DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

The Contribution of Rare De Novo and Inherited Coding Variants in Neurodevelopmental Disorders

Citation
Kosmicki, Jack. 2019. The Contribution of Rare De Novo and Inherited Coding Variants in Neurodevelopmental Disorders. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029831

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

The contribution of rare *de novo* and inherited coding variants in neurodevelopmental disorders

A dissertation presented

by

Jack Alphonse Kosmicki

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biomedical informatics

Harvard University

Cambridge, Massachusetts

May 2019

© 2019 Jack Alphonse Kosmicki

All rights reserved.

The contribution of rare *de novo* and inherited coding variants in neurodevelopmental disorders

Abstract

High-throughput sequencing technologies allowed for studying rare (allele frequency <5%) genetic variation previously inaccessible through genotyping arrays used in genome-wide association studies. While each rare variant explains a negligible amount of heritability, they can potentially better identify trait-associated genes. In this dissertation, we identified rare *de novo* and inherited coding variation via exome sequencing to discover genes in complex neurodevelopmental traits.

First, we found that ~1/3 previously identified *de novo* variants were present as standing variation in the Exome Aggregation Consortium's cohort of 60,706 adults; these recurrent *de novo* variants in aggregate did not contribute to risk for neurodevelopmental disorders. We further used a loss-of-function (LoF)-intolerance metric, pLI, to identify a subset of LoF-intolerant genes that contained the observed signal of associated *de novo* protein truncating variants (PTVs) in neurodevelopmental disorders. LoF-intolerant genes also carried a modest excess of inherited PTVs; though the strongest *de novo* impacted genes contributed little to this, suggesting the excess of inherited risk resides lower-penetrant genes.

Working with the Autism Sequencing consortium, we analyzed rare *de novo* and inherited variants from the largest exome sequencing study of autism spectrum disorders (ASDs) to date (35,584 samples) to discover 26 Bonferroni significant genes and upwards of 102 genes (FDR<0.1). Comparing the frequency of deleterious *de novo* variants in ascertained ASD and ascertained intellectual disability / developmental disorders (ID/DD) samples, half of the identified genes conferred more risk to ID/DD than ASD and these two groups of genes have different phenotypic outcomes.

Finally, aggregating genetic and phenotypic data for ID/DD, ASD, and congenital heart disease (CHD) individuals, we evaluated the effect of severe ID/DD on both *de novo* variant frequencies and gene discovery in ASD and CHD. Within ascertained ID/DD, comorbid ASD and CHD does not affect either the *de novo* variant frequency or the number of significant genes, but the converse was not true: ID/DD increased both the *de novo* variant frequency and the number of Bonferroni significant genes discovered in ASD and CHD.

Table of Contents

Acknowledgements

It's impossible to convey how thankful I am that Mark Daly took me on as a graduate student. While he might not agree, he took a big risk in mentoring me, especially after I struggled in both rounds of preliminary qualifying exam. Despite coming into graduate school without any knowledge or experience in the two core foundations of his lab, human genetics and statistics, Mark still took the necessary time and effort to teach me the fundamental concepts, answer my endless stream of questions (e.g., when would you use a chi-square test instead of Fisher's exact test?), provide me with useful reference materials (*The Lady Tasting Tea* by David Salsburg was one of the best), and instill a strong sense of statistical rigor that helped cover lost ground. Ultimately, all the analyses in this dissertation would not be possible without Mark's patience, guidance, enthusiasm, and support over these past seven years.

I was also extremely privileged to have so many other mentors within and outside of ATGU. I first met Dennis Wall as a summer intern at Harvard and spent two summers working in his lab. He was instrumental in helping me throughout the graduate school application process, introduced me to Mark, taught me to see the big picture, and provided invaluable advice and perspective throughout graduate school. Elise Robinson provided additional support and guidance, especially early on while she was a post-doc with Mark, and continued to do so after becoming an instructor and later an assistant professor at Harvard School of Public Health. Ben Neale was always an amazing font of knowledge about statistics and human genetics. Daniel MacArthur, who also served on both my PQE and DAC, created both ExAC and gnomAD that enabled a lot of my research to be possible. It has been a pleasure to watch all four advance in their respective scientific careers and see their labs grow and mature.

Beyond the incredible amount of data and guidance that Mark provided me, he also built a collaborative, friendly environment full of amazing people in ATGU. I was privileged to share office space with a number of young talented people: Kaitlin Samocha, Nikita Artomov, Sherif Gerges, Jackie Goldstein, and Henrike Heyne. In particular, Kaitlin Samocha, my "big sister", has become a fantastic friend; it was a privilege to learn from and work with her. I could not have asked for a better graduate student role model than her. I have to thank F. Kyle Satterstrom for being a fantastic collaborator as we worked with the Autism Sequencing Consortium. The vast majority of people in ATGU are post-docs so it has been wonderful to share the experience with the few graduate students in the lab: Beryl Cummings and later, Sherif Gerges and Masahiro Kanai. Sherif has been a fantastic addition to the lab and has become a wonderful friend; I know he has a bright future ahead of him. A special thanks to our admins, Jill Doucette, Beth Raynard, and Carla Hammond, who made sure everything in the lab ran smoothly and in a timely fashion.

Graduate school has been a very long and winding road full of challenges, successes, and despairs. But it becomes much easier with a lot of help from friends. In particular, I must thank my roommate (or flat-mate as the Europeans in lab prefer to say) throughout the entirety of graduate school, David Fronk, for always being there through thick and thin as we traveled through graduate school together. I appreciated Joseph Timpona's cheerful and optimistic attitude and companionship in seeing the new Star Wars movies on opening night. My best friends from college, Johnny and Laura Conner, made a huge effort to write me letters and postcards throughout these seven years; opening up my mailbox to find a letter from these two kept me going and lifted my spirits after a long day. Caitlin Nichols has provided me with countless laughs and stories and was a great collaborator as well. Lastly, thanks to the other four

BIG PhD students in my cohort: Jeremiah Wala, Dustin Griesemer, Kamil Slowikowski, and Yu-Han Hsu for making the first two years of classes and the rest of graduate school little bit more bearable while our program learned the ropes.

Finally, I would not be here today without the support from my family. My parents, Susan and Peter Kosmicki, encouraged and supported me to pursue my passion, and I am so thankful for their ability and wisdom to put events and situations in perspective, especially during the most difficult times. I'm so thankful to my two siblings, Elizabeth and Joseph, who were always available to chat. A sincere thank you to my family for all their help throughout this entire journey; I couldn't have done it without you.

Chapter 1

Introduction

Portions of this chapter was previously published as:

Kosmicki, J.A. *et al*. Discovery of rare variants for complex phenotypes. *Hum Genet* **135** 625-634 (2016).

Overview

The goal of statistical and medical genetics is to understand the genetic basis of human traits. However, this lofty goal is complicated by the fact that not all traits have the same genetic architecture (i.e., the number of genetic loci, effect size and minor allele frequency [MAF] distribution), as well as the fact that the environment also contributes to many such traits[1]. As such, traits are placed in groups based on similar genetic architectures. Mendelian (named after the Austrian monk, Gregor Mendel) traits are governed by a single locus with large effects. On the opposite end of spectrum are complex traits (e.g., height, type II diabetes, schizophrenia) for which a large number of distinct genetic loci influence the phenotypic variability.

Restriction fragment length polymorphism genotyping coupled with linkage mapping was one of the earliest methods to identify trait-associated genetic loci[2,3]. Linkage mapping required collecting large pedigrees with both affected and unaffected members. Genetic loci were genotyped in all members in the pedigree(s), and researchers followed the segregation of these loci with the segregation of the trait under the assumption that the trait-associated loci would follow the trait's inheritance pattern. As such, linkage mapping worked very well for traits caused by large effect variants such as the trinucleotide repeat $(CAG)_n$ in *HTT*[4] located on 4p16.3[5] that causes Huntington's Disease as well as ΔF508 in *CFTR* in cystic fibrosis[6].

While linkage mapping successfully identified the genetic components of Mendelian traits, it was by and large unsuccessful for complex traits[7]. For complex traits, genome-wide association studies (GWAS) have become the standard approach to understand their genetic architecture. In 2005, the first successful GWAS in a complex trait, age-related macular degeneration, identified two single-nucleotide polymorphisms (SNPs) in the intron of *CFH* that surpassed the study-wide significant threshold of 4.8×10^{-7} with 96 cases and 50 controls[8]. As

with many early successful studies of their kind, not all subsequent GWAS would be as successful (especially at similar sample sizes in the hundreds). To illustrate, one of the earliest schizophrenia GWAS in 2009 with 3,322 cases and 3,587 controls didn't find any genome-wide significant loci[9], but a later study in 2011 identified five genome-wide significant loci with 9,394 cases and 12,462 controls[10]. The largest study in 2014 identified 108 loci with 36,989 cases and 113,075 controls[11]. With the continued drop in genotyping costs, meta-analysis of GWAS have reached hundreds of thousands of samples enabling sufficient power to detect small effects at common single nucleotide variants (i.e., those with a minor allele frequency [MAF] ≥ 5%). These hypothesis-free genome-wide scans delivered many novel discoveries, including some particularly unexpected results such as implicating the hippocampus and limbic system in body mass index[12], autophagy in Crohn's Disease[13], and the complement system in age-related macular degeneration[14]. To date, GWAS have been used to study over 3,200 traits such as post-traumatic stress disorder[15], coffee consumption[16], hoarding[17], and type II diabetes[18]. To date, the catalogue of genome-wide significant associations contains over 126,000 variants[19].

Rare variants

Current genotyping arrays commonly used in GWAS capture most common variants through imputation, but have limited capture of variants below a 5% MAF. With decreased costs, and development of high throughput next generation sequencing (NGS) technologies to sequence both the exome (~1% of the genome that covers protein coding genes)[20] and the entire genome, researchers could for the first time identify all variation in the genome, but most notably rare variants and perform rare variant association studies (RVAS). With time and larger sample sizes, the MAF definition of rare variation has shifted from less than 5% for the earliest GWAS

to 0.5% or even 0.1% - with the term, low-frequency variant, being applied to the MAF range of 5% - 0.5%[21]. Furthermore, population based whole genome sequencing studies (WGS) such as the 1000 Genomes[22] and UK10K Project[23], verified that most genetic variation is rare. What's more, at current sample sizes the majority of variants are singletons, meaning that only one copy of the minor allele is observed in the entire sample (**Figure 1.1**). Beyond capturing SNPs, NGS technologies also capture insertions/deletions (indels) of nucleotides, as well as more complicated structural variation such as copy-number variants (CNVs) and large-scale inversions or deletions. Current sequencing technologies capture almost all SNPs, but accurate detection of indels and structural variants still poses a challenge.

Whole exome sequencing (WES) and WGS identified many causal genes for Mendelian and monogenic disorders[24-33]. Part of the initial motivation for looking at rare variants in complex traits came from targeted candidate gene studies that discovered rare coding variants of large effects. For example, rare coding variants in *NOD2* were linked to risk of Crohn's Disease[34], and rare variants in *PCSK9* and *ABCA1* were found to have large effects on low-density lipoprotein (LDL)-cholesterol and high-density lipoprotein (HDL)-levels respectively[35-37]. Furthermore, successfully translating the discovery of *PCSK9* to a therapeutic intervention has demonstrated the potential of taking rare variant association through to clinical application[38,39].

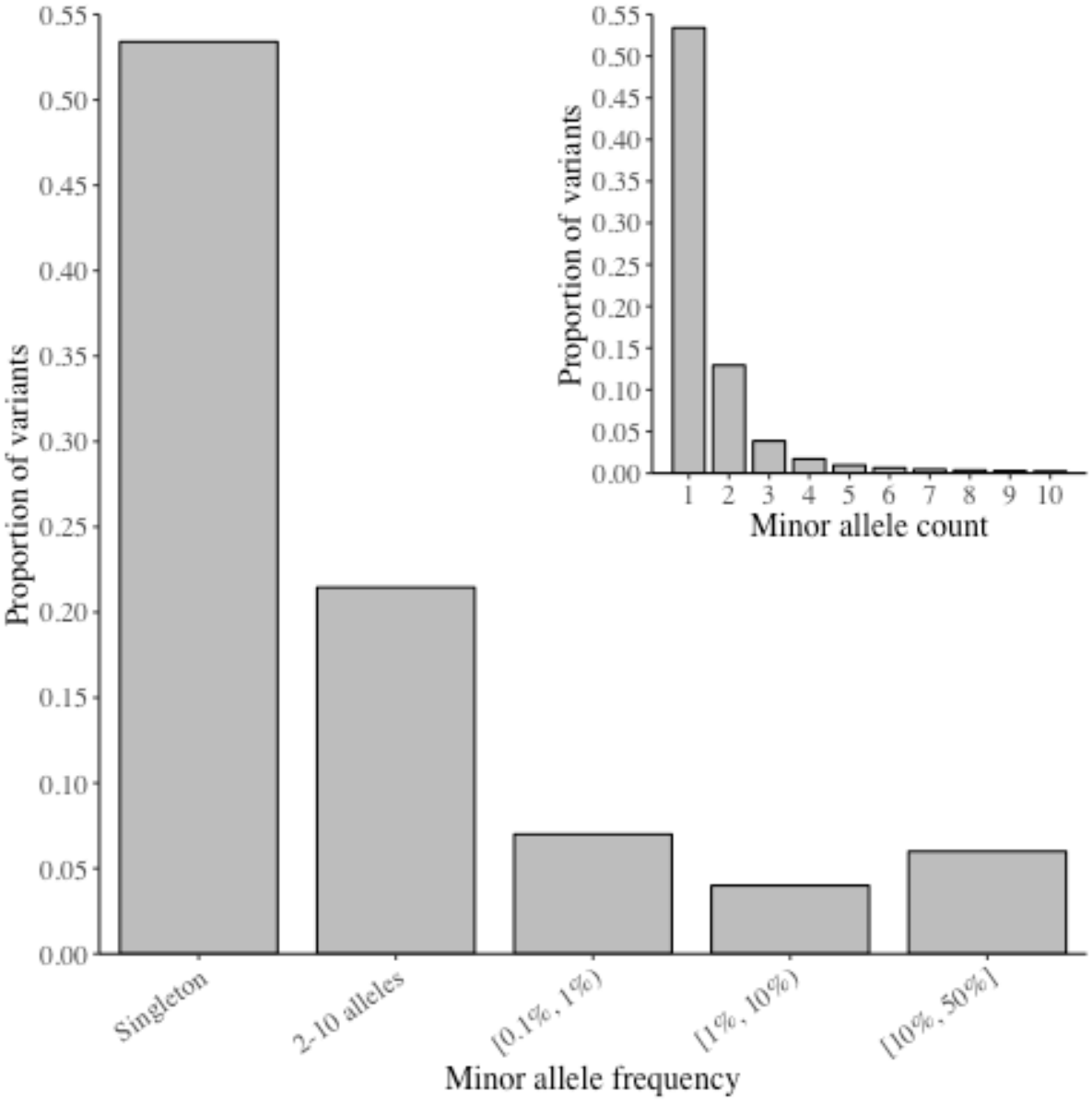


Figure 1.1: Minor allele frequency distribution from exome sequencing of 2883 individuals of Swedish ancestry. The vast majority of variants are rare (MAF < 0.1%) with 53% observed only once (singleton). The inset figure expands out the fraction of variants observed at minor allele counts 1-10.

While WGS is a powerful approach that enables the unbiased survey of genetic variants genome-wide, it has two main limitations. First, the costs of sequencing are still considerable, resulting in smaller samples for any one study. Second, as described above, interpreting the functional consequences of non-coding variants remains an ongoing challenge. Nevertheless, as

costs continue to decline and technologies improve, WGS will likely be the standard approach for genetic investigation. However, the single most important factor in driving discovery in genetic studies is sample size, meaning that more cost-effective approaches for large samples may successfully identify significant loci more rapidly.

In contrast to WGS, WES targets the capture of the protein coding regions (~1% of the genome). While WES is more expensive than genotyping arrays, it remains considerably less expensive than WGS. This cost-reduction enables larger sample sizes and therefore higher-powered studies. Prior research suggests the exome will have more rare variants than the non-coding region, because the coding region has a 10-fold greater selection coefficient than evolutionarily conserved non-coding regions[40,41], and those variants are more directly actionable[42]. This makes the discovery of associations from WES more likely to inform our understanding of the pathology of disease as well as increase the likelihood of identifying viable therapeutic targets. Furthermore, our ability to interpret the functional impact of coding variants far outstrips our understanding of noncoding variation, meaning that extracting biological insight is much more straightforward (although not without its own challenges). All together, these properties of the coding region increase power to identify novel associations as well as provide a better interpretation of those associations. Nevertheless, WGS projects may likely have a longer shelf life than WES projects.

Burden and variance components tests

For individual rare variants, not enough copies of the minor allele are present to achieve sufficient levels of evidence to be convincingly associated in single marker analysis[43]. To address this issue, grouping and burden tests have long been proposed in the analysis of rare

variants[37,44-48]. These groupings aim to ensure that there are enough individuals carrying a rare variant to perform an association test. There are two main classes of group-wise tests: burden tests, where the rare variants in a region are assumed to have the same direction of effect and variance component tests which allow for effects in opposite directions.

The most straightforward of the gene-based tests[44,45,49-51], burden tests function by comparing the number or *burden* (i.e., sum) of variants in cases and controls. These tests collapse variants within a gene or a defined region of the genome into a single score and test for association between the score and the trait of interest. One can simply consider all variants in the pre-specified grouping and apply either a threshold (0 or 1) or a weight based on their functional category and/or MAF in the model. However, burden tests are limited by the assumption that all variants act in the same direction (i.e., all risk or all protective) and as such, lose power if there is a mixture of both protective and risk conferring variants in the same gene.

Variance-component tests[52,53], most notably the sequence based kernel association (SKAT)[54] or C-alpha[46] (which is a special case of SKAT), were designed to address this issue in which a gene may possess a mixture of risk and protective variants. They test for association by evaluating whether individuals that carry the same rare variant tend to be more similar phenotypically. By assessing the *distribution* of variants, rather than their combined additive effect, these tests are robust to instances where the rare variants affect phenotype in different directions[55]. Thus, variance-component tests are more powerful than burden tests if there is a mixture of both risk and protective variation. However, variance component tests lose power compared to burden tests when the majority of variants are in the same direction.

<u>Which region to test</u>

One of the central questions in RVAS, especially for WGS, is what regional definitions should be used to group rare variants in an association-testing framework. The most common, and arguably most intuitive, choice is to aggregate variants across a gene. This is particularly appealing in exome sequencing studies where genetic variation is being captured specifically at genes. This gene-based approach can be expanded to include particular functional classes (e.g., DNase hypersensitivity sites, nonsense variants), all genes within a pathway, or all genes within a gene set. In the context of WGS however, the majority of rare variants fall outside of genes, and the decision of which regions to group them over for testing becomes less clear. In this case, one could group variants by class of regulatory element such as promoter, enhancer, or transcription factor binding site. One challenge with grouping in this manner is that regulatory elements tend to be small (100-200bp) and thus require more samples to achieve the same power as when testing a whole gene[42]. Another way to consider aggregating rare variants, especially in the case of the noncoding region, is to use a sliding window of a specified genomic length[56]. However, determining the optimal size for a sliding window is tricky, as there is a tradeoff between using a few large windows which incurs a smaller multiple hypothesis testing burden, but comes at the cost of including variants that might be functionally unimportant or have negligible effect sizes to using a lot of small windows with a higher multiple testing burden. The UK10K study applied this technique with a window size of 3kb to test 31 different traits for noncoding associations, but this analysis did not return any significant associations[23].

Once a specified region is chosen, one must determine which variants within that region to include in the analysis. Each individual variant will either increase the probability of having the disease (risk-conferring), or decrease it (protective), or have no effect on risk (neutral). Ideally, we would only include the risk-conferring variants, or alternatively the protective

variants, since including neutral variants will reduce power. However, this information is typically not known, so the challenge is to balance the chance of including the risk-conferring (or protective) variants and excluding neutral variants.

Gene level testing

When considering gene level analyses, one of the most natural approaches is to restrict to only variants predicted to truncate the protein[57] or ablate it through nonsense-mediated decay[58]. Four different functional categories fit in this group: frameshift, splice donor, splice acceptor, and nonsense variants. Collectively, these variants are referred to by a variety of descriptions: loss-of-function (LoF), likely gene disrupting (LGD), or protein truncating variants (PTVs[58]); we will use the term PTV for the remainder of this dissertation. One of the most attractive features of PTVs is the expectation that all the variants will act in the same direction. However, most genes in the genome are strongly conserved, meaning that natural selection keeps PTVs rare, and thus large sample sizes are necessary to observe a sufficient number of rare alleles to test for association with the trait of interest.

One possible way to increase power without increasing sample size is to also include missense variants. However, the classification of missense variants into risk, neutral, and protective is challenging. A variety of different computational approaches for pathogenicity prediction of missense mutations have been proposed, such as SIFT[59], PolyPhen2[60], MutationTaster[61], MPC[62], among others[63,64]. Each of these tools leverages different indicators of deleteriousness for missense mutations; some measure conservation (e.g., GERP++[65], SIFT[59], phyloP[66]), while others evaluate the functional effect of alternate amino acids on protein structure (PolyPhen2[60]). Given the differences in information source, the predictions of

deleteriousness often differ. Additionally, the various datasets used for training and testing these tools differ in how they define pathogenic or neutral variants, which further contributes to the inconsistency across tools[64]. Regardless of the particular annotation method adopted, the resulting set of variants will likely contain a mixture of both risk and neutral variants.

Population stratification

For case-control and cohort association studies, population stratification is a major source of type I error[67-69]; principal components analysis (PCA) and linear mixed models (LMMs) have been applied with great success in correcting for these confounders[70]. PCA-based correction assumes a smooth distribution of MAF over ancestry or geographical space, which is appropriate in the space of common variation. However, this approach may not be appropriate for rare variation as the MAFs may be sharply localized and geographically clustered due to the fact that they have recently arisen, thus violating this assumption[71]. One proposed method to correct for stratification in RVAS is Fast-LLM-Select[72], which performs feature selection on the variants, retaining only those that are phenotypically informative to use in constructing the generalized relationship matrix (GRM). Nevertheless, Fast-LLM-Select loses power when causal variants are geographically clustered[72,73].

Family studies

In 1987, Falk and Rubinstein proposed a study design using trios (father, mother, and child) as a way to control for population substructure and admixture[74]. Family-based studies avoid the problems of population stratification because the child is perfectly controlled by their parents. Additionally, they enable the interrogation of both inherited and *de novo* variation.

Their primary disadvantage is that trio-based studies are harder to recruit for, as they require all three family members to obtain a single data point. As such, issues of non-paternity are prohibitively expensive as the data from the remaining members of the family is of reduced value. For family studies, two main analytic approaches are available: within family tests, (e.g., the transmission disequilibrium test [TDT]) and *de novo* (i.e., newly arising mutations).

