335R) in one family, and a maternally inherited missense variant (A680V) that was absent from three unaffected siblings in another family.

It is important to note that the individual with a *de novo* mutation in *STAU2* also had a *de novo* nonsense variant in *RBM48* (RNA binding motif protein 48), however this gene has a pLI score of 0, indicating that it is tolerant of LOF variants and such variants are unlikely to have a severe effect on phenotype.

Several genes with previous support for involvement in autism or intellectual disability were found to contain rare *de novo* or inherited variants in our cohort. For example, *USP54* (Ubiquitin Specific Peptidase 54) contained a *de novo* nonsense variant in one individual. Rare, missense variants in *CHD8* (Chromodomain Helicase DNA Binding Protein 8) were inherited from the mother in three independent families. This gene encodes a chromatin remodeller and has been found to be recurrently mutated in ASD (Bernier *et al*., 2014; Neale *et al*., 2012; O'Roak, Vives, Fu, *et al*., 2012; O'Roak, Vives, Girirajan, *et al*., 2012; Talkowski *et al*., 2012). However, given that rare missense or LOF variants are seen in *CHD8* in 3.12% of the general population, our results are not statistically significant.

To identify novel gene pathways that were enriched in the genes disrupted by the *de novo* mutations, we tested the mutation burden in all the gene-sets listed in the Gene Ontology. We found a significant enrichment of variants in pathways involved in "homophilic cell adhesion via plasma membrane adhesion molecules", "cell-cell signaling", and "nervous system development"; molecular functions that were enriched include “ATP-dependent microtubule motor activity” and “calcium ion binding”. Our results are largely consistent with previous findings (Wen *et al*., 2016). Genes in the pathway “non-integrin membrane-ECM interactions”, which is known to play a central role in central nervous system development and synaptic plasticity (Kerrisk *et al*., 2014), were also significantly enriched in our dataset.

Under the assumption that different genes harboring suspected causative mutations for the same disorder often physically interact, we next considered whether there was evidence of protein-protein interactions (PPIs) using DAPPLE (Rossin *et al*., 2011). The set of 32 unique genes with *de novo* SVs, LOF or missense variants analyzed resulted in one network of direct PPIs encoded by 7 of these genes. (Figure 3.2a) The largest network has the gene *SPTBN1* as its hub; using DAVID, it is enriched for terms including Hippo signaling pathway (BHC $P = 4.2 \times 10^{-6}$) and MAPK cascade (BHC $P = 7.4 \times 10^{-4}$) (Supplementary Table S3.3). Of note, the genes *Discs, large homolog 4 (Drosophila)* (*DLG4*), *Mitogen-activated protein kinase 1* (*MAPK1*), *Catenin (cadherin-associated protein), beta 1, 88kDa* (*CTNNB1*), *mitogen-activated protein kinase 3* (*MAPK3*) and *Glutamate receptor, ionotropic, N-methyl D-aspartate 1* (*GRIN1*) are drawn into the gene network by virtue of their interactions by two other genes in the network. These genes have been previously associated with ASD (Abrahams *et al*., 2013).

By including genes that fit a recessive mode of inheritance, we also analyzed a union set of 264 unique genes using DAPPLE. A single network was derived with an FDR ≤0.01 ($P = 0.15$ for direct interactions; $P = 0.03$ for indirect interactions) that included 97 genes from the three gene lists: 14/30 (46.7%) genes containing *de novo* LOF or missense variants; 1/2 (50.0%) *de novo* SV genes; and 82/232 (35.3%) genes containing rare, homozygous LOF or damaging missense variants, or possible compound heterozygous variants. (Figure 3.2b) Analysis of overrepresented terms in the entire network and subnetworks identified functional themes related to calcium, actin binding, ATP binding, DNA repair, SH3 domain, ECM-receptor interaction, and focal adhesion. Many of these functional terms have been previously connected with autism (Oron and Elliott, 2017; Wen *et al*., 2016). Thus, despite little overlap between genes in our dataset and known ASD genes, the overlap of significantly enriched molecular functions and biological pathways in the resulting networks suggests that the effects of distinct mutations might converge in previously known ASD pathways (Rossin *et al*., 2011).

3.5 Discussion

Using whole genome sequencing of individuals and families with ASD, we provide further evidence for the extreme genetic heterogeneity underlying ASD. In the small number of trios sequenced, we did not observe *de novo* variants in known ASD genes. However, the fact that genes containing rare inherited and *de novo* variants converge in known ASD-associated protein-protein interaction networks suggests that our study may provide evidence for novel ASD genes that currently are missing statistical support from previous studies.

Among novel candidate ASD genes, our findings provide statistical support for *STAU2* playing a role in the etiology of ASD. Staufen proteins play a role during both the early differentiation of neurons and in the synaptic plasticity of mature neurons, and Stau2 regulates mRNA stability, translation, and localisation of mRNA (Heraud-Farlow and Kiebler, 2014). mRNA targets include *RGS4* and calmodulin proteins, which have been implicated in schizophrenia and autism, respectively (Buckholtz *et al*., 2007; Levitt *et al*., 2006; Oron and Elliott, 2017; Schmunk and Gargus, 2013), and other neuronal targets, such as proteins in the G protein-coupled receptor pathway, dopaminergic and serotonergic pathways (Heraud-Farlow *et al*., 2013). The serotonergic, dopaminergic, and cholinergic pathways are involved in synaptic functions and are known to be enriched in genes involved in ASD (David, Enard, Ozturk *et al*., 2016). Due to the discovery of multiple rare *de novo* and inherited variants in *STAU2*, the protein's role in mRNA transport within the neuron, and the fact that an mRNA target has already been associated with ASD, we considered *STAU2* as a candidate susceptibility gene for ASD, and functional studies will be required to validate the effect of the variants we observed.

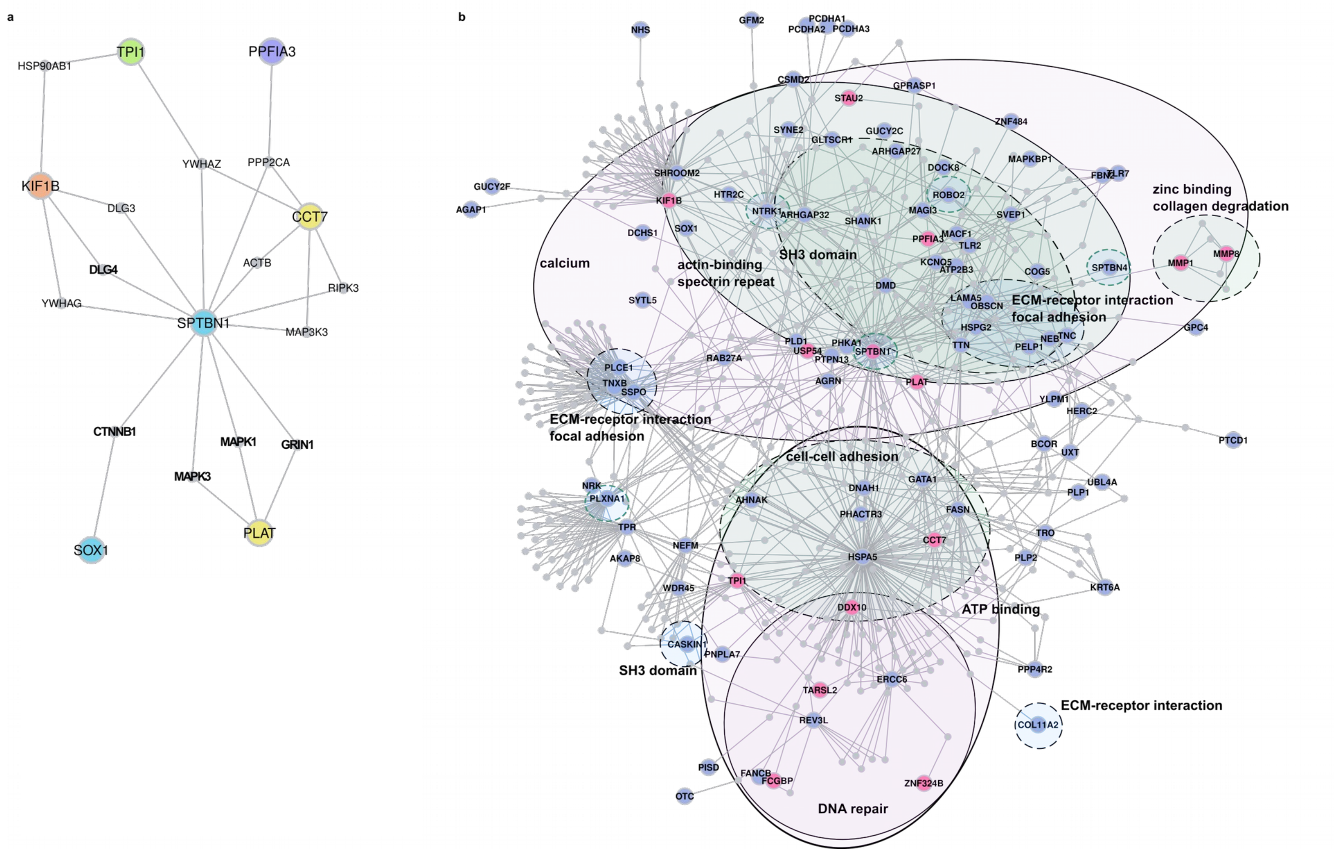


Figure 3.2. Protein-protein interaction networks between genes

(a) 32 genes with *de novo* LOF mutations, missense mutations or SVs were submitted as seed to form a DAPPLE PPI network. The seed genes are shown in colored circles, and genes with the same color were mutated in the same individual. Protein-protein interactions are shown as gray lines (edges) and additional genes are pulled into the network to form indirect connections. The gene names in bold are previously suggested ASD genes (SFARI gene score <= 4). (b) The analysis in (a) was repeated using 264 genes that included the previous gene set, as well as genes that contained rare homozygous or possible compound heterozygous variants. All seed genes were submitted to DAVID and select top gene ontology (GO) terms are shown labeling clusters that contain most of the genes with the GO term. Clusters with a solid outline were statistically significant in the entire network or the circled subnetwork (Benjamini Hochberg corrected P-value < 0.05). Genes with similar GO terms were not always near each other in the PPI network; as an example, genes annotated with the term "axon guidance" are circled with a turquoise dashed line.

Of the 18 trios sequenced, one case can be considered “solved”. In one proband, a *de novo* duplication event was observed in the gene *Lysine Methyltransferase 2C* (*KMT2C; MLL3)*, which was identified as a gene strongly enriched for variants likely to affect ASD risk with a false discovery rate (FDR) of <0.1 (De Rubeis *et al*., 2014), and the gene has prior support from multiple ASD studies (Iossifov *et al*., 2014; O’Roak, Vives, Fu, *et al*., 2012; Yuen *et al*., 2017). Mutations in this gene have been found to cause Kleefstra syndrome (OMIM ID: 610253) (Kleefstra *et al*., 2012; Koemans *et al*., 2017), which is characterized by mental retardation without speech development, hypotonia, and characteristic facial features including microcephaly, brachycephaly, hypertelorism, synophrys, midface hypoplasia, and eversion of the lower lip. This phenotype is consistent with the phenotype of the affected child.

The family-centered approach to data analysis highlights the information that can be gleaned from small, well-phenotyped and well-sequenced datasets. We also demonstrate that as reference datasets become increasingly larger, the need for trio sequencing diminishes: variants that are absent from the population can be considered to be possibly *de novo* variants, and this variant set is also enriched in genes known to cause ASD. There have been several large ASD studies to date, but most are focused on identifying *de novo* LOF mutations and structural variants in recurring genes. By incorporating data from not only trios and case-control individuals as but also complex family structures, and taking into account the model that unaffected mothers may carry causal ASD variants, we were able to identify several candidate novel genes that may have otherwise been overlooked.

3.6 Supplementary Tables

Supplementary Table 3.S1: PCR primer sequences used for Sanger validation of de novo mutations.

Chr:Pos (hg38)	Gene	Forward primer (5'->3')	Reverse primer (5'->3')
6:31951593	CFB	TGTGCTACAAGTGCCCAAGG	ACATGCGGTCCTAAGGTGAG
2:236167476	GBX2	AGCAGTCCGTTTCTCCCAGA	AACTTCGACAAGGCGGAGG
2:42444447	KCNG3	CAATGAAGTGACGGGCAAGC	TTCTGTTGCGTGCATCCTGG
4:153585595	KIAA0922	GTCTCTGGCGCTAGCTTTCA	TCCTGGTCCCTTTGAGCTCTT
1:10326139	KIF1B	CCAACTATTCCTCTCTCCCTGGC	GGGATCGGTCGTAAAACGGG
7:152181665-152375113	KMT2C	CCATTTTGCTTACAGACATACCTC	GCCCTATAGCAGCTGGAACC
Control	KMT2C	CCAGGTGAAAACACTTGCGG	CCCCCATGATGGTAGTTTTCTTCC
11:102795207	MMP1	GCAGATCACAAAGAAGAACTGACA	ACAAGGGCCTGGTGGTCTTT
8:42178946	PLAT	TACAGGTGGAGTTGCCTGTG	TGATGGTTCTCCCCTTTCAGTG
19:49133126	PPFIA3	AGGTAGGTGGCAGGGACTT	CTCCTACCCCTCCAGATTCACG
7:92532429	RBM48	AGGAAAGGGGAGGCTCTACG	GAGTGCCTGACACACAGGAG
13:112068504-112068570	SOX1	AGTACAGCCCCATCTCCAAC	GCTCCGACTTCACCAGAGAG
2:230401408	SP140L	GAGTCGTGTGCTGTGTCTCAT	CAGGAAGAGCCACCTACCTTG
2:54612259-54612260	SPTBN1	GGCACATGTGACTAGGCAGT	GTCGGCCTAAGTTCTGAGCA
8:73673196	STAU2	CATGCACAATGCCCAAGGTC	TCATACAGGTGGTCCATCCA
15:101705701	TARSL2	CTTGTAATGGCCCCCACTGC	TGTATCCCTCGTGTCTCAGCA
10:73516727	USP54	CTCCCTTCTGATCTTTCCCTG	AACTCTGGGCTGCCTAATGG
18:76905313	ZNF236	GCGTGCTCAAAGCACACATC	ACACTCCCCACCATCTACTGG
19:58455837	ZNF324B	AGACCTTCACAGAGTACCGGG	CACACTGTGCGCAAGCATAA

Supplementary Table 3.S2: Rare de novo and inherited variants observed in individuals with ASD. Shown are rare *de novo* variants in trios, and ultra-rare heterozygous LOF variants (MAF < 1E-5 in the GnomAD exomes dataset) that were either known to be inherited from the mother or inheritance could not be confirmed as *de novo* due to lack of sequencing data from both parents.

See material attached with the dissertation.

3.7 Bibliography

Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., et al. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Molecular autism*, *4*(1), 36. doi:10.1186/2040-2392-4-36

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, *21*(6), 974–984. doi:10.1101/gr.114876.110

Banerjee-Basu, S., & Packer, A. (2010). SFARI Gene: an evolving database for the autism research community. *Disease models & mechanisms*, *3*(3–4), 133–5. doi:10.1242/dmm.005439

Basu, S. N., Kollu, R., & Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. *Nucleic Acids Research*, *37*(Database issue), D832-836. doi:10.1093/nar/gkn835

Bernier, R., Golzio, C., Xiong, B., Stessman, H. A., Coe, B. P., Penn, O., et al. (2014). Disruptive CHD8 Mutations Define a Subtype of Autism Early in Development. *Cell*, *158*(2), 263–276. doi:10.1016/j.cell.2014.06.017

Betancur, C. (2011). Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Brain Research*, *1380*, 42–77. doi:10.1016/j.brainres.2010.11.078

Bonaglia, M. C., Giorda, R., Beri, S., De Agostini, C., Novara, F., Fichera, M., et al. (2011). Molecular Mechanisms Generating and Stabilizing Terminal 22q13 Deletions in 44 Subjects with Phelan/McDermid Syndrome. *PLoS Genetics*, *7*(7). doi:10.1371/journal.pgen.1002173

Brandler, W. M., Antaki, D., Gujral, M., Noor, A., Rosanio, G., Chapman, T. R., et al. (2016). Frequency and Complexity of De Novo Structural Mutation in Autism. *The American Journal of Human Genetics*, *98*(4), 667–679. doi:10.1016/j.ajhg.2016.02.018

Buckholtz, J. W., Meyer-Lindenberg, A., Honea, R. A., Straub, R. E., Pezawas, L., Egan, M. F., et al. (2007). Allelic Variation in RGS4 Impacts Functional and Structural Connectivity in the Human Brain. *Journal of Neuroscience*, *27*(7), 1584–1593. doi:10.1523/JNEUROSCI.5112-06.2007

Butler, M. G., Rafi, S. K., & Manzardo, A. M. (2015). High-resolution chromosome ideogram representation of currently recognized genes for autism spectrum disorders. *International Journal of Molecular Sciences*, *16*(3), 6464–6495. doi:10.3390/ijms16036464

Bylund, M., Andersson, E., Novitch, B. G., Muhr, J. (2003). Vertebrate neurogenesis is counteracted by Sox1-3 activity. *Nature Neuroscience,* 6, 1162–1168. doi:10.1038/nn1131

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., et al. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*(8), 1220–1222. doi:10.1093/bioinformatics/btv710

Daniels, A.M., Rosenberg, R.E., Anderson, C., Law, J.K., Marvin, A.R., Law, P.A. (2012). Verification of parent-report of child autism spectrum disorder diagnosis to a web-based autism registry. *Journal of Autism and Developmental Disorders*, 42(2), 256–265. doi: 10.1007/s10803-011-1236-7

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Ercument Cicek, A., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, *515*(7526), 209–215. doi:10.1038/nature13772

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W. M., Willemsen, M. H., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, *511*(7509), 344–347. doi:10.1038/nature13394

Heraud-Farlow, J. E., Sharangdhar, T., Li, X., Pfeifer, P., Tauber, S., Orozco, D., et al. (2013). Staufen2 Regulates Neuronal Target RNAs. *Cell Reports*, *5*(6), 1511–1518. doi:10.1016/j.celrep.2013.11.039

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, *515*(7526), 216–221. doi:10.1038/nature13908

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron*, *74*(2), 285–299. doi:10.1016/j.neuron.2012.04.009