TDT

The most commonly used association test in family designs[75] is the transmission disequilibrium test (TDT)[76]. The TDT can be thought of as a family-based case-control association procedure, in which the control is not a random unaffected individual but the alleles the affected child could have inherited but did not (i.e. a pseudo-control). The TDT boils down to testing whether the frequency of transmitted alleles (case) is the same as alleles not transmitted to the affected child (control) from a heterozygous parent and uses McNemar's chi-squared test statistic[77] to determine *P*-values. Because the TDT relies on the variant allele having a 50% chance of being transmitted or untransmitted, parents who are homozygous variant are not used as the transmission is guaranteed.

Arguably the greatest advantage of the TDT is that it is free from population stratification as the control (i.e. the untransmitted allele) is sampled from within the same family as the case. The TDT assumes Mendelian inheritance (i.e. that each allele is equally likely to be transmitted), and that a variant more often transmitted than not to the affected offspring indicates a disease-associated locus that is linked with the marker. Thus, both linkage and association are required to reject the null hypothesis; this dual hypothesis shields the TDT from population stratification. A recent study by Elansary and colleagues found that the TDT can produce false positive

associations with X-linked variants near the pseudo-autosomal region for traits with sex-limited expression and when the allele frequencies of the locus differs between the X and Y chromosomes. These false positive associations arise because transmission is not equally likely in both sexes: fathers transmit the Y allele to their sons and the X allele to their daughters. These false positives can be fixed by considering only maternal transmissions and removing trios in which the father and mother are both heterozygous at these sites[78].

De novo tests

The scenario where studying *de novo* mutations for gene discovery is most effective is when the selective pressure against mutations is extremely strong and the effect size is quite large. Strong selective pressure means that when deleterious mutations arise, they are rapidly removed from the population, keeping the frequency of those mutations in the population extremely low. For instance, Hutchinson-Gilford progeria syndrome is a rare genetic disorder (incidence of ~1 in 4 million[79]) marked by accelerated aging, scleroderma, and hair loss with an average lifespan of 13 years[79]. As these affected individuals do not live long enough to reproduce, the disorder is most commonly caused by *de novo* missense variants in *LMNA*, some of which create a cryptic splice site that leads to a truncated protein[80]. Beyond childhood lethal disorders such as Hutchinson-Gilford progeria syndrome, *de novo* variants were successfully used to identify the causal genes in Mendelian disorders such as achondroplasia[81], Bohring-Opitz syndrome[82], Kabuki Syndrome[27], KGB syndrome[83], Miller syndrome[28], and Schinzel-Giedion Syndrome[84].

Germline *de novo* mutations originate during DNA replication in both mitosis and the first half of meiosis. Due to differences in the male and female germline, *de novo* mutations are

more often paternal in origin[85]. While oocytes are created once and very early in a women's life with a fixed 23 genome replications, spermatogonial stem cells are replicated every year after puberty throughout a man's life[86]. Thus, the germ line of a 20-year-old male has undergone ~160 genome replications, rising to ~610 genome replications by the time the male reaches 40-years-old[87-89]. The fact that *de novo* mutations accumulate in the male germline as men age results in an increased risk of bearing children with genetic disorders caused by such mutations. In 1912, W. Weinberg observed that sporadic cases of achondroplasia occurred more often in the last-born child[90] and J.B.S. Haldane discovered in 1947 that the hemophilia-associate gene's mutation rate was higher in men[91]. Furthermore, risk for achondroplasia[92] and other disorders[87] were noted to increase with paternal age.

The key to analyzing *de novo* variation is to understand the mutability of each potential mutation site in the genome. Across the genome, the mutation rate has been show to vary as a function of a large number of factors including replication timing[93-95], nucleosome position[96], local base context[97,98], and other large-scale phenomena[99]. While the chance of mutation at any one gene is extremely rare (typically 2×10^{-4}), we are all expected to carry ~75-100 *de novo* variants on average[100,101]. In order to have sufficient power to test such variants for association without knowledge of whether the variant is *de novo* or without an expectation of how many *de novo* variants would be observed by chance, very large sample sizes would be required. To illustrate, ~100,000 samples are required to detect a gene in which *de novo* PTVs confer a 20-fold increase in risk[42]. Building a mutation rate model for *de novo* variant analysis dramatically improves the gene discovery power because one can compare the number of observed *de novo* variants to what would be expected if *de novo* variants were randomly distributed across the genome[98]. Prior to the work of the Samocha and colleagues[98], studies of *de novo* variation

assumed the presence of at least 2 deleterious (i.e., missense, PTV) coding *de novo* variants were sufficient for association[102-106]. With a null mutation model for each gene and each variant class, studies could statistically evaluate their findings (**Table 1.1)**. One will notice that while most studies followed the tried-and-true example laid down by GWAS of using a strict Bonferroni significance threshold, some (particularly studies of ASD) opt for a more permissive false discovery rate (FDR) with a varying (and arbitrary) cutoff to report more genes despite less confidence in each individual association.

Table 1.1: Genes discovered in complex traits via *de novo* variation. Sample size indicates the number of trios (mother, father, child). Abbreviations: ASD: autism spectrum disorder, CHD: congenital heart disease, DD: developmental delay, FDR: false discovery rate, ID: intellectual disability, NDD: neurodevelopmental disorders, O/E: observed vs. expected, TADA: transmission and *de novo* association

Paper	Phenotype	Sample size	Number of genes	Method	Significance
Samocha 2014[98]	ASD	1078	3	O/E[98]	Bonferroni ($P<10^{-6}$)
De Rubeis 2014[107]	ASD	2270	33	TADA[108]	FDR < 0.1
De Rubeis 2014[107]	ASD	2270	107	TADA[108]	FDR < 0.3
Sanders 2015[109]	ASD	3981	65	TADA[108]	FDR < 0.1
DDD 2015[110]	DD	1133	21	O/E[98]	Bonferroni ($P<10^{-6}$)
DDD 2015[110]	DD-meta	3477	31	O/E[98]	Bonferroni ($P<10^{-6}$)
Homsy 2015[111]	CHD	1213	3	O/E[98]	Bonferroni ($P<10^{-7}$)
Sifrim 2016[112]	Severe CHD	1365	11	O/E[98]	Bonferroni ($P<10^{-6}$)
Lelieveld 2016[113]	ID/DD	2104	10	O/E[98]	FDR < 0.05
DDD 2017[114]	DD-meta	7580	93	O/E[98]	Bonferroni ($P<10^{-7}$)
Kosmicki 2017[115]	ASD	3981	7	O/E[98]	Bonferroni ($P<10^{-6}$)
Yuen 2017[116]	ASD	5326	54	O/E[98]	FDR < 0.15
Jin 2017[11]	CHD	2645	7	O/E[98]	Bonferroni ($P<10^{-6}$)
Heyne 2018[117]	NDD w/epilepsy	1942	33	O/E[98]	Bonferroni ($P<10^{-6}$)
Satterstrom 2018[118]	ASD	6430	102	TADA[108]	FDR < 0.1

Neurodevelopmental disorders

Neurodevelopmental disorders encompass a broad swath of disorders including, but not limited to: autism spectrum disorder (ASD), intellectual disability (ID), developmental delay disorders (DD), epilepsy, attention deficit hyperactivity disorder, tic disorders including

Tourette's syndrome, and motor disorders[119]. Many neurodevelopmental disorders occur early in childhood, have reduced fecundity[120], and are highly heritable[121]. In the next two sections, we will describe two neurodevelopmental disorders that we focused on in this dissertation: autism spectrum disorder (ASD) and intellectual disability (ID).

Genetics of autism spectrum disorders

First described by Leo Kanner in 1943[122], ASD is a phenotypically heterogeneous disorder that currently encompasses a range of disorders that are characterized by impairments in two core domains: (1) communication and social interaction and (2) restricted interests and repetitive behaviors[123]. The clinical and diagnostic features of ASD has expanded over the years[124] and with that change came increases in the prevalence of ASD, now at 1:59. With the changing diagnostic criteria came changes in the comorbidity space in which ASD was most often diagnosed in individuals with comorbid ID/DD, but now this same segment comprises ~30% of the ASD population. ASD has a strong male to female bias with on average four males to every one female, but this ratio changes based on the IQ of the individual. For the most severe ASD individuals, the male to female ratio drops to 2:1, while those on the high end of the IQ distribution reaches 7:1[125-128].

One way to measure selection against a trait is to study the fecundity or number of offspring individuals with a given trait bare. Traits under positive selection will have higher than average fecundity, while traits under negative selection will have lower than average fecundity and the magnitude of the difference is related to the strength of selection (*s*). In a Swedish birth cohort, Power and colleagues measured the fecundity of individuals with a variety of different psychiatric traits, one of which was ASD. They found that both males and females with ASD

had fewer children compared to their unaffected relatives and the fecundity of males and females were not the same with males having 75% fewer children and females having 48% fewer[129]. Because of the reduced fecundity in individuals with ASD, negative selection will keep the MAF of genetic variants with large effect very low. However, because ASD is a phenotypically heterogeneous disorder, it may very well be that the fecundity also differs based on the severity of the phenotype as individuals with severe to profound ID rarely produce offspring[130].

Early twin studies demonstrated that ASD is a highly heritable trait with heritability estimates as high as 80%[121,131,132], indicating a large genetic component. However, identifying the genetic basis of ASD proved challenging. Early ASD linkage association studies failed to identify large swaths of associated loci, largely due to ASD's polygenic genetic architecture[133], but managed to locate the causal genes for genetic syndromes (of which ASD was one, among many, phenotypic outcomes) including Rett syndrome[134], tuberous sclerosis[135], and fragile X syndrome[136]. Outside of two loci that managed to replicate across studies, *NRXN1*[137] and *SHANK3*[138], most linkage studies produced lengthy lists of candidate genes for researchers to compile into "high-confidence" gene sets[139] that never replicated[98]. In a similar fashion, multiple underpowered ASD GWAS either failed to identify genome-wide significant loci or found loci that also never replicated[140-142].

The advent and adoption of CNV genotyping arrays allowed for the discovery of recurrent, large-scale *de novo* CNVs. Beginning in 2009, Jonathan Sebat and colleagues performed a genome-wide survey of CNVs in 264 families and observed a 10-fold excess of *de novo* CNVs in simplex families of ASD (meaning 1 affected child with ASD and no other affected family members) and a 3-fold excess in multiplex ASD families (multiple children with ASD)[143]. While they were underpowered to identify any specific CNVs, they demonstrated that

de novo CNVs contributed to ASD risk. This initial discovery was replicated in larger samples while also identifying specific loci including 1q21.1[144], 7q11.23[144,145], 15q11.2-13.1[144-147], 15q13.2-13.3[144], 16p11.2[144-147], and 22q11.2[144,146,147]. Although these large CNVs provided insight into underlying trait biology and genetic architecture, CNVs are often incompletely penetrant, rarely implicate a single gene, and confer risk to multiple traits[143]. Given that many of these structural variants were *de novo* in origin, multiple groups, including ours, hypothesized that like *de novo* CNVs, *de novo* coding single nucleotide variants (SNVs) would also contribute to risk – with the added benefit of implicating specific genes. It was fortuitous that sequencing entire families to identify *de novo* SNVs in an unbiased, genome-wide fashion was now feasible with the advent of whole exome sequencing[20]. Given that *de novo* CNVs were mores strongly enriched in simplex ASD families, some groups specifically targeted simplex ASD families for their trio-based exome sequencing studies[148] – although we later found no difference in the frequency of *de novo* SNVs between simplex and multiplex ASD families (Chapter 4). As we discussed earlier, these trio-based exome sequencing studies[102,105,149,150] were enormously successful at identifying genes with sample sizes as small as 175 trios, virtually unheard of for complex traits in the era of GWAS and case-control RVAS. Thus, these studies ushered in a new approach for ASD genetics.

The genetics of intellectual disability

The second neurodevelopmental disorder we examined in this dissertation is intellectual disability (ID), an early-onset disorder with a worldwide prevalence of ~1%[151] characterized by significant deficiencies in adaptive behavior and cognitive functioning before 18-years-of-age. ID is formally defined as an intellectual quotient (IQ) < 70[123], and the severity of ID varies based

on IQ and is split into bins denoted as mild, moderate, severe, and profound. The majority of individuals with ID are identified early in childhood because of observed developmental delays in crawling, sitting, walking, and speaking[130]. Both genetic and environmental influences contribute to ID; potential environmental risk factors include birth complications, lack of oxygen, severe malnutrition, infections, and maternal alcohol consumption during pregnancy[130]. ID occurs both by itself (referred to as isolated-ID) as well as in conjunction with many other disorders such as congenital heart disease, ASD, developmental disorders, epilepsy, and neuromuscular deficits (e.g. sensory/motor neuropathy, ataxia, muscular dystrophy, spastic paraplegia). The fact that ID is often comorbid with other disorders can create issues with genetic studies focusing primarily on ID as well as studies of comorbid disorders[117]. The frequency of sporadic ID cases is positively correlated with increasing severity, suggesting *de novo* or recessive contributions on top of the overall inherited genetic liability. As such, most rare variant association studies of ID tend to focus on the severe and profound cases[104,113,152-154].

In 1959, the first genetic association for ID[155] was discovered: trisomy 21 - Down syndrome, which is currently the most common cause of ID comprising 15% of cases[156]. Ten years later, Lubs and colleagues discovered the next genetic association for ID with a marker on the X-chromosome for fragile X syndrome[157], and it took another 22 years before the causal gene (*FMR1*) was discovered[158]. Fragile X syndrome currently accounts for 0.5% of ID cases[159]. With these two discoveries coupled with the cytogenic banding technologies, large-scale chromosomal aberrations were discovered and altogether comprise another 15% of ID cases[156].

Beginning in the 1990s, the X-chromosome became the focus of ID studies due to a combination of factors: the causal gene for fragile X syndrome resided on the X-chromosome[157], the hypothesis that the X-chromosome was partially responsible for the elevated frequency in ID

among males[160], and lastly, larger (and therefore, more powered) linkage studies could be performed with affected males[160]. This exploration reached its zenith in 2009 when Tarpey and colleagues performed targeted sequencing of all the exons of the X-chromosome and reported nine novel X-linked ID genes (albeit without rigorous statistical evidence)[161]. There are now over 100 X-linked ID genes and while no individual gene explains even 0.1% of ID, collectively they account for 10% of ID in males[162].

Homozygosity mapping using SNP microarrays with Sanger sequencing of candidate genes for follow-up studies were used to discovery over 300 recessive genes on the autosomes[130]. However, ~97% of these recessive genes were pleiotropic in nature with ID as one of the many phenotypic outcomes. Only a handful of recessive genes solely cause ID[163]. As with ASD, the introduction of microarrays enabled genome-wide identification of ID-associated CNVs[164-167] with higher resolution than was previously possible with the light microscope[168]. Roughly 10% of these CNVs were *de novo* in origin and the number of genes affected by CNVs was positively correlated with increasing severity of ID[169].

Energized by the discovery of *de novo* CNVs, the development of WES, and prior successful efforts in identifying causal genes in rare syndromes via WES, trio-based WES in severe ID as well as severe developmental disorders (of which ID was in nearly all of the individuals)[110] was carried out. As with ASD, these studies were enormously successful and future studies are reaching more than 32,000 trios (unpublished).

Summary

When I began my PhD, the first trio-based exome sequencing studies of neurodevelopmental disorders had just been published and the promising early results suggested

larger sample sizes could both identify more genes and provide deeper biological insights into these disorders. Furthermore, rare variant association studies were still in their infancy and the contribution of rare variants to complex traits was very much unclear. In this dissertation, we sought to explore the contribution of rare *de novo* and inherited coding variation in neurodevelopmental disorders and use this genetic variation to identify neurodevelopmental disorder associated genes.

In chapter 2, we investigated the role of recurrent mutations in the ExAC database using published *de novo* variants in ASD, ID/DD, congenital heart disease, and schizophrenia. We observed that ~1/3 de novo variants were present as standing variation in 60,706 individuals in ExAC and that these *de novo* variants were not associated with neurodevelopmental risk. At the gene level, we applied a recently developed constraint method, pLI, to identify genes intolerant to PTVs and these highly constrained genes contained the previously observed enrichment. Using ExAC as a variant-level filter and pLI as the gene-level filter, we observed for the first time a significant enrichment in both inherited and case-control PTVs; as expected, this enrichment was not as strong as *de novo* PTVs.

In chapter 3, we co-led the largest exome sequencing study of ASD to date, with more than 32,000 samples from 31 sampling sources. Using a Bayesian framework that incorporated both *de novo* and case-control variation and leveraged gene and regional constraint, we discovered 26 Bonferroni significant genes and 102 genes (FDR<0.1) associated with ASD. Meta-analyzing our results with published *de novo* variants from 5264 intellectual disability / developmental disorder (ID/DD) trios indicated that 49 of the 102 ASD-associated were more strongly associated with ID/DD than ASD, as evidenced by a higher frequency of *de novo* variants in ascertained ID/DD individuals than ASD individuals. We further demonstrated that

these 49 ID/DD-preferential genes were markedly different from the ASD-preferential genes in terms of the degree of negative selection and phenotypic presentation.

Lastly in chapter 4, we delved into the genetic architecture of ASD, ID/DD, and congenital heart disease to evaluate how the comorbidity landscape of each disorder influenced the frequency of *de novo* coding SNVs and each ascertained trait's power for gene discovery. In contrast to previous studies, we failed to observe any evidence of an oligogenic model for ASD via *de novo* variants and also failed to observe any difference in the frequency of *de novo* SNVs between simplex and multiplex ASD families. Furthermore, given that many genes are identified across each of these separate ascertainments, we statistically evaluated the degree of phenotypic specificity for each of these genes.

References

1. Lakhani, C.M. *et al.* Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nature Genetics* **51**, 327-334 (2019).

2. Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**, 314-31 (1980).

3. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* **11**, 241-7 (1995).

4. MacDonald, M.E. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971-983 (1993).

5. Gusella, J.F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234-8 (1983).

6. Riordan, J.R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066-73 (1989).

7. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881-8 (2008).

8. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-9 (2005).

9. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-52 (2009).

10. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43**, 969--976 (2011).

11. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).

12. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).

13. Rioux, J.D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* **39**, 596-604 (2007).

14. Edwards, A.O. *et al.* Complement Factor H Polymorphism and Age-Related Macular Degeneration. *Science* **308**, 421-424 (2005).

15. Ashley-Koch, A.E. *et al.* Genome-wide association study of posttraumatic stress disorder in a cohort of Iraq-Afghanistan era veterans. *J Affect Disord* **184**, 225-34 (2015).

16. Pirastu, N. *et al.* Non-additive genome-wide association scan reveals a new gene associated with habitual coffee consumption. *Scientific Reports* **6**, 31590 (2016).

17. Perroud, N. *et al.* Genome-wide association study of hoarding traits. *Am J Med Genet B Neuropsychiatr Genet* **156**, 240-2 (2011).

18. Replication, D.I.G. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234-244 (2014).

19. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).

20. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272--276 (2009).

21. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186 (2017).

22. Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).

23. The, U.K.K.C. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).

24. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 19096--19101 (2009).

25. Bolze, A. *et al.* Whole-exome-sequencing-based discovery of human FADD deficiency. *Am J Hum Genet* **87**, 873-81 (2010).

26. Byun, M. *et al.* Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. *J Exp Med* **207**, 2307-12 (2010).

27. Ng, S.B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42**, 790-3 (2010).

28. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* **42**, 30--35 (2010).

29. Teer, J.K. & Mullikin, J.C. Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics* **19**, R145-R151 (2010).

30. Walsh, T. *et al.* Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *Am J Hum Genet* **87**, 90-4 (2010).

31. Wang, J.L. *et al.* TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain* **133**, 3510-8 (2010).

32. Puente, X.S. *et al.* Exome sequencing and functional analysis identifies BANF1 mutation as the cause of a hereditary progeroid syndrome. *Am J Hum Genet* **88**, 650-6 (2011).

33. Simpson, M.A. *et al.* Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. *Nat Genet* **43**, 303-5 (2011).

34. Rivas, M.A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* **43**, 1066-73 (2011).

35. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* **37**, 161-5 (2005).

36. Cohen, J.C., Boerwinkle, E., Mosley, T.H., Jr. & Hobbs, H.H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354**, 1264-72 (2006).

37. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869-72 (2004).

38. Roth, E.M., McKenney, J.M., Hanotin, C., Asset, G. & Stein, E.A. Atorvastatin with or without an Antibody to PCSK9 in Primary Hypercholesterolemia. *New England Journal of Medicine* **367**, 1891-1900 (2012).

39. Stein, E.A. *et al.* Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N Engl J Med* **366**, 1108-18 (2012).

40. Chen, C.T., Wang, J.C. & Cohen, B.A. The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet* **80**, 692-704 (2007).

41. Kryukov, G.V., Schmidt, S. & Sunyaev, S. Small fitness effect of mutations in highly conserved non-coding regions. *Hum Mol Genet* **14**, 2221-9 (2005).

42. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-64 (2014).

43. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).

44. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311-21 (2008).

45. Madsen, B.E. & Browning, S.R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet* **5**, e1000384 (2009).

46. Neale, B.M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet* **7**, e1001322 (2011).

47. Neale, B.M. & Sham, P.C. The future of association studies: gene-based analysis and replication. *Am J Hum Genet* **75**, 353-62 (2004).

48. Terwilliger, J.D. & Ott, J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered* **42**, 337-46 (1992).

49. Asimit, J.L., Day-Williams, A.G., Morris, A.P. & Zeggini, E. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Hum Hered* **73**, 84-94 (2012).

50. Morgenthaler, S. & Thilly, W.G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **615**, 28-56 (2007).

51. Morris, A.P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* **34**, 188-93 (2010).

52. Auer, P.L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* **7**, 16 (2015).

53. Bansal, V., Libiger, O., Torkamani, A. & Schork, N.J. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* **11**, 773-85 (2010).

54. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).

55. Moutsianas, L. *et al.* The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLoS Genet* **11**, e1005165 (2015).

56. Psaty, B.M. *et al.* Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet* **2**, 73-80 (2009).

57. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).

58. Rivas, M.A. *et al.* Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666-9 (2015).

59. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res* **11**, 863-74 (2001).

60. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).

61. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-2 (2014).

62. Samocha, K.E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* (2017).

63. Sunyaev, S.R. Inferring causality and functional significance of human coding DNA variants. *Hum Mol Genet* **21**, R10-7 (2012).

64. Grimm, D.G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat* **36**, 513-23 (2015).

65. Davydov, E.V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).

66. Cooper, G.M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**, 628-640 (2011).

67. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037-48 (1994).

68. Pritchard, J.K. & Donnelly, P. Case-control studies of association in structured or admixed populations. *Theor Popul Biol* **60**, 227-37 (2001).

69. Knowler, W.C., Williams, R.C., Pettitt, D.J. & Steinberg, A.G. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* **43**, 520-6 (1988).

70. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).

71. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* **44**, 243-6 (2012).