Kerrisk, M. E., Cingolani, L. A., & Koleske, A. J. (2014). ECM receptors in neuronal structure, synaptic plasticity, and behavior. *Progress in Brain Research*, *214*, 101–131. doi:10.1016/B978-0-444-63486-3.00005-0

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *advance online publication*. doi:10.1038/ng.2892

Kleefstra, T., Kramer, J. M., Neveling, K., Willemsen, M. H., Koemans, T. S., Vissers, L. E. L. M., et al. (2012). Disruption of an EHMT1-Associated Chromatin-Modification Module Causes Intellectual Disability. *The American Journal of Human Genetics*, *91*(1), 73–82. doi:10.1016/j.ajhg.2012.05.003

Koemans, T. S., Kleefstra, T., Chubak, M. C., Stone, M. H., Reijnders, M. R. F., de Munnik, S., et al. (2017). Functional convergence of histone methyltransferases EHMT1 and KMT2C involved in intellectual disability and autism spectrum disorder. *PLoS Genetics*, *13*(10), e1006864

Lee, H., Marvin, A. R., Watson, T., Piggot, J., Law, J. K., Law, P. A., et al. (2010). Accuracy of phenotyping of autistic children based on internet implemented parent report. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *153B*(6), n/a-n/a. doi:10.1002/ajmg.b.31103

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. doi:10.1038/nature19057

Levitt, P., Ebert, P., Mirnics, K., Nimgaonkar, V. L., & Lewis, D. A. (2006). Making the Case for a Candidate Vulnerability Gene in Schizophrenia: Convergent Evidence for Regulator of G-Protein Signaling 4 (RGS4). *Biological Psychiatry*, *60*(6), 534–537. doi:10.1016/j.biopsych.2006.04.028

Lim, E. T., Raychaudhuri, S., Sanders, S. J., Stevens, C., Sabo, A., MacArthur, D. G., et al. (2013). Rare Complete Knockouts in Humans: Population Distribution and Significant Role in Autism Spectrum Disorders. *Neuron*, *77*(2), 235–242. doi:10.1016/j.neuron.2012.12.029

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, *37*(3), 235–241. doi:10.1002/humu.22932

MacLean, H. E., Favaloro, J. M., Warne, G. L., & Zajac, J. D. (2006). Double-strand DNA break repair with replication slippage on two strands: a novel mechanism of deletion formation. *Human Mutation*, *27*(5), 483–489. doi:10.1002/humu.20327

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., et al. (2008). Structural Variation of Chromosomes in Autism Spectrum Disorder. *American Journal of Human Genetics*, *82*(2), 477–488. doi:10.1016/j.ajhg.2007.12.009

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., & Thomas, P. D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, *45*(D1), D183–D189. doi:10.1093/nar/gkw1138

Mosca, E., Bersanelli, M., Gnocchi, M., Moscatelli, M., Castellani, G., Milanesi, L., & Mezzelani, A. (2017). Network Diffusion-Based Prioritization of Autism Risk Genes Identifies Significantly Connected Gene Modules. *Frontiers in Genetics*, *8*, 129. doi:10.3389/fgene.2017.00129

Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, *485*(7397), 242–245. doi:10.1038/nature11011

O'Roak, B. J., Vives, L., Fu, W., Egertson, J. D., Stanaway, I. B., Phelps, I. G., et al. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*, *338*(6114), 1619–1622. doi:10.1126/science.1227764

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, *485*(7397), 246–250. doi:10.1038/nature10989

Oron, O., & Elliott, E. (2017). Delineating the Common Biological Pathways Perturbed by ASD's Genetic Etiology: Lessons from Network-Based Studies. *International journal of molecular sciences*, *18*(4). doi:10.3390/ijms18040828

Rao, A. R., & Nelson, S. F. (2017). Calculating the statistical significance of rare variants causal for Mendelian and complex disorders. *doi.org*, 103218. doi:10.1101/103218

Ritvo, E. R., Spence, M. A., Freeman, B. J., Mason-Brothers, A., Mo, A., & Marazita, M. L. (1985). Evidence for autosomal recessive inheritance in 46 families with multiple incidences of autism. *American Journal of Psychiatry*, *142*(2), 187–192. doi:10.1176/ajp.142.2.187

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. doi:10.1038/nbt.1754

Ronemus, M., Iossifov, I., Levy, D., & Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. *Nature Reviews Genetics*, *15*(2), 133–141. doi:10.1038/nrg3585

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., et al. (2011). Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS Genetics*, *7*(1). doi:10.1371/journal.pgen.1001273

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, *46*(9), 944–950. doi:10.1038/ng.3050

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, *485*(7397), 237–241. doi:10.1038/nature10945

Schmunk, G., & Gargus, J. J. (2013). Channelopathy pathogenesis in autism spectrum disorders. *Frontiers in genetics*, *4*, 222. doi:10.3389/fgene.2013.00222

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., et al. (2007). Strong Association of De Novo Copy Number Mutations with Autism. *Science*, *316*(5823), 445–449. doi:10.1126/science.1138659

Talkowski, M. E., Rosenfeld, J. A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., et al. (2012). Sequencing Chromosomal Abnormalities Reveals Neurodevelopmental Loci that Confer Risk across Diagnostic Boundaries. *Cell*, *149*(3), 525–537. doi:10.1016/j.cell.2012.03.028

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. doi:10.1093/bib/bbs017

VarSeq. (n.d.). Bozeman, MT: Golden Helix, Inc.

Vorstman, J. A. S., Parr, J. R., Moreno-De-Luca, D., Anney, R. J. L., Nurnberger, J. I., & Hallmayer, J. F. (2017). Autism genetics: opportunities and challenges for clinical translation. *Nature Reviews. Genetics*, *18*(6), 362–376. doi:10.1038/nrg.2017.4

Wen, Y., Alshikho, M. J., & Herbert, M. R. (2016). Pathway Network Analyses for Autism Reveal Multisystem Involvement, Major Overlaps with Other Diseases and Convergence upon MAPK and Calcium Signaling. *PLOS ONE*, *11*(4), e0153329. doi:10.1371/journal.pone.0153329

Yu, T. W., Chahrour, M. H., Coulter, M. E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., et al. (2013). Using whole-exome sequencing to identify inherited causes of autism. *Neuron*, *77*(2), 259–273. doi:10.1016/j.neuron.2012.11.002

Yuen, R. K. C., Merico, D., Bookman, M., Howe, J. L., Thiruvahindrapuram, B., Patel, R. V., et al. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience*, *advance online publication*. doi:10.1038/nn.4524

Yuen, R. K. C., Thiruvahindrapuram, B., Merico, D., Walker, S., Tammimies, K., Hoang, N., et al. (2015). Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature Medicine*, *21*(2), 185–191. doi:10.1038/nm.3792

Zhao, X., Leotta, A., Kustanovich, V., Lajonchere, C., Geschwind, D. H., Law, K., et al. (2007). A unified genetic theory for sporadic and inherited autism. *Proceedings of the National Academy of Sciences*, *104*(31), 12831–12836. doi:10.1073/pnas.0705803104

Chapter 4

A tool for calculating mutational burden of genes causal for Mendelian and complex disorders

4.1 Abstract

With the expanding use of next-gen sequencing (NGS) to diagnose the thousands of rare Mendelian genetic diseases, it is critical to be able to interpret individual DNA variation. To calculate the significance of finding a rare protein-altering variant in a given gene, one must know the frequency of seeing a variant in the general population that is at least as damaging as the variant in question.

We developed a general method to better interpret the likelihood that a rare variant is disease causing if observed in a given gene or genic region mapping to a described protein domain, using genome-wide information from a large control sample. Based on data from 2,504 individuals in the 1000 Genomes Project dataset, we calculated the number of individuals who have a rare variant in a given gene for numerous filtering threshold scenarios, which may be useful for vetting rare variants causal for disease. Additionally, we calculated mutational burden data for the number of individuals with rare variants in genic regions mapping to protein domains.

We describe how to apply the mutational burden data for use in predictive genomics and predict whether a person will develop a disease given their genotype, and describe how to calculate the statistical significance of observing rare variants in a given proportion of independent, affected individuals. We present SORVA, a web tool that allows users to browse

the mutational burden dataset. Finally, we demonstrate that using our dataset to rank genes based on intolerance for variation, the ranking correlates well with pLI scores derived from the Exome Aggregation Consortium (ExAC) dataset ($\rho = 0.515$), with the benefit that the scores are directly interpretable.

In conclusion, we have presented a strategy that is useful for vetting candidate genes from NGS studies and allows researchers to provide support for variants in a given gene or protein domain that may be candidates for follow-up studies.

4.2 Introduction

Whole-exome sequencing has enabled the identification of causal genes responsible for causing hundreds of rare, Mendelian disorders in just a few years; however, there remain hundreds, if not thousands, more to be uncovered. The genetic basis has been determined for 4,803 of the rare diseases, whereas the number of disease phenotypes with a known or suspected Mendelian basis lies close to 6,419 based on data in Online Mendelian Inheritance in Man (OMIM) (2015). Next-gen sequencing (NGS) studies are certain to uncover many disease-phenotype relationships in the near future, but for cases involving rare diseases with limited sample sizes, determining causality between phenotypes and novel genes, and distinguishing true pathogenic variants from rare benign variants remains a challenge. Often disease causality of a given rare variant is only clear when additional affected individuals with similar rare variants in the same gene are identified, which can take years to occur due to the rarity of these disorders. Thus, improvements in determining disease causality or likely pathogenicity would greatly enhance efforts to prioritize genes and gene variants for further molecular analysis, even if only a single affected individual was identified.

Variants identified through broad based NGS technologies are typically classified as pathogenic, likely pathogenic, variant of uncertain significance (VUS) or likely benign according to multiple criteria, largely based on prior knowledge about the specific variant. Novel variants are evaluated individually and placed into discrete categories if they meet complex combinations of criteria, which include thresholds for allele frequency, segregation, number of affected unrelated individuals, and known functional relevance (Dorschner *et al.*, 2013; Amendola *et al.*, 2015). For example, a variant would be deemed pathogenic if the allele frequency threshold falls below a given threshold and the variant segregates with a disorder in at least two unrelated affected families, or if other criteria are met. In brief, variants are evaluated individually based on variant-specific annotations.

An additional source of information that would aid in variant prioritization would be a gene-specific annotation describing mutational burden in the overall population. To illustrate, consider a gene that has very few functional variants in the general population, and several unrelated patients were found to carry distinct protein-altering, rare missense or potential loss-of-function (LOF) variants in the given gene and within a highly conserved protein domain. Under a model for a rare Mendelian disorder caused by highly penetrant variants, we assume that common variants cannot be considered causal, and rare variants in genes intolerant of mutations are deemed highly suspicious of being causal for disease even if no other information is known about the variants. Therefore, knowing the population-wide mutational burden of a given gene for rare variants would be informative.

While there are gene-ranking methods based on other parameters (Gill *et al.*, 2014), recently several gene-level ranking systems have emerged based on measures for intolerance to mutations in the general population. The Residual Variation Intolerance Score (RVIS) generates a score based on the frequencies of observed common coding variants compared to the total number of observed variants in the same gene or protein domain (Petrovski *et al.*, 2013; Gussow *et al.*, 2016). A second ranking system, in addition to these parameters, also incorporates the frequency at which genes are found to be affected by rare, likely functional variants, and their findings suggest that disease associations to genes which frequently contain variants, termed as FLAGS, should be evaluated with extra caution (Shyr *et al.*, 2014). Next, the Exome Aggregation Consortium (ExAC) dataset provides missense *Z* scores that describe the degree to which a gene is depleted of missense variants compared to expected values based on the frequency of synonymous variants, and provides pLI scores that describe probabilities of a gene being LOF intolerant (Samocha *et al.*, 2014; Lek *et al.*, 2016). Of these two metrics, pLI is less correlated with coding sequence length and outperforms the *Z* score as an intolerance metric (Lek *et al.*, 2016). Another method, EvoTol, combines genic intolerance with evolutionary conservation of whole protein sequences or their constituent protein domains to prioritize disease-causing genes, and extends the RVIS method by leveraging the information on protein sequence evolution to identify genes where the number of mutations that are likely to be damaging based on evolutionary protein information is higher than expected (Rakham *et al.*, 2015). Although these methods may be useful in ranking genes and prioritizing variants in order to highlight those in genes that frequently contain variants, neither results in a score that is directly interpretable in order to calculate statistics about NGS findings and determine the significance of seeing a variant in a given number of affected individuals.

One tool that calculates a *P*-value of finding a true association through clinical exome sequencing, RD-Match (Akle *et al.*, 2015), allows researchers to calculate the probability of finding phenotypically similar individuals who share variants in a gene through systems such as Matchmaker Exchange. The tool incorporates the probability of an individual having a rare, nonsynonymous variant in a gene by taking the sum of the allele frequencies of all rare (MAF < 0.1%) nonsynonymous variants annotated in ExAC (Lek *et al.*, 2016). With higher MAF thresholds and large population sizes, this is problematic because an individual may have multiple variants in a gene that frequently contains rare variation, causing one to overestimate the fraction of the population carrying rare variants in the gene, hence the fixed, low MAF threshold. Furthermore, this tool is applicable to studies in which the affected individuals are selected based on phenotype as well as the prior knowledge that they share rare variants in a given gene. Finally, RD-Match does not allow researchers to customize variant filtering thresholds according to the disease model with regards to minor allele frequency or predicted consequence such as LOF or missense variant.

Another method that calculates the significance of NGS findings, the Transmission And De novo Association test (TADA), is a Bayesian model that combines data from *de novo* mutations, inherited variants in families, and variants in cases and controls in a population (He *et al.*, 2013). This method has been used to identify risk-conferring genes in whole-exome sequencing studies of autism spectrum disorders and neurodevelopmental delay (De Rubeis *et al.*, 2014; Sanders *et al.*, 2015; Berko *et al.*, 2017). While TADA analysis has proven to be a critical first step in the development of quantitative methods to assess risk genes, it is restricted to integrating trio and case-control data and is unable to leverage information from large reference datasets, and therefore, it cannot be used for calculating the *P*-value of findings in smaller studies.

Here we describe a method, named SORVA for Significance Of Rare VAriants, for ranking genes based on mutational burden. In addition to incorporating information from variant allele frequencies, we use population-derived data to precompute an unbiased, easily interpretable score, which allows one to calculate the significance of observing rare variants in a given gene in unrelated, affected individuals. For example, one may then answer the question: what is the probability of observing missense variants in three out of ten unrelated affected individuals, given that only one in a thousand individuals in the general population carry a missense variant in the gene? Essentially, a model can be constructed to estimate the probability of drawing *n* unrelated families with similar biallelic genotypes by chance from the general population (Akawi *et al.*, 2015). Conversely, if one has a large list of variants of unknown significance, the significance level may be useful in prioritizing variants within the same category of pathogenicity, and in improving the interpretation of variants in studies of Mendelian genetic disorders.

4.3 Materials and Methods

4.3.1 Datasets

Genomic data and allele frequencies for calculating *a priori* probabilities of observing a variant within a gene were obtained from the 1000 Genomes Project (phase 3 variant set) (The 1000 Genomes Project Consortium, 2012). This variant set contains 2,504 individuals from 26 populations in Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), and the Americas (AMR).

4.3.2 Bioinformatics pipeline

Genomic annotations were assigned to each variation using *SNP & Variation Suite (SVS)* v8.1 (Bozeman, MT) with the following parameters: gene set Ensembl release 75 (Cunningham *et al.*, 2015), human genome version GRCh37.p13. Variants were filtered for coding mutations that result in a change in the amino acid sequence (e.g. missense, nonsense and frameshift mutations), or mutations that reside within a splice site junction (intronic distance of 2 base pairs). Biallelic data were recoded based on an additive model to correct for MAF of variants on the X chromosome for male samples, using a script in SVS. Variants were then filtered for minor allele frequency thresholds of MAF < 5%, < 1%, < 0.5%, < 0.1% and < 0.05%, based on allelic frequency within the dataset. For each filtered list of variants, we collapsed variants by gene and performed the following two scenarios: 1) an individual was counted as having a rare variant in a gene if the variant mapped to any transcript of a gene; 2) we counted the number of variants in a given gene per individual, i.e. if an individual carried two rare mutations within a gene, they were counted twice. In a separate analysis, we collapsed variants by protein domains obtained from Interpro (Mitchell *et al.*, 2015) using the Ensembl API (McLaren *et al.*, 2010). Finally, we repeated each analysis using a subset of the 1000 Genomes Project data grouped according to superpopulation. Variant collapsing methods were performed using a custom Python script run by SVS, and an individual was counted as having a rare variant in a gene if the variant mapped to any transcript of a gene.

In addition to replicating the analysis for gene versus protein domain, for each population, and for each MAF threshold, we also repeated the calculations for multiple categories of predicted variant consequence on the protein transcript. The two categories were 1) nonsynonymous variants or those predicted to be more severe by Ensembl (Cunningham *et al.*, 2015), briefly nonsynonymous or LOF variants, and 2) potential LOF variants (includes splice site, protein truncation stop codon gain mutations, and frameshift indels).

4.3.3 Comparison of disease gene categories

To determine whether our results show concordance with studies identifying essential genes critical for the survival of a human, we compared the number of individuals with rare, deleterious mutations between gene lists containing essential human genes, those known to cause Mendelian diseases, and control genes, defined as genes not included in either category. We considered genes to be essential human genes if they were determined as such in at least one of the following two studies. The first essential human gene set is defined as 'core' essential genes that are required for fitness of cells from both the HAP1 and KBM7 cell lines, determined through extensive mutagenesis in near-haploid human cells (N=1734) (Blomen *et al.*, 2015). The second essential human gene set consists of genes essential to four screened cell lines, KBM7, K562, Raji and Jiyoye, determined using the CRISPR system. From the latter set, we selected genes with an adjusted *P*-value CRISPR score< 0.4025 for each cell line (N=1878), which is the set of genes that the authors determined to be essential for optimal proliferation in their screen, although the precise set would depend on the score cutoff chosen (Wang *et al.*, 2015).

To identify genes known to cause Mendelian disease, we parsed data from Online Mendelian Inheritance in Man (OMIM) (2015) and identified phenotype descriptions with known molecular basis. We parsed the genotype description field for the gene name and the following phrases: 'caused by heterozygous/homozygous mutation', 'autosomal recessive', 'autosomal dominant', 'X-linked', ' on chromosome X', and categorized genes as autosomal recessive (AR) (N=655), autosomal dominant (AD) (N=785), and X-linked (XL) (N=126) accordingly.

4.3.4 Comparison of gene ranking methods

Genic mutational intolerance scores were obtained from four previous studies and included the Residual Variation Intolerance Score (RVIS) (Petrovski *et al.*, 2013), scores from Shyr *et al.* 2014 (FLAGS), pLI scores based on the ExAC dataset (Samocha *et al.*, 2014; Lek *et al.*, 2016), and EvoTol scores (Rackham *et al.*, 2015). We considered 15,266 genes that were found in all four datasets, as well as ours, and ranked genes based on scores obtained using each method. Spearman's rho test (Erich 1975; Conover 1980) was used to measure the size and statistical significance of the association between the rankings obtained from ExAC and those obtained by RVIS, FLAGS and SORVA methods. This test measures the strength and direction of association between two ranked variables.

In order to assess the performances of all five methods when prioritizing putative disease genes and plot receiver operating characteristic (ROC) curves, we used the sets of OMIM genes described earlier. We filtered the OMIM gene sets to overlap the 15,266 genes that were scored by all five methods. Genes were ranked according to each metric and a count of the number of disease-causing genes that would be found at each percentile are reported. In order to show the baseline prediction, the result of randomly assigning a percentile to each gene is also shown. SORVA genes were ranked according to the number of 1000 Genomes Project individuals who were heterozygous or homozygous for rare (MAF<0.005) LOF variants in a given gene, and ties between genes were resolved based on the number or individuals who have rare (MAF<0.005) LOF or missense variants in a gene, and finally less rare (MAF<0.05) LOF or missense variants.

4.3.5 Calculating depletion of variants in protein domains

We performed two analyses: first, we calculated whether protein domains in a gene were depleted of variation compared to the rest of the gene, and second, we calculated whether there were any types of protein domains that were depleted of variation in general across the entire genome.

First, for each protein domain mapping within a gene, we calculated whether domains were depleted of variation compared to the rest of the gene. Depletion was calculated as: (number of variants per individual in protein domain / number of variants per individual in exonic region of a gene × length of protein domain / length of transcript). A value of 1 is expected by chance, and a small value indicates protein domains most intolerant towards mutations. We then calculated the *P*-value of obtaining such a depletion score using the binomial cumulative density function, under the assumption that each site is equally likely to be mutated. This *P*-value is then "PHRED-scaled" by expressing the rank in order of magnitude terms rather than the precise rank itself. High scaled scores indicate that a protein domain is depleted of rare (MAF < 0.5%) mutations compared to the rest of the gene, hence protein domains with high scores tend to be enriched for highly mutated genes. Our goal was to determine mutational burden at a finer resolution than the previous gene-based analysis enabled, and not to simply determine highly mutated genes as previously done. Therefore, we filtered out protein domains that span more than 50% of the length of the transcript. We also filtered out genes with no observed mutations. These filtering steps resulted in 7,828 genes remaining.

Next, we calculated whether there were any types of protein domains that were depleted of variation in general across the entire genome. We weighted each gene with instances of the protein domain equally. In other words, if a gene had multiple instances of a protein domain, we first calculated the mean number of heterozygous rare (MAF <=0.5%) LOF variants observed (in the entire dataset of 2,504 individuals) in either protein domain within the gene. Next, we calculated the mean and variance of the means for each gene.

To determine whether a protein domain was well covered by sequencing, we calculated the mean coverage of an instance of a protein domain in the 1000 Genomes Project sample HG00096. We calculated depth of coverage from phase 3 exome alignment data using GATK and custom code, which is available at https://github.com/alizrrao/DepthOfCoveragePerInterval.

4.3.6 Observing rare variants in unrelated individuals

Given the number of individuals in the population who have a potentially damaging variant anywhere in a gene, we calculated the significance of seeing the observed number of independent individuals carrying a rare variant in a gene, given the number of cases sequenced. Specifically, let p be the *a priori* probability that an individual has a heterozygous or homozygous mutation in a gene. Then, the P-value equals the probability of seeing X or more individuals with a heterozygous or homozygous variant, out of n independent individuals, where $Pr(X = k)$ is the probability mass function for the random variable X with binomial distribution $B(n, p)$. This can be calculated as

$$P_{X,n}=\sum_{k=X}^{n} Pr(X=k)=\sum_{k=X}^{n} \binom{n}{k} p^{k}(1-p)^{n-k}$$

.

4.3.7 Availability of data and materials

Gene-based mutational burden datasets and the webtool are available for querying at the SORVA website, https://sorva.genome.ucla.edu.

Standalone software and datasets are freely available for download at https://github.com/alizrrao/sorva.

The 1000 Genomes Project datasets analysed during the current study are available in the International Genome Sample Resource (IGSR), http://www.internationalgenome.org/data.

4.4 Results

To generate a mutational burden dataset that would aid in prioritizing candidate genes and variants from NGS studies, we calculated the frequency of observing a variant in each gene in an individual within the population by using a large control dataset and collapsing variants in exonic regions of each gene. Calculations are based on data from 2,504 individuals in the 1000 Genomes Project phase 3 dataset, which includes targeted exome sequencing data (mean depth = 65.7×) from individuals from five "superpopulations" (European, African, East Asian, South Asian, and ad-mixed American). We repeated the analysis for variants filtered according to various minor allele frequency and protein consequence thresholds that researchers may use when filtering variants. First, we filtered out common variants that met various minor allele frequency (MAF) thresholds used in the literature and others: 5%, 1%, 0.5%, 0.1% and 0.05%. We then filtered rare variants according to two scenarios before collapsing variants across genes: 1) we included all protein-altering variants, i.e. those that cause a nonsynonymous change in the protein transcript or have a more deleterious consequence, and 2) we filtered for potential loss-of-function (LOF) variants, i.e. splice site, stop codon gain and frameshift variants.

Below, we present general findings in population and molecular genetics that can be gleaned from the dataset, and illustrate how the dataset can be used as a control group to vet candidate genes and variants.

4.4.1 Population differences

Of 18,877 genes that are in the union of the Ensembl and RefSeq gene sets, most genes contained heterozygous or homozygous missense variants in individuals in all populations; only 2.3% contain no rare variants (MAF < 5%), and 1.0% of genes have an identified variant in only a single population. Lowering our MAF threshold does not decrease the number of genes much. Many genes do not contain any rare LOF variants in the 1000 Genomes Project data, and filtering variants to include only LOF variants reduces the number of genes containing variants in the dataset to 9641, or 51.1% of genes in the dataset. (Figure 4.1) These results demonstrate

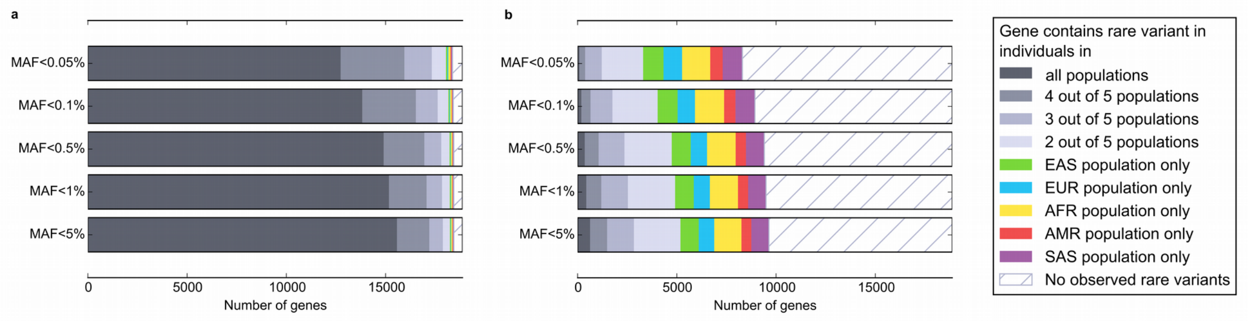


Figure 4.1. The proportion of genes (n=18877) containing rare variation in individuals in various populations. A gene was considered mutated if at least one individual was heterozygous or homozygous for an uncommon or rare (MAF < 5%) variant anywhere in the gene. Variants were filtered by predicted consequence for **(a)** protein-altering (missense or potential loss-of-function) variants, or **(b)** potential loss-of-function variants only. Abbreviations: EUR, European. AFR, African. EAS, East Asian. SAS, South Asian. AMR, ad-mixed American.

that choosing the correct MAF threshold is not nearly as important as identifying the correct protein consequence threshold to use when filtering variants. For instance, including all missense variants when LOF variants are generally causal for a given disease would reduce power to detect the gene associated with the disease.

The number of individuals who carried a heterozygous or homozygous variant in a given gene was generally higher in the African population compared to other populations (Figure 4.2a), which is expected given that African individuals are observed to have up to three times as many low-frequency variants as those of European or East Asian origin (The 1000 Genomes Project Consortium, 2012), which reflects ancestral bottlenecks in non-African populations (Marth *et al.*,

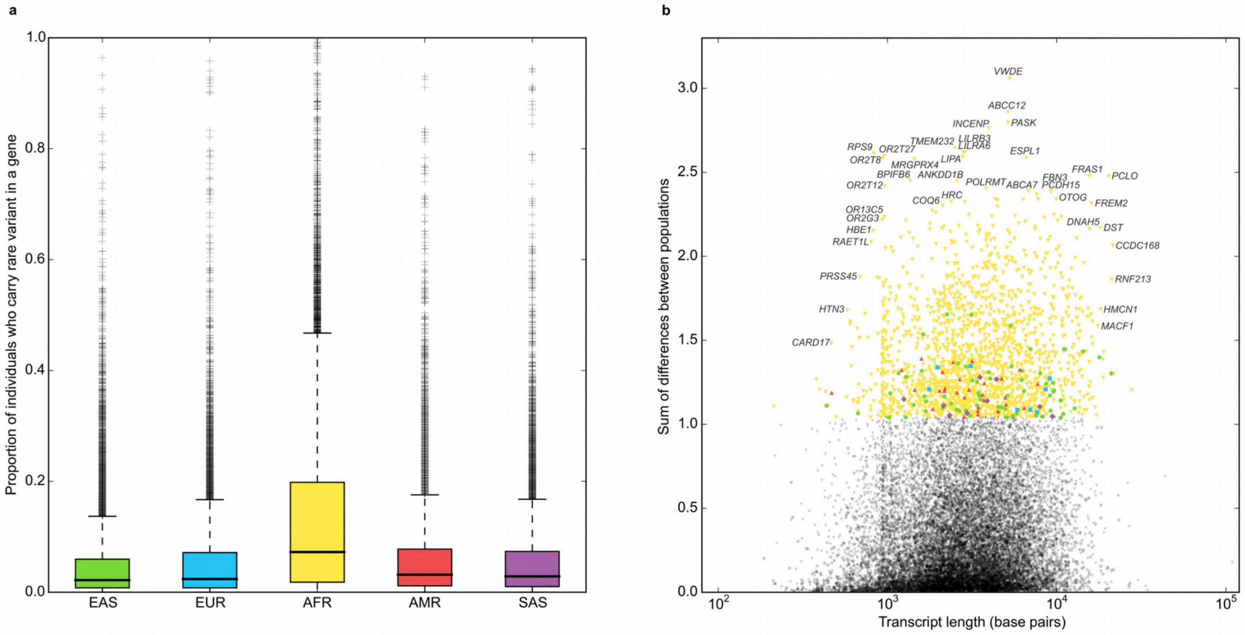


Figure 4.2. Population differences between the number of individuals mutated for a gene between populations. (**a**) Each data point in the histogram represents the proportion of individuals within a population who are heterozygous or homozygous for an uncommon (MAF < 5%) missense variant in a given gene. (**b**) The number of individuals carrying uncommon variants in a gene differs between populations. We plotted the variance of the count for each gene and colored high-variance genes to denote which population differed most from the mean.

2003). Conversely, regarding genes for which the number of individuals with a rare variant in the gene differed between populations, the genes having the greatest difference between populations tended to diverge most in the African population. (Figure 4.2b) Genes whose mutational burden diverges most between populations are significantly enriched for a large number of biological functional terms, including glycoprotein, olfactory transduction and sensory perception, cell adhesion, various repeats, basement membrane and extracellular matrix part, cadherin, microtubule motor activity, immunoglobulin and EGF-like domain. It is important to note differences between populations, because, in many cases, researchers would be advised to use control populations similar to their study population. However, if a gene is associated with a severe, childhood-onset disorder in one population, it is likely to be associated with disease in other populations, as well, and knowledge that a gene frequently contains variation in African populations would be useful in prioritizing candidate genes even if one is studying variation in another population. In this case, such information would point towards reduced likelihood for disease association.

4.4.2 Properties of known disease genes

To determine whether calculating the frequency of individuals who have a rare variant in a given gene in the general population may be helpful in determining which genes are more likely to cause disease, we compared the counts between multiple categories of genes: a) “essential” genes, defined as genes essential for cell survival in human cell lines, b) genes in which variants are known to cause autosomal dominant disorders, c) genes in which variants are known to cause autosomal recessive disorders, d) genes in which variants are known to cause X-linked disorders, and e) all other genes. As expected, fewer individuals carry rare, protein-altering or LOF variants in genes known to cause Mendelian disorders compared to other genes, and genes associated with X-linked disorders tend to be least tolerant of mutations (Figure 4.3; Supplementary Figure 4.S1). Although frequency counts overlapped between gene categories for every variant filtering threshold, clusters were most differentiated when plotting the proportion of individuals who are heterozygous for rare LOF variants in a gene. Furthermore, the differentiation between clusters increased as the MAF threshold became more stringent, as the datasets became enriched for deleterious variants that can only subsist at a low allele frequency in a population due to selective pressure. (Supplementary Figure 4.S1)

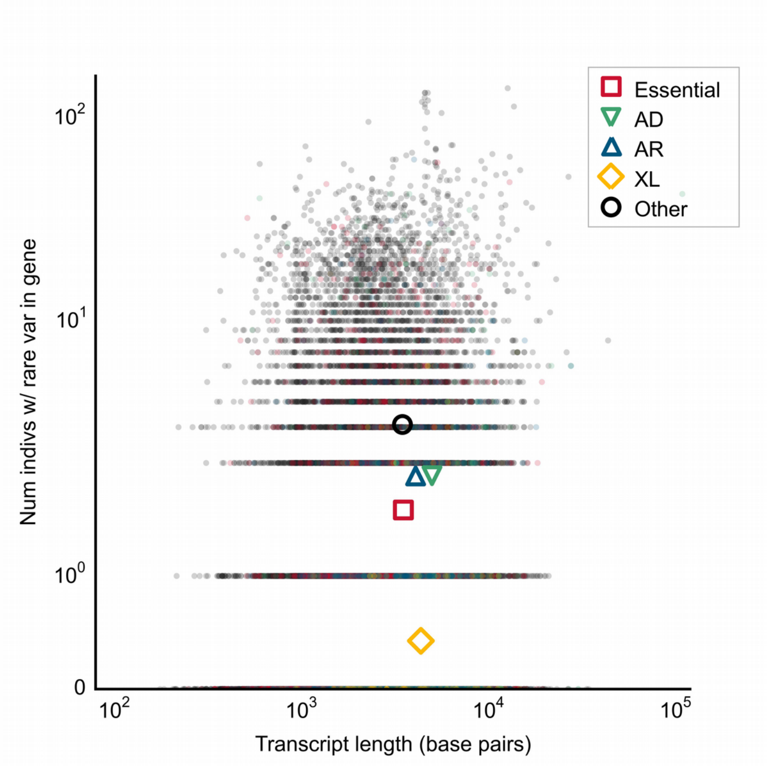


Figure 4.3. The number of individuals heterozygous for a rare (MAF < 0.5%) potentially LOF mutation in a gene. Each data point represents a single gene, mutated in the aggregate population (n=2504 individuals). Genes are grouped according to whether they are an essential gene, or are known to cause autosomal dominant, autosomal recessive or X-linked disease. Colored shapes indicate the centroids of each group. Abbreviations: nonsyn, nonsynonymous. LOF, loss-of-function. AD, autosomal dominant. AR, autosomal recessive. XL, X-linked.

Previous research suggests that 2.0% of adults of European ancestry and 1.1% of adults of African ancestry can be expected to have actionable highly penetrant pathogenic (including novel expected pathogenic) or likely pathogenic single-nucleotide variants (SNVs) in 112 medically actionable genes (Amendola *et al.*, 2015). If we look for rare variants in 1000 Genomes Project individuals—benign as well as pathogenic variants—, we find that a larger proportion of individuals, 5.8% of European individuals and 3.3% of African individuals, are heterozygous or homozygous for extremely rare (MAF < 0.0005) LOF variants in these 112 genes, highlighting the large number of benign variants that are found in the population at low allele frequencies and should be filtered out by manual curation.

4.4.3 Depletion of variants in regions mapping to specific protein domains

It has been suggested previously that collapsing variants by protein domain could lead to improved gene-based intolerance scoring systems, as certain regions of the gene could be much more constrained than others (Petrovski *et al.*, 2013). We incorporated data for 322,772 protein domains from Interpro (Mitchell *et al.*, 2015) and calculated the average number of individuals who have a variant in any given type of protein domain (Supplementary Table 3.S1), after filtering for rare (MAF < 0.5%), heterozygous LOF variants. Protein domains that are highly constrained, well covered during exome sequencing and rarely contain variants despite their large size include the Family A G protein-coupled receptor-like protein domain (Superfamily: SSF81321), which is found in 660 genes and has a mean length of 965 base pairs; none of the 2,504 individuals carry rare variants in the region mapping to this protein domain. Other highly constrained protein domains that occur throughout the human genome include Glutamic acid-rich region profile (PfScan: PS50313), Proline-rich region profile (PfScan:PS50099), Immunoglobulin (Superfamily: SSF48726), and Cysteine-rich region profile (PfScan: PS50311). (Supplementary Table 4.S1) If an NGS study finds that affected individuals have rare variants in variation intolerant protein domains such as those listed, the variants would become highly suspicious of being causal.