72. Listgarten, J., Lippert, C. & Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nat Genet* **45**, 470-1 (2013).

73. Mathieson, I. & McVean, G. Reply to: "FaST-LMM-Select for addressing confounding from spatial structure and rare variants". *Nat Genet* **45**, 471-471 (2013).

74. Falk, C.T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* **51**, 227-33 (1987).

75. He, Z. *et al.* Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am J Hum Genet* **94**, 33-46 (2014).

76. Spielman, R.S., McGinnis, R.E. & Ewens, W.J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**, 506-516 (1993).

77. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153-157 (1947).

78. Elansary, M. *et al.* On the use of the transmission disequilibrium test to detect pseudo-autosomal variants affecting traits with sex-limited expression. *Anim Genet* **46**, 395-402 (2015).

79. Hennekam, R.C.M. Hutchinson–Gilford progeria syndrome: Review of the phenotype. *American Journal of Medical Genetics Part A* **140A**, 2603-2624 (2006).

80. Eriksson, M. *et al.* Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* **423**, 293-8 (2003).

81. Bellus, G.A. *et al.* Achondroplasia is defined by recurrent G380R mutations of FGFR3. *Am J Hum Genet* **56**, 368-73 (1995).

82. Hoischen, A. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat Genet* **43**, 729-31 (2011).

83. Sirmaci, A. *et al.* Mutations in ANKRD11 cause KBG syndrome, characterized by intellectual disability, skeletal malformations, and macrodontia. *American journal of human genetics* **89**, 289--294 (2011).

84. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet* **42**, 483-5 (2010).

85. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471--475 (2012).

86. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat Genet* **48**, 126-133 (2016).

87. Risch, N., Reich, E.W., Wishnick, M.M. & McCarthy, J.G. Spontaneous mutation and parental age in humans. *American journal of human genetics* **41**, 218-248 (1987).

88. Vogel, F. & Rathenberg, R. Spontaneous mutation in man. *Adv Hum Genet* **5**, 223-318 (1975).

89. Crow, J.F. The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* **1**, 40-47 (2000).

90. Weinberg, W. Zur Vererbung des Zwergwuchses. *Arch. Rassen-u. Gesel. Biolog.* **9**, 710-718 (1912).

91. Haldane, J.B.S. The Mutation Rate of the Gene for Haemophilia, and its Segregation Ratios in Males and Females. *Annals of Eugenics* **13**, 262-271 (1947).

92. Penrose, L.S. Parental age and mutation. *Lancet* **269**, 312-3 (1955).

93. Hardison, R.C. *et al.* Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* **13**, 13-26 (2003).

94. Hellmann, I. *et al.* Why do human diversity levels vary at a megabase scale? *Genome Res* **15**, 1222-31 (2005).

95. Lercher, M.J. & Hurst, L.D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18**, 337-40 (2002).

96. Pich, O. *et al.* Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell* **175**, 1074-1087.e18 (2018).

97. Coulondre, C., Miller, J.H., Farabaugh, P.J. & Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274**, 775-780 (1978).

98. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950 (2014).

99. Ellegren, H., Smith, N.G. & Webster, M.T. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev* **13**, 562-8 (2003).

100. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**, 712-4 (2011).

101. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Human Mutation* **21**, 12-27 (2003).

102. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-246 (2012).

103. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).

104. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674-82 (2012).

105. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).

106. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-3 (2013).

107. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).

108. He, X. *et al.* Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet* **9**, e1003671 (2013).

109. Sanders, S.J. *et al.* Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215-33 (2015).

110. Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).

111. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-1266 (2015).

112. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat Genet* **48**, 1060-1065 (2016).

113. Lelieveld, S.H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nature Neuroscience* **19**, 1194-1196 (2016).

114. Deciphering Developmental Disorders, S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438 (2017).

115. Kosmicki, J.A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat Genet* **49**, 504-510 (2017).

116. Yuen, R.K. *et al.* Genome-wide characteristics of de novo mutations in autism. *NPJ Genom Med* **1**, 160271-1602710 (2016).

117. Heyne, H.O. *et al.* De novo variants in neurodevelopmental disorders with epilepsy. *Nature Genetics* **50**, 1048-1053 (2018).

118. Satterstrom, F.K. *et al.* Novel genes for autism implicate both excitatory and inhibitory cell lineages in risk. *bioRxiv*, 484113 (2018).

119. Hu, W.F., Chahrour, M.H. & Walsh, C.A. The Diverse Genetic Landscape of Neurodevelopmental Disorders. *Annual Review of Genomics and Human Genetics* **15**, 195-213 (2014).

120. Power, R.A., Kyaga, S., Uher, R. & et al. FEcundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**, 22-30 (2013).

121. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* **46**, 881-885 (2014).

122. Kanner, L. Autistic disturbances of affective contact. *Nervous Child* **2**, 217-250 (1943).

123. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*, (American Psychiatric Publishing, Washington, DC, 2013).

124. Berg, J.M. & Geschwind, D.H. Autism genetics: searching for specificity and convergence. *Genome Biology* **13**, 247 (2012).

125. Baio, J. *et al.* Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *MMWR. Surveillance Summaries* **67**, 1--23 (2018).

126. Baron-Cohen, S. *et al.* Why are autism spectrum conditions more prevalent in males? *PLoS Biol* **9**, e1001081 (2011).

127. Developmental, D.M.N.S.Y. & Investigators, P. Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002)* **63**, 1 (2014).

128. Kirkovski, M., Enticott, P.G. & Fitzgerald, P.B. A review of the role of female gender in autism spectrum disorders. *Journal of Autism and Developmental Disorders* **43**, 2584--2603 (2013).

129. Power, R.A. *et al.* Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**, 22-30 (2013).

130. Vissers, L.E., Gilissen, C. & Veltman, J.A. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet* **17**, 9-18 (2016).

131. Ronald, A. & Hoekstra, R.A. Autism spectrum disorders and autistic traits: A decade of new twin studies. *Am J Med Genet B Neuropsychiatr Genet* **156B**, 255--274 (2011).

132. Steffenburg, S. *et al.* A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. *Journal of child psychology and psychiatry, and allied disciplines* **30**, 405--416 (1989).

133. Risch, N. *et al.* A genomic screen of autism: evidence for a multilocus etiology. *American journal of human genetics* **65**, 493--507 (1999).

134. Amir, R.E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185-8 (1999).

135. Hunt, A. & Shepherd, C. A prevalence study of autism in tuberous sclerosis. *J Autism Dev Disord* **23**, 323-39 (1993).

136. Brown, W.T. *et al.* Association of fragile X syndrome with autism. *Lancet* **1**, 100 (1982).

137. Kim, H.G. *et al.* Disruption of neurexin 1 associated with autism spectrum disorder. *Am J Hum Genet* **82**, 199--207 (2008).

138. Durand, C.M. *et al.* Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nature genetics* **39**, 25--27 (2007).

139. Betancur, C. Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Brain Research* **1380**, 42--77 (2011).

140. Ma, D. *et al.* A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann Hum Genet* **73**, 263--273 (2009).

141. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528--533 (2009).

142. Weiss, L. & Arking, D.a. A genome-wide linkage and association scan reveal novel loci for autism. *Nature* **461**, 802--808 (2009).

143. Sebat, J., Levy, D.L. & McCarthy, S.E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet* **25**, 528--535 (2009).

144. Sanders, Stephan J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams Syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011).

145. Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886-897 (2011).

146. Marshall, C.R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82**, 477-488 (2008).

147. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-372 (2010).

148. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192-195 (2010).

149. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).

150. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246--250 (2012).

151. Maulik, P.K., Mascarenhas, M.N., Mathers, C.D., Dua, T. & Saxena, S. Prevalence of intellectual disability: a meta-analysis of population-based studies. *Res Dev Disabil* **32**, 419-36 (2011).

152. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* **367**, 1921-9 (2012).

153. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344-7 (2014).

154. Hamdan, F.F. *et al.* De novo mutations in moderate or severe intellectual disability. *PLoS Genet* **10**, e1004772 (2014).

155. Lejeune, J., Gauthier, M. & Turpin, R. Les chromosomes humains en culture de tissus. Vol. 248 602--603 (1959).

156. van Karnebeek, C.D., Jansweijer, M.C., Leenders, A.G., Offringa, M. & Hennekam, R.C. Diagnostic investigations in individuals with mental retardation: a systematic literature review of their usefulness. *Eur J Hum Genet* **13**, 6-25 (2005).

157. Lubs, H.A. A marker X chromosome. *American journal of human genetics* **21**, 231-244 (1969).

158. Verkerk, A.J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905--914 (1991).

159. Coffee, B. *et al.* Incidence of fragile X syndrome by newborn screening for methylated FMR1 DNA. *Am J Hum Genet* **85**, 503--514 (2009).

160. Leonard, H. & Wen, X. The epidemiology of mental retardation: challenges and opportunities in the new millennium. *Ment Retard Dev Disabil Res Rev* **8**, 117-34 (2002).

161. Tarpey, P.S. *et al.* A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet* **41**, 535--543 (2009).

162. Lubs, H.A., Stevenson, R.E. & Schwartz, C.E. Fragile X and X-linked intellectual disability: four decades of discovery. *Am J Hum Genet* **90**, 579-90 (2012).

163. Miller, D.T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* **86**, 749-64 (2010).

164. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nature Genetics* **43**, 838 (2011).

165. de Vries, B.B.A. *et al.* Diagnostic genome profiling in mental retardation. *American journal of human genetics* **77**, 606-616 (2005).

166. Shaw-Smith, C. *et al.* Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J Med Genet* **41**, 241-8 (2004).

167. Wagenstaller, J. *et al.* Copy-number variations measured by single-nucleotide-polymorphism oligonucleotide arrays in patients with mental retardation. *American journal of human genetics* **81**, 768-779 (2007).

168. Albertson, D.G. & Pinkel, D. Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet* **12 Spec No 2**, R145-52 (2003).

169. Girirajan, S. *et al.* Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *N Engl J Med* **367**, 1321-31 (2012).

Chapter 2

Refining the role of *de novo* protein truncating variants in neurodevelopmental disorders using population reference samples

This chapter was previously published as:

Kosmicki, J.A. *et al*. Refining the role of *de novo* protein truncating variants in neurodevelopmental disorders using population reference samples. *Nat Genet* **49** 504-510 (2017).

Abstract

Recent research has uncovered a significant role for *de novo* variation in neurodevelopmental disorders. Using aggregated data from 9246 families with autism spectrum disorder, intellectual disability, or developmental delay, we show ~1/3 of *de novo* variants are independently observed as standing variation in the Exome Aggregation Consortium's cohort of 60,706 adults, and these *de novo* variants do not contribute to neurodevelopmental risk. We further use a loss-of-function (LoF)-intolerance metric, pLI, to identify a subset of LoF-intolerant genes that contain the observed signal of associated *de novo* protein truncating variants (PTVs) in neurodevelopmental disorders. LoF-intolerant genes also carry a modest excess of inherited PTVs; though the strongest *de novo* impacted genes contribute little to this, suggesting the excess of inherited risk resides lower-penetrant genes. These findings illustrate the importance of population-based reference cohorts for the interpretation of candidate pathogenic variants, even for analyses of complex diseases and *de novo* variation.

Introduction

Autism spectrum disorders (ASDs) are a phenotypically heterogeneous group of heritable disorders that affect ~1 in 68 individuals in the United States[1]. While estimates of the common variant (heritable) contribution toward ASD liability are upwards of 50%[2-4], few specific risk variants have been identified, in part because ASD GWAS sample sizes to date remain limited. Conversely, the field made substantial progress understanding the genetic etiology of ASD via analysis of *de novo* (newly arising) variation using exome sequencing of parent-offspring trios[5-10]. Severe intellectual disability and developmental delay (ID/DD) are considerably less heritable than ASDs[11] (though frequently comorbid) and have demonstrated a stronger

contribution from *de novo* frameshift, splice acceptor, splice donor, and nonsense variants (collectively termed protein truncating variants [PTVs])[12-14]. Furthermore, ASD cases with comorbid ID show a significantly higher rate of *de novo* PTVs than those with normal or above average IQ[6,9,15-17], while higher IQ cases have a stronger family history of neuropsychiatric disease[15], suggesting a greater heritable contribution.

De novo variants comprise a unique component of the genetic architecture of human disease since, having not yet passed through a single generation, any heterozygous variants with complete or near-complete elimination of reproductive fitness must reside almost exclusively in this category. Despite prior evidence of mutational recurrence[18] (i.e., the same mutation occurring *de novo* in multiple individuals), most studies implicitly assumed each *de novo* variant was novel, in line with Kimura's infinite sites model[19], and thereafter analyzed *de novo* variants genome-wide without respect to their allele frequency in the population. However, the mutation rate is not uniform across the genome, with some regions and sites experiencing higher mutation rates than others (e.g., CpG sites[20]). A classic example comes from achondroplasia, in which the same G-to-C and G-to-T variant at a CpG site was observed *de novo* in 150 and 3 families, respectively[18]. As such, it should not be surprising to observe a *de novo* variant at a given site and also observe the same variant (defined as one with the same chromosome, position, reference, and alternate allele) present as standing variation in the population.

Given the strong selective pressure on neurodevelopmental disorders[21-23], we expect most highly deleterious (high-risk conferring) *de novo* PTVs will linger in the population for at most a few generations. Thus, the collective frequency of such variants in the population will approximate their mutation rate. Individual PTVs tolerated to be seen in relatively healthy adults, and more generally PTVs in genes that tolerate the survival of such variants in the

population, may be less likely to contribute significant risk to such phenotypes, and are therefore permitted by natural selection to reach allele frequencies orders of magnitude larger than those of highly deleterious variants. Given the current size of the human population (~7 billion), and the expectation of one *de novo* variant per exome (1 in ~30 million bases), every non-embryonic lethal coding mutation is likely present as a *de novo* variant at least once in the human population. This reasoning, along with the availability of large exome sequencing reference databases, motivated our interest in searching for variants observed *de novo* in trio sequencing studies that are also present as standing variation in the human population, indicating a recurrent mutation. We will herein refer to these *de novo* variants that are also observed as standing variation in the population as class 2 *de novo* variants, with the remaining *de novo* variants referred to as class 1 *de novo* variants (i.e., observed only *de novo*; **Figure 2.1**).

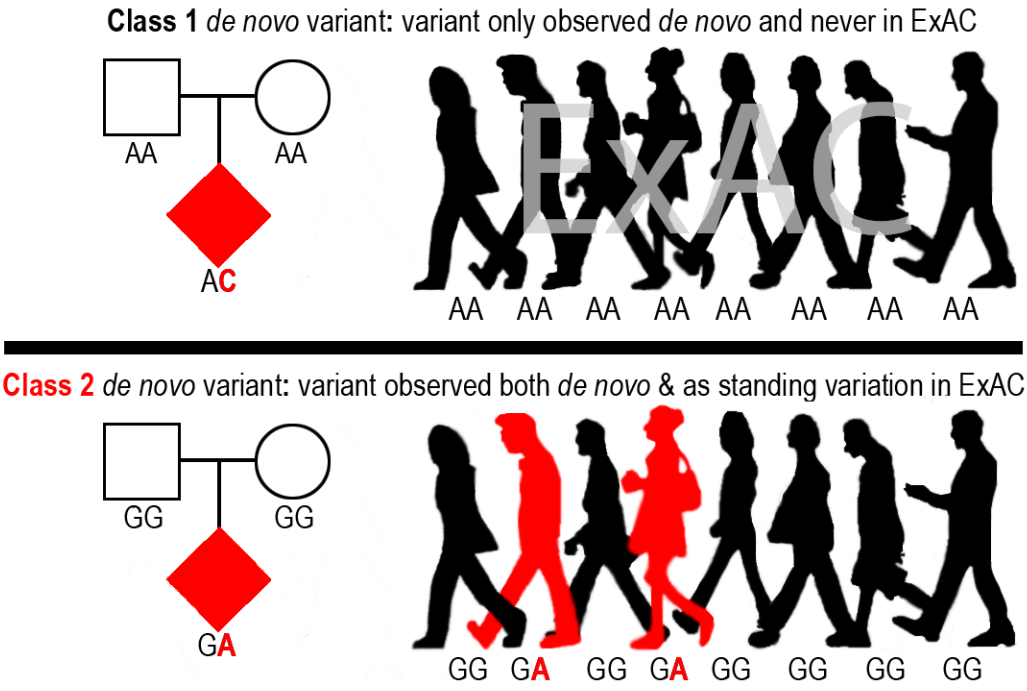


Figure 2.1: Illustration of class 1 and class 2 *de novo* variants with the genotypes of each variant for 8 of the 60,706 individuals in ExAC.

With the release of the Exome Aggregation Consortium's (ExAC) dataset of 60,706 adult individuals without severe developmental abnormalities[24], we can now empirically investigate the rate and relative pathogenicity of recurrent mutations. While there have been many studies examining *de novo* variation in human disease, the success in ASD and ID/DD, coupled with the large sample sizes published to date, led us to focus our evaluation on these phenotypes.

Results

Class 2 *de novo* variation

We first asked how many of the 10,093 variants observed *de novo* in ID/DD cases[13], ASD cases, and unaffected ASD siblings are also observed as standing variation in the 60,706 reference exomes from ExAC[24] (**Figure 2.1**). We found that 1854 ASD (31.66%), 841 unaffected ASD sibling (33.05%), and 410 ID/DD (24.23%) *de novo* variants are observed as standing variation in one or more ExAC individuals (class 2 *de novo* variants) (**Figure 2.2A**). When we removed the 15,330 exomes originating from psychiatric cohorts (many of which are controls), the rate of class 2 *de novo* variation drops to 28.47% (±1.03%, 95% CI), a rate statistically indistinguishable from the expected recurrence rate of 28.13% (±0.42%, 95% CI; binomial test P=0.45; **Figure 2.2B**). We found similar rates of class 2 *de novo* variants in published trio studies of schizophrenia[25] and congenital heart disease[26,27]. While the presence of class 2 *de novo* variants is not a novel observation[18,25], the rate is approximately three times larger than previous estimates[25] owing to significantly larger reference datasets (**Figure 2.2B**).

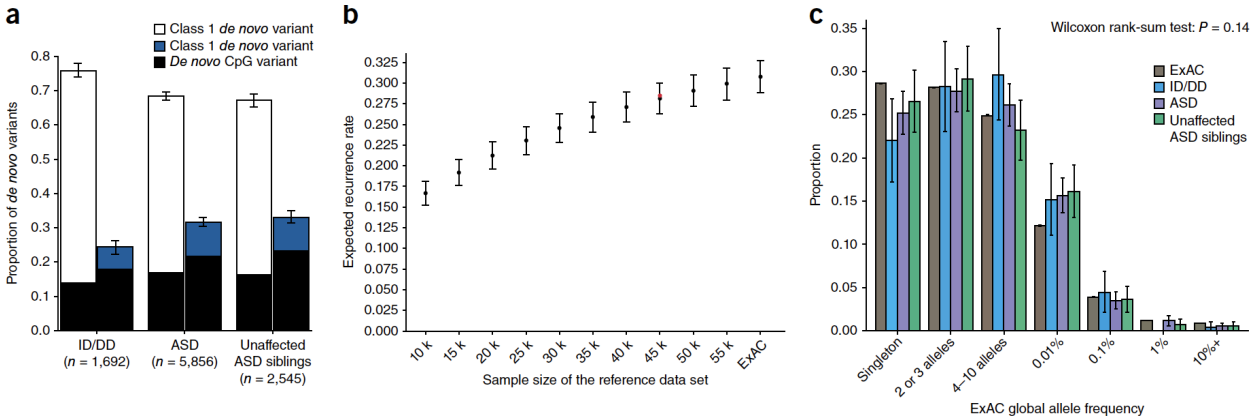


Figure 2.2: Properties of class 2 *de novo* variants. (**a**) The proportion of *de novo* variants across each cohort split between class 1 (left) and class 2 (right) with CpG variants marked in black. Class 2 *de novo* variants are strongly enriched for CpG variants ($P < 10^{-20}$). (**b**) Expected recurrence rate (rate of class 2 *de novo* variants across ID/DD, ASD, and unaffected ASD siblings) given the sample size of the reference dataset. The red dot indicates the observed recurrence rate of the non-psychiatric version of ExAC. (**c**) Allele frequency distribution of class 2 *de novo* CpGs by cohort with the matching distribution of CpG variants in ExAC for comparison. Allele frequency distributions do not significantly differ ($P = 0.14$; Wilcoxon rank sum test). Error bars represent 95% confidence intervals throughout (**a**) – (**c**).

We first sought confirmation that the observed recurrence rate – the proportion of variants observed both *de novo* and as standing variation in the population – was technically sound and not the result of some undetected contamination or missed heterozygote calls in parents (i.e., false *de novo* variants). Five secondary analyses strongly support the technical validity of this work. 1) In line with previous publications, class 2 *de novo* variants, regardless of their functional impact, are enriched at CpG sites as compared to class 1 *de novo* variants ($P < 1x10^{-20}$; Fisher's exact test). 2) As the exomes used to call *de novo* variants in De Rubeis et al. (2014) were used in the joint calling of ExAC, and many were sequenced at the same center as the majority of ExAC samples, it is possible that false class 2 *de novo* variants could be elevated in this dataset via contamination or joint calling artifacts. However, we observe no difference in the rate of class 2 *de novo* variation between De Rubeis et al. (2014) and Iossifov, O'Roak, Sanders, Ronemus et al. (2014) (P=0.10; Fisher's exact test). 3) The frequency distribution of class 2 *de novo* variants should be skewed dramatically upward towards common variation if

contamination or missed parental heterozygotes were contributing; however, the ExAC frequency of class 2 *de novo* variants at CpGs were indistinguishable from all such variants in ExAC compared to variants drawn at random with the same annotation and CpG content (P=0.14; Wilcoxon rank sum test; **Figure 2.2C**). 4) In fact, a subset of synonymous variants experimentally validated in the ASD studies showed nearly the same recurrence rate as the overall set, most definitively establishing that these mutations indeed arose independently (P=0.60; Fisher's exact test). 5) Lastly, it is well documented that mutation rate increases with paternal age, thus rates of both class 1 and class 2 *de novo* variants should show association with paternal age if both classes were genuine *de novo* variants. Indeed, for the 1861 unaffected ASD siblings with available paternal age information, rates of both class 1 and class 2 *de novo* variants are associated with increasing paternal age (class 1: β=0.002, P=4.11x10^{-9}; class 2: β=0.0009, P=1.06x10^{-5}; linear regression).