We also calculated whether specific genes contain protein domains that are significantly depleted of variation, given the frequency of variants in the gene overall. Filtering out protein domains in genes with no variants and those with missing information reduced the dataset to 67,138 protein domains in 7,004 genes. 77 protein domains in 26 genes were significantly depleted of variation compared to the rest of the gene. Specifically, the number of rare (MAF < 0.5%), heterozygous LOF variants per individual in the protein domains were significantly lower than expected after correcting for multiple testing by the number of genes. (Figure 3.4) Functional enrichment analysis in DAVID revealed that the most significant biological functions in the gene list were related to tubulin-tyrosine ligase activity (P=0.015), and G-protein coupled receptor, rhodopsin-like superfamily (P=0.05). Depletion values for all protein domains may be found in Supplementary Table 4.S2. Information about whether a protein domain is significantly depleted of variation despite being in a gene with frequently observed variation, or conversely, whether it is enriched for rare variants, may be useful in distinguishing between pathogenic and benign rare variants within genes containing regions under different degrees of evolutionary constraint.

4.4.4 Significance of findings in studies of rare genetic disorders involving independent individuals

Below, we present methods for calculating the significance of observing a given variant in a given gene, in studies of independent, unrelated individuals. In the simplest case, a study involving a single family, calculating the *P*-value is relatively simple. Consider a case of a severe, pediatric-onset Mendelian disorder, in which both parents and the affected child are

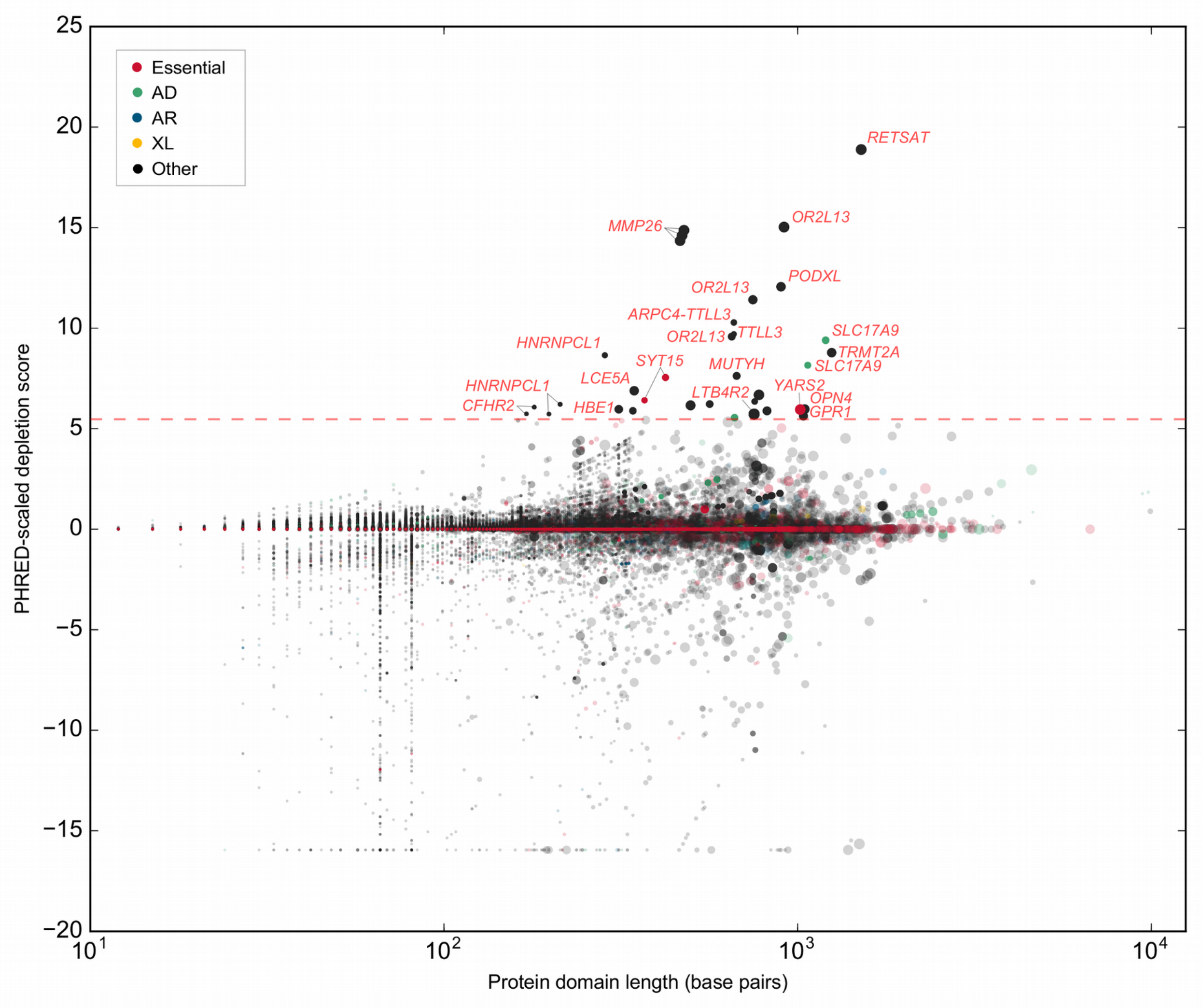


Figure 4.4. Depletion of rare, heterozygous LOF variants in regions mapping to protein domains. We plotted scaled protein domain depletion scores for each domain mapping within a gene; high scaled scores indicate that a protein domain is depleted of rare (MAF < 0.5%) mutations compared to the rest of the gene. Darkened points above the red dashed line represent protein domains that are significantly depleted of mutations after correcting for the number of genes remaining after filtering. Larger points indicate protein domains with a greater length in proportion to the transcript length. Points are colored if the protein domain is within a gene that is an essential human gene or is causal for a Mendelian disorder. Abbreviations: AD, autosomal dominant. AR,

autosomal recessive. XL, X-linked.

sequenced to identify the causal variant. If only *de novo* variants are identified within a putative gene, one can easily estimate the probability of at least one *de novo* mutation occurring in a gene by random chance; one could multiply the per-base mutation rate by the length of the gene transcript and make adjustments to account for CpG content related variation in mutation rates (Supplementary Methods).

In studies that identify both *de novo* and inherited variants, calculating the significance of a variant is more complex. First, we generalize the equation for calculating the significance of observing a *de novo* mutation in a gene for studies involving multiple unrelated individuals. The *P*-value of observing independent *de novo* events in the same gene in *s* out of *n* individuals is

$P = 1 - BinomCDF(s - 1, n, l_{tx}dc)$

if multiple individuals are sequenced, where l_{tx} is the length of the transcript in nucleotide bases and *d* is the mean rate of *de novo* single-nucleotide variants (SNVs) arising per nucleotide per generation, estimated to have a lower bound of 1.2×10^{-8} per site per generation (Campbell and Eichler, 2013; Conrad *et al.*, 2011; Veltman and Brunner, 2012). The parameter *c* is the fraction of *de novo* events that meet our protein consequence threshold. It is predicted that 2.85% of *de novo* events are splice site altering or nonsense events, and 70.64% of de novo events are protein-altering, i.e. missense or LOF (Kryukov *et al.*, 2007); these may be used as the respective values for *c* depending on the variant filtering criteria used.

Consider the following example. Clinical exome sequencing (CES) in four independent families identified *de novo* nonsense mutations in *KAT6A* in all probands displaying significant developmental delay, microcephaly, and dysmorphism (Arboleda *et al.*, 2015). *De novo* nonsense mutations arising in this gene in all four individuals is highly unlikely by chance (*P* =

2.66×10^{-12}), and the statistical findings would support *KAT6A* as highly suspicious for causing the disorder. Further experiments and the identification of multiple other affected individuals by a separate study (Tham *et al.*, 2015) confirmed this result.

If inherited variants are also observed in a gene, calculating the statistical significance of findings requires incorporating information about the number of individuals who carry a variant in the particular gene in the general population. The frequencies of the number of individuals who contain rare variants in a given gene or protein domain for various filtering thresholds may be queried through our online database called SORVA (https://sorva.genome.ucla.edu). (Supplementary Figure 4.S2) Researchers can select the variant filtering thresholds identical to those used in hard filtering variants in a given study. Minor allele frequency thresholds range from 5%, useful for studies involving more common, complex disorders where less stringent filtering criteria are used, to 0.05% for studies involving extremely rare disorders. Then, knowing the expected number of individuals who carry a variant in the gene or protein domain in question, one can calculate the significance of seeing the observed number of singletons (variants observed in a single individual) as follows.

Let f_{hom} be the fraction of individuals in the general population with a homozygous variant in a gene or protein domain. Then, the P-value of seeing k individuals with a homozygous variant, out of n total unrelated individuals is

$$P_{k,n} = 1 - BinomCDF(k - 1, n, f_{hom})$$

where BinomCDF denotes the binomial cumulative distribution function.

If the affected individuals are heterozygous for the putative variants, the P-value is

$$P_{k,n} = 1 - BinomCDF(k - 1, n, f_{both})$$

where f_{both} is the probability of an individual having either a heterozygous or homozygous variant in the gene of interest.

The *a priori* probability *p* can be queried from the SORVA dataset online, and standalone computer software for obtaining *p* and calculating the *P*-value based on the methods described herein is also available on the SORVA website (https://sorva.genome.ucla.edu/).

4.4.5 Challenges of evaluating findings in large-scale studies of complex disorders

In complex disorders where most of the genes contributing to risk remain unknown, our dataset may be used to provide additional evidence supporting novel gene findings. As an example, several large-scale whole-exome sequencing (WES) studies have been carried out to-date in trios and quads to elucidate causal genes underlying autism spectrum disorders (ASD) (Iossifov *et al.*, 1012; Neale *et al.*, 2012; O'Roak *et al*., 2012a; O'Roak *et al.*, 2012b; Sanders *et al.*, 2012; Yuen *et al.*, 2015). However, genes identified as containing *de novo* variants rarely overlap between studies, raising the question of how many genes are truly causal and how likely genes are to be identified as associated with autism by chance in these studies as well as others. We assessed the number of individuals carrying rare (MAF<0.1%), heterozygous LOF variants in 1145 genes cumulatively associated with ASD by more than a dozen studies, meta-analyses and reviews (Vorstman *et al.*, 2005; Kumar & Christian, 2009; Betancur *et al.*, 2011; Miles 2011; Vieland *et al.*, 2011; Davis *et al.*, 2012; Kou *et al.*, 2012; Li *et al.*, 2012; Michaelson *et al*., 2012; Novarino *et al.*, 2012; O'Roak *et al.*, 2012b; Koshimizu *et al.*, 2013; Yu *et al.*, 2013; De Rubeis *et al.*, 2014; Jeste & Geschwind 2014; Liu *et al*., 2014; Toma *et al.,* 2014; Butler *et al.*, 2015; Lee *et al.*, 2015; Turner *et al.*, 2015). There was no significant difference between the distribution of values and that of all genes, and assuming that truly causal genes are more intolerant of rare LOF variants, our findings support the hypothesis that many genes could have been randomly associated with the disorder. (Figure 4.5, Supplementary Table 4.S3) Furthermore, there are 19 putative autism genes in which >0.5% of individuals carry rare, LOF variants. These genes are likely to be false positives, because no single gene contributes to a large proportion of autism cases. Our results highlight the need to perform statistical validation of findings involving genes associated with complex disorders.

Appropriately, several WES studies on ASD calculate the significance of their findings. For example, Sanders *et al.* (2012) demonstrate in a study which identifies *de novo* coding mutations

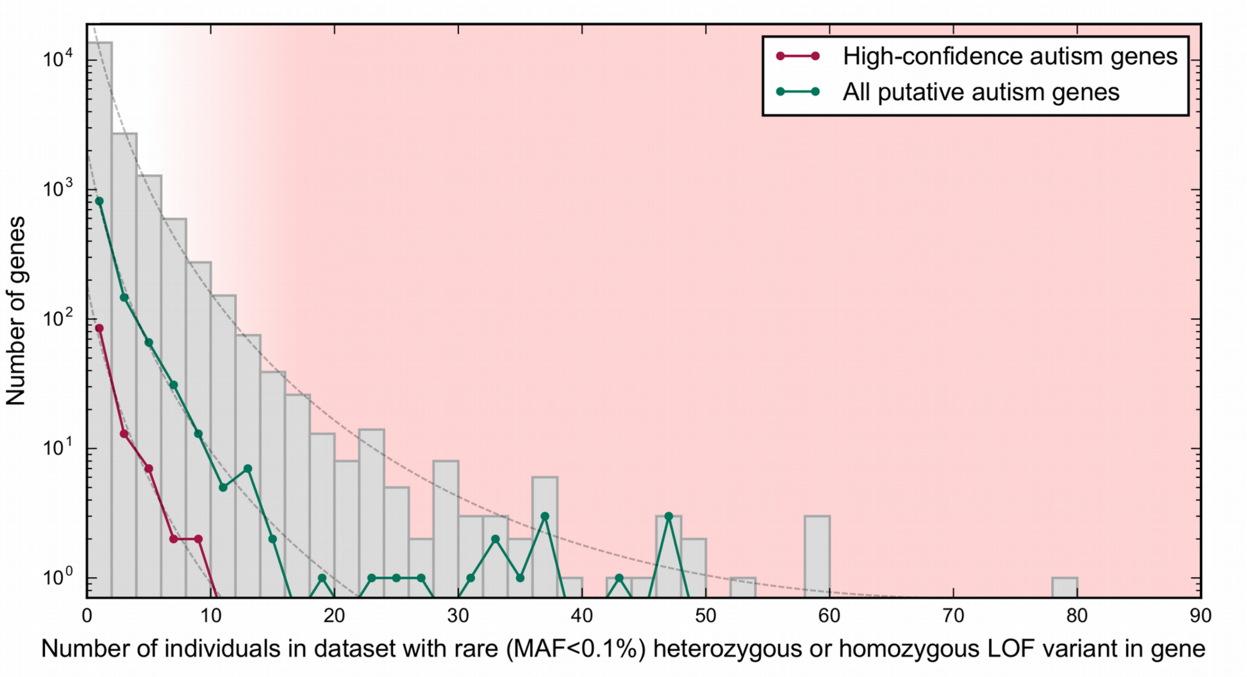


Figure 4.5. Histogram of the number of individuals with rare LOF variants in putative autism genes. The distribution of the number of individuals with a rare variant (MAF < 0.1%) in all genes is nearly identical to the distribution for putative autism genes (N=1145) and high-confidence autism genes (N=109) (dashed lines), suggesting that the genes may have been associated with autism by chance. Genes that frequently contain rare LOF variants in the population (red shaded region) are unlikely to be causal for ASD.

in 928 individuals that finding two independent *de novo* mutations in a single gene is highly unlikely by chance, and this occurring is viewed as evidence for association between ASD and the gene *SCN2A* (sodium channel, voltage-gated, type II, α subunit). Neale *et al.* (2012) also consider the probability of seeing two independent *de novo* mutations in a single gene when evaluating their findings. Iossifov *et al.* (2012) demonstrates that disrupted genes are significantly enriched for FRMP-associated function; however, they also highlight several individual non-FRMP-associated genes based on their plausibility to cause an ASD phenotype but make no attempt at applying statistics when considering these. In fact, *de novo* mutations in genes may have arisen in these genes by chance (Iossifov *et al.*, 2012). This example highlights the challenges faced when evaluating genes associated with complex disorders such as ASD, and the importance of presenting supporting evidence for findings using statistics or follow-up studies before a gene can be established as a high-confidence ASD gene.

4.4.6 Applications in predictive genomics

If a genetic disease is associated with the presence of variants in a given gene, information about the variants in the gene in affected individuals and in population controls can be used to more accurately assess the probability that a person will develop a disease given their genotype.

Consider a randomly chosen person from the general population who is undergoing prenatal genetic testing. Define A as the event that their child will be born with a disease, and B as the event that the child carries a rare, LOF variant in a given gene associated with the disease. For many heterogeneic Mendelian disorders, studies of large cohorts provide information regarding the relative contribution of individual causative genes and the genotype–phenotype correlations, giving us the conditional probability $P(B|A)$. The term $P(A)$ can be defined as the disease incidence, and the value of $P(B)$, or the proportion of individuals carrying a rare, LOF variant in the gene, can be queried from our dataset. Then, according to Bayes' theorem

$$P(A|B) = [\ P(B|A) \times P(A)\ ] / P(B)$$

we can calculate that the probability that the child will have the disorder. The following example illustrates such an application.

Consider that prenatal testing identified that a fetus is compound heterozygous for novel variants in the gene *POMGNT1*, which suggests a possible phenotype of congenital muscular dystrophy (CMD). It is known that 53% of patients with CMD have homozygous or compound heterozygous variants in one of six known CMD genes, 10% have homozygous or compound heterozygous variants in *POMGNT1*, and the incidence of CMD is estimated to be 1:21,500 (Sparks *et al*., 1993; Mercuri *et al.*, 2009). Since most mutations observed in affected individuals are novel and are not found in healthy population controls, we will assume a low MAF threshold of 0.1% for variant filtering. At this threshold, 2 out of 2,504 individuals (0.08 %) in our dataset have a rare protein-altering variant in the gene *POMGNT1*, therefore $P(B)$=0.0008, and we calculate that the positive predictive value (PPV), the probability that the child will have the disease given a positive test result, is roughly 1.0%. Using this method, sensitivity, the probability $P(B|A)$, is quite low (10%); whereas specificity is high (1-P(B) = 99.9%). If we aggregate data for all known CMD genes, we can increase sensitivity to 53% with a negligible decrease in specificity, due to the fact that the other CMD genes contains very few, in any variants in our dataset. This example highlights that sensitivity greatly depends on the proportion of cases that can be explained by variants in a given set of genes. This type of analysis thus has implications for interpretation of broad NGS-based prenatal testing and can be extrapolated as well to preconception testing and risk to potential children.

It is important to note that the extreme numbers involved—the very low prevalence of a disorder and in many cases, the fact that no individual in the 1000 Genomes Project dataset had been observed with variants in a gene, i.e. the lack of previous false-positive results—make it difficult to compute the PPV. A previous study suggests that the latter "zero numerator" problem can be solved using a Bayesian approach that incorporates a prior distribution describing the initial uncertainty about the false-positive rate (Smith *et al*., 2000). Alternatively, the number of rare LOF variants observed in a gene has been published as part of the ExAC dataset, which contains information about 60,706 individuals (Lek *et al.*, 2016). Although only nonsense or splice site variants were included in the LOF classification, and they only include values for a single MAF threshold of 0.1%, the number can be used a rough estimate for *f*. Furthermore, if even the ExAC count is zero, we can assume that *f* is less than 1/60706, or 3/60706 if we are being conservative.

To summarize, for monogenic disorders and disorders where there exist detailed phenotype-genotype correlation data, our dataset will provide the denominator in the equation to calculate the probability that an individual with a rare variant in a known disease gene will have a rare genetic disorder. As further research uncovers novel gene-disease associations, and as we increase the size of the public dataset from which $P(B)$ values can be calculated, we can update expected false-positive rates and calculating PPVs will become increasingly accurate. As illustrated, our methods will be be useful for applications in predictive genomics, including prenatal testing and testing for late-onset genetic disorders.

4.4.7 Comparison to other gene ranking methods

The rankings of frequencies at which a gene contains rare, deleterious variants is comparable to previously published gene ranking methods for prioritizing variants. The list of genes sorted and ranked according to the number of individuals carrying rare (MAF < 0.5%) heterozygous, loss-of-function variants correlates well with genes ranked based on pLI scores, which describe the probability that a gene is intolerant of LOF variation (ρ = 0.515) (Samocha *et al.*, 2014; Lek *et al.*, 2016). These scores were derived from the ExAC dataset consisting of exome sequencing data from 60,706 individuals. The order in which ExAC pLI score ranks genes correlates more closely with SORVA rankings than rankings based on EvoTol (Rackham *et al.*, 2015) (ρ = 0.400), RVIS (Petrovski *et al.*, 2013) (ρ = -0.157) and FLAGS (Shyr *et al.*, 2014) (ρ = 0.278) methods.

We compare methods in their ability to prioritize disease-causing genes from the Online Mendelian Inheritance in Man (OMIM) database (2015). pLI scores, EvoTol, and RVIS outperform SORVA for known autosomal dominant disease genes, however all methods perform similarly for autosomal recessive genes, and SORVA outperforms EvoTol, RVIS, and FLAGS for genes known to cause X-linked disorders. (See Supplementary Figure 3.S3 for receiver operating characteristic (ROC) curves.)

4.5 Discussion

We demonstrated the utility of using mutational burden data to aid in prioritizing exonic variants in genes and known protein domains *in silico*. Other metrics such as gene constraint pLI scores and EvoTol rankings (Samocha *et al.*, 2014; Rackham *et al.*, 2015) are also appropriate for prioritizing genes by their likelihood of causing genetic disorders, but our scores are directly interpretable and can be used to calculate the statistical significance of findings when the study involves sequencing multiple independent individuals.

Although there was some variation between the frequency of individuals with a rare variant in a given gene between populations, and selecting a comparable population to a study would be ideal when calculating variant significance, this restriction is not necessary. To illustrate, if individuals in the African population frequently carry LOF variants in a gene but this does not hold true for another population that more closely matches the study population, one may nevertheless consider the gene to be less likely to cause a rare Mendelian disorder.

A limitation of this method of ranking genes is that genes are prioritized on the basis of their likelihood of being involved in disease in general rather than in the specific disease of interest (Gill *et al*., 2014). On the other hand, this can be viewed as a benefit in the sense that results are unbiased and do not depend on previously existing annotations, which would bias rankings to prefer known and well-studied genes. This bias is a known issue in the interpretation of clinical variants (Wang *et al.*, 2014). To illustrate, Bell *et al.* (2011) discovered that an unexpected proportion (27%) of literature-annotated disease variants in recessive disease-causing genes were incorrect, and Piton *et al.* (2013) estimated that 25% of X-linked intellectual disability genes are incorrect or require further review based on allele frequency estimates that have become more accurate with the availability of large-scale sequencing datasets. Disease genes that are incorrectly annotated as disease-causing may explain the lack of difference between the average number of individuals carrying variants in genes causal for autosomal dominant and autosomal recessive genes. One would expect decreased counts for autosomal dominant disease genes due to stronger purifying selection among deleterious variants that arise in these genes, where a single variant may be sufficient to cause disease (Blekhman *et al.*, 2008). Another possibility is that the sample size may be too small to include a sufficient number of individuals who are carriers for rare, deleterious variants in recessive disease genes.

Future improvements to our methods would include increasing the amount of genetic information from unaffected individuals. Our results suggest that for most applications, low MAF thresholds should be used to achieve power to detect genes associated with disease; however, at thresholds of MAF < 0.0005, most genes will lack any data; e.g. there will be no individuals observed who are carriers of LOF variants. The SORVA dataset is useful in its current state with data from a relatively small number of individuals, but increasing the population size by several orders of magnitude will increase the utility of the application. The recently approved Precision Medicine Initiative will fund sequencing and data collection from 1 million or more Americans and make the data accessible to qualified researchers, and the methods described in this manuscript could be applied to this larger dataset and contribute towards the aim of this initiative to generate knowledge applicable to the whole range of health and disease (Collins & Varmus, 2015).

Additional improvements would include incorporating additional information regarding specific categories of variants, such as the degree to which stop codon gain (also know as nonsense) variants in a gene are constrained to the end of the gene. Knowing whether an essential gene is highly intolerant of nonsense mutations in only certain regions of the gene would allow one to lower the priority of nonsense variants in mutationally tolerant regions when evaluating variants *in silico*. For example, Li *et al*. (2015) exclude stop-gain variants occurring in the terminal gene exon and those that do not affect all transcripts of a gene when evaluating deleterious LOF mutations in a large cohort of individuals. The limitation to providing individual-level mutational burden counts at such a high level of granularity is that researchers will be restricted to following the same methods of filtering and annotating variants. This would be problematic because, by default, many commonly-used software pipelines do not annotate variants with the information about the proportion of transcript truncated (SNP & Variation Suite, Bozeman, MT; Wang *et al.*, 2010; McLaren *et al.*, 2010; Yandell *et al.*, 2011; Habegger *et al.*, 2012; Lucas *et al.*, 2012). Selecting variant filtering thresholds in SORVA that are identical to those used in one's study is essential in having comparable data with which to calculate variant significance. For this reason, we also did not filter missense variants based on annotations from commonly tools such as SIFT (Kumar *et al.*, 2009), PolyPhen-2 (Adzhubei *et al.*, 2010), and CADD (Kircher *et al.*, 2014), which provide an interpretation of mutation impacts.

Finally, future developments would include developing the statistics to calculate statistical significance of findings in the case that related individuals, such as sibpairs, trios or distantly related individuals, are sequenced.

4.6 Conclusions

Our methods provide a score for prioritizing variants within a gene that is unbiased and directly interpretable. Restricted by the sample size of our dataset, we provide limited population-level data, and adding more data will greatly improve the utility of our method. However, even in its current state, SORVA is useful for determining whether genes and known protein domains are depleted of rare variation and vetting candidate variants from NGS studies.

4.7 Supplementary Methods

This section describes deriving the equation for calculating variant significance.

4.7.1. Calculating significance of a de novo variant in a single-family study

Consider a case of a severe, pediatric-onset Mendelian disorder, in which both parents and the affected child are sequenced to identify the causal variant. If only *de novo* variants are identified within a putative gene, making a rough estimate of the *P*-value is relatively simple, as we need to calculate the probability of at least one *de novo* mutation occurring in a gene by random chance. The probability of two or more *de novo* mutations appearing in the same gene is close to zero and can thus be omitted. Ignoring variations in per-base mutation rates (Campbell and Eichler, 2013):

$$P \approx l_{tx} dc$$

where l_{tx} is the length of the transcript in nucleotide bases and d is the mean rate of *de novo* single-nucleotide variants (SNVs) arising per nucleotide per generation. The genome-wide mutation rate is estimated to have a lower bound of 1.2×10^{-8} per site per generation (Campbell and Eichler, 2013; Conrad *et al.*, 2011; Veltman and Brunner, 2012), although sequencing technologies used in the studies are biased against GC-rich DNA, and the rate of mutation at CpG dinucleotides has been observed to be 10- to 18-fold the rate of non-CpG dinucleotides (Campbell *et al.*, 2012; Kondrashov, 2003; Kong *et al.*, 2012; Lynch, 2010). Therefore, the CpG content of a gene should also be considered when determining the parameter *d*. The parameter *c* is the fraction of *de novo* events that meet our protein consequence threshold. It is predicted that 2.85% of *de novo* events are splice site altering or nonsense events, and 70.64% of de novo events are protein-altering, i.e. missense or LOF (Kryukov *et al.*, 2007); these may be used as the respective values for c depending on the variant filtering criteria used.

If the sample size is one and all genes are considered equally likely to cause disease *a priori*, then the P-value may not be significant after correcting for the number of genes in the human genome; hence, follow-up studies are still required in such cases. To illustrate, a whole-exome sequencing study of a single pair of identical twins with autism and seizures identified a de novo missense variant in *KCND2* (Lee *et al.*, 2014), which has a 5,331 base transcript, and the variant was confirmed by functional studies to support causality between the variant and phenotype. The uncorrected P-value would be calculated as $4.5 \times 10{-}5$, which is significant despite the small sample size. However, the P-value corrected for the number of sequenced genes—24,000 to be ultra-conservative—is not significant. In this case, the authors bolstered their study by performing functional studies. Generally, observing a *de novo* variant in an "N=1" study will not be significant, and the relative P-values of genes containing rare variation would be used to either prioritize genes to perform functional studies on, or to identify additional individuals with undiagnosed diseases who carry variants in the same gene, as previous studies have done (Chong *et al.*, 2016; Takenouchi *et al.*, 2016).

4.7.2. Observing homozygous variants in unrelated individuals

Let f_{hom} be the *a priori* fraction of individuals in the population that have a rare, homozygous variant in a given gene. Assume that we sequence n singletons and find that k of these individuals have a variant in the gene. The random variable X is the number of times an individual is seen with a homozygous variant in the gene ("successes") out of *n* individuals sequenced ("independent trials"), and $X \sim$ Binom(n, π) where π is the parameter corresponding to the probability of success on any trial. Let H_0: $\pi \leq f_{hom}$ be the null hypothesis of no association between the phenotype and an individual being homozygous for a variant in the gene. Let H_1: $\pi > f_{hom}$ be the alternative hypothesis that we see a greater number of individuals with a homozygous variant in the gene than expected. The probability of getting exactly *k* successes is:

$$P(X=k)=\binom{n}{k} f_{hom}^{k}(1-f_{hom})^{n-k} \; .$$

The one-sided p-value is the probability of observing at least *k* successes and can be expressed as:

$$P(X \geq k)=P(X>k-1)=1-P(X \leq k-1)=1-BinomCDF(k-1,n,f_{hom})$$

where BinomCDF denotes the binomial cumulative distribution function.

4.7.3. Observing heterozygous variants in unrelated individuals

Let's assume that we observe heterozygous variants in the gene of interest. A *P* value is the probability of obtaining an effect at least as extreme as the one observed, assuming the truth of the null hypothesis. The effect that is at least as extreme as the one observed is equivalent to seeing at least as many individuals who are heterozygous *or homozygous* for a variant in the gene of interest. Therefore, the p-value becomes

$$P(X \geq k)=1-BinomCDF(k-1,n,f_{both})$$

where f_{both} is the fraction of individuals in the population that have either a heterozygous or homozygous variant in the gene of interest.

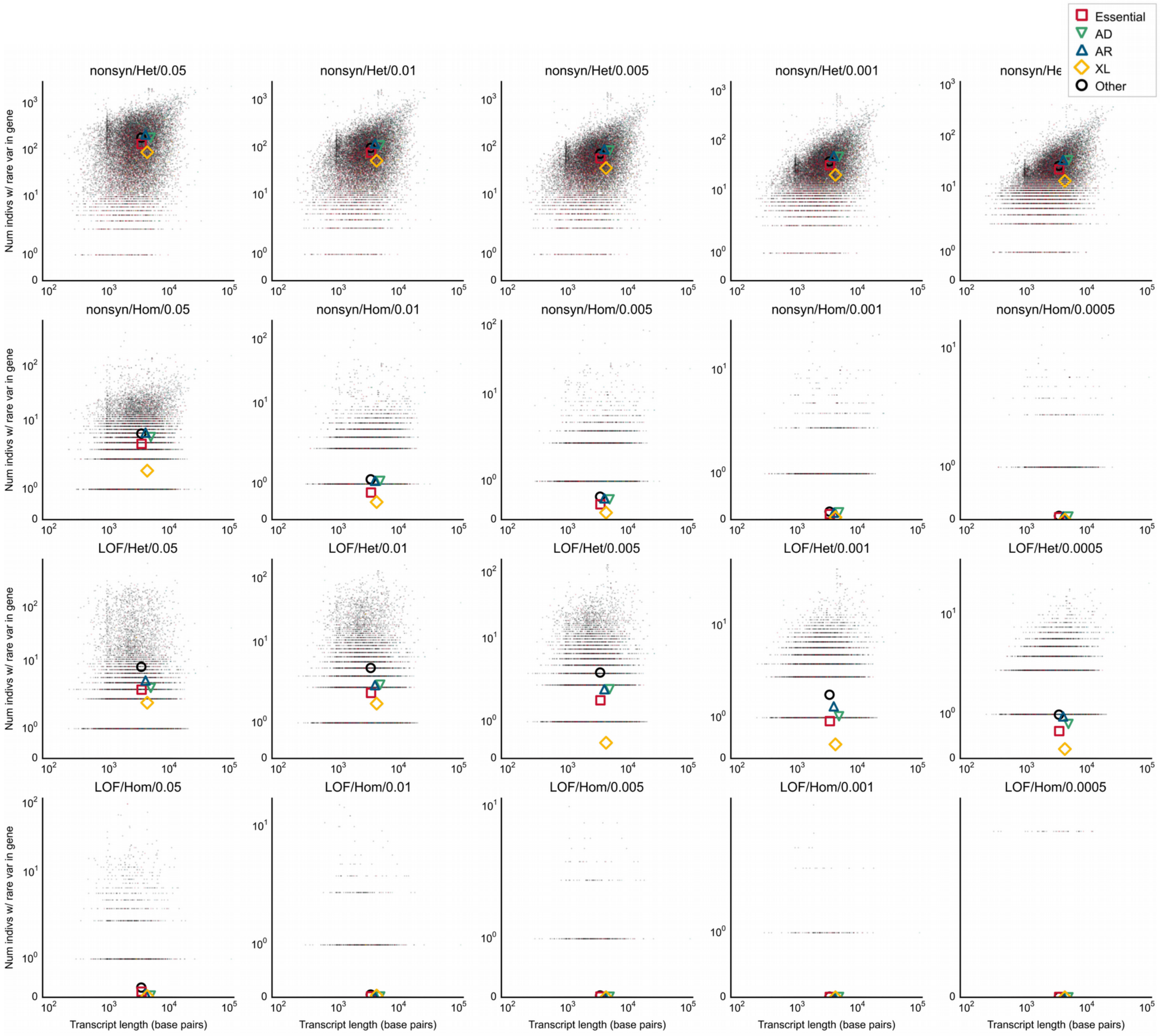
4.7.4. Adjusting for cases where $f = 0$

When calculating the significance of observing homozygous or heterozygous variants in a gene for which no individuals have been observed with a rare variant in the population ($f_{hom} = 0$ or $f_{both} = 0$, respectively), then the calculated *P* value would mistakenly always seem significant. Therefore, for these cases, we set f to a very small number, arbitrarily to the pseudocount

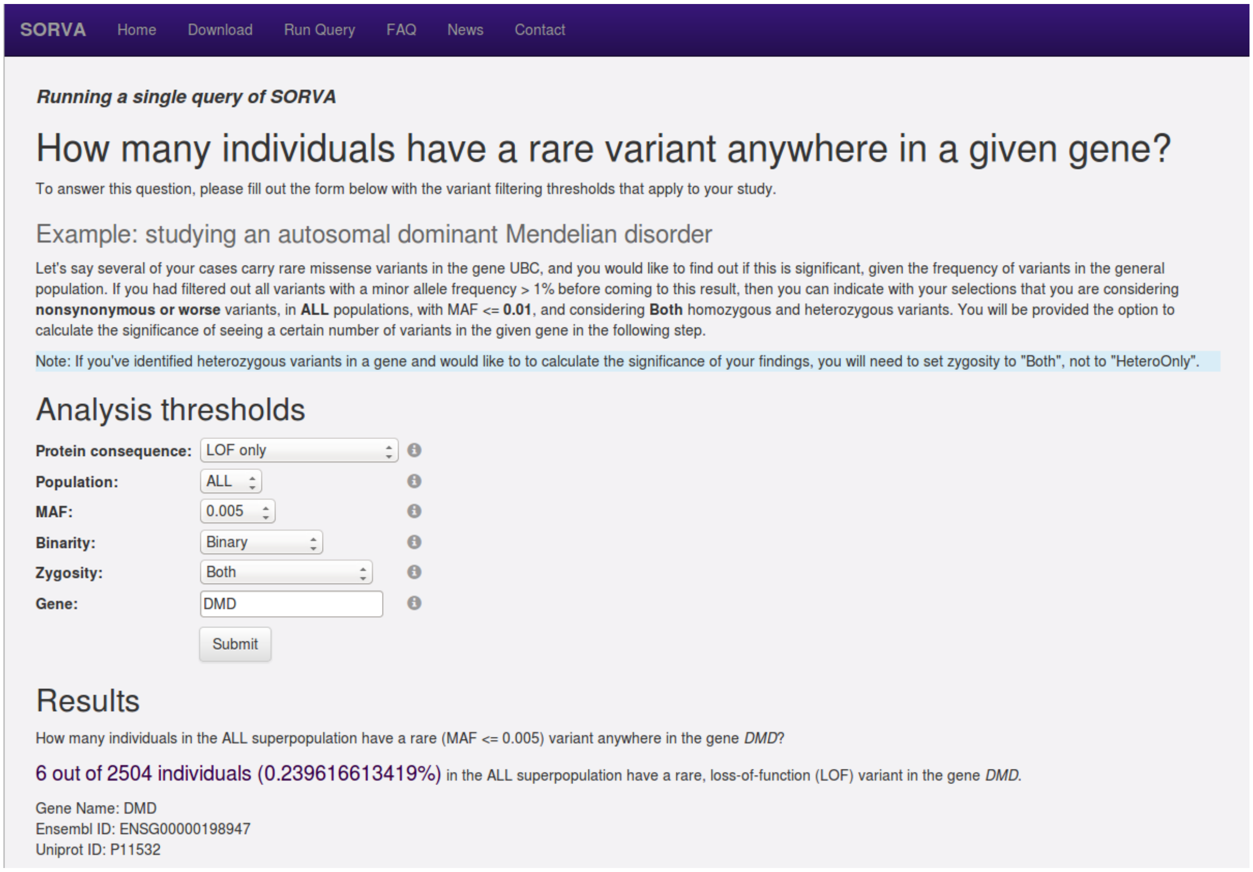
$$f = \frac{1}{2N}$$

where N is the population size from which f was originally derived. The implicit assumption is that if we were to sequence twice the number of samples, we may observe a single individual with a variant in the gene. For genes that are very rarely mutated, the value of f will still be overestimated, resulting in a conservative calculation of the *P*-value.