We then sought to determine whether class 1 and class 2 *de novo* variants contribute equally to ASD and ID/DD risk. As a control for the comparison of functional variants, rates of both class 1 and class 2 *de novo* synonymous variants are equivalent across ASD, ID/DD, and unaffected ASD siblings (**Figure 2.3A**) and remain unchanged when we removed the psychiatric cohorts within ExAC. Thus, collectively neither class 1 nor class 2 *de novo* synonymous variants show association with ASD or ID/DD, consistent with previous reports that as a class, *de novo* synonymous variants do not contribute to risk[5-10]. While previous reports implicated *de novo* PTVs as significant risk factors for ASD[5,6,15,16] and ID/DD[13], the class 2 *de novo* subset of PTVs show no such association for either ASD (0.015 per case vs. 0.023 per unaffected ASD sibling; P=0.98; one-sided Poisson exact test[28]) or ID/DD (0.016 per case vs. 0.023 per unaffected ASD sibling; P=0.94; one-sided Poisson exact test), with slightly higher rates in

unaffected ASD siblings (**Figure 2.3B**). By contrast, after removing class 2 *de novo* PTVs, class 1 *de novo* PTVs are significantly more enriched in individuals with ASD (0.13 per case) and ID/DD (0.19 per case) as opposed to unaffected ASD siblings (0.07 per control) (ASD vs. control: rate ratio =1.83; P=6.07x10^{-12}, ID/DD vs. control: rate ratio=2.61; P=6.31x10^{-21}; one-sided Poisson exact test). The lack of excess case burden in class 2 *de novo* variants was consistent with what would be expected if such variants did not contribute to ASD and ID/DD risk. However, to ensure we were not losing causal variants by removing all *de novo* variants found in ExAC, we tested the class 2 *de novo* PTVs at three ExAC allele frequency (AF) thresholds: singletons (1 allele in ExAC), AF < 0.0001, and AF < 0.001. We found no significant difference between the rate of class 2 *de novo* PTVs between individuals with ID/DD or ASD as compared to unaffected ASD siblings at any threshold (**Figure 2.3C**). Furthermore, these results remain consistent regardless of whether the psychiatric exomes in ExAC are retained or excluded, demonstrating they are not driving the observed associations. Thus, the data provides no evidence to suggest these class 2 *de novo* variants contribute to the previously observed enrichment of *de novo* variation in ASD and ID/DD cases, and removing those variants present in ExAC increases the effect size for *de novo* PTVs in ASD and ID/DD. Moving forward, we focus our analyses solely on variation absent from ExAC.

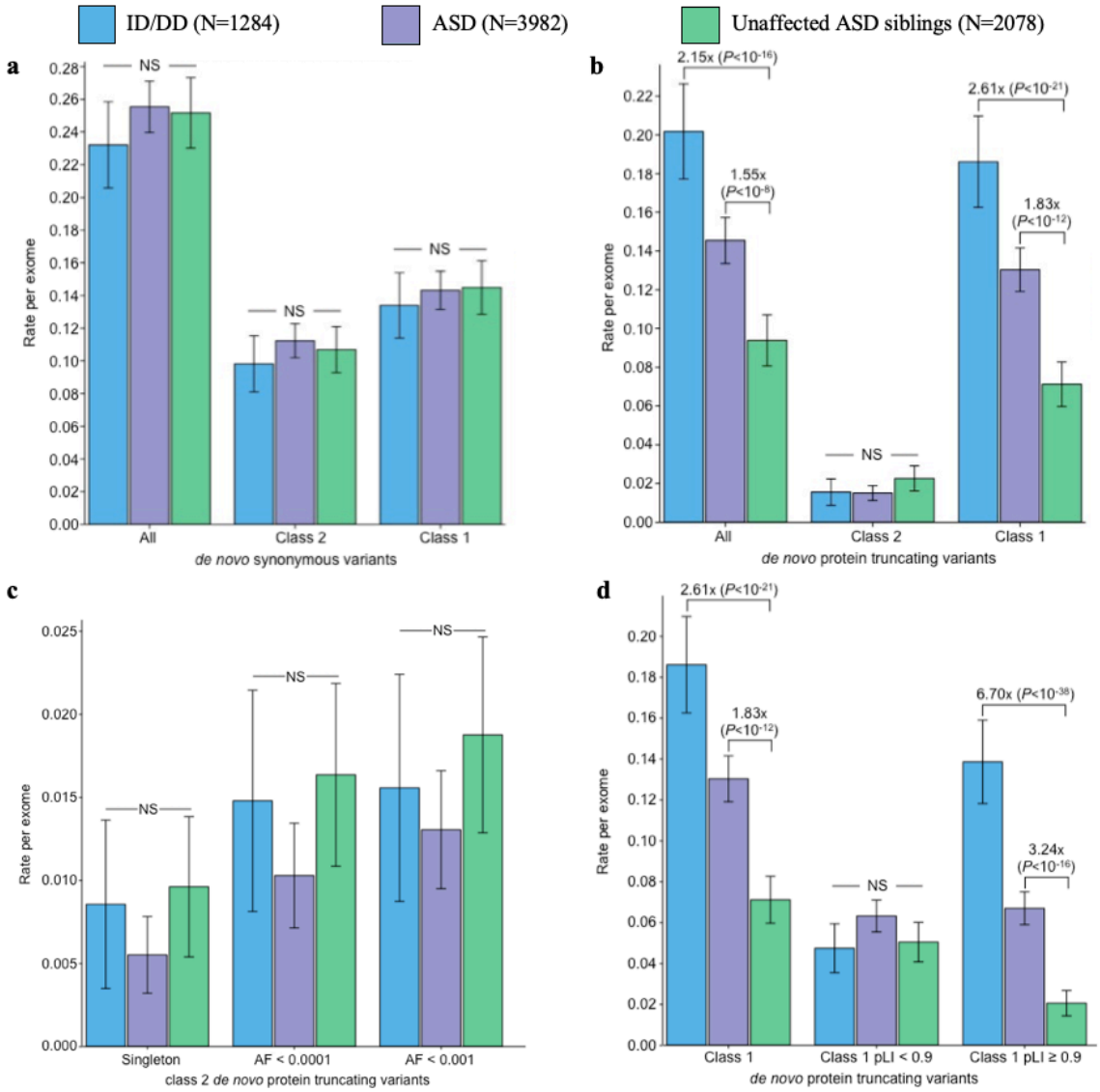


Figure 2.3: Partitioning the rate of *de novo* variants per exome by class 1, class 2, and pLI. Within each grouping, the rate – variants per individual – is shown for ID/DD (left), ASD (middle), and unaffected ASD siblings (right) with the number of individuals labeled in the legends. (**a**) Rate of *de novo* synonymous variants per exome partitioned into class 2 (middle) and class 1 (right). No significant difference was observed for any grouping of *de novo* synonymous variants. (**b**) Rate of *de novo* PTVs per exome partitioned into class 2 (middle) and class 1 (right). Only class 1 *de novo* PTVs in ID/DD and ASD show association when compared to unaffected ASD siblings. (**c**) Rate of class 2 *de novo* PTVs broken by different ExAC global allele frequency (AF) thresholds: singleton (observed once; left), AF < 0.0001 (middle), and AF < 0.001 (right). (**d**) Rate of class 1 *de novo* PTVs partitioned into class 1 *de novo* PTVs in pLI ≥ 0.9 genes (right), and class 1 *de novo* PTVs in pLI < 0.9 genes (middle). The entire observed association for *de novo* PTVs resides in class 1 *de novo* PTVs in pLI ≥ 0.9 genes. For all such analyses, the rate ratio and significance were calculated by comparing the rate for ID/DD and ASD to the rate in unaffected ASD siblings using a two-sided, two-sample Poisson exact test[28] for synonymous variants and one-sided, two-sample for the remainder (Materials and Methods). Error bars represent 95% confidence intervals throughout (**a**) – (**d**).

Gene level analyses

Since observed risk to ASD or ID/DD was carried only by *de novo* variants absent from the standing variation of ExAC, we next sought to extend this concept by evaluating whether the overall rate of PTVs per gene in ExAC provided a similar guide to which ASD and ID/DD variants were relevant. Specifically, we investigated whether the gene-level constraint metric, pLI[16] (probability of loss-of-function intolerance), could improve our ability to decipher which class 1 *de novo* PTVs confer the most risk to ASD and ID/DD (Materials and Methods). Using the same threshold as Lek et al. (2016), we used a threshold of pLI ≥0.9 to define loss-of-function (LoF)-intolerant genes and investigated whether individuals with ASD had an increased burden of class 1 *de novo* PTVs in such genes. When we restricted to solely class 1 *de novo* PTVs in LoF-intolerant genes, we observed a significant excess in individuals with ASD (0.067 per exome) compared to their unaffected siblings (0.021 per exome; rate ratio=3.24; P=3.14x10^{-16}; one-sided Poisson exact test). For individuals with ID/DD, the rate of class 1 *de novo* PTVs in LoF-intolerant genes becomes more striking, with a rate of 0.139 per exome, resulting in a 6.70 rate ratio when compared to the control group of unaffected ASD siblings (P=6.34x10^{-38}; one-sided Poisson exact test). By contrast, the rate of class 1 *de novo* PTVs in LoF-tolerant genes (pLI <0.9) show no difference between individuals with ASD (0.063 vs. 0.051; P=0.06; two-sided Poisson test), or individuals with ID/DD (0.048 vs. 0.051; P=0.75; two-sided Poisson exact test; **Figure 2.3D**) when compared to unaffected ASD siblings. Again, results remain unchanged when we exclude the ExAC psychiatric samples. The same trend is observed in congenital heart disease[26,27] and schizophrenia[25]: no association among *de novo* PTVs present in ExAC (congenital heart disease: P= 0.93; schizophrenia: P=0.93; one-sided Poisson exact test) or in LoF-tolerant genes (congenital heart disease: P=0.28; SCZ: P=0.67; one-sided Poisson

exact test). Class 1 *de novo* PTVs in LoF-intolerant genes carried the observed association in congenital heart disease (rate ratio=2.31; P=2.12x10^{-5}; one-sided Poisson exact test) and schizophrenia (rate ratio=1.76; P=0.04; one-sided Poisson exact test), although the latter does not survive Bonferroni correction. Hence, all detectable *de novo* PTV signal in these phenotypes can be localized to 18% of genes with clear intolerance to PTVs in ExAC, with, consequently, substantially amplified rate ratios in this gene set.

Recent studies inferred the presence of multiple *de novo* PTVs in the same gene as evidence of contribution to ASD risk[5-10]. Of the 51 genes with ≥2 *de novo* PTVs, only 38 are absent in controls. This not only reinforces the point that the mere observation of multiple *de novo* PTVs in a gene is not sufficient to define that gene as important[5,16], but also provides an opportunity to explore whether the pLI metric can refine the identification of specific genes. In fact, 32 of the 38 case-only genes, but only 5 of 13 control-only or case-control hit genes, are LoF-intolerant, a highly significant difference (OR=8.07; P=0.003; Fisher's exact test) that greatly refines the list of genes to be pursued as likely ASD contributors.

Phenotypic associations for class 1 *de novo* PTVs in LoF-intolerant genes

While enrichment of *de novo* PTVs is one of the hallmarks of ASD *de novo* studies[5-10,15,16], another consistent finding is an increased burden of these variants among females with ASD[6,15] and in ASD individuals with low full-scale IQ (FSIQ)[6,15,16]. We investigated whether these hallmarks were present in the 6.55% of ASD cases with a class 1 *de novo* PTV in LoF-intolerant genes (pLI ≥0.9). Indeed, females are overrepresented in the subset (12.26% of females; 5.80% males; P=1.75x10^{-5}; Fisher's exact test). Importantly, for the 6.86% of ASD cases with a class 2 *de novo* PTV or a class 1 *de novo* PTV in a LoF-tolerant gene (pLI <0.9),

there is no difference between the sexes, with 6.86% of females and 6.83% of males falling in this category (P=1; Fisher's exact test). Furthermore, class 2 *de novo* PTVs and class 1 *de novo* PTVs in LoF-tolerant genes show no association with FSIQ (β=-0.001; P=0.76; Poisson regression), while class 1 *de novo* PTVs in LoF-intolerant genes predominately explain the skewing towards lower FSIQ (β=-0.023; $P=7\times10^{-8}$; Poisson regression; **Figure 2.4A**). Given these observations, we split the ASD class 1 *de novo* PTV signal in LoF-intolerant genes by sex and intellectual disability status (Materials and Methods). Females with comorbid ASD and intellectual disability have the highest rate of class 1 *de novo* PTVs in LoF-intolerant genes (rate ratio=8.71; $P=2.73\times10^{-12}$; one-sided Poisson exact test). Despite the overwhelming enrichment in females and individuals with comorbid ASD and intellectual disability, males with ASD without intellectual disability still show enrichment of class 1 *de novo* PTVs in LoF-intolerant genes (rate ratio=2.95; $P=1.31\times10^{-9}$; one-sided Poisson exact test; **Figure 2.4B**). These secondary analyses strongly support the implication of the primary analysis: that collectively, class 2 *de novo* PTVs and class 1 *de novo* PTVs in LoF-tolerant genes have little to no association to ASD or ID/DD and no observable phenotypic impact on those cases carrying them. By contrast, the class 1 *de novo* variants occurring in LoF-intolerant genes contain the association signal and phenotypic skewing observed to date.

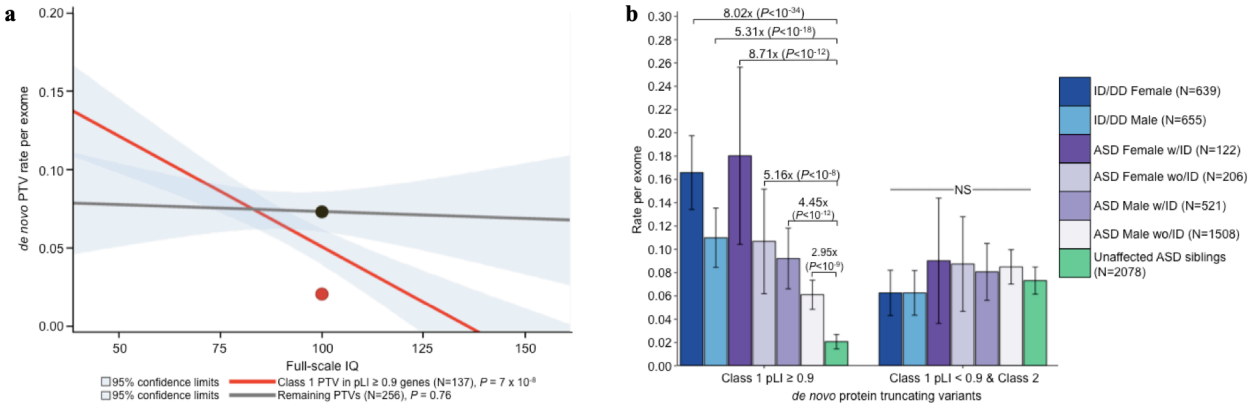


Figure 2.4: Phenotypic associations for ASD de novo PTVs. (**a**) IQ distribution of class 1 *de novo* PTVs in pLI ≥ 0.9 genes (red) and remaining *de novo* PTVs (class 2 and class 1 pLI < 0.9; grey) in 393 individuals with ASD with a measured full-scale IQ. Dots indicate the rate in unaffected ASD siblings for their respective categories of de novo PTVs. P-values calculated using Poisson regression. Only class 1 *de novo* PTVs show association with full-scale IQ. (**b**) Rate of class 1 *de novo* PTVs (left set) and the remaining *de novo* PTVs (class 2 & class 1 in LoF-tolerant genes, right set) in ID/DD (left two bars) and ASD (middle four bars) split by sex and ID with the number of individuals labeled in the legends. Error bars represent 95% confidence intervals, and P-values were calculated using one-sided, two-sample Poisson exact tests comparing to unaffected ASD siblings.

Inherited variation

As the effect size for *de novo* PTVs increased after removing those variants present in ExAC, we postulated a similar increase could be obtained from rare inherited PTVs. Under the assumption that risk-conferring variants should be transmitted more often to individuals with ASD, we tested for transmission disequilibrium in a cohort of 4319 trios with an affected proband (Materials and Methods). Without filtering by pLI or presence/absence status in ExAC, singleton PTVs, as a class, showed no over-transmission (P=0.31; binomial test). After removing all of the variants present in ExAC or in a LoF-tolerant gene (pLI <0.9), we found a modest excess of transmitted PTVs in ASD cases (rate ratio=1.16; P=9.85x10^{-3}; binomial test). As with all previous analyses, this result is virtually identical when the psychiatric cohorts in ExAC are removed (rate ratio=1.14; P=0.02; binomial test). While there are far more inherited

PTVs than *de novo* PTVs, the inherited variant effect size (1.16 rate ratio) is paradoxically minute by comparison to that of *de novo* PTVs (3.24 rate ratio).

Despite the different effect sizes between *de novo* and inherited PTVs, the data does not suggest the two classes of variation differ in penetrance. Instead, the data suggest the excess of inherited PTVs resides in a different set of genes than those implicated by *de novo* variation. Specifically, the largest *de novo* variant excess resides in a limited and extremely penetrant set of genes that do not contribute substantially to inherited PTV counts. If we consider the 11 genes with ≥3 class 1 *de novo* PTVs in ASD cases and none in controls (47 *de novo* PTVs in total), all 11 are intolerant of truncating variation (pLI ≥0.9) (**Table 2.1**). These variants confer risk to particularly severe outcomes: of the cases with available IQ data, 14 of the 29 individuals have IQ below 70 or were unable to complete a traditional IQ test[15], while only 27% of all ASD individuals with available IQ data in this study fall into this group (P=0.008; Fisher's exact test). Across this same gene set, there are only 4 inherited PTVs (from a total of 5 observed in the parents of the 4,319 ASD trios). Of those 4 inherited PTVs, only the inherited frameshift in *CHD8* bore evidence of mosaic transmission (P=5.49x10^{-3}; binomial test) suggesting it may have arisen post-zygotically and not carried by a parent. This ratio – that 80-90% of the observed variants are *de novo* rather than inherited in ASD cases – indicates enormous selective pressure against mutations in these genes, far greater than the direct selection against ASD in general (**Table 2.1**). Indeed, despite ascertaining these 11 genes based on those with the most class 1 *de novo* PTVs in ASD, we observe a higher rate of *de novo* PTVs in these same genes in the ID/DD studies (37 mutations in 1284 cases). This underscores that selection against these variants likely arises from more severe and widespread impact on neurodevelopment and cognition. Despite the

minor contribution of inherited variation in these genes, some insights from studying families may be particularly useful.

We investigated whether any *de novo* variants were observed in the transmission data (i.e., a variant that was both inherited and *de novo* in separate unrelated families). We observed 164 transmitted variants and 66 untransmitted variants that were also observed *de novo* in individuals with ASD and their unaffected siblings, respectively. Of these 164 transmitted and 66 untransmitted variants, 23 and 14 were absent in ExAC. Among the 23 transmitted variants absent from ExAC, two variant were of particular interest, a nonsense variant in *ANK2*, and a probably damaging missense variant in *RGL1*. *ANK2* is a gene previously implicated for risk for ASD due to having multiple *de novo* PTV mutations[14,15] (**Table 2.1**). In *RGL1*, a Ral guanyl-nucleotide exchange factor, the PolyPhen2[16] probably-damaging missense variant was transmitted to an affected ASD proband in three separate unrelated trios, and observed *de novo* in a fourth ASD trio. Thus, we now have observed 4 instances of this specific probably-damaging missense variant in *RGL1* in individuals with ASD and none in 60,706 individuals in ExAC.

Table 2.1: Top 12 genes with ≥ 3 class 1 *de novo* PTVs in individuals with ASD. Twelve genes with ≥ 3 class 1 *de novo* PTVs in 3982 individuals with ASD. Additionally, for each gene, we have listed the number of class 1 *de novo* PTVs in 2078 unaffected ASD siblings and in 1284 individuals with ID/DD, as well as the number of singleton, LOFTEE high-confidence PTVs absent from ExAC that were transmitted (T) or untransmitted (U) to 4319 individuals with ASD and present in 404 cases of ASD and 3654 population controls. *P*-values represent the Poisson probability of observing more than the expected number of class 1 *de novo* PTVs (Materials and Methods). ID/DD, intellectual disability / developmental delay; ASD, autism spectrum disorder; PTV, protein truncating variant; pLI, probability of loss-of-function intolerance

	Class 1 *de novo* PTVs			Inherited		Case-control			
Gene	ASD	Unaffected ASD siblings	ID/DD	T	U	Case	Control	pLI	*P*-value
CHD8	7	0	0	1	0	0	0	1	3.70E-13
ARID1B	5	0	11	0	0	0	0	1	1.07E-08
DYRK1A	5	0	2	0	0	0	0	0.9996	2.46E-11
SYNGAP1	5	0	9	0	0	0	0	1	2.47E-10
ADNP	4	0	4	0	0	1	0	0.9989	3.93E-09
ANK2	4	0	0	1	1	0	0	1	7.07E-06
DSCAM	4	0	0	2	0	1	0	1	3.62E-07
SCN2A	4	0	7	0	0	1	0	1	1.25E-06
ASH1L	3	0	0	0	0	0	0	1	1.67E-04
CHD2	3	0	2	0	0	0	0	1	7.81E-05
KDM5B	3	2	0	5	1	0	0	5.09E-05	7.22E-05
POGZ	3	0	2	0	0	2	0	1	3.12E-05

Case-control analysis

Having observed a significant enrichment in both *de novo* and inherited PTVs absent from ExAC in LoF-intolerant genes (pLI ≥0.9), we applied this same methodology to case-control cohorts. Given that the variation present in a single individual will be a combination of *de novo* (both somatic and germline) and inherited variation, we expect to see an effect size for PTVs intermediate between that of the *de novo* and inherited PTVs absent from ExAC in LoF-intolerant genes. Using a published cohort of 404 ASD cases and 3654 controls from Sweden[5], we first analyzed the rate of singleton synonymous variants as a control for further analyses. We found no case-control difference among those present/absent from ExAC (*P*=0.59; Fisher's exact

test). Turning to the PTV category, we observe a slight excess of singleton PTVs in cases with ASD (917 PTVs in 404 cases) compared to controls (7259 PTVs in 3654 controls; OR=1.16; P=3.13x10^{-5}; Fisher's exact test). This signal increases once we remove all singleton PTVs present in ExAC or in LoF-tolerant genes, providing the first instance of an exome-wide excess of PTVs demonstrated in ASD without the use of trios (128 PTVs in 404 cases, 447 PTVs in 3654 controls; 2.63 OR; P=1.37x10^{-18}; Fisher's exact test). Consistent with the previous *de novo* and inherited analyses, no signal exists for the remaining 7601 singleton PTVs (OR=1.06; P=0.11; Fisher's exact test). Lastly, removing the psychiatric cohorts from ExAC results in a 2.42 OR for singleton PTVs absent from ExAC in LoF-intolerant genes (133 PTVs in 404 cases, 506 PTVs in 3654 controls; P=1.06x10^{-16}; Fisher's exact test).

Discussion

Here we demonstrated that ~1/3 of *de novo* variants identified in neurodevelopmental disease cohorts are also present as standing variation in ExAC, indicating the presence of widespread mutational recurrence. Reinforcing this, we demonstrated that these class 2 *de novo* variants are enriched for more mutable CpG sites. Most importantly, however, these class 2 *de novo* variants confer no detectable risk to ID/DD and ASDs, and eliminating them from our analysis improved all genetic and phenotypic associations by removing the "noise" of benign variation.