4.8 Supplementary Figures



Supplementary Figure 4.S1: Number of individuals carrying a rare variant in a gene under various filtering thresholds. Each data point represents a single gene which contains a variant in the aggregate population (n=2504 individuals). Calculations were repeated using multiple variant filtering thresholds to determine the scenario that most differentiates between essential genes, those known to cause autosomal dominant, autosomal recessive or X-linked disease, and other genes. We varied filters for type of variant ('LOF or missense' or 'LOF only'), zygosity (Het or Hom) and MAF threshold. Colored shapes indicate the centroids of each group of genes. Abbreviations: LOF, loss-of-function; nonsyn, nonsynonymous or LOF; het, heterozygous; hom, homozygous; ess, essential; AD, autosomal dominant; AR, autosomal recessive; XL, X-linked.



Running a single query of SORVA

How many individuals have a rare variant anywhere in a given gene?

To answer this question, please fill out the form below with the variant filtering thresholds that apply to your study.

Example: studying an autosomal dominant Mendelian disorder

Let's say several of your cases carry rare missense variants in the gene UBC, and you would like to find out if this is significant, given the frequency of variants in the general population. If you had filtered out all variants with a minor allele frequency > 1% before coming to this result, then you can indicate with your selections that you are considering **nonsynonymous or worse** variants, in **ALL** populations, with MAF <= **0.01**, and considering **Both** homozygous and heterozygous variants. You will be provided the option to calculate the significance of seeing a certain number of variants in the given gene in the following step.

Note: If you've identified heterozygous variants in a gene and would like to to calculate the significance of your findings, you will need to set zygosity to "Both", not to "HeteroOnly".

Analysis thresholds

Protein consequence: LOF only
Population: ALL
MAF: 0.005
Binarity: Binary
Zygosity: Both
Gene: DMD

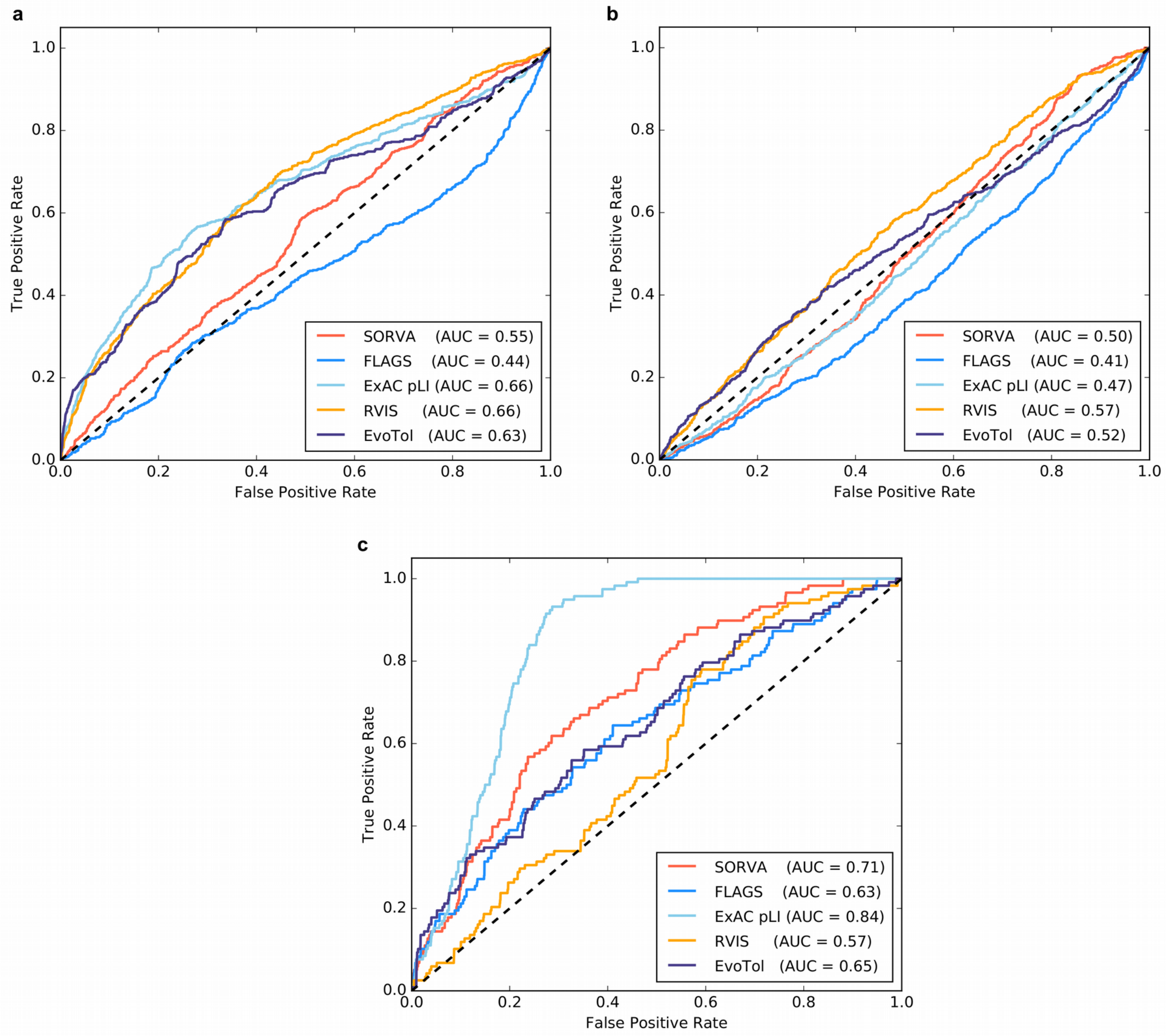
Results

How many individuals in the ALL superpopulation have a rare (MAF <= 0.005) variant anywhere in the gene *DMD*?

6 out of 2504 individuals (0.239616613419%) in the ALL superpopulation have a rare, loss-of-function (LOF) variant in the gene *DMD*.

Gene Name: DMD
Ensembl ID: ENSG00000198947
Uniprot ID: P11532

Supplementary Figure 4.S2: Screenshot of an example query run on SORVA. Users can select variant filtering thresholds such as population, MAF cutoff, zygosity and whether to consider only LOF variants or missense variants. Output includes the number of individuals who carry a rare variant in the gene and in any protein domain that maps to the gene.



Supplementary Figure 4.S3: ROC curves for the selection of known disease-causing genes from gene rankings. Comparison between gene ranking metrics from SORVA, FLAGS, ExAC pLI score, RVIS, and EvoTol using the OMIM database, showing the cumulative percentage plots for the residual scores for three OMIM gene lists. The OMIM gene categories are **(a)** autosomal dominant disease causing (N=681), **(b)** autosomal recessive disease causing (N=556), and **(c)** X-linked disease causing (N=118). SORVA were based on the number of 1000 Genomes Project individuals who were heterozygous or homozygous for rare (MAF<0.005) LOF variants in a given gene. Dashed lines indicate control. Abbreviations: ROC, Receiver Operating Characteristic; AUC, area under the curve, LOF, loss-of-function.

4.9 Supplementary Tables

Legends are provided below for supplementary material attached with the dissertation.

Supplementary Table 4.S1: Mean number of individuals mutated for different types of protein domains. We calculated the mean number of individuals (out of 2,504 individuals) who carried mutations in a given type of protein domains, averaging per gene.

Supplementary Table 4.S2: Variant depletion scores for all protein domain in every gene. For each instance of a protein domain in a gene, we calculated variant depletion scores to identify regions within a gene that may be under differing degrees of evolutionary constraint.

Supplementary Table 4.S3: List of candidate autism genes. Genes listed were used to produce Figure 5.

4.10 Bibliography

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, **7**(4), 248–249. doi:10.1038/nmeth0410-248

Akawi, N., McRae, J., Ansari, M., Balasubramanian, M., Blyth, M., Brady, A. F., Clayton, S., Cole, T., Deshpande, C., Fitzgerald, T. W., Foulds, N., Francis, R., Gabriel, G., Gerety, S. S., Goodship, J., Hobson, E., Jones, W. D., Joss, S., King, D., Klena, N., Kumar, A., Lees, M., Lelliott, C., Lord, J., McMullan, D., O'Regan, M., Osio, D., Piombo, V., et al. (2015). Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nature Genetics*, **47**(11), 1363–1369. doi:10.1038/ng.3410

Akle, S., Chun, S., Jordan, D. M., and Cassa, C. A. (2015). Mitigating false-positive associations in rare disease gene discovery. *Human Mutation*, **36**(10), 998–1003. doi:10.1002/humu.22847

Amendola, L. M., Dorschner, M. O., Robertson, P. D., Salama, J. S., Hart, R., Shirts, B. H., Murray, M. L., Tokita, M. J., Gallego, C. J., Kim, D. S., Bennett, J. T., Crosslin, D. R., Ranchalis, J., Jones, K. L., Rosenthal, E. A., Jarvik, E. R., Itsara, A., Turner, E. H., Herman, D. S., Schleit, J., Burt, A., Jamal, S. M., Abrudan, J. L., Johnson, A. D., Conlin, L. K., Dulik, M. C., Santani, A., Metterville, D. R., et al. (2015). Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Research*, **25**(3), 305–315. doi:10.1101/gr.183483.114

Arboleda, V. A., Lee, H., Dorrani, N., Zadeh, N., Willis, M., Macmurdo, C. F., Manning, M. A., Kwan, A., Hudgins, L., Barthelemy, F., Miceli, M. C., Quintero-Rivera, F., Kantarci, S., Strom, S. P., Deignan, J. L., Grody, W. W., Vilain, E., and Nelson, S. F. (2015). De novo nonsense mutations in KAT6A, a lysine acetyl-transferase gene, cause a syndrome including microcephaly and global developmental delay. *American Journal of Human Genetics*, **96**(3), 498–506. doi:10.1016/j.ajhg.2015.01.017

Arboleda, V. A., Lee, H., Parnaik, R., Fleming, A., Banerjee, A., Ferraz-de-Souza, B., Délot, E. C., Rodriguez-Fernandez, I. A., Braslavsky, D., Bergadá, I., Dell'Angelica, E. C., Nelson, S. F., Martinez-Agosto, J. A., Achermann, J. C., and Vilain, E. (2012). Mutations in the PCNA-binding domain of CDKN1C cause IMAGe syndrome. *Nature Genetics*, **44**(7), 788–792. doi:10.1038/ng.2275

Baasch, A.-L., Hüning, I., Gilissen, C., Klepper, J., Veltman, J. A., Gillessen-Kaesbach, G., Hoischen, A., and Lohmann, K. (2014). Exome sequencing identifies a de novo SCN2A mutation in a patient with intractable seizures, severe intellectual disability, optic atrophy, muscular hypotonia, and brain abnormalities. *Epilepsia*, **55**(4), e25–e29. doi:10.1111/epi.12554

Bagnall, R. D., Molloy, L. K., Kalman, J. M., and Semsarian, C. (2014). Exome sequencing identifies a mutation in the ACTN2 gene in a family with idiopathic ventricular fibrillation, left ventricular noncompaction, and sudden death. *BMC Medical Genetics*, **15**, 99. doi:10.1186/s12881-014-0099-0

Bayram, Y., Aydin, H., Gambin, T., Akdemir, Z. C., Atik, M. M., Karaca, E., Karaman, A., Pehlivan, D., Jhangiani, S. N., Gibbs, R. A., and Lupski, J. R. (2015). Exome sequencing identifies a homozygous C5orf42 variant in a Turkish kindred with oral-facial-digital syndrome type VI. *American Journal of Medical Genetics Part A*, **167**(9), 2132–2137. doi:10.1002/ajmg.a.37092

Belaya, K., Cruz, P. M. R., Liu, W. W., Maxwell, S., McGowan, S., Farrugia, M. E., Petty, R., Walls, T. J., Sedghi, M., Basiri, K., Yue, W. W., Sarkozy, A., Bertoli, M., Pitt, M., Kennett, R., Schaefer, A., Bushby, K., Parton, M., Lochmüller, H., Palace, J., Muntoni, F., and Beeson, D. (2015). Mutations in GMPPB cause congenital myasthenic syndrome and bridge myasthenic disorders with dystroglycanopathies. *Brain*, **138**(9), 2493–2504. doi:10.1093/brain/awv185

Bell, C. J., Dinwiddie, D. L., Miller, N. A., Hateley, S. L., Ganusova, E. E., Mudge, J., Langley, R. J., Zhang, L., Lee, C. C., Schilkey, F. D., Sheth, V., Woodward, J. E., Peckham, H. E., Schroth, G. P., Kim, R. W., and Kingsmore, S. F. (2011). Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science Translational Medicine*, **3**(65), 65ra4-65ra4. doi:10.1126/scitranslmed.3001756

Benson, D. W., Wang, D. W., Dyment, M., Knilans, T. K., Fish, F. A., Strieper, M. J., Rhodes, T. H., and George, A. L. (2003). Congenital sick sinus syndrome caused by recessive mutations in the cardiac sodium channel gene (SCN5A). *The Journal of Clinical Investigation*, **112**(7), 1019–1028. doi:10.1172/JCI18062

Berko, E. R., Cho, M. T., Eng, C., Shao, Y., Sweetser, D. A., Waxler, J., Robin, N. H., Brewer, F., Donkervoort, S., Mohassel, P., Bönnemann, C. G., Bialer, M., Moore, C., Wolfe, L. A., Tifft, C. J., Shen, Y., Retterer, K., Millan, F., and Chung, W. K. (2017). De novo missense variants in HECW2 are associated with neurodevelopmental delay and hypotonia. *Journal of Medical Genetics*, **54**(2), 84–86. doi:10.1136/jmedgenet-2016-103943

Betancur, C. (2011). Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Brain Research*, **1380**, 42–77. doi:10.1016/j.brainres.2010.11.078

Bilgüvar, K., Öztürk, A. K., Louvi, A., Kwan, K. Y., Choi, M., Tatlı, B., Yalnızoğlu, D., Tüysüz, B., Çağlayan, A. O., Gökben, S., Kaymakçalan, H., Barak, T., Bakırcıoğlu, M., Yasuno, K., Ho, W., Sanders, S., Zhu, Y., Yılmaz, S., Dinçer, A., Johnson, M. H., Bronen, R. A., Koçer, N., Per, H., Mane, S., Pamir, M. N., Yalçınkaya, C., Kumandaş, S., Topçu, M., et al. (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature*, **467**(7312), 207–210. doi:10.1038/nature09327

Blekhman, R., Man, O., Herrmann, L., Boyko, A. R., Indap, A., Kosiol, C., Bustamante, C. D., Teshima, K. M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. *Current Biology*, **18**(12), 883–889. doi:10.1016/j.cub.2008.04.074

Blomen, V. A., Májek, P., Jae, L. T., Bigenzahn, J. W., Nieuwenhuis, J., Staring, J., Sacco, R., Diemen, F. R. van, Olk, N., Stukalov, A., Marceau, C., Janssen, H., Carette, J. E., Bennett, K. L., Colinge, J., Superti-Furga, G., and Brummelkamp, T. R. (2015). Gene essentiality and synthetic lethality in haploid human cells. *Science*, aac7557. doi:10.1126/science.aac7557

Butler, M. G., Rafi, S. K., and Manzardo, A. M. (2015). High-resolution chromosome ideogram representation of currently recognized genes for autism spectrum disorders. *International Journal of Molecular Sciences*, **16**(3), 6464–6495. doi:10.3390/ijms16036464

Campbell, C. D., Chong, J. X., Malig, M., Ko, A., Dumont, B. L., Han, L., Vives, L., O'Roak, B. J., Sudmant, P. H., Shendure, J., Abney, M., Ober, C., and Eichler, E. E. (2012). Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics*, **44**(11), 1277–1281. doi:10.1038/ng.2418

Campbell, C. D., and Eichler, E. E. (2013). Properties and rates of germline mutations in humans. *Trends in genetics : TIG*, **29**(10), 575–584. doi:10.1016/j.tig.2013.04.005

Campeau, P. M., Kasperaviciute, D., Lu, J. T., Burrage, L. C., Kim, C., Hori, M., Powell, B. R., Stewart, F., Félix, T. M., van den Ende, J., Wisniewska, M., Kayserili, H., Rump, P., Nampoothiri, S., Aftimos, S., Mey, A., Nair, L. D. V., Begleiter, M. L., De Bie, I., Meenakshi, G., Murray, M. L., Repetto, G. M., Golabi, M., Blair, E., Male, A., Giuliano, F., Kariminejad, A., Newman, W. G., et al. (2014). The genetic basis of DOORS syndrome: an exome-sequencing study. *The Lancet Neurology*, **13**(1), 44–58. doi:10.1016/S1474-4422(13)70265-5

Chen, Y.-Z., Matsushita, M. M., Robertson, P., Rieder, M., Girirajan, S., Antonacci, F., Lipe, H., Eichler, E. E., Nickerson, D. A., Bird, T. D., and Raskind, W. H. (2012). Autosomal dominant familial dyskinesia and facial myokymia: single exome sequencing identifies a mutation in adenylyl cyclase 5. *Archives of neurology*, **69**(5), 630–635. doi:10.1001/archneurol.2012.54

Chong, J. X., Yu, J.-H., Lorentzen, P., Park, K. M., Jamal, S. M., Tabor, H. K., Rauch, A., Saenz, M. S., Boltshauser, E., Patterson, K. E., Nickerson, D. A., Bamshad, M. J., and Genomics, U. of W. C. for M. (2016). Gene discovery for Mendelian conditions via social networking: de novo variants in KDM1A cause developmental delay and distinctive facial features. *Genetics in Medicine*, **18**(8), 788–795. doi:10.1038/gim.2015.161

Chudasama, K. K., Winnay, J., Johansson, S., Claudi, T., König, R., Haldorsen, I., Johansson, B., Woo, J. R., Aarskog, D., Sagen, J. V., Kahn, C. R., Molven, A., and Njølstad, P. R. (2013). SHORT syndrome with partial lipodystrophy due to impaired phosphatidylinositol 3 kinase signaling. *The American Journal of Human Genetics*, **93**(1), 150–157. doi:10.1016/j.ajhg.2013.05.023

Collins, F. S., and Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, **372**(9), 793–795. doi:10.1056/NEJMp1500523

Comino-Méndez, I., Gracia-Aznárez, F. J., Schiavi, F., Landa, I., Leandro-García, L. J., Letón, R., Honrado, E., Ramos-Medina, R., Caronia, D., Pita, G., Gómez-Graña, Á., de Cubas, A.

A., Inglada-Pérez, L., Maliszewska, A., Taschin, E., Bobisse, S., Pica, G., Loli, P., Hernández-Lavado, R., Díaz, J. A., Gómez-Morales, M., González-Neira, A., Roncador, G., Rodríguez-Antona, C., Benítez, J., Mannelli, M., Opocher, G., Robledo, M., and Cascón, A. (2011). Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nature Genetics*, **43**(7), 663–667. doi:10.1038/ng.861

Conrad, D. F., Keebler, J. E. M., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., Zilversmit, M., Cartwright, R., Rouleau, G. A., Daly, M., Stone, E. A., Hurles, M. E., Awadalla, P., and 1000 Genomes Project. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, **43**(7), 712–714. doi:10.1038/ng.862

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., et al. (2015). Ensembl 2015. *Nucleic Acids Research*, **43**(D1), D662–D669. doi:10.1093/nar/gku1010

Davis, L. K., Gamazon, E. R., Kistner-Griffin, E., Badner, J. A., Liu, C., Cook, E. H., Sutcliffe, J. S., and Cox, N. J. (2012). Loci nominally associated with autism from genome-wide analysis show enrichment of brain expression quantitative trait loci but not lymphoblastoid cell line expression quantitative trait loci. *Molecular Autism*, **3**(1), 3. doi:10.1186/2040-2392-3-3

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Ercument Cicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., Singh, T., Klei, L., Kosmicki, J., Fu, S.-C., Aleksic, B., Biscaldi, M., Bolton, P. F., Brownfeld, J. M., Cai, J., Campbell, N. G., Carracedo, A., Chahrour, M. H., Chiocchetti, A. G., Coon, H., Crawford, E. L., Crooks, L., Curran, S. R., Dawson, G., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, **515**(7526), 209–215. doi:10.1038/nature13772

Deardorff, M. A., Bando, M., Nakato, R., Watrin, E., Itoh, T., Minamino, M., Saitoh, K., Komata, M., Katou, Y., Clark, D., Cole, K. E., Baere, E. D., Decroos, C., Donato, N. D., Ernst, S., Francey, L. J., Gyftodimou, Y., Hirashima, K., Hullings, M., Ishikawa, Y., Jaulin, C., Kaur, M., Kiyono, T., Lombardi, P. M., Magnaghi-Jaulin, L., Mortier, G. R., Nozaki, N., Petersen, M. B., et al. (2012). HDAC8 mutations in Cornelia de Lange Syndrome affect the cohesin acetylation cycle. *Nature*, **489**(7415), 313–317. doi:10.1038/nature11316

Deardorff, M. A., Kaur, M., Yaeger, D., Rampuria, A., Korolev, S., Pie, J., Gil-Rodríguez, C., Arnedo, M., Loeys, B., Kline, A. D., Wilson, M., Lillquist, K., Siu, V., Ramos, F. J., Musio, A., Jackson, L. S., Dorsett, D., and Krantz, I. D. (2007). Mutations in cohesin complex members SMC3 and SMC1A cause a mild variant of Cornelia de Lange syndrome with predominant mental retardation. *American Journal of Human Genetics*, **80**(3), 485–494.

Dorschner, M. O., Amendola, L. M., Turner, E. H., Robertson, P. D., Shirts, B. H., Gallego, C. J., Bennett, R. L., Jones, K. L., Tokita, M. J., Bennett, J. T., Kim, J. H., Rosenthal, E. A., Kim, D. S., National Heart, L., Tabor, H. K., Bamshad, M. J., Motulsky, A. G., Scott, C. R., Pritchard, C. C., Walsh, T., Burke, W., Raskind, W. H., Byers, P., Hisama, F. M., Nickerson, D. A., and Jarvik, G. P. (2013). Actionable, pathogenic incidental findings in 1,000 participants' exomes. *The American Journal of Human Genetics*, **93**(4), 631–640. doi:10.1016/j.ajhg.2013.08.006

Dyment, D. A., Smith, A. C., Alcantara, D., Schwartzentruber, J. A., Basel-Vanagaite, L., Curry, C. J., Temple, I. K., Reardon, W., Mansour, S., Haq, M. R., Gilbert, R., Lehmann, O. J., Vanstone, M. R., Beaulieu, C. L., Majewski, J., Bulman, D. E., O'Driscoll, M., Boycott, K. M., and Innes, A. M. (2013). Mutations in PIK3R1 cause SHORT syndrome. *The American Journal of Human Genetics*, **93**(1), 158–166. doi:10.1016/j.ajhg.2013.06.005

Erich L. Lehmann. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Oakland, Calif.: Holden-Day. http://www.springer.com/us/book/9780387352121. Accessed 7 December 2015

Gill, N., Singh, S., and Aseri, T. C. (2014). Computational disease gene prioritization: an appraisal. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **21**(6), 456–465. doi:10.1089/cmb.2013.0158

Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S., and Goldstein, D. B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biology*, **17**, 9. doi:10.1186/s13059-016-0869-4

Habegger, L., Balasubramanian, S., Chen, D. Z., Khurana, E., Sboner, A., Harmanci, A., Rozowsky, J., Clarke, D., Snyder, M., and Gerstein, M. (2012). VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics*, **28**(17), 2267–2269. doi:10.1093/bioinformatics/bts368

He, X., Sanders, S. J., Liu, L., De Rubeis, S., Lim, E. T., Sutcliffe, J. S., Schellenberg, G. D., Gibbs, R. A., Daly, M. J., Buxbaum, J. D., State, M. W., Devlin, B., and Roeder, K. (2013). Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. *PLoS Genet*, **9**(8), e1003671. doi:10.1371/journal.pgen.1003671

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y., Narzisi, G., Leotta, A., Kendall, J., Grabowska, E., Ma, B., Marks, S., Rodgers, L., Stepansky, A., Troge, J., Andrews, P., Bekritsky, M., Pradhan, K., Ghiban, E., Kramer, M., Parla, J., Demeter, R., Fulton, L. L., Fulton, R. S., Magrini, V. J., Ye, K., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron*, **74**(2), 285–299. doi:10.1016/j.neuron.2012.04.009

Jeste, S. S., and Geschwind, D. H. (2014). Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nature Reviews Neurology*, **10**(2), 74–81. doi:10.1038/nrneurol.2013.278

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, **46,** 310–315. doi:10.1038/ng.2892

Kondrashov, A. S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human Mutation*, **21**(1), 12–27. doi:10.1002/humu.10147

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W. S. W., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., et al. (2012). Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature*, **488**(7412), 471–475. doi:10.1038/nature11396

Koshimizu, E., Miyatake, S., Okamoto, N., Nakashima, M., Tsurusaki, Y., Miyake, N., Saitsu, H., and Matsumoto, N. (2013). Performance comparison of bench-top next generation sequencers using microdroplet PCR-based enrichment for targeted sequencing in patients with autism spectrum disorder. *PLOS ONE*, **8**(9), e74167. doi:10.1371/journal.pone.0074167

Kou, Y., Betancur, C., Xu, H., Buxbaum, J. D., and Ma'ayan, A. (2012). Network- and attribute-based classifiers can prioritize genes and pathways for autism spectrum disorders and intellectual disability. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, **160C**(2), 130–142. doi:10.1002/ajmg.c.31330

Kryukov, G. V., Pennacchio, L. A., and Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *American Journal of Human Genetics*, **80**(4), 727–739.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, **4**(7), 1073–1081. doi:10.1038/nprot.2009.86

Kumar, R. A., and Christian, S. L. (2009). Genetics of autism spectrum disorders. *Current Neurology and Neuroscience Reports*, **9**(3), 188–197. doi:10.1007/s11910-009-0029-2

Lee, H., Graham, J. M., Rimoin, D. L., Lachman, R. S., Krejci, P., Tompson, S. W., Nelson, S. F., Krakow, D., and Cohn, D. H. (2012). Exome sequencing identifies PDE4D mutations in acrodysostosis. *American Journal of Human Genetics*, **90**(4), 746–751. doi:10.1016/j.ajhg.2012.03.004

Lee, H., Lin, M. A., Kornblum, H. I., Papazian, D. M., and Nelson, S. F. (2014). Exome sequencing identifies de novo gain of function missense mutation in KCND2 in identical twins with autism and seizures that slows potassium channel inactivation. *Human Molecular Genetics*, **23**(13), 3481–3489. doi:10.1093/hmg/ddu056

Lee, M. S., Kim, Y. J., Kim, E. J., and Lee, M. J. (2015). Overlap of autism spectrum disorder and glucose transporter 1 deficiency syndrome associated with a heterozygous deletion at the 1p34.2 region. *Journal of the Neurological Sciences*, **356**(1), 212–214. doi:10.1016/j.jns.2015.06.041

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**(7616), 285–291. doi:10.1038/nature19057

Li, A. H., Morrison, A. C., Kovar, C., Cupples, L. A., Brody, J. A., Polfus, L. M., Yu, B., Metcalf, G., Muzny, D., Veeraraghavan, N., Liu, X., Lumley, T., Mosley, T. H., Gibbs, R. A., and Boerwinkle, E. (2015). Analysis of loss-of-function variants and 20 risk factor phenotypes in 8,554 individuals identifies loci influencing chronic disease. *Nature Genetics*, **47**(6), 640–642. doi:10.1038/ng.3270

Li, X., Zou, H., and Brown, W. T. (2012). Genes associated with autism spectrum disorder. *Brain Research Bulletin*, **88**(6), 543–552. doi:10.1016/j.brainresbull.2012.05.017

Liu, L., Lei, J., Sanders, S. J., Willsey, A. J., Kou, Y., Cicek, A. E., Klei, L., Lu, C., He, X., Li, M., Muhle, R. A., Ma'ayan, A., Noonan, J. P., Šestan, N., McFadden, K. A., State, M. W., Buxbaum, J. D., Devlin, B., and Roeder, K. (2014). DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Molecular Autism*, **5**, 22. doi:10.1186/2040-2392-5-22

Lopez, E., Thauvin-Robinet, C., Reversade, B., Khartoufi, N. E., Devisme, L., Holder, M., Ansart-Franquet, H., Avila, M., Lacombe, D., Kleinfinger, P., Kaori, I., Takanashi, J.-I., Merrer, M. L., Martinovic, J., Noël, C., Shboul, M., Ho, L., Güven, Y., Razavi, F., Burglen, L., Gigot, N., Darmency-Stamboul, V., Thevenon, J., Aral, B., Kayserili, H., Huet, F., Lyonnet, S., Caignec, C. L., et al. (2013). C5orf42 is the major gene responsible for OFD syndrome type VI. *Human Genetics*, **133**(3), 367–377. doi:10.1007/s00439-013-1385-1

Lucas, F. A. S., Wang, G., Scheet, P., and Peng, B. (2012). Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics*, **28**(3), 421–422. doi:10.1093/bioinformatics/btr667

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences*, **107**(3), 961–968. doi:10.1073/pnas.0912629107

Marth, G., Schuler, G., Yeh, R., Davenport, R., Agarwala, R., Church, D., Wheelan, S., Baker, J., Ward, M., Kholodov, M., Phan, L., Czabarka, E., Murvai, J., Cutler, D., Wooding, S., Rogers, A., Chakravarti, A., Harpending, H. C., Kwok, P.-Y., and Sherry, S. T. (2003). Sequence variations in the public human genome data reflect a bottlenecked population history. *Proceedings of the National Academy of Sciences*, **100**(1), 376–381. doi:10.1073/pnas.222673099

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**(16), 2069–2070. doi:10.1093/bioinformatics/btq330

Mercuri, E., Messina, S., Bruno, C., Mora, M., Pegoraro, E., Comi, G. P., D'Amico, A., Aiello, C., Biancheri, R., Berardinelli, A., Boffi, P., Cassandrini, D., Laverda, A., Moggio, M., Morandi, L., Moroni, I., Pane, M., Pezzani, R., Pichiecchio, A., Pini, A., Minetti, C., Mongini, T., Mottarelli, E., Ricci, E., Ruggieri, A., Saredi, S., Scuderi, C., Tessa, A., et al. (2009). Congenital muscular dystrophies with defective glycosylation of dystroglycan A population study. *Neurology*, **72**(21), 1802–1809. doi:10.1212/01.wnl.0000346518.68110.60

Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., Wu, W., Corominas, R., Peoples, Á., Koren, A., Gore, A., Kang, S., Lin, G. N., Estabillo, J., Gadomski, T., Singh, B., Zhang, K., Akshoomoff, N., Corsello, C., McCarroll, S., Iakoucheva, L. M., Li, Y., Wang, J., and Sebat, J. (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, **151**(7), 1431–1442. doi:10.1016/j.cell.2012.11.019

Miles, J. H. (2011). Autism spectrum disorders—A genetics review. *Genetics in Medicine*, **13**(4), 278–294. doi:10.1097/GIM.0b013e3181ff67ba

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.-Y., Bateman, A., Punta, M., Attwood, T. K., Sigrist, C. J. A., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., et al. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, **43**(D1), D213–D221. doi:10.1093/nar/gku1243

Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., Polak, P., Yoon, S., Maguire, J., Crawford, E. L., Campbell, N. G., Geller, E. T., Valladares, O., Schafer, C., Liu, H., Zhao, T., Cai, G., Lihm, J., Dannenfelser, R., Jabado, O., Peralta, Z., Nagaswamy, U., Muzny, D., Reid, J. G., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, **485**(7397), 242–245. doi:10.1038/nature11011

Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., Beck, A. E., Tabor, H. K., Cooper, G. M., Mefford, H. C., Lee, C., Turner, E. H., Smith, J. D., Rieder, M. J., Yoshiura, K., Matsumoto, N., Ohta, T., Niikawa, N., Nickerson, D. A., Bamshad, M. J., and Shendure, J. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics*, **42**(9), 790–793. doi:10.1038/ng.646

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., and Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, **42**(1), 30–35. doi:10.1038/ng.499

Novarino, G., El-Fishawy, P., Kayserili, H., Meguid, N. A., Scott, E. M., Schroth, J., Silhavy, J. L., Kara, M., Khalil, R. O., Ben-Omran, T., Ercan-Sencicek, A. G., Hashish, A. F., Sanders, S. J., Gupta, A. R., Hashem, H. S., Matern, D., Gabriel, S., Sweetman, L., Rahimi, Y., Harris, R. A., State, M. W., and Gleeson, J. G. (2012). Mutations in BCKD-kinase lead to a potentially treatable form of autism with epilepsy. *Science*, **338**(6105), 394–397. doi:10.1126/science.1224631

Online Mendelian Inheritance in Man, OMIM®. (n.d.). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), Jan 2015. http://omim.org/

O'Roak, B. J., Vives, L., Fu, W., Egertson, J. D., Stanaway, I. B., Phelps, I. G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J. B., Turner, E. H., Levy, R., O'Day, D. R., Krumm, N., Coe, B. P., Martin, B. K., Borenstein, E., Nickerson, D. A., Mefford, H. C., Doherty, D., Akey, J. M., Bernier, R., Eichler, E. E., and Shendure, J. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*, **338**(6114), 1619–1622. doi:10.1126/science.1227764

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., Levy, R., Ko, A., Lee, C., Smith, J. D., Turner, E. H., Stanaway, I. B., Vernot, B., Malig, M., Baker, C., Reilly, B., Akey, J. M., Borenstein, E., Rieder, M. J., Nickerson, D. A., Bernier, R., Shendure, J., and Eichler, E. E. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, **485**(7397), 246–250. doi:10.1038/nature10989

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., and Goldstein, D. B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*, **9**(8), e1003709. doi:10.1371/journal.pgen.1003709

Piton, A., Redin, C., and Mandel, J.-L. (2013). XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. *The American Journal of Human Genetics*, **93**(2), 368–383. doi:10.1016/j.ajhg.2013.06.013

Rackham, O. J. L., Shihab, H. A., Johnson, M. R., and Petretto, E. (2015). EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. *Nucleic Acids Research*, **43**(5), e33. doi:10.1093/nar/gku1322

Raza, M. H., Mattera, R., Morell, R., Sainz, E., Rahn, R., Gutierrez, J., Paris, E., Root, J., Solomon, B., Brewer, C., Basra, M. A. R., Khan, S., Riazuddin, S., Braun, A., Bonifacino, J. S., and Drayna, D. (2015). Association between rare variants in AP4E1, a component of intracellular trafficking, and persistent stuttering. *American Journal of Human Genetics*, **97**(5), 715–725. doi:10.1016/j.ajhg.2015.10.007

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook Jr, E. H., Gibbs, R. A., Schellenberg, G. D., Sutcliffe, J. S., Devlin, B., Roeder, K., Neale, B. M., and Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, **46**(9), 944–950. doi:10.1038/ng.3050

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., Murtha, M. T., Bal, V. H., Bishop, S. L., Dong, S., Goldberg, A. P., Jinlu, C., Keaney, J. F., Klei, L., Mandell, J. D., Moreno-De-Luca, D., Poultney, C. S., Robinson, E. B., Smith, L., Solli-Nowlan, T., Su, M. Y., Teran, N. A., Walker, M. F., Werling, D. M., Beaudet, A. L., Cantor, R. M., Fombonne, E., Geschwind, D. H., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*, **87**(6), 1215–1233. doi:10.1016/j.neuron.2015.09.016