We further refined the class 1 *de novo* PTV association using a gene-level intolerance metric (pLI) developed using the ExAC resource and identified that all detectable mutational excess resided in 18% of genes most strongly and recognizably intolerant of truncating mutation. Specifically, 13.5% (±2.0%, 95% CI) of individuals with ID/DD and 6.55% (±0.8%, 95% CI) of

individuals with ASD, but only 2.1% (±0.6%, 95% CI) of controls, have a *de novo* PTV absent from ExAC and present in a gene with a very low burden of PTVs in ExAC (pLI ≥0.9). ASD cases with such a variant are more likely to be female and/or have intellectual disability than the overall ASD population. For the remaining 93.45% of the ASD cohort, we fail to observe any meaningful phenotypic difference (i.e., IQ or sex) between the 6.86% of individuals with and the 86.59% of individuals without a class 2 *de novo* PTV or a class 1 *de novo* PTV in a LoF-tolerant gene. These results, taken together with an overall lack of excess case burden, suggest that collectively, neither class 2 nor class 1 *de novo* PTVs in LoF-tolerant genes (pLI <0.9) appear to confer significant risk toward ASD. Thus, we have refined the role of *de novo* protein truncating variation in ASD, confining the signal to a smaller subset of patients than previously described[6,29].

This analysis framework, operating at the variant level, also enabled a careful examination of inherited variation in ASD. While ASD is highly heritable[3], few analyses[30] have demonstrated specific inherited components. By removing inherited PTVs present in ExAC or in LoF-tolerant genes, we discovered a modest signal of over-transmitted PTVs, in line with previous reports[30]. The vast majority of inherited PTVs appear to affect genes that have yet to show signal from *de novo* variation, with only 1% residing in the strongest associated genes, indicating the inherited variants reside in genes with a somewhat weaker selective pressure against them. Ultimately, however, as these variants occur in 15.4% of cases but carry only a 1.16-fold increased risk as a group, they explain little of the overall heritability (<1% of the variance in liability).

Given the current size of ExAC and the general scarcity of truncating variants, the pLI metric for constraint against loss-of-function variation does not yet provide precise resolution of

the selection coefficient acting on PTVs in that gene. That is, even a pLI ≥0.9 does not guarantee a selection coefficient sufficiently high to ensure the vast majority of variation is *de novo* rather than inherited. In fact, selection coefficients for pLI ≥0.9 genes range from 0.1–0.5 (where the majority of variation will be inherited), all the way to selection coefficients approaching 1, in which the variants are almost completely reproductively null. Only larger reference panels will enable refining these estimates, articulating a gradient from the strongest genes we currently flag (e.g., the 11 genes with ≥3 *de novo* PTVs in ASD and none in controls that make their contribution almost entirely through penetrant, single-generation *de novo* variation) to those genes we have yet to define clearly that will make their contribution largely through inherited, albeit less penetrant, variation. The significant expansion of exome sequencing in ASD, alongside larger reference panels from which to draw more precise inferences about selective pressure against variation in each gene, will allow us to fill in the genetic architecture of ASDs in the region of the effect size spectrum between severe *de novo* variation at one end and common variation at the other.

ExAC currently has 15,330 individuals from psychiatric cohorts, with the schizophrenia cohort being the largest[24]. Given the shared genetics between ASD and schizophrenia[2,5,16,17,25,31,32] , it is reasonable to hypothesize that the psychiatric cohorts within ExAC could influence our analyses. As we have shown however, removing the psychiatric cohorts within ExAC does not change our results. In fact, of the 16 *de novo* PTVs in LoF-intolerant genes that were also variant in ExAC, only two reside solely in the 15,330 individuals from the psychiatric cohorts (*CUX2* in ASD, *LARP1* in unaffected ASD siblings). This number being so small is in retrospect not surprising because it is so unusual to observe a deleterious variant both *de novo* and present as standing variation in individuals with the same ascertained

phenotype, let alone in different ascertained phenotypes. The *ANK2* nonsense variant was the only such instance of the same deleterious variant being *de novo* in one ASD trio and inherited in another.

While we use ASDs and ID/DD here to explore this framework, it can certainly be applied toward any trait. However, this framework is optimally powered in traits governed by genes under strong selection, as it will remove *de novo* variants that are more common when examined in the context of a larger reference population. Our results reinforce the point that not all *de novo* variants are rare and contribute to risk, while highlighting the tremendous value of large population sequence resources even for the interpretation of *de novo* variation and complex disease. This is especially important in the case of clinical sequencing, in which the paradigm has unfortunately become that if a protein-altering *de novo* variant is present in the gene of interest, then it is often considered the causal variant[33,34]. Clearly, not all *de novo* variants are equal, and not all *de novo* variants in a gene contribute to risk in the same way.

Materials and Methods

Datasets and data processing

Two versions of the Exome Aggregation Consortium (ExAC) database were used in this analysis: the full version of ExAC (N = 60,706) and the non-psychiatric version of ExAC (N = 45,376). The non-psychiatric version of ExAC has the following cohorts removed: Bulgarian trios (N = 461), sequencing in Suomi (N = 948), Swedish schizophrenia & bipolar studies (N = 12,119), schizophrenia trios from Taiwan (N = 1505), and Tourette syndrome association international consortium for genomics (N = 297). We used a combined set of 8401 published *de novo* variants from 3982 probands with ASD and 2078 of their unaffected siblings from two

recent large-scale exome sequencing studies: de Rubeis *et al* (N_{ASD} = 1474, $N_{unaffected\_sib}$ = 267)[5], Iossifov, O'Roak, Sanders, Ronemus *et al* (N_{ASD} = 2508, $N_{unaffected\_sib}$ = 1911)[6]. We also used 1692 *de novo* variants from 1284 probands published in studies of intellectual disability (ID) (de Ligt *et al*: N = 100[12], Rauch *et al*: N = 51[14]) and developmental delay (DD) (DDD: N = 1133)[35]. *De novo* variants from congenital heart disease[26,27] and schizophrenia[25] were also downloaded for additional confirmation of the recurrent mutation rate. Details of the sequencing and *de novo* calling can be found in the referenced publications.

To ensure uniformity in variant representation and annotation across datasets and with respect to the ExAC reference database[36], we created a standardized variant representation through a Python implementation of vt normalize[37] and re-annotated all variants using Variant Effect Predictor (VEP)[38] version 81 with GENCODE v19 on GRCh37. VEP provided the Ensembl Gene IDs, gene symbol, the Ensembl Transcript ID for use in determining canonical transcripts, as well as PolyPhen2 and SIFT scores. We used the canonical transcript when possible for cases when the variant resided in multiple transcripts, and the most deleterious annotation in cases of multiple canonical transcripts. If no canonical transcript was available, the most deleterious annotation was used.

Determining class 1 or class 2 *de novo* variants

De novo variants were classified as class 1 or class 2 based on their respective absence or presence in ExAC. Presence or absence in ExAC was defined if the variant had the same chromosome, position, reference, and alternate allele in both files. Due to the heterogeneous nature of ExAC, and the different capture arrays used in the original exome sequencing studies incorporated into ExAC, we elected to use all of the variants in ExAC, not just those with a

PASS status in the GATK variant calling filter. For insertions/deletions, we took a conservative stance that they must match exactly (i.e., a subset was not sufficient). To illustrate, if a *de novo* variant on chromosome 5 at position 77242526 has a reference allele of AGATG and a *de novo* alternate allele where four nucleotides are deleted (AGATG to A), we would not say that variant is present in ExAC if there was another variant at the same genomic position in ExAC where only the first two of these nucleotides are deleted (AGA to A). Lastly, for variants outside of the proportion of the genome covered by ExAC, we considered them to be class 1 *de novo* variants – as expected, none of these variants reside in the coding region (**Table 2.2**).

Table 2.2: Variants residing in regions not covered by ExAC per functional class and cohort

Functional class	ASD	Unaffected ASD siblings	ID/DD
3-prime UTR variant	4	1	1
5-prime UTR variant	3	2	0
Downstream gene variant	4	3	3
Intergenic variant	5	5	14
Intron variant	55	37	42
Non-coding region transcript exon variant	6	4	2
Regulatory region variant	2	1	10
Upstream gene variant	1	6	4
Splice region variant	0	0	3
Total	80	59	79

Variant calling for transmission and case-control analysis

We used the Genome Analysis Toolkit (GATK v3.1-144) to recall a dataset of 22,144 exomes from the Autism Sequencing Consortium (ASC)[39] & Simons Simplex Collection (SSC)[40] sequencing efforts. This call set contained 4319 complete trios (including all those from which the published and validated *de novo* mutations were identified), which we used to evaluate inherited variation, and a published case-control dataset of individuals of Swedish ancestry (404 individuals with ASD and 3564 controls)[5]. We applied a series of quality control filters on the genotype data, using the genome-wide transmission rate as a guide for filter inclusion/exclusion.

More specifically, we calibrated various genotyping filters such that synonymous singleton variants – where the alternative allele was seen in only one parent in the dataset – was transmitted at a rate of 50%, because we expect, as a class, synonymous variants to be transmitted 50% of the time. As with the ExAC analysis[36], we found GATK's default Variant Quality Score Recalibration (VQSR) too restrictive due to the bias toward common sites. In order to reduce the number of singleton variants being filtered out, we recalibrated the Variant Quality Score Log Odds (VQSLOD) threshold from -1.49 to -1.754, dropping the singleton synonymous transmission rate from 51.1% to 49.9998%. Additional filtering was done at the individual level, in which we required a minimum read depth of 10 and a minimum GQ and PL of 25 for each individual's variant call. We also applied an allele balance filter specific for each of the three genotypes (homozygous reference, heterozygous, homozygous alternate), where allele balance is defined as the number of alternate reads divided by the total number of reads. We required the allele balance for homozygous reference individuals to be less than 0.1, allele balance for heterozygous individuals to be between 0.3 and 0.7, and the allele balance for homozygous alternate individuals to be greater than 0.9. Calls that did not pass these filters were set to missing. Lastly, for the transmission analysis, we removed variants in which more than 20% of families failed one of our filters. For the case-control analysis, we removed variants in which more than 5% of families failed one of our filters.

On the use of the Poisson exact test for comparing rates of *de novo* variation between two samples

As with many other papers[6,8,41-43], we too were interested in testing whether the rate of a given class of *de novo* variation was significantly different between our cohorts of individuals

with ASD or ID/DD as compared to unaffected ASD siblings. As the number of *de novo* variants per individual follows a Poisson distribution[8], we tested $H_A : \lambda_1 \neq \lambda_2$ vs. $H_0 : \lambda_1 = \lambda_2$, where λ_i is the rate of a given class of *de novo* variation in group *i*, using the Poisson exact test (also known as the *C*-test)[28]. Note: we could not compare the rates to expectation, because the expectations published in Samocha et al., (2014) are for all *de novo* variants, not just *de novo* variants present/absent from ExAC. An important consequence of our hypothesis test is that effect sizes are reported as rate ratios, which is simply the quotient of the two rates. While more commonly reported, odds ratios require Bernoulli random variables (e.g., an individual either harbors or does not harbor a *de novo* variant), and as such, would be incorrect given the hypothesis we are testing. Had we been interested in testing for a significant difference between the proportion of individuals harboring a *de novo* PTV, then an odds ratio would be appropriate (and Fisher's exact test would suffice in this case). Thus, only in using the Poisson exact test could we reject the null hypothesis that the rate of *de novo* PTVs is the same between individuals with ASD and their unaffected siblings and find evidence that individuals with ASD have a higher rate of *de novo* PTVs than their unaffected siblings. The difference between the two tests is a subtle, but important one.

<u>On the use of pLI (probability of loss-of-function intolerance)</u>

Using the observed and expected number of PTVs per gene in the ExAC dataset, we developed a metric to evaluate a gene's apparent intolerance to such variation[24]. Briefly, the probability of loss-of-function intolerance (pLI) was computed using an EM algorithm that assigned genes to one of three categories: fully tolerant (in which PTVs are presumed neutral and occur, like synonymous variants, at rates proportional to the mutation rate), "recessive-like"

(showing PTV depletion similar to known severe autosomal recessive diseases) and "haploinsufficient-like" (showing PTV depletion similar to established severe haploinsufficiencies). pLI is the posterior probability that a gene resides in the last, most loss-of-function intolerant, category. See section 4 of the supplement of Lek, et al. (2016) for more details.

Phenotype analysis

Full-scale deviation IQ scores were measured using several tests including the Differential Ability Scales, the Wechsler Intelligence Scale for Children, and the Wechsler Abbreviated Scale of Intelligence. IQ has previously been associated with *de novo* PTV rate in the SSC[6,15,44]. In this analysis, we used Poisson regression to estimate the relationship between the rate of each of class 1 and class 2 PTVs and proband full-scale deviation IQ.

Calculating the expected number of class 2 *de novo* variants in a reference database

For a set of r *de novo* variants, each with the same allele count, K, in ExAC, we can estimate the number of those variants still observed at least once in a subset of size n using the hypergeometric distribution (**Figure 2.5**). That is to say, how many of those same sites will still be present as standing variation in a down-sampled version of ExAC? Specifically,

$$\text{expected count} = \mathrm{r}\big(1 - \mathrm{P}(\mathrm{k} = 0)\big) = \mathrm{r}\left(1 - \frac{\binom{\mathrm{K}}{0}\binom{\mathrm{N}-\mathrm{K}}{\mathrm{n}-0}}{\binom{\mathrm{N}}{\mathrm{n}}}\right) = \mathrm{r}\left(1 - \frac{\binom{\mathrm{N}-\mathrm{K}}{\mathrm{n}}}{\binom{\mathrm{N}}{\mathrm{n}}}\right)$$

where k is approximately$\mathrm{hypergeometric}\,(\mathrm{N}, \mathrm{K}, \mathrm{n})$, and N is the number of chromosomes in the current version of ExAC (N=121,412). This only holds when each down-sampled set of ExAC preserves the ancestry proportions of the total sample.

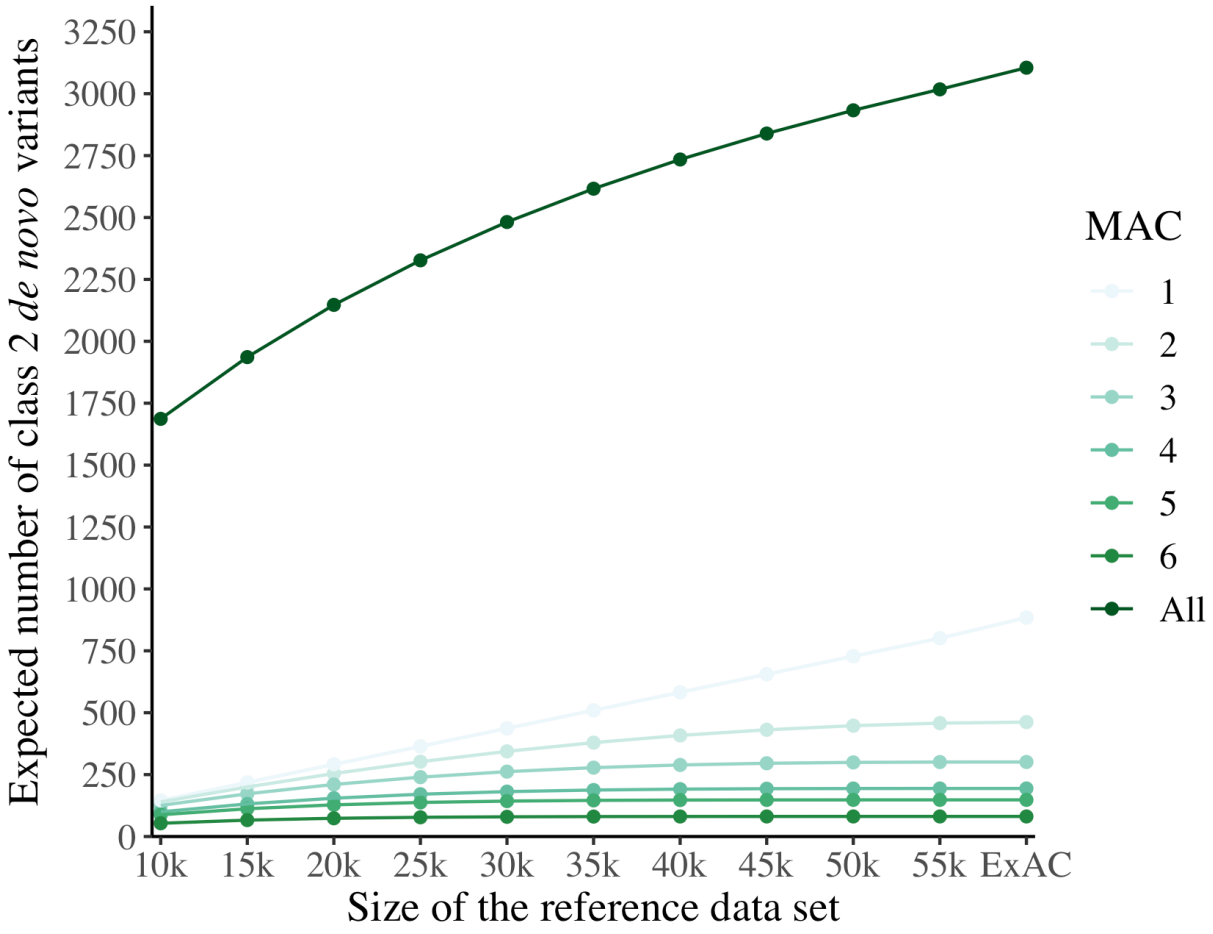


Figure 2.5: Recurrence rate is a function of allele frequency and reference-population size. Expected number of discovered class 2 *de novo* variants by size of the reference dataset, partitioned based on the number of copies of the variant currently present in ExAC. The number of *de novo* variants found in the standing population is a function of the sample size of the reference dataset and the current estimated minor allele count (MAC).

Calculating mutation rates for class 1 and class 2 *de novo* PTVs

Samocha et al. calculated per gene mutation rates for ALL synonymous, missense, and PTVs, not for those present/absent in ExAC. If we are interested in comparing the rate of class 1 *de novo* PTVs to the expected depth-corrected mutation rate for class 1 *de novo* PTVs, we can roughly calculate it. For a given gene, we can derive the class 1 and class 2 PTV mutation rate by breaking down the overall mutation rate for PTVs, denoted as $\hat{\mu}_{PTV}$, using equation (1)

$$\hat{\mu}_{PTV} = \hat{\mu}_{class\ 1\ PTV} + \hat{\mu}_{class\ 2\ PTV} \quad (1)$$

In case the logic behind equation 1 isn't completely clear, it may help to point out that the number of class 1 and class 2 PTVs is equal to the total number of PTVs. Now, Samocha et al. provides us with $\hat{\mu}_{PTV}$, so all we need to do is calculate $\hat{\mu}_{class\ 1\ PTV}$ and $\hat{\mu}_{class\ 2\ PTV}$. Given all of the PTVs in ExAC, and the probability of each trinucleotide-to-trinucleotide mutation, we can calculate $\hat{\mu}_{class\ 2\ PTV}$ using equation (2)

$$\hat{\mu}_{class\ 2\ PTV} = \sum_{PTV_i}^{PTV_n} \hat{\mu}_{SNPi} \quad (2)$$

where i indexes the n PTVs for a given gene present in ExAC, and $\hat{\mu}_{SNPi}$ is the mutation rate of that specific trinucleotide substitution that creates a PTV. With $\hat{\mu}_{class\ 2\ PTV}$ calculated, $\hat{\mu}_{class\ 1\ PTV}$ follows from equation 1. However, these per gene $\hat{\mu}_{PTV}$ calculations do not account for sequencing depth. Correcting for depth of sequencing becomes tricky, as the depth of sequencing varies between studies and will not necessarily be the same as the depth of sequencing for ExAC. However, we can roughly approximate the depth-corrected $\hat{\mu}_{class\ 2\ PTV}$ for each gene using the following equation under the assumption that the fraction of the raw mutability from class 2 $\left(i.e.,\ \frac{\hat{\mu}_{class\ 2\ PTV}}{\hat{\mu}_{PTV}}\right)$ is equal to the fraction of the class 2 depth-corrected mutability $\left(i.e.,\ \frac{\hat{\mu}_{class\ 2\ PTV,depth\ corrected}}{\hat{\mu}_{PTV,depth\ corrected}}\right)$

$$\hat{\mu}_{class\ 2\ PTV,depth\ corrected} = \hat{\mu}_{PTV,depth\ corrected}\left(\frac{\hat{\mu}_{class\ 2\ PTV}}{\hat{\mu}_{PTV}}\right) \quad (3)$$

The depth corrected $\hat{\mu}_{class\ 1\ PTV}$ follows using the same logic as we used in equation (1).

A note on semantics: *de novo* mutation vs. *de novo* variant

The two terms – *de novo* mutation and *de novo* variant – can be interchangeable. While one might admittedly consider the choice of which term to use largely a matter of taste, the word mutation can have different meanings; it can refer to both the process of nucleotide change, as

well as the end product. Thus, the sentence, "a mutation creates a mutation*"* is a grammatically correct sentence, albeit an unlikely one to be heard. In order to avoid any ambiguity throughout this paper, we use the term *mutation* to refer to the biological process and *variant* to refer to the corresponding change in the DNA. As such, we would say, "a *de novo* mutation creates a *de novo* variant". This definition enables the use of the phrase, recurrent *de novo* mutation, to be logical, whereas a *de novo* variant cannot by our definition be recurrent.

Author Contributions

Jack Kosmicki: method design, data analyses (exceptions below), writing

Mark Daly: method design, writing, overall guidance

Dennis Wall: overall guidance

Elise Robinson: IQ regression (Figure 2.3a)

Kaitlin Samocha: created pLI

Monkol Lek and *Daniel MacArthur:* created the ExAC resource

References

1. Developmental, D.M.N.S.Y. & Investigators, P. Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002)* **63**, 1 (2014).

2. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-94 (2013).

3. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* **46**, 881-5 (2014).

4. Klei, L. *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Mol Autism* **3**, 9 (2012).

5. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).

6. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).

7. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).

8. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).

9. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).

10. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).

11. Reichenberg, A. *et al.* Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. *Proceedings of the National Academy of Sciences* **113**, 1098-1103 (2016).

12. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* **367**, 1921-9 (2012).

13. Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).

14. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674-82 (2012).

15. Robinson, E.B. *et al.* Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proc Natl Acad Sci U S A* **111**, 15161-5 (2014).

16. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).

17. Robinson, E.B. *et al.* Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat Genet* **advance online publication**(2016).

18. Bellus, G.A. *et al.* Achondroplasia is defined by recurrent G380R mutations of FGFR3. *Am J Hum Genet* **56**, 368-73 (1995).

19. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893-903 (1969).

20. Coulondre, C., Miller, J.H., Farabaugh, P.J. & Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274**, 775-780 (1978).