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., Ercan-Sencicek, A. G., DiLullo, N. M., Parikshak, N. N., Stein, J. L., Walker, M. F., Ober, G. T., Teran, N. A., Song, Y., El-Fishawy, P., Murtha, R. C., Choi, M., Overton, J. D., Bjornson, R. D., Carriero, N. J., Meyer, K. A., Bilguvar, K., Mane, S. M., Šestan, N., Lifton, R. P., Günel, M., Roeder, K., Geschwind, D. H., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, **485**(7397), 237–241. doi:10.1038/nature10945

Shyr, C., Tarailo-Graovac, M., Gottlieb, M., Lee, J. J., Karnebeek, C. van, and Wasserman, W. W. (2014). FLAGS, frequently mutated genes in public exomes. *BMC Medical Genomics*, **7**(1), 64. doi:10.1186/s12920-014-0064-y

Smith, J. E., Winkler, R. L., and Fryback, D. G. (2000). The first positive: Computing positive predictive value at the extremes. *Annals of Internal Medicine*, **132**(10), 804–809. doi:10.7326/0003-4819-132-10-200005160-00008

SNP & Variation Suite ™ (Version 8.1). (Available from http://www.goldenhelix.com). Bozeman, MT: Golden Helix, Inc. http://www.goldenhelix.com

Sparks, S., Quijano-Roy, S., Harper, A., Rutkowski, A., Gordon, E., Hoffman, E. P., and Pegoraro, E. (1993). Congenital muscular dystrophy overview. In R. A. Pagon, M. P. Adam, H. H. Ardinger, S. E. Wallace, A. Amemiya, L. J. Bean, T. D. Bird, C.-T. Fong, H. C. Mefford, R. J. Smith, and K. Stephens (Eds.), *GeneReviews(®)*. Seattle (WA): University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK1291/. Accessed 3 May 2016

Stránecký, V., Hoischen, A., Hartmannová, H., Zaki, M. S., Chaudhary, A., Zudaire, E., Nosková, L., Barešová, V., Přistoupilová, A., Hodaňová, K., Sovová, J., Hůlková, H., Piherová, L., Hehir-Kwa, J. Y., de Silva, D., Senanayake, M. P., Farrag, S., Zeman, J., Martásek, P., Baxová, A., Afifi, H. H., St Croix, B., Brunner, H. G., Temtamy, S., and Kmoch, S. (2013). Mutations in ANTXR1 cause GAPO syndrome. *American Journal of Human Genetics*, **92**(5), 792–799. doi:10.1016/j.ajhg.2013.03.023

Stray-Pedersen, A., Backe, P. H., Sorte, H. S., Mørkrid, L., Chokshi, N. Y., Erichsen, H. C., Gambin, T., Elgstøen, K. B. P., Bjørås, M., Wlodarski, M. W., Krüger, M., Jhangiani, S. N., Muzny, D. M., Patel, A., Raymond, K. M., Sasa, G. S., Krance, R. A., Martinez, C. A., Abraham, S. M., Speckmann, C., Ehl, S., Hall, P., Forbes, L. R., Merckoll, E., Westvik, J., Nishimura, G., Rustad, C. F., Abrahamsen, T. G., et al. (2014). PGM3 mutations cause a congenital disorder of glycosylation with severe immunodeficiency and skeletal dysplasia. *American Journal of Human Genetics*, **95**(1), 96–107. doi:10.1016/j.ajhg.2014.05.007

Takenouchi, T., Miura, K., Uehara, T., Mizuno, S., and Kosaki, K. (2016). Establishing SON in 21q22.11 as a cause a new syndromic form of intellectual disability: Possible contribution to Braddock–Carey syndrome phenotype. *American Journal of Medical Genetics Part A*, n/a-n/a. doi:10.1002/ajmg.a.37761

Tham, E., Lindstrand, A., Santani, A., Malmgren, H., Nesbitt, A., Dubbs, H. A., Zackai, E. H., Parker, M. J., Millan, F., Rosenbaum, K., Wilson, G. N., and Nordgren, A. (2015).

Dominant mutations in KAT6A cause intellectual disability with recognizable syndromic features. *The American Journal of Human Genetics*, **96**(3), 507–513. doi:10.1016/j.ajhg.2015.01.016

The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56–65. doi:10.1038/nature11632

Toma, C., Torrico, B., Hervás, A., Valdés-Mas, R., Tristán-Noguero, A., Padillo, V., Maristany, M., Salgado, M., Arenas, C., Puente, X. S., Bayés, M., and Cormand, B. (2014). Exome sequencing in multiplex autism families suggests a major role for heterozygous truncating mutations. *Molecular Psychiatry*, **19**(7), 784–790. doi:10.1038/mp.2013.106

Turner, T. N., Sharma, K., Oh, E. C., Liu, Y. P., Collins, R. L., Sosa, M. X., Auer, D. R., Brand, H., Sanders, S. J., Moreno-De-Luca, D., Pihur, V., Plona, T., Pike, K., Soppet, D. R., Smith, M. W., Cheung, S. W., Martin, C. L., State, M. W., Talkowski, M. E., Cook, E., Huganir, R., Katsanis, N., and Chakravarti, A. (2015). Loss of δ-catenin function in severe autism. *Nature*, **520**(7545), 51–56. doi:10.1038/nature14186

Tuz, K., Bachmann-Gagescu, R., O'Day, D. R., Hua, K., Isabella, C. R., Phelps, I. G., Stolarski, A. E., O'Roak, B. J., Dempsey, J. C., Lourenco, C., Alswaid, A., Bönnemann, C. G., Medne, L., Nampoothiri, S., Stark, Z., Leventer, R. J., Topçu, M., Cansu, A., Jagadeesh, S., Done, S., Ishak, G. E., Glass, I. A., Shendure, J., Neuhauss, S. C. F., Haldeman-Englert, C. R., Doherty, D., and Ferland, R. J. (2014). Mutations in CSPP1 cause primary cilia abnormalities and Joubert syndrome with or without Jeune asphyxiating thoracic dystrophy. *American Journal of Human Genetics*, **94**(1), 62–72. doi:10.1016/j.ajhg.2013.11.019

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C. K., Stephens, P., Davies, H., Jones, D., Lin, M.-L., Teague, J., Bignell, G., Butler, A., Cho, J., Dalgliesh, G. L., Galappaththige, D., Greenman, C., Hardy, C., Jia, M., Latimer, C., Lau, K. W., Marshall, J., McLaren, S., Menzies, A., Mudie, L., Stebbings, L., Largaespada, D. A., Wessels, L. F. A., Richard, S., et al. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, **469**(7331), 539–542. doi:10.1038/nature09639

Veltman, J. A., and Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics*, **13**(8), 565–575. doi:10.1038/nrg3241

Vieland, V. J., Hallmayer, J., Huang, Y., Pagnamenta, A. T., Pinto, D., Khan, H., Monaco, A. P., Paterson, A. D., Scherer, S. W., Sutcliffe, J. S., Szatmari, P., and (agp), T. A. G. P. (2011). Novel method for combined linkage and genome-wide association analysis finds evidence of distinct genetic architecture for two subtypes of autism. *Journal of Neurodevelopmental Disorders*, **3**(2), 113–123. doi:10.1007/s11689-011-9072-9

Vorstman, J. a. S., Staal, W. G., van Daalen, E., van Engeland, H., Hochstenbach, P. F. R., and Franke, L. (2005). Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism. *Molecular Psychiatry*, **11**(1), 18–28. doi:10.1038/sj.mp.4001757

Vulto-van Silfhout, A. T., Rajamanickam, S., Jensik, P. J., Vergult, S., de Rocker, N., Newhall, K. J., Raghavan, R., Reardon, S. N., Jarrett, K., McIntyre, T., Bulinski, J., Ownby, S. L., Huggenvik, J. I., McKnight, G. S., Rose, G. M., Cai, X., Willaert, A., Zweier, C., Endele, S., de Ligt, J., van Bon, B. W. M., Lugtenberg, D., de Vries, P. F., Veltman, J. A., van Bokhoven, H., Brunner, H. G., Rauch, A., de Brouwer, A. P. M., et al. (2014). Mutations affecting the SAND domain of DEAF1 cause intellectual disability with severe speech impairment and behavioral problems. *The American Journal of Human Genetics*, **94**(5), 649–661. doi:10.1016/j.ajhg.2014.03.013

Walsh, T., Shahin, H., Elkan-Miller, T., Lee, M. K., Thornton, A. M., Roeb, W., Abu Rayyan, A., Loulus, S., Avraham, K. B., King, M.-C., and Kanaan, M. (2010). Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *The American Journal of Human Genetics*, **87**(1), 90–94. doi:10.1016/j.ajhg.2010.05.010

Wang, J., and Shen, Y. (2014). When a “disease-causing mutation” is not a pathogenic variant. *Clinical Chemistry*, **60**(5), 711–713. doi:10.1373/clinchem.2013.215947

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, **38**(16), e164–e164. doi:10.1093/nar/gkq603

Wang, S.-K., Choi, M., Richardson, A. S., Reid, B. M., Lin, B. P., Wang, S. J., Kim, J.-W., Simmer, J. P., and Hu, J. C.-C. (2014). ITGB6 loss-of-function mutations cause autosomal recessive amelogenesis imperfecta. *Human Molecular Genetics*, **23**(8), 2157–2163. doi:10.1093/hmg/ddt611

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., and Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science*, aac7041. doi:10.1126/science.aac7041

Willer, T., Lee, H., Lommel, M., Yoshida-Moriguchi, T., de Bernabe, D. B. V., Venzke, D., Cirak, S., Schachter, H., Vajsar, J., Voit, T., Muntoni, F., Loder, A. S., Dobyns, W. B., Winder, T. L., Strahl, S., Mathews, K. D., Nelson, S. F., Moore, S. A., and Campbell, K. P. (2012). ISPD loss-of-function mutations disrupt dystroglycan O-mannosylation and cause Walker-Warburg syndrome. *Nature Genetics*, **44**(5), 575–580. doi:10.1038/ng.2252

WJ Conover. (1980). *Practical Nonparametric Statistics* (2nd ed.). New York: John Wiley. http://as.wiley.com/WileyCDA/WileyTitle/productCd-0471160687.html. Accessed 7 December 2015

Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L. B., and Reese, M. G. (2011). A probabilistic disease-gene finder for personal genomes. *Genome Research*, **21**(9), 1529–1542. doi:10.1101/gr.123158.111

Yu, T. W., Chahrour, M. H., Coulter, M. E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., Schmitz-Abe, K., Harmin, D. A., Adli, M., Malik, A. N., D'Gama, A. M., Lim, E. T., Sanders, S. J., Mochida, G. H., Partlow, J. N., Sunu, C. M., Felie, J. M., Rodriguez, J., Nasir, R. H., Ware, J., Joseph, R. M., Hill, R. S., Kwan, B. Y., Al-Saffar, M., Mukaddes, N. M., Hashmi, A., Balkhy, S., Gascon, G. G., et al. (2013). Using whole-exome sequencing to identify inherited causes of autism. *Neuron*, **77**(2), 259–273. doi:10.1016/j.neuron.2012.11.002

Yuen, R. K. C., Thiruvahindrapuram, B., Merico, D., Walker, S., Tammimies, K., Hoang, N., Chrysler, C., Nalpathamkalam, T., Pellecchia, G., Liu, Y., Gazzellone, M. J., D'Abate, L., Deneault, E., Howe, J. L., Liu, R. S. C., Thompson, A., Zarrei, M., Uddin, M., Marshall, C. R., Ring, R. H., Zwaigenbaum, L., Ray, P. N., Weksberg, R., Carter, M. T., Fernandez, B. A., Roberts, W., Szatmari, P., and Scherer, S. W. (2015). Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature Medicine*, **21**(2), 185–191. doi:10.1038/nm.3792

Chapter 5

Conclusion

5.1 Summary

In conclusion, I have shown that by sequencing cases with bipolar disorder and ASD, two complex disorders, and analyzing findings using a model that takes into account family structure, novel candidate genes can be identified from studies even with modest sample sizes. First, I demonstrated that WES in a small number of BD families was successful in identifying multiple rare, deleterious variants in genes consistent with a plausible biological role. I identified 14 rare and likely damaging mutations that segregated with the disorder. To note, a patient that also had a seizure disorder carried a mutation in the gene *PRICKLE1*, a known gene for progressive myoclonic epilepsy 1B. Also, multiple genes containing rare, protein-altering variants had GTPase-activating function. GTPases are a target of lithium, a drug frequently used to treat BD, and have been suggested to play a role in BD (Akula *et al.*, 2014; Lachman and Papolos, 1989). Although individual gene findings did not meet statistical thresholds for significance, further research may provide further support as larger BD cohorts are sequenced.

In the third chapter, I presented findings from a WGS study of 188 individuals, which included 71 individuals affected with ASD. Consistent with previous studies, our results highlight the extreme genetic heterogeneity in ASD, and although we identified few rare LOF or *de novo* variants in known ASD genes, enriched gene pathways included cell adhesion, cell-cell signaling and nervous system development, which provides support consistent with previous findings. Using SORVA, I also identified several genes significantly enriched for rare missense or LOF variants in our dataset, given the number of individuals in the general population with rare variants in these genes. The first, *STAU2*, plays a role during both the early differentiation of neurons and in the synaptic plasticity of mature neurons (Heraud-Farlow and Kiebler, 2014) and is a promising candidate for follow-up studies. The second, *PPFIA3*, is a member of the LAR protein-tyrosine phosphatase-interacting protein (liprin) family, and liprins are known to be important for axon guidance (Spangler and Hoogenraad, 2007).

It is important to highlight the need to provide additional support for NGS findings when functional validation is not feasible due to resources and ethical or technical difficulties. Otherwise, genes that are frequently mutated or studied tend to be the genes most often highlighted by NGS studies (Shyr *et al.*, 2014), leading to future studies repeatedly highlighting the same genes in candidate gene lists. To address this issue, in the fourth chapter I presented a precomputed dataset of mutational burden in all genes and known protein domains, derived from the 1000 Genomes Project dataset (The 1000 Genomes Project Consortium, 2015). This dataset, named SORVA, can be useful for ranking variants in genes and known protein domains. In addition, I proposed applications in predictive genomics, i.e. calculating the probability that an individual with a rare variant in a known disease gene will have a rare genetic disorder. To full realize SORVA's clinical utility, the underlying dataset can be recalculated using larger reference datasets such as GnomAD for future updates (Lek *et al.*, 2016). Furthermore, as the genetic basis underlying Mendelian and complex diseases continue to be revealed, the utility of these methods in predictive genomics will increase, as well.

5.2 Recommendations

In the past decade, next generation sequencing studies have become popular, and with ever-larger studies underway, the pace of sequencing does not appear to be slowing down. For example, one effort to decipher ASD genetics, Simons Foundation Powering Autism Research for Knowledge (SPARK), aims to build a cohort of 50,000 individuals with ASD over the next 3 years. Another effort led by Ambry Genetics, AmbryShare, strives to enroll and sequence 10,000 patients with ASD. As sample sizes has grown by orders of magnitude, recruitment methods have changed, as well, and large studies including SPARK and IAN Genetics use web-based recruitment and collect phenotypic data reported by patients as opposed to clinicians (Lee *et al.* 2010). Large WES studies stemming from efforts such as these have and will continue to provide value due to their unbiased approach towards disease gene discovery. However, future studies would benefit from integrating findings with epigenetic findings and gene-environment interactions, and following up results with functional validation to test the biological impact of identified variants (Chahrour *et al.*, 2016). At the same time, studies across the wide range of human diseases will benefit from increases in the size of public reference datasets to determine which variants are ultra-rare in different populations, and from more complete protein-protein interaction networks and pathways to reveal how disparate genetic findings converge and cause shared phenotypes between families. Finally, while functional studies are often lacking, *in silico* data from previously published datasets must be used to provide support for NGS findings, whether for specific genes as shown for autism in the third chapter, or for identifying pathways involved in diseases such as bipolar disorder, as shown in the second chapter. In conclusion, the methods and findings in this thesis are impactful and can be useful for motivating analysis and interpretations of future sequencing datasets in complex neuropsychiatric diseases.

5.3 Bibliography

Akula, N., Barb, J., Jiang, X., Wendland, J. R., Choi, K. H., Sen, S. K., Hou, L., Chen, D. T. W., Laje, G., Johnson, K., Lipska, B. K., Kleinman, J. E., Corrada-Bravo, H., Detera-Wadleigh, S., Munson, P. J., and McMahon, F. J. (2014). RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder. *Molecular Psychiatry*, **19**(11), 1179–1185. doi:10.1038/mp.2013.170

Chahrour, M., O'Roak, B. J., Santini, E., Samaco, R. C., Kleiman, R. J., and Manzini, M. C. (2016). Current Perspectives in Autism Spectrum Disorder: From Genes to Therapy. *The Journal of Neuroscience*, **36**(45), 11402–11410. doi:10.1523/JNEUROSCI.2335-16.2016

Heraud-Farlow, J. E., and Kiebler, M. A. (2014). The multifunctional Staufen proteins: conserved roles from neurogenesis to synaptic plasticity. *Trends in Neurosciences*, **37**(9), 470–479. doi:10.1016/j.tins.2014.05.009

Lachman, H. M., and Papolos, D. F. (1989). Abnormal signal transduction: A hypothetical model for bipolar affective disorder. *Life Sciences*, **45**(16), 1413–1426. doi:10.1016/0024-3205(89)90031-3

Lee, H., Marvin, A. R., Watson, T., Piggot, J., Law, J. K., Law, P. A., et al. (2010). Accuracy of phenotyping of autistic children based on internet implemented parent report. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *153B*(6), 1119–1126. doi:10.1002/ajmg.b.31103

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**(7616), 285–291. doi:10.1038/nature19057

Shyr, C., Tarailo-Graovac, M., Gottlieb, M., Lee, J. J., Karnebeek, C. van, and Wasserman, W. W. (2014). FLAGS, frequently mutated genes in public exomes. *BMC Medical Genomics*, **7**(1), 64. doi:10.1186/s12920-014-0064-y

Spangler, S. A., and Hoogenraad, C. C. (2007). Liprin-alpha proteins: scaffold molecules for synapse maturation. *Biochemical Society transactions*, **35**(Pt 5), 1278–82. doi:10.1042/BST0351278

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74. doi:10.1038/nature15393