21. Haukka, J., Suvisaari, J. & Lonnqvist, J. Fertility of patients with schizophrenia, their siblings, and the general population: a cohort study from 1950 to 1959 in Finland. *Am J Psychiatry* **160**, 460-3 (2003).

22. Laursen, T.M. & Munk-Olsen, T. Reproductive patterns in psychotic patients. *Schizophr Res* **121**, 234-40 (2010).

23. Power, R.A., Kyaga, S., Uher, R. & et al. FEcundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**, 22-30 (2013).

24. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).

25. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-84 (2014).

26. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-1266 (2015).

27. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-3 (2013).

28. Przyborowski, J. & Wilenski, H. Homogeneity of Results in Testing Samples from Poisson Series: With an Application to Testing Clover Seed for Dodder. *Biometrika* **31**, 313-323 (1940).

29. Picoraro, J. & Chung, W. Delineation of New Disorders and Phenotypic Expansion of Known Disorders Through Whole Exome Sequencing. *Current Genetic Medicine Reports* **3**, 209-218 (2015).

30. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat Genet* **47**, 582-588 (2015).

31. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-90 (2014).

32. Singh, T. *et al.* Rare schizophrenia risk variants are enriched in genes shared with neurodevelopmental disorders. *bioRxiv* (2016).

33. Akawi, N. *et al.* Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat Genet* **47**, 1363-1369 (2015).

34. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405-24 (2015).

35. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).

36. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* (2015).

37. Tan, A., Abecasis, G.R. & Kang, H.M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202-4 (2015).

38. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).

39. Buxbaum, J.D. *et al.* The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* **76**, 1052-6 (2012).

40. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192-195 (2010).

41. Ben-David, E. & Shifman, S. Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. *Mol Psychiatry* **18**, 1054-6 (2013).

42. Takata, A., Ionita-Laza, I., Gogos, J.A., Xu, B. & Karayiorgou, M. De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia. *Neuron* **89**, 940-7 (2016).

43. Lelieveld, S.H. *et al.* Meta-analysis of 2,104 trios provides support for 10 novel candidate genes for intellectual disability. *bioRxiv* (2016).

44. Sanders, S.J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-33 (2015).

Chapter 3

Discovery and characterization of 102 genes associated with autism spectrum disorder from exome sequencing of 35,584 individuals

Work presented in this chapter will be published as part of:

Satterstrom, F.K., Kosmicki, J.A., Wang, J. *et al*. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *bioRxiv*, 484113 (2018).

Abstract

We present the largest exome sequencing study of autism spectrum disorder (ASD) to date (n=35,584 total samples, 11,986 with ASD). Using an enhanced Bayesian framework to integrate *de novo* and case-control rare variation, we identify 102 risk genes at a false discovery rate ≤ 0.1. Of these genes, 49 show higher frequencies of disruptive *de novo* variants in individuals ascertained for severe neurodevelopmental delay, while 53 show higher frequencies in individuals ascertained for ASD; comparing ASD cases with mutations in these groups reveals phenotypic differences.

Introduction

Autism spectrum disorder (ASD), a childhood-onset neurodevelopmental condition characterized by deficits in social communication and restricted, repetitive patterns of behavior or interests, affects more than 1% of individuals[1]. Multiple studies have demonstrated high heritability, much of it due to common variation[2], although rare inherited and *de novo* variants are major contributors to individual risk[3-5]. When this rare variation disrupts a gene in individuals with ASD more often than expected by chance, it implicates that gene in risk[6]. ASD risk genes, in turn, provide insight into the underpinnings of ASD, both individually[7,8] and *en masse* [3,5,9]. However, fundamental questions about the altered neurodevelopment and altered neurophysiology in ASD—including when it occurs, where, and in what cell types—remain poorly resolved.

Here we present the largest exome sequencing study in ASD to date. Through an international collaborative effort and the willingness of thousands of participating families, we assembled a cohort of 35,584 samples, including 11,986 with ASD. We introduce an enhanced

Bayesian analytic framework, which incorporates recently developed gene- and variant-level scores of evolutionary constraint of genetic variation, and we use it to identify 102 ASD-associated genes (FDR ≤ 0.1). Because ASD is often one of a constellation of symptoms of neurodevelopmental delay (NDD), we identify subsets of the 102 ASD-associated genes that have disruptive *de novo* variants more often in NDD-ascertained or ASD-ascertained cohorts. Together, these insights form an important step forward in elucidating the neurobiology of ASD.

Results

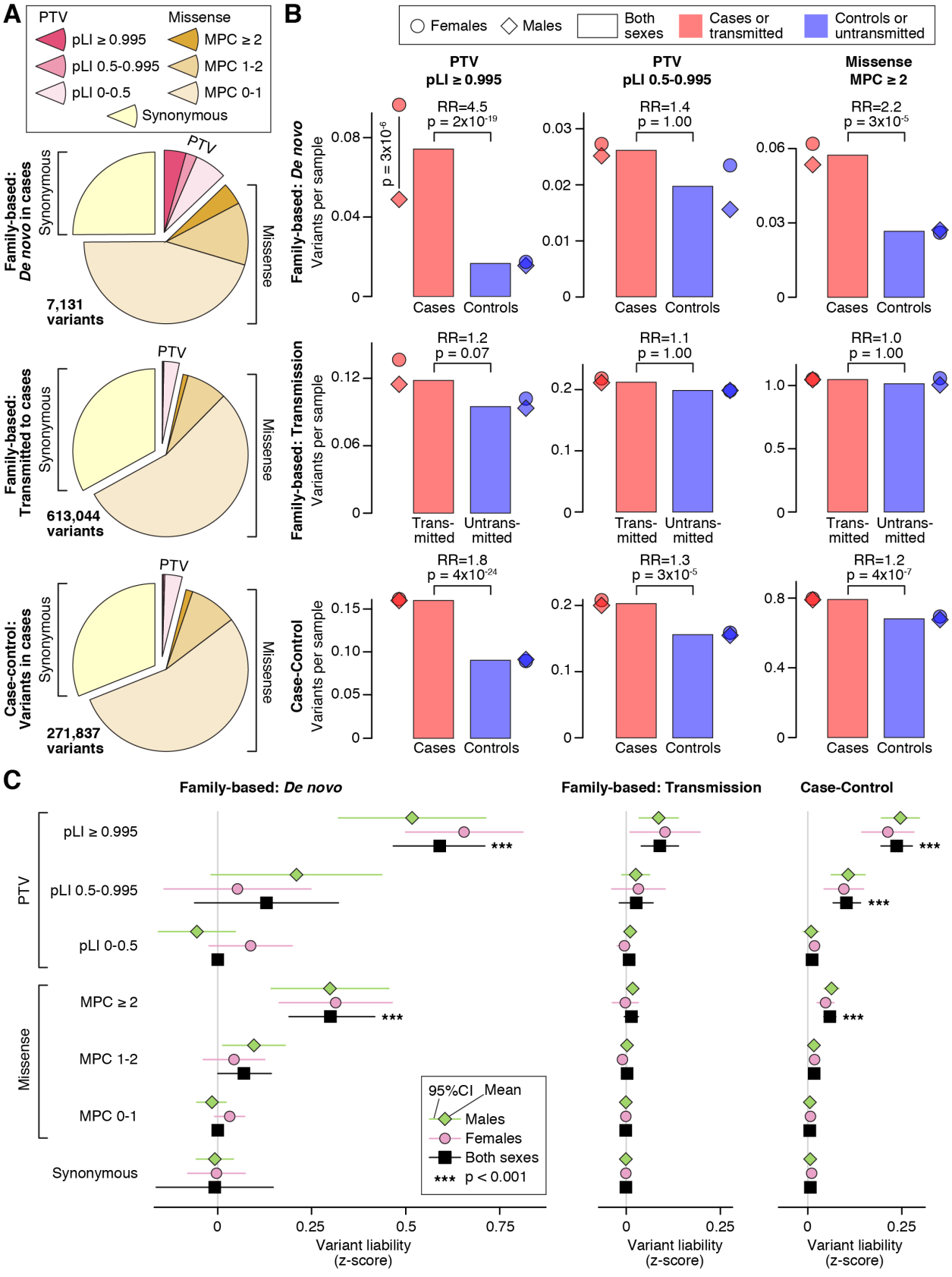
We analyzed whole-exome sequence data from 35,584 samples that passed our quality control procedures (Materials and Methods). This included 21,219 family-based samples (6,430 ASD cases, 2,179 unaffected controls, and both of their parents) and 14,365 case-control samples (5,556 ASD cases, 8,809 controls). Of these, 17,462 samples were either newly sequenced by our consortium (6,197 samples: 1,908 probands with parents; 274 ASD cases; 25 controls) or newly incorporated (11,265 samples: 416 probands with parents; 4,811 ASD cases and 5,214 controls from the Danish iPSYCH study[10]).

From this cohort, we identified a set of 9,345 rare *de novo* variants in protein-coding exons (allele frequency ≤ 0.1% in our dataset as well as in the non-psychiatric subsets of the reference databases ExAC and gnomAD, with 63% of probands and 59% of unaffected offspring carrying at least one such rare coding *de novo* variant—4,073 out of 6,430 and 1,294 out of 2,179, respectively). For rare inherited and case-control analyses, we included variants with an allele count no greater than five in our dataset and in the non-psychiatric subset of ExAC[11,12].

Impact of genetic variants on ASD risk

The differential burden of genetic variants carried by cases versus controls reflects the average liability they impart for ASD. For example, because protein-truncating variants (PTVs, consisting of nonsense, frameshift, and essential splice site variants) show a greater difference in burden between ASD cases and controls than missense variants, their average impact on liability must be larger[6]. Recent analyses have shown that measures of functional severity, such as the "probability of loss-of-function intolerance" (pLI) score[11,12] and the integrated "missense badness, PolyPhen-2, constraint" (MPC) score[13], can further delineate variant classes with higher burden. Therefore, we divided the list of rare autosomal genetic variants into seven tiers of predicted functional severity—three tiers for PTVs by pLI score (≥0.995, 0.5-0.995, 0-0.5), in order of decreasing expected impact; three tiers for missense variants by MPC score (≥2, 1-2, 0-1), also in order of decreasing impact; and a single tier for synonymous variants, expected to have minimal impact. We further divided the variants by their inheritance pattern: *de novo*, inherited, and case-control. Unlike inherited variants, *de novo* mutations are exposed to minimal selective pressure and have the potential to mediate substantial risk to disorders that limit fecundity, including ASD[14]. This expectation is borne out by the substantially higher proportions of all three PTV tiers and the two most severe missense variant tiers in *de novo* variants compared to inherited variants (**Figure 3.1A**).

Figure 3.1: Distribution of rare autosomal protein-coding variants in ASD cases and controls. **A**, The proportion of rare autosomal genetic variants split by predicted functional consequences, represented by color, is displayed for family-based data (split into *de novo* and inherited variants) and case-control data. PTVs and missense variants are split into three tiers of predicted functional severity, represented by shade, based on the pLI and MPC metrics, respectively. **B**, The relative difference in variant frequency (i.e. burden) between ASD cases and controls (top and bottom) or transmitted and untransmitted parental variants (middle) is shown for the top two tiers of functional severity for PTVs (left and center) and the top tier of functional severity for missense variants (right). Next to the bar plot, the same data are shown divided by sex. **C**, The relative difference in variant frequency shown in 'B' is converted to a trait liability z-score, split by the same subsets used in 'A'. For context, a z-score of 2.18 would shift an individual from the population mean to the top 1.69% of the population (equivalent to an ASD threshold based on 1 in 68 children[15]). No significant difference in liability was observed between males and females for any analysis. Statistical tests: B, C: Binomial Exact Test (BET) for most contrasts; exceptions were "both" and "case-control", for which Fisher's method for combining BET p-values for each sex and, for case-control, each population, was used; *P*-values corrected for 168 tests are shown.


A
PTV
pLI ≥ 0.995
pLI 0.5-0.995
pLI 0-0.5
Missense
MPC ≥ 2
MPC 1-2
MPC 0-1
Synonymous
Family-based: De novo in cases
7,131 variants
Family-based: Transmitted to cases
613,044 variants
Case-control: Variants in cases
271,837 variants
B
Females
Males
Both sexes
Cases or transmitted
Controls or untransmitted
PTV pLI ≥ 0.995
PTV pLI 0.5-0.995
Missense MPC ≥ 2
Family-based: De novo
Variants per sample
RR=4.5
p = 2x10-19
p = 3x10-6
RR=1.4
p = 1.00
RR=2.2
p = 3x10-5
Cases
Controls
Family-based: Transmission
RR=1.2
p = 0.07
RR=1.1
p = 1.00
RR=1.0
p = 1.00
Trans-mitted
Untrans-mitted
Case-Control
RR=1.8
p = 4x10-24
RR=1.3
p = 3x10-5
RR=1.2
p = 4x10-7
C
Family-based: De novo
Family-based: Transmission
Case-Control
PTV
pLI ≥ 0.995
pLI 0.5-0.995
pLI 0-0.5
Missense
MPC ≥ 2
MPC 1-2
MPC 0-1
Synonymous
***
95%CI
Mean
Males
Females
Both sexes
*** p < 0.001
Variant liability (z-score)

Comparing probands to unaffected siblings, we observe a 3.5-fold enrichment of *de novo* PTVs in the 1,447 autosomal genes with a pLI ≥ 0.995 (366 in 6,430 cases versus 35 in 2,179 controls; 0.057 vs. 0.016 variants per sample (vps); $P = 4x10^{-17}$, two-sided, two-sample Poisson exact test; **Figure 3.1B**). A less pronounced difference is observed for rare inherited PTVs in these genes, with a 1.2-fold enrichment of transmitted versus untransmitted alleles (695 transmitted versus 557 untransmitted in 5,869 parents; 0.12 vs. 0.10 vps; $P = 0.07$, binomial exact test; **Figure 3.1B**). The relative burden in the case-control data falls between the estimates for *de novo* and inherited data in these most severe PTVs, with a 1.8-fold enrichment in cases versus controls (874 in 5,556 cases versus 759 in 8,809 controls; 0.16 vs. 0.09 vps; $P = 4x10^{-24}$, binomial exact test; **Figure 3.1B**). Analysis of the middle tier of PTVs (0.5 ≤ pLI < 0.995) shows a similar, but muted, pattern (**Figure 3.1B**), while the lowest tier of PTVs (pLI < 0.5) shows no case enrichment.

De novo missense variants are observed more frequently than *de novo* PTVs and, *en masse*, they show only marginal enrichment over the rate expected by chance[3] (**Figure 3.1**). However, the most severe *de novo* missense variants (MPC ≥ 2) occur at a frequency similar to *de novo* PTVs, and we observe a 2.1-fold case enrichment (354 in 6,430 cases versus 58 in 2,179 controls; 0.055 vs. 0.027 vps; $P = 3x10^{-8}$, two-sided, two-sample Poisson exact test; **Figure 3.1B**), with a consistent 1.2-fold enrichment in the case-control data (4,277 in 5,556 cases versus 6,149 in 8,809 controls; 0.80 vs. 0.68 vps; $P = 4x10^{-7}$, binomial exact test; **Figure 3.1B**). Of note, in the *de novo* data, this top tier of missense variation shows stronger enrichment in cases than the middle tier of PTVs. The other two tiers of missense variation are not significantly enriched in cases.

Sex differences in ASD risk

The prevalence of ASD is higher in males than females. In line with previous observations of females with ASD carrying a higher genetic burden than males[3], we observe a 2-fold enrichment of *de novo* PTVs in highly constrained genes in affected females (N=1,097) versus affected males (N=5,333) ($P = 3x10^{-6}$, two-sided Poisson exact test; **Figure 3.1B**). This result is consistent with the female protective effect (FPE) model, which postulates that females require an increased genetic load (in this case, high-liability PTVs) to reach the threshold for a diagnosis[16]. The converse hypothesis is that risk variation has larger effects in males than in females so that females require a higher genetic burden to reach the same diagnostic threshold as males; however, across all classes of genetic variants, we observed no significant sex differences in trait liability, consistent with the FPE model (Materials and Methods; **Figure 3.1C**). In the absence of sex-specific differences in liability, we estimated the liability z-scores for different classes of variants across both sexes together (**Figure 3.1C**) and leveraged them to enhance gene discovery.

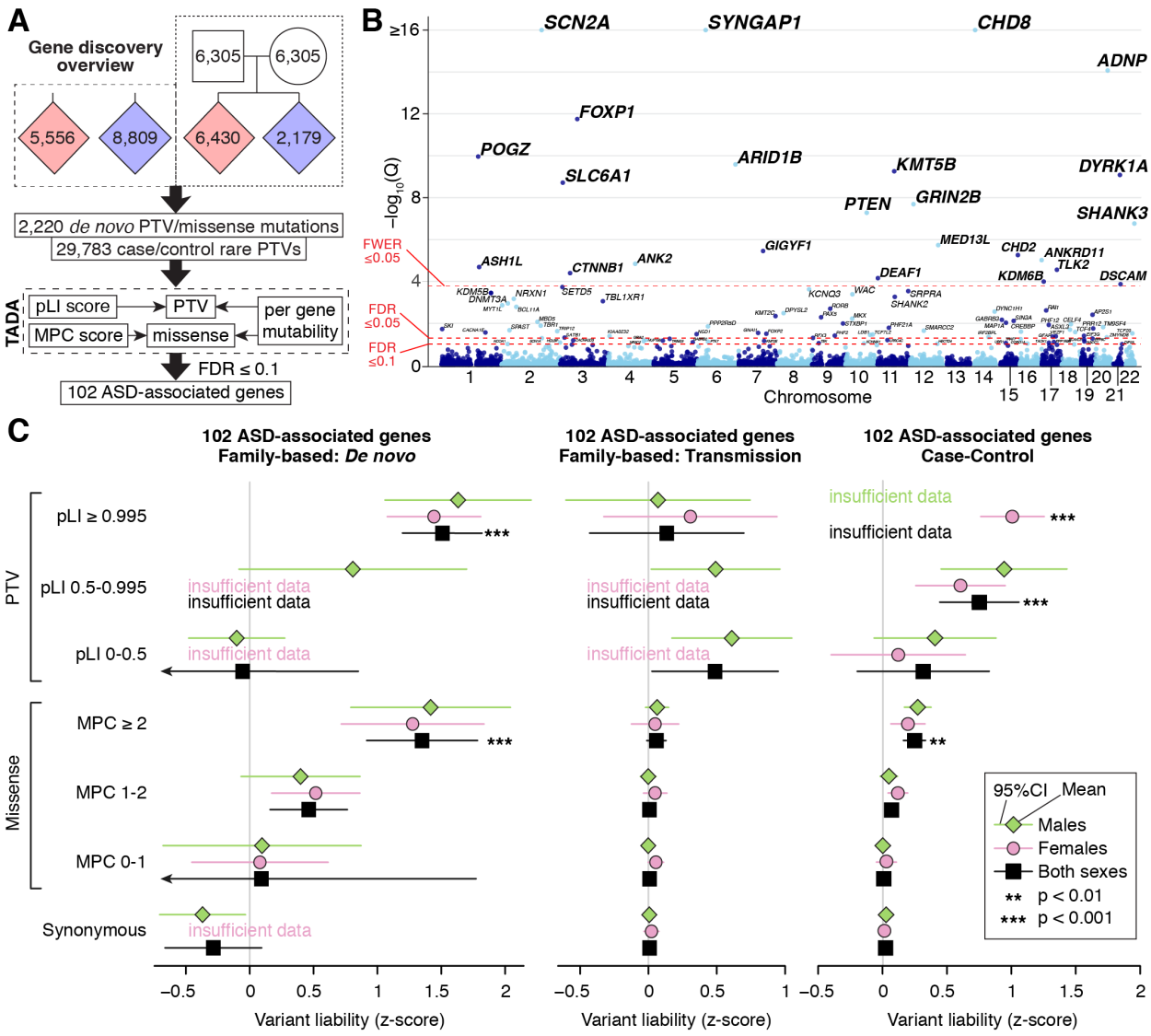


Figure 3.2. Gene discovery in the ASC cohort. A, WES data from 35,584 samples are entered into a Bayesian analysis framework (TADA) that incorporates pLI score for PTVs and MPC score for missense variants. **B**, The model identifies 102 autosomal genes associated with ASD at a false discovery rate (FDR) threshold of ≤ 0.1, which is shown on the y-axis of this Manhattan plot with each point representing a gene. Of these, 78 exceed the threshold of FDR ≤ 0.05 and 26 exceed the threshold family-wise error rate (FWER) ≤ 0.05. **C**, Repeating our ASD trait liability analysis (Figure 1C) restricted to variants observed within the 102 ASD-associated genes only. Statistical tests: B, TADA; C, Binomial Exact Test (BET) for most contrasts; exceptions were "both" and "case-control", for which Fisher's method for combining BET P-values for each sex and, for case-control, each population, was used; P-values corrected for 168 tests are shown.

Gene discovery

In previous risk gene discovery efforts, we used the Transmitted And *De novo* Association (TADA) model[6] to integrate protein-truncating and missense variants that are *de novo*, inherited, or from case-control populations and to stratify autosomal genes by FDR for association. Here, we update the TADA model to include pLI score as a continuous metric for PTVs, and MPC score as a two-tiered metric (≥2, 1-2) for missense variants (Materials and Methods). From family data we include *de novo* PTVs as well as *de novo* missense variants, while for case-control we include only PTVs; we do not include inherited variants due to the limited liabilities observed (**Figure 3.1C**). These modifications result in an enhanced TADA model with greater sensitivity and accuracy than the original model (**Figure 3.2A**; Materials and Methods).

Our refined TADA model identifies 102 ASD risk genes at FDR ≤ 0.1, of which 78 meet the more stringent threshold of FDR ≤ 0.05, with 26 significant after Bonferroni correction (**Figure 3.2B**). Of the 102 ASD-associated genes, 60 were not discovered by our earlier analyses[3-5], including 31 that have not been implicated in autosomal dominant neurodevelopmental disorders and were not significantly enriched for *de novo* and/or rare variants in previous studies, and that can therefore be considered novel. The patterns of liability seen for these 102 genes are similar to that seen over all genes, although the effects of variants are uniformly larger, as would be expected for this selected list of putative risk genes that would be enriched for true risk variants.

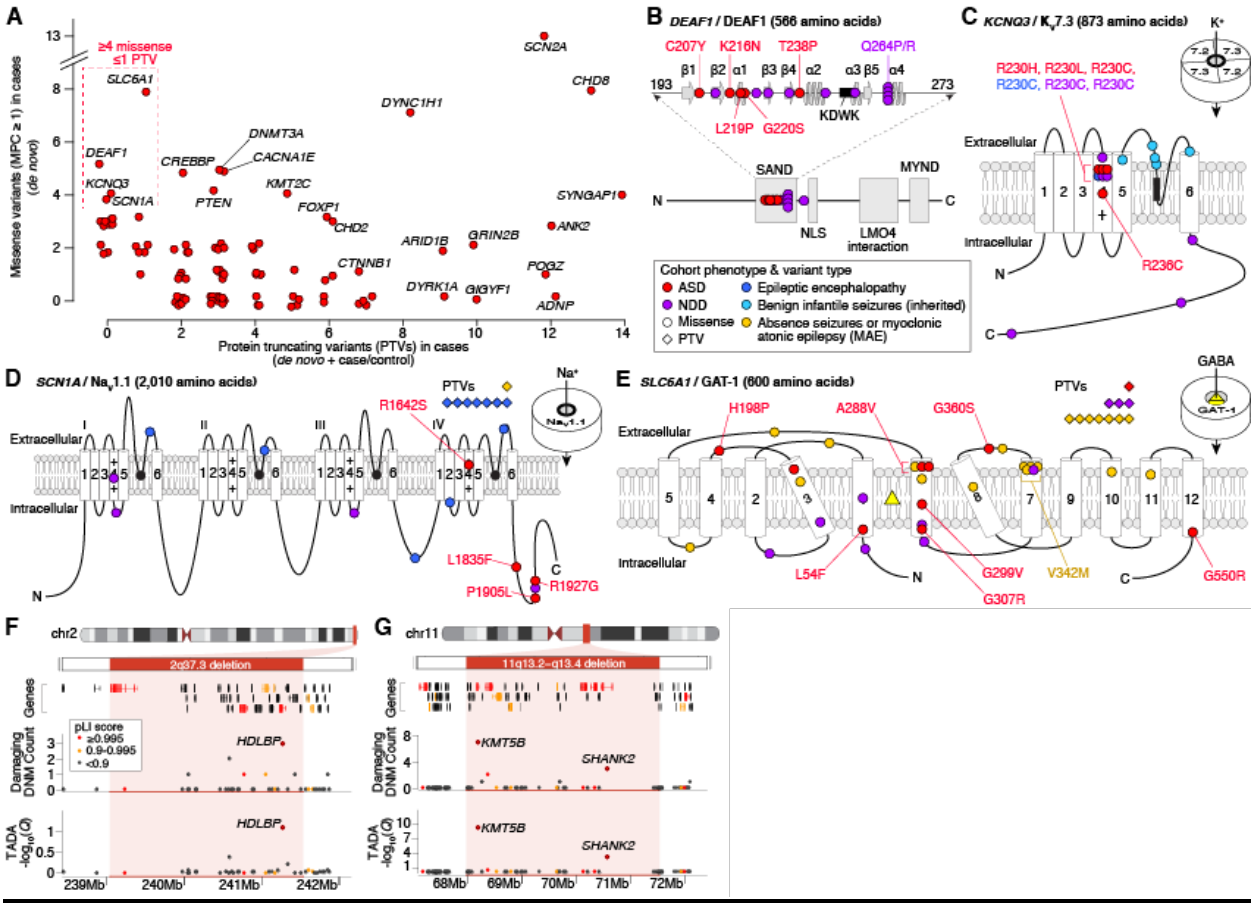


Figure 3.3: Genetic characterization of ASD genes. **A**, Count of PTVs versus missense variants (MPC ≥ 1) in cases for each ASD-associated gene (red points, selected genes labeled). These counts reflect the data used by TADA for association analysis: *de novo* and case/control data for PTVs; only *de novo* for missense. **B**, Location of ASD *de novo* missense variants in *DEAF1*. The five ASD mutations (marked in red) are in the SAND DNA-binding domain (amino acids 193-273, spirals show alpha helices, arrows show beta sheets, *KDWK* is the DNA-binding motif) alongside ten NDD variants, several reduce DNA binding, including Q264P and Q264R[17-19]. **C**, Location of ASD missense variants in *KCNQ3*. All four ASD variants resided in the voltage sensor (fourth of six transmembrane domains), with three in the same residue (R230), including the gain-of-function R230C mutation observed in NDD[19]. Five inherited variants observed in benign infantile seizures reside in the pore loop[20,21]. **D**, Location of ASD missense variants in *SCN1A*, alongside 17 NDD and epilepsy *de novo* variants[19]. **E**, Location of ASD missense variants in *SLC6A1*, alongside 31 NDD and epilepsy *de novo* variants[19,22]. **F**, Subtelomeric 2q37 deletions are associated with facial dysmorphisms, brachydactyly, high BMI, neurodevelopmental delay, and ASD[23]. While three genes within the locus have a pLI score ≥ 0.995, only *HDLBP* is associated with ASD. **G**, Deletions at the 11q13.2−q13.4 locus have been observed in NDD, ASD, and otodental dysplasia[24,25]. Five genes within the locus have a pLI score ≥ 0.995, including two ASD genes: *KMT5B* and *SHANK2*. Statistical tests: F, G, TADA

Patterns of mutations in ASD genes

Within the set of observed mutations, the ratio of PTVs to missense mutations varies substantially between genes (**Figure 3.3A**). Some genes, such as *ADNP*, reach our association threshold through PTVs alone, amongst which three genes have a significant excess of PTVs, relative to missense mutations, in the current dataset, based on gene mutability: *SYNGAP1*, *DYRK1A*, and *ARID1B* ($P < 0.0005$, binomial test). Because of the increased cohort size and availability of the MPC metric, we are also able for the first time to associate genes with ASD based primarily on *de novo* missense variation. We therefore examined four genes with four or more *de novo* missense variants (MPC ≥ 1) in ASD cases and one or no PTVs: *DEAF1*, *KCNQ3*, *SCN1A*, and *SLC6A1* (**Figure 3.3A**).

We observe five *de novo* missense variants and no PTVs in *DEAF1*, which encodes a self-dimerizing transcription factor involved in neuronal differentiation[26]. All five missense variants reside in the SAND domain (**Figure 3.3B**), which is critical for both dimerization and DNA binding[26,27]. A similar pattern of SAND domain missense enrichment is observed in individuals with intellectual disability, speech delay, and behavioral abnormalities[17-19].

Four *de novo* missense variants and no PTVs are observed in *KCNQ3*, which encodes a subunit of a neuronal voltage-gated potassium channel (**Figure 3.3C**). All four variants modify arginine residues in the voltage-sensing fourth transmembrane domain, with three at a single residue previously characterized as gain-of-function in NDD (R230C, **Figure 3.3C**)[28]. These data suggest gain-of-function mutations in *KCNQ3* also confer risk to ASD.

SCN1A encodes a voltage-gated sodium channel and has been associated, predominantly through PTVs, with Dravet syndrome[29], a form of progressive epileptic encephalopathy which often meets diagnostic criteria for ASD[30]. We observe four *de novo* missense variants and no

PTVs in *SCN1A* (**Figure 3.3A**), with three located in the C-terminus (**Figure 3.3D**), and all four cases are reported to have seizures.

The gene *SLC6A1* encodes a voltage-gated GABA transporter and has been associated with developmental delay and cognitive impairment[19,31], as well as myoclonic atonic epilepsy and absence seizures[22]. Here, we extend the phenotypic spectrum to include ASD, through the observation of eight *de novo* missense variants and one PTV, all in cases (**Figure 3.3E**). Four of these variants reside in the sixth transmembrane domain, with one recurring in two independent cases (A288V). Five of the six cases with available information on history of seizure had seizures, and all four cases with available data on cognitive performance have intellectual disability.

ASD genes within recurrent copy number variants (CNVs)

Large CNVs represent another important source of risk for ASD[32], but these genomic disorder (GD) segments can include dozens of genes, complicating the identification of driver gene(s) within these regions. We sought to determine whether the 102 ASD genes could nominate driver genes within GD regions. We first curated a consensus GD list from nine sources, totaling 823 protein-coding genes in 51 autosomal GD loci associated with ASD or ASD-related phenotypes, including NDD.

Within the 51 GDs, 12 GD loci encompassed 13 ASD-associated genes, which is greater than expected by chance when simultaneously controlling for number of genes, PTV mutation rate, and brain expression levels per gene (2.3-fold increase; $P = 2.3 \times 10^{-3}$, permutation). These 12 GD loci divided into three groups: 1) the overlapping ASD gene matched the consensus driver gene, e.g., *SHANK3* for Phelan-McDermid syndrome[33]; 2) an ASD gene emerged that did

not match the previously predicted driver gene(s) within the region, such as *HDLBP* at 2q37.3 (**Figure 3.3F**), where *HDAC4* has been hypothesized as a driver gene in some a analyses[34]; and 3) no previous driver gene had been established within the GD locus, such as *BCL11A* at 2p15-p16.1. One GD locus, 11q13.2-q13.4, had two of our 102 genes (*SHANK2* and *KMT5B*, **Figure 3.3G**), highlighting that GDs can result from risk conferred by multiple genes, potentially including genes with small effect sizes that we are underpowered to detect.

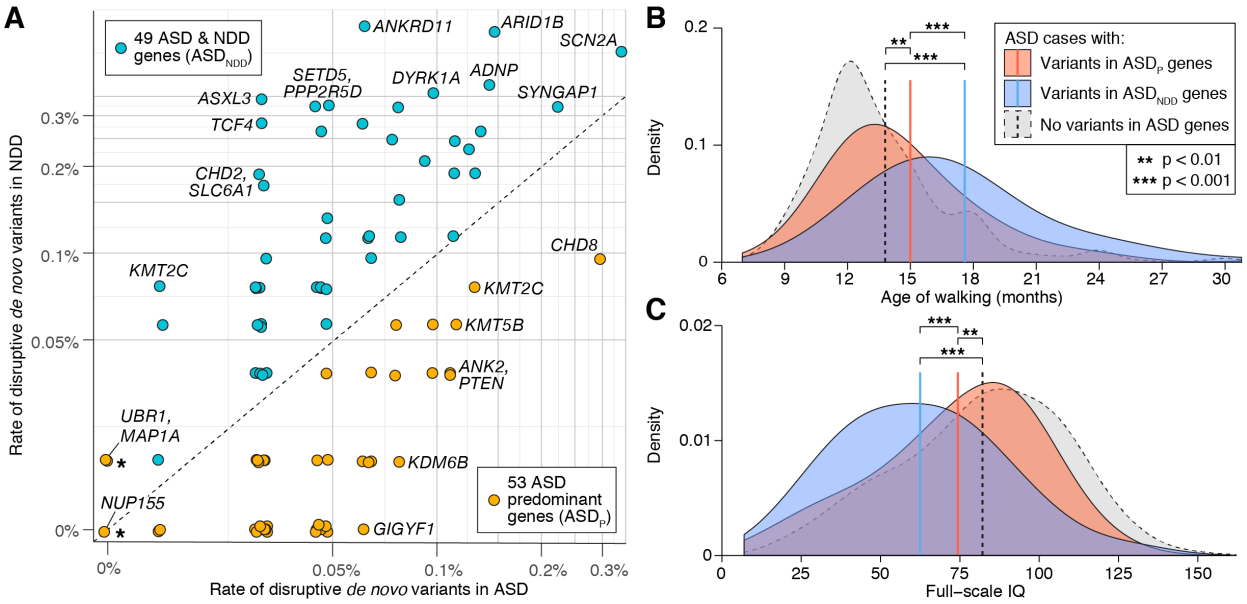


Figure 3.4: Phenotypic and functional categories of ASD-associated genes. **A**, The frequency of disruptive *de novo* variants (e.g. PTVs or missense variants with MPC ≥ 1) in ASD-ascertained and NDD-ascertained cohorts is shown for the 102 ASD-associated genes (selected genes labeled). Fifty genes with a higher frequency in ASD are designated ASD-predominant (ASD_P), while the 49 genes more frequently mutated in NDD are designated as ASD_{NDD}. Three genes marked with a star (*UBR1*, *MAP1A*, and *NUP155*) are included in the ASD_P category on the basis of case-control data, which are not shown in this plot. **B**, ASD cases with disruptive *de novo* variants in ASD genes show delayed walking compared to ASD cases without such *de novo* variants, and the effect is greater for those with disruptive *de novo* variants in ASD_{NDD} genes. **C**, Similarly, cases with disruptive de novo variants in ASD_{NDD} genes and, to a lesser extent, ASD_P genes have a lower full-scale IQ than other ASD cases. Statistical tests: B, C, t-test.

<u>Relationship between ASD and other neurodevelopmental disorders</u>

Family studies yield high heritability estimates in ASD[35], but comparable estimates of heritability in severe NDD are lower[36]. Consistent with these observations, exome studies identify a higher frequency of disruptive *de novo* variants in severe NDD than in ASD[31]. Because of the 30% co-morbidity between ASD subjects and intellectual disability/NDD, it is unsurprising that many genes are associated with both disorders[37]. Distinguishing genes that, when disrupted, lead to ASD more frequently than NDD may shed new light on how atypical neurodevelopment maps onto the core deficits of ASD.

To partition the 102 ASD genes in this manner, we compiled data from 5,264 trios ascertained for severe NDD and compared the relative frequency, *R*, of disruptive *de novo* variants (which we define as PTVs or missense variants with MPC ≥ 1) in ASD- or NDD-ascertained trios. Genes with $R > 1$ were classified as ASD-predominant (ASD_P, 50 genes), while those with $R < 1$ were classified as ASD with NDD (ASD_{NDD}, 49 genes). An additional three genes were assigned to the ASD_P group on the basis of case-control data, totaling 53 ASD_P genes (**Figure 3.4A**). For this partition, we then evaluated transmission of rare PTVs (relative frequency < 0.001) from parents to their affected offspring: for ASD_P genes, 44 such PTVs were transmitted and 18 were not ($P = 0.001$, transmission disequilibrium test [TDT]), whereas, for ASD_{NDD} genes, 14 were transmitted and 8 were not ($P = 0.29$; TDT). The frequency of PTVs in parents is significantly greater in ASD_P genes (1.17 per gene) than in ASD_{NDD} genes (0.45 per gene; $P = 6.6 \times 10^{-6}$, binomial test), while the frequency of *de novo* PTVs in probands is not markedly different between the two groups (95 in ASD_P genes, 121 in ASD_{NDD} genes, $P = 0.07$, binomial test with probability of success = 0.503 [PTV in ASD_P gene]). The paucity of inherited PTVs in ASD_{NDD} genes is consistent with greater selective pressure acting against disruptive variants in these genes.

Consistent with this partition, ASD subjects who carry disruptive *de novo* variants in ASD_{NDD} genes walk 2.6 ± 1.2 months later (**Figure 3.4B**; $P = 2.3x10^{-5}$, two-sided, two-sample t-test, df=251) and have an IQ 11.9 ± 6.0 points lower (**Figure 3.4C**; $P = 1.1x10^{-4}$, two-sided, two-sample t-test, df=278), on average, than ASD subjects with disruptive *de novo* variants in ASD_P genes. Both sets of subjects differ significantly from the rest of the cohort with respect to IQ and age of walking (**Figures 3.4B**, **3.4C**). Thus, the data support some overall distinction between the genes identified in ASD and NDD *en masse*, although our current analyses are not powered for variant-level or gene-level resolution.

Discussion

By characterizing rare *de novo* and inherited coding variation from 35,584 individuals, including 11,986 ASD cases, we implicate 102 genes in risk for ASD at FDR ≤ 0.1, of which 31 are novel risk genes. Notably, analyses of this set of risk genes lead to novel genetic, phenotypic, and functional findings. Evidence for several of the genes is driven by missense variants, including confirmed gain-of-function mutations in the potassium channel *KCNQ3* and patterns that may similarly reflect gain-of-function or altered function in *DEAF1*, *SCN1A*, and *SLC6A1*. Further, we strengthen evidence for driver genes in genomic disorder loci and we propose a new driver gene (*BCL11A*) for the recurrent CNV at 2p15-p16.1.

We perform a genetic partition between genes predominantly conferring liability for ASD (ASD_P) and genes imparting risk to both ASD and NDD (ASD_{NDD}). Two lines of evidence support the partition: first, cognitive impairment and motor delay are more frequently observed in our subjects—all ascertained for ASD—with mutations in ASD_{NDD} than in ASD_P genes; second, we find that inherited variation plays a lesser role in ASD_{NDD} than in ASD_P genes.

Together, these observations indicate that ASD-associated genes are distributed across a spectrum of phenotypes and selective pressure. At one extreme, gene haploinsufficiency leads to global developmental delay, with impaired cognitive, social, and gross motor skills leading to strong negative selection (e.g. *ANKRD11*, *ARID1B*). At the other extreme, gene haploinsufficiency leads to ASD, but there is a more modest involvement of other developmental phenotypes and selective pressure (e.g. *GIGYF1*, *ANK2*). This distinction has important ramifications for clinicians, geneticists, and neuroscientists, because it suggests that clearly delineating the impact of these genes across neurodevelopmental dimensions may offer a route to deconvolve the social dysfunction and repetitive behaviors that define ASD from more general neurodevelopmental impairment. Larger cohorts will be required to reliably identify specific genes as being enriched in ASD compared to NDD. ASD must arise by phenotypic convergence amongst these diverse neurobiological trajectories, and further dissecting the nature of this convergence, especially in the genes that we have identified herein, is likely to hold the key to understanding the neurobiology that underlies the ASD phenotype.

Materials and methods

Samples

The Autism Sequencing Consortium is a large-scale genomic consortium collecting and sequencing cohorts worldwide[38]. The analysis presented here drew from 35,584 samples collected from 32 distinct sample sets. These include cohorts sequenced by the Autism Sequencing Consortium (ASC) and published in our first[3] or second study[39] (Germany, Japan, PAGES, Pittsburgh, Seaver, Spain, TASC, and UCSF), as well as new collections (Boston, Brazil, CHARGE, Chicago, Hong Kong, Miami, Portugal, Rome, Siena, Turin, UC Irvine, and

Utah), with a total of 6,197 newly collected and sequenced samples included in our final analysis. We also sequenced samples from the Autism Genetic Resource Exchange (AGRE), the Boston Autism Consortium, two sites in Finland, and Swedish controls from epidemiological studies in schizophrenia and bipolar disorder. We imported exome sequence data from the Simons Simplex Collection[4], as well as an unpublished Norwegian cohort, and included them in our dataset alongside ASC-sequenced samples. In addition, we incorporated published *de novo* variants from the UK10K consortium, the University of Pennsylvania, Vanderbilt University, and a collection of samples from the Middle East. Finally, we integrated gene-level variant counts from autism cases and matched controls from the iPSYCH research initiative[10].

The bulk of new ASC samples were sequenced at the Broad Institute on Illumina HiSeq sequencers using the Illumina Nextera exome capture kit. The remainder were sequenced at three other sites: the University of California, San Franciso (N=495), the Sanger Institute (N=443), and Johns Hopkins University (N=302), all using similar methods. Each sample's sequencing reads were aggregated into a BAM file and processed through a pipeline based on the Picard set of software tools. The BWA aligner mapped reads onto the human genome build 37 (hg19). Single nucleotide polymorphism (SNPs) and insertions / deletions (indels) were jointly called across all samples using Genome Analysis Toolkit (GATK[40]) HaplotypeCaller package version 3.4. Variant call accuracy was estimated using the GATK Variant Quality Score Recalibration (VQSR) approach. The VCF file was produced by the Broad sequencing and calling pipeline with GATK version 3.4 (g3c929b0) and was itself VCF format v4.1.

Dataset QC

The VCF file, containing approximately 29,000 exomes, was loaded into Hail 0.1 (http://hail.is; https://github.com/hail-is/hail) to perform basic quality control steps. Multi-allelic sites were split into bi-allelic sites and each variant was then annotated with the Variant Effect Predictor (VEP)[41] by prioritizing coding canonical transcripts. VEP assigned properties such as gene name and consequence to each variant.

To check the accuracy of reported pedigree information, relatedness was calculated between each pair of samples using Hail's ibd() function and sex was imputed for each sample using Hail's impute_sex() function. The relatedness values were input into the program PRIMUS, which inferred pedigree structure for every related group of samples. Combined with the imputed sex, these inferred pedigrees were compared to reported pedigrees and checked for discrepancies. Obvious errors in reporting were fixed (e.g., swapped mother and father labels in the same family, or swapped parent/proband labels in the same trio), and samples with a discrepancy that could not be resolved (~200) were dropped. Parents without a child in the dataset (~250) were also dropped, resulting in 28,547 samples and 5,420,608 variants.

During a first round of variant quality control (QC), low-complexity regions were removed (110,963 variants), as were SNPs that failed variant quality score recalibration (VQSR, 265,130 variants), leaving 5,044,515 variants. For genotype QC, several genotype filters were applied: we filtered calls with a depth less than 10 or greater than 1000; for homozygous reference calls, we filtered genotypes with less than 90% of the read depth supporting the reference allele or with a genotype quality less than 25; for homozygous variant calls, we filtered genotypes with less than 90% of the read depth supporting the alternate allele or with a phred-scaled likelihood (PL) of being homozygous reference less than 25; and for heterozygous calls, we filtered genotypes with less than 90% of the read depth supporting either the reference or

alternate allele, with a PL of being homozygous reference less than 25, with less than 25% of the read depth supporting the alternate allele (i.e. an allele balance less than 0.25), or with a probability of the allele balance (calculated from a binomial distribution centered on 0.5) less than $1x10^{-9}$. We additionally filtered any heterozygous call in the X or Y non-pseudoautosomal regions in a sample that imputed as male. For samples imputed as female, calls from the Y chromosome were removed. After applying these filters and removing sites that were no longer variant, the dataset contained 28,547 samples and 4,755,048 variants.

Next, we applied sample quality control filters, removing samples with estimated contamination levels >7.5% (20 samples) or chimeric reads >7.5% (121 samples). Stratifying samples into 18 different groups (by exome capture/year/cohort/sequencing center), samples were filtered if their call rate was greater than 3 standard deviations from the group mean (300 samples). Duplicate samples were then removed (761 samples), as were samples for which the imputed sex did not match the reported sex (59 samples). Following these sample filters, family structures were re-evaluated: if one or more parents of a proband had been filtered, the proband was reclassified as a case and the remaining parent (if any) was dropped; if the proband had an unaffected sibling, the sibling was kept as a "sibling of case;" if one or more parents were filtered and no proband remained, then data for remaining family members were dropped; second degree or greater relatives of probands were also dropped. After applying these rules, the dataset contained 5833 complete families, with 5924 affected probands, 2007 unaffected offspring, 5834 fathers, and 5833 mothers (one family contained two probands, two fathers, and one mother).

The dataset also contained 2388 cases, 106 siblings of cases, and 4324 controls, none of whom were part of a complete trio. From these categories, we filtered one of each related pair of samples (although each case was allowed to keep 1 sibling in the event this became interesting

for future study). We defined related samples as a pair of samples with a KING[42] kinship value of 0.1 or greater, approximately corresponding to a PI_HAT value of 0.2 or greater. Following this filtering, the dataset contained 2353 cases, 100 siblings of cases, and 4316 controls, for a total of 26,367 samples.

After filtering sites that were no longer variant, there were 4,605,130 variants. For a second round of variant QC, variants with call rate <10% (17,083 variants) or a Hardy-Weinberg equilibrium p-value less than $1x10^{-12}$ (27,862 variants) were filtered, leaving 26,367 samples and 4,560,185 variants. This dataset was then used as the starting point for the *de novo*, inherited, and case-control workflows. Most of the remaining samples were ultimately used in our TADA analysis, but some were subject to additional filtration during these workflows.

Tallying of variant classes

De novo variants were called from the 26,367-sample dataset described above, including 5924 affected probands and 2007 unaffected offspring. After filtering any genotype with a GQ < 25, *de novo* variants were called using the de_novo() function of Hail 0.1, which implements the caller used in previous ASC work (https://github.com/ksamocha/de_novo_scripts). Population allele frequencies for variants were obtained from the non-psychiatric subset of gnomAD (http://gnomad.broadinstitute.org/), and these frequencies were used as the input priors. As additional parameters, parents' homozygous reference genotypes were required to have no more than 3% of reads supporting the alternate allele, children's heterozygous calls were required to have at least 30% of reads supporting the alternate allele, and the ratio of child read depth to parental read depth was required to be at least 0.3.

This process identified 44,562 *de novo* variants (26,577 distinct variants) in the 7931 children in the dataset. Of the 7931 children, 519 were part of a whole genome sequencing project[43], and we added 168 *de novo* variants called in these samples from the whole genome sequencing that were not called in the exome sequencing. We also incorporated 338 previously published and validated *de novo* variants in our samples that were not identified by our caller[11]. Thus, in total, we had 45,068 *de novo* variants (27,083 distinct variants) in 7931 children. For QC on the *de novo* variants, we retained variants if they were high confidence as indicated by the calling algorithm, medium confidence and a singleton in the dataset, or previously experimentally validated (removed 20,862 calls). To filter calls stemming from cell line artifacts, an allele balance of at least 0.4 was required for samples from immortalized cell lines (773 probands and 40 siblings) (removed 2171 calls). Next, a call was removed if it had an allele frequency >0.1% in our dataset, in ExAC (r0.3, non-psychiatric subset, http://exac.broadinstitute.org/), or in gnomAD (non-psychiatric subset) (removed 5068 calls). Calls were removed if they appeared more than twice (removed 403 calls) and were then limited to one variant per person per gene (removed 570 calls), retaining calls with the most severe consequence when selecting which one to keep. Finally, samples whose DNA source was whole blood or saliva were dropped if they had more than 7 coding variants (removed 20/5143 probands and 13/1967 unaffected children), and samples whose DNA source was immortalized cell lines were dropped if they had more than 5 coding calls (filtered 35/773 probands and 1/40 unaffected children). Retained were 14,569 *de novo* variant calls from 5869 probands and 1993 unaffected children. To maximize power, we supplemented this set with 933 and 287 published *de novo* variants in 561 probands and 186 siblings[3,5,11,44], respectively, for whom original sequence data were not available.

Inherited variation

QC for inherited variants began with the dataset of 26,367 samples and 4,560,185 variants. Any genotype call with a GQ < 25 was removed and heterozygous genotypes were required to have an allele balance ≥ 0.3. Variants were required to have a call rate ≥90%, insertions and deletions were required to pass VQSR, and SNPs were required to have a VQSLOD (variant quality score log odds) ≥ -2.085. The VQSLOD threshold for SNPs was determined by identifying the threshold at which synonymous variants with an allele count of 1 amongst parents in the dataset were transmitted to the child 50% of the time, as described previously[11,12]. Protein-truncating variants were required to be high confidence ("HC") by the LOFTEE plugin for VEP and to have no LOFTEE flags other than "SINGLE_EXON".

For purposes of gene-level counts, variants were tallied in the 5869 probands and 1993 unaffected children who passed *de novo* QC. Variants were required to have an allele count ≤ 5 in the combined parents, cases, and controls (18,153 people) in our dataset, as well as an allele count ≤ 5 in the non-psychiatric subset of ExAC.

Case-control variation

Variants in ASC cases and controls were QC'd in the same way as inherited variants. For purposes of gene-level counts, variants were again required to have an allele count ≤ 5 in the 18,153 combined parents, cases, and controls in the dataset, as well as an allele count ≤ 5 in the non-psychiatric subset of ExAC.

To ensure well-matched cases and controls, probable ancestry was calculated by merging our raw dataset with genotypes from the 1000 Genomes Project and conducting principal

components analysis (PCA) in Hail on a set of ~5000 common SNPs. A naive Bayes classifier was trained (using the naiveBayes function from the R package e1071) on the 1000 Genomes samples labeled as either European or East Asian and used to predict which of our samples fell into those populations. Synonymous rates were well-matched between cases and controls from the Swedish contributing site which were classified European (745 cases and 3595 controls), as well as between cases and controls from the Japanese contributing site which were classified East Asian (196 cases and 298 controls). For inclusion in TADA, we counted variants from the 4340 Swedish samples. Overall variant rates were higher in the Japanese samples than the Swedish samples, possibly because our filtering was based on allele counts in ExAC, and ExAC has less representation from East Asian samples than European ones.

Analysis of variant classes

To model a qualitative trait—in this case, the presence or absence of ASD—using standard quantitative genetics concepts, we imagine that there is an unobserved, normally distributed variable called "liability" that determines whether or not an individual is diagnosed with ASD. We assume that liability, L, has mean 0 and variance 1 in the general population. Individuals with L greater than some threshold t are diagnosed with ASD and individuals with $L < t$ are considered "typical". Under this model, the prevalence difference between males and females is viewed as a difference in thresholds for males and females. For a male to be diagnosed with ASD, his liability must be larger than t_m. For a female to be diagnosed with ASD her liability must be larger than t_f. Since ASD is more common in males than females, we conclude that $t_m < t_f$. For all that follows we will assume that the prevalence of ASD, Ψ_m, is 1 in 42 in males (implying $t_m \sim 1.98$), and the prevalence of ASD is 1 in 189 females, Ψ_f (implying $t_f \sim$

2.56). We model ASD+ID similarly, but with lower prevalence than all ASD (male prevalence 0.00499, and female 0.00138).

When considering the effects of individual alleles on liability, we employ an elaboration to the standard quantitative genetics model, which is sometimes called the "mixed model of inheritance". We assume that individual alleles make additive contributions to liability, so that for some allele, A_1, individuals with 0 copies of the allele have mean $-\mu$, variance 1 liability, but individuals with 1 copy have mean $\alpha - \mu$, variance 1, and individuals with 2 copies have mean $2\alpha - \mu$, variance 1 liability. Assuming Hardy-Weinberg equilibrium for genotypes, and the frequency of A_1 equaling p, $\mu = 2\alpha p^2 + \alpha 2pq = 2\alpha p$. Here μ is a normalizing factor to ensure the overall population has mean liability 0.

For several of our analyses we are interested in the effect, α, for variants of a particular type in a collection of genes, for instance *de novo* PTVs in genes with pLI scores > 0.9. If a variant is individually exceptionally rare, we have virtually no power to estimate its individual effect size, but over a large collection of such variants average properties are estimable. To do so, we model the entire collection of variants as if there were a single allele with frequency equal to the sum of the individual variant frequencies. This approach makes little sense for common variants, but for sufficiently rare variants, where single individuals seldom harbor more than one, this is a reasonable and helpful approximation. For some variant types, however, such as silent variants, the count of alleles can be substantial. For this reason, rather than standardize by $2N$, where N is the number of subjects, we standardize by $2NM$, where $M = 17{,}484$ is the number of genes analyzed herein. This standardization has no material impact on calculations of parameters of interest. To distinguish between cases and controls, we write N_{ca} and N_{co} respectively.

Thus, for each type of variant we are interested in studying, *de novo* PTV mutations, say, we count the number of observations of this class of variant in cases (our probands in trios), and the number of observations of this class of variant in controls (our siblings in trios). For a given type of variant, V, we call $Pr\{V|D\}$ the frequency of this type of variant in cases (observed number of variants divided by $2NM$), and $Pr\{V|\neg D\}$ the corresponding value in controls. We make these calculations separately in males and females, which we denote as $Pr\{V_m|D_m\}$, $Pr\{V_f|D_f\}$, $Pr\{V_m|\neg D_m\}$, and $Pr\{V_f|\neg D_f\}$ where the m and f subscripts distinguish male and females. The overall frequency of the variant class can be found by

$$Pr\{V_g\} = Pr\{V_g|D_g\}\Psi_g + Pr\{V_g|\neg D_g\}(1 - \Psi_g)$$

where g can be either f or m, for females and males, respectively. From this the Penetrance (probability of disease given variant) of the variant class can be found immediately by Bayes rule

$$Pr\{D_g|V_g\} = \frac{Pr\{V_g|D_g\}\Psi_g}{Pr\{V_g\}}.$$

To find the average effect, α_{V_g}, of this variant class we note

$$Pr\{D_g|V_g\} = \int_{t_g-\alpha_{V_g}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

Thus, we can find the effect size by inverting a standard normal cumulative distribution, $\Phi(x)$

$$\alpha_{V_g} = t_g - \Phi^{-1}(1 - Pr\{D_g|V_g\}).$$

Empirically, the relative risk for the variant type is calculated as $Pr\{V_m|D\}/Pr\{V|\neg D\}$ for the contrast of cases versus controls. To assess whether or not there is any difference in this variant class between cases and controls, we perform an exact Binomial Test on the underlying observed counts, where the probability of success is given by $N_{ca}/(N_{ca} + N_{co})$. The odds ratio is

computed from four observations, the number of variants of the risk class in cases, a; the number of variants of the risk class in controls, b; the number of alleles not in the risk class in cases $2N_{ca}M - a$; and the parallel calculation for controls, $2N_{co}M - b$.

To estimate a confidence interval of α_{V_g}, we note that in a very formal sense α_{V_g} is the average effect on the liability scale of the variant. Were we able to observe those effects directly, we could have calculated the observed mean and standard error of those effects. Because we cannot observe liability directly here, we infer the standard error of α_{V_g} by the following procedure: map the p-value from the Binomial test, described above, onto an equivalent z-value from the normal distribution, z; then α_{V_g}/z is a reasonable estimator for the standard error of the estimator for α_{V_g}.

Calculations for "All Genes" and for "Other Genes" were performed separately for males and females and also separately for the PAGES and DBS samples. Inherited analysis calculations were also separated by male / female and by proband / sibling. To combine effects between males and females, we took inverse-variance weighted averages of male and female effect sizes. We performed analogous calculations for the populations of case-control samples. For these calculations for the 102 ASD genes, however, because the counts of events were often small, we combined data over males and females and over PAGES and DBS samples to compute overall parameters (i.e., performed mega- versus meta-analysis). When parameters could not be estimated, this is noted as NA.

Transmission And De novo Association test (TADA)

Published analyses of WES data using TADA have evaluated two categories of rare variation, namely protein truncating variants (PTV; i.e., frameshift, stop gained, splice site

acceptor or donor mutation) and probably damaging missense according to PolyPhen-2 (Mis3)[45], in the context of three categories of inheritance pattern: *de novo*, inherited, and case-control. TADA requires a mutational model[46] which accounts for gene size and sequence composition to obtain an expectation for mutations per gene, given sample size. It treats all PTV mutations within a gene as equivalent, although their impact on risk is allowed to vary across genes and inheritance patterns (likewise for Mis3). TADA first computes a gene-specific Bayes factor for each mutation category and inheritance pattern, and then it multiplies these Bayes factors to generate a statistic that summarizes all evidence of association for each gene. The total Bayes factor is finally converted to a q-value to control FDR[3]. As a Bayesian model, TADA requires prior parameters or hyperparameters, namely the fraction of genes in the genome affecting risk, thus far taken to be 0.05, and γ, the relative risk for a particular mutation category. See He et al. (2013) for estimators.

Evaluating TADA and False Discovery Rate (FDR)

For downstream analysis it is critical to ensure reliable performance of TADA so that risk gene lists, such as those with FDR $< 10\%$, are properly calibrated. Such guarantees are straightforward to prove in many settings[47]. In the WES setting, however, and especially for the relatively discrete counts of *de novo* events, a demonstration that the FDR rate holds is warranted. It is worth noting that, even though there are many genes that contain no mutations, the mutation rate is gene-specific and varies with gene length. Consequently, with the exception of the genes with a signal, the p-values from the TADA analysis of PTV and Mis3 mutations are almost uniformly distributed.

To evaluate the validity of the FDR framework in the context of TADA analysis, we conduct "empirical-known signal experiments" (EKSE). The idea is to perform TADA analyses in which the true signal is known *a priori*. To make the simulation as real as possible, it is performed using real *de novo* mutation counts as a base. These mutations are chosen to carry no detectable signal (i.e., mimicking the null distribution because they are believed to be non-functional). Simulated signals for association are then generated for randomly selected genes. Once the data are generated, TADA is used to analyze them, and the resulting FDR and other features of the method are examined.

EKSE Simulations to Assess the Properties of FDR.

For these empirically-known signal experiments, we let synonymous variants play the role of Mis3 (denoted as Mis3new) and Mis1 play the role of PTV (denoted as PTVnew). Signals are layered onto genes that are randomly chosen. Below is the detailed procedure:

1. Divide all 17,484 genes into 20 bins of equal size. Let $b = 1, \ldots, 20$.
2. For each of the 20 bins, iteratively generate a signal for all genes in the bin; the remaining 19 bins, with no signal, represent the null genes. The extra signal in the ith gene for both new kinds of *de novo* variants is simulated using $X_i|\gamma_i \sim \text{Poisson}(2\mu_i(\gamma_i - 1)N)$, where $\gamma_i \sim \text{Gamma}(\gamma, \beta)$. The hyperparameters are selected to yield signals similar to the real data: $\beta = 0.2$ and γ is set to be 2.4 and 5.4 for Mis3new and PTVnew respectively, and $N = 6430$. The "-1" is to account for the observed *de novo* variants already included from the real data. The simulated *de novo* events are added to the observed Mis3new and PTVnew to create each of the 20 data sets.
3. Perform TADA analysis for each of the 20 datasets.

4. Display the resulting q-FDR curves for $b = 1, \ldots, 20$, and q-FDR averaged over b.

Pure Simulations to Assess the Properties of FDR.

This simulation is closely related to EKSE. The only difference is that the null mutations are generated randomly from a multinomial distribution instead of adopted directly from the Syn and Mis1 variants. The procedure is described below.

1. Randomly sample a fraction of all 17,484 genes as signal genes, denoted as set S. We set the fraction as $\pi = 0.05$. The number of trios is $N = 6430$.
2. For both two new types of variants, Mis3new and PTVnew, the mutations of all the genes are randomly generated from a multinomial distribution, $\mathbf{X} \sim \text{Multinom}(\text{M}, \mathbf{p})$, where the probability vector $\mathbf{p}$ is proportional to $\mathbf{p} = \{\mu_i \gamma_i\}_{i=1,\ldots,17484}$, where $\gamma_i \sim \text{Gamma}(\gamma, \beta)$ if $i \in S$, otherwise equals 1. The total number of mutations is $M = 2N \sum_{i=1}^{17484} \mu_i \gamma_i$. The mutation rates of Mis3new are taken from Syn, and the mutation rates of PTVnew are taken from Mis1. The hyperparameters γ, β are set to be the same as in EKSE.
3. Perform TADA analysis on the two generated types of variants. Display the resulting q-FDR curve.
4. Repeat steps 1-3 100 times.

For $q < 0.1$ the average curve follows the diagonal line (roughly), which indicates that the actual FDR is well controlled in the region of primary interest. We do detect a slight bump in the actual FDR for $q > 0.1$. To understand this deviation, we compared the observed counts for synonymous (Mis3new) and Mis1 (PTVnew) to simulated counts generated from the model.

The distribution of the number of genes with synonymous counts ≥3 and Mis1 counts ≥2 is contrasted with the observed counts. The contrasts show that there is a slight excess of

multiple hits in the observed counts compared to the model. Adding counts of synonymous and Mis1 mutations we obtain a single distribution of mutations per gene and find that there is an excess of counts of 0, 2, 3, and >3 and a relative lack of counts of 1; overall the counts are fairly similar, but they differ significantly from expectations (chi-square $P = 0.012$). The 8 null genes with the strongest TADA signal are *GNS, LRRFIP1, GALC, GRN, MYH9, FOXK2, AP1B1*, and *UNC45B*, and these are the genes that contribute to the bump in the FDR. However, none of these genes are significant ($q < 0.1$) in the EKSE analysis or in the actual data analysis of Mis3 and PTV mutations. From this EKSE experiment we conclude that the TADA model does not perfectly capture reality and the actual FDR deviates slightly from reported value for values of q > 0.1. This deviation is likely due to inexact estimates of the per gene mutation rate.

TADA relies on a mutation rate model for genes, which is an estimated quantity. Hence, we evaluate the impact of misspecification of mutation rates. To quantify the deviation from the expected null distribution due to mutation rate misspecification, we use the theory of genomic control[48], specifically estimating the inflation factor λ_{GC}. In this experiment we randomly select 10-50% of genes and artificially make the nominal mutation rates increasingly lower than their true mutation rates. This will make the observed mutation count larger than the expected count for a subset of genes. The result is that test statistics for association will tend to be increased for some genes, and the larger the discrepancy, the larger the set of test statistics that do not follow the expected null distribution. The genomic control factor, based on the z-statistics from the TADA analysis, quantifies this inflation. As expected, the genomic control factor increases as more genes are analyzed with lower nominal than true mutation rates. The inflation for λ_{GC} is modest, however, even for these fairly notable misspecifications of the mutation rates.

Because TADA is a Bayesian method it is more natural to use FDR than a Family-Wise Error Rate (FWER) cutoff to determine significance. In this gene discovery setting it is informative to compare the numbers of true discoveries (TD), false discoveries (FD), and FDR for different p-value and FDR thresholds and to examine the impact of model mis-specifications on FDR. We measure discrepancies via the genomic control factor (λ_{GC}). We simulate the Z-value of 20,000 genes, 5% with a signal from $N(\mu, \lambda_{GC})$ and 95% from the null $N(0, \lambda_{GC})$, where λ_{GC} varies from 1 to 1.2. The value of μ is chosen to be 2 to approximately mimic the real data. Based on 1,000 replications, we calculate the average TD, FD, and FDR for a Bonferroni adjusted p-value threshold and different FDR thresholds. As expected, FWER has considerably fewer FD but also notably fewer TD than FDR, and the observed FDR is well calibrated when $\lambda_{GC} = 1$. (For $\lambda_{GC} = 1$, TD = 5, 52, 113, 334, and FD = 0.1, 3, 13, and 144 for the four thresholds examined. In each case the error rate is controlled at the expected rate.) However, as λ_{GC} increases the actual FDR increases rapidly, especially for larger q-values. In contrast, FWER is fairly well controlled even for model discrepancies.

A more powerful TADA model

TADA required input of several parameters, most notably the relative risk, γ. To estimate the relative risk for a category of mutations, we used the burden-relative risk relationship derived in He et al. (2013): $\gamma = 1 + (\lambda - 1)/\pi$, where $\pi = 0.05$ is the estimated fraction of risk genes and the burden λ is calculated by comparing mutation counts in probands and unaffected siblings. Because differences in sequencing depths and variant calling procedures may lead to systematic differences in mutation rates, we normalized the counts using synonymous mutations

counts. Let x and S be the number of mutations in the category of interest and compare the counts in cases (cs) and controls (cn) as $\lambda = (x_{cs}S_{cn})/(x_{cn}S_{cs})$.

Previous TADA analyses[3,5] used two annotation categories, PTV and Mis3. Here we developed a more powerful version of TADA, which used additional annotation information. For clarity we labeled the original version TADA^0 and the refined model TADA$^+$.

Recent studies have refined our understanding of what variation was likely to be meaningful for risk in two ways. Regarding PTVs, Kosmicki et al. (2017) demonstrated that signals carried by PTVs involve a subset of genes that are evolutionarily constrained. For these genes, the population tends to have far fewer PTVs than would be expected based on gene size, base-pair content and evolutionary models. This constraint feature of genes is embodied in pLI (the probability of being loss-of-function [aka PTV] intolerant)[12], which is a metric ranging from zero to one, with a larger pLI representing a greater dearth of PTV variation. Kosmicki et al. (2017) found that genes with pLI > 0.9 tend to harbor most of the ASD association signal from PTVs. In this work, we modeled the relative risk (γ) of *de novo* PTVs as a continuous function of pLI. We created seven bins of data and fit a logistic curve to the data. The dots are the data and the black line is the fitted curve. Then we computed error bars based on the 95% prediction interval around the fitted curve. In the upcoming implementation, we truncated γ at the null value of one.

More refined information was also available for missense variants. Samocha et al. (2017) recently introduced the MPC score, a missense deleteriousness metric composed of "Missense badness", PolyPhen-2[45], and Constraint. This metric also used the concept of evolutionary constraint and seeks to quantify the degree of constraint for all missense variation in the genome. To determine how MPC might be used in TADA, we computed the average relative risk (γ, the

hyperparameters for TADA) for a moving window of MPC in the ASC data. Using a window over probands' missense variants ordered by MPC score, and with a width of 7.5% of the variants, we obtained the curve showing the average relative risk as a function of MPC score. Three levels of γ naturally emerged from this relationship, with the first level (MPC < 1) being close to marginal relative risk and two levels showing evidence for excess burden in ASD. Based on the nature of these results, we chose to group missense mutations into two categories for TADA, using established thresholds of MPC[13]: $1 \leq$ MPC < 2 (MisA) and MPC ≥ 2 (MisB). Note that missense variation with MPC < 1 was treated as benign. The relative risk for each of the two missense categories was computed directly from the data (He et al., 2013) as γ_{MisA} = 4.18 and γ_{MisB} = 22.15.

Besides the *de novo* variants, we also considered PTVs from case-control data by aggregating the iPSYCH (Danish) data and PAGES (Swedish) data. Following the same procedure as for *de novo* PTVs, within seven bins, we estimated the relative risks for the two case-control datasets separately and combined them with a precision weight. We then fit a logistic curve using the seven points to smooth γ as a continuous function of pLI. In the TADA analysis, we treated γ of each gene as fixed for case-control data to achieve closed-form solutions and thus facilitate the computation.

These analyses defined three categories of mutation potentially meaningful for risk. The gene-specific mutation rates for PTVs and missense variants have been reported previously[12], and we further estimated the mutation rates for MisA and MisB. With mutation rates and hyperparameters estimated above, the refined TADA model can be applied to the data to identify risk genes for ASD. To allow for more variability in the prior for γ, we set $\beta = 0.2$.

To resolve an emerging issue with the model's Bayes factor (BF) values, we implemented a floor adjustment that imposes a lower bound of 1 on all BF. The issue is that for some genes with larger mutation rates and zero *de novo* MisB mutations, the MisB BF is << 1. Multiplying this with the other evidence rendered those genes not significant. Indeed, with the mutation rates provided and the high relative risk of MisB, the model clearly expected to observe at least one *de novo* MisB variant. (This happened for other categories as well, but most notably for MisB.) We assumed the problem was heterogeneity of genes—some genes with a *de novo* PTV do not have MisB mutations in the data, even though these mutations are expected. It did not make sense to have the observation of no mutations drive the model. To circumvent the problem, we made a modification of the method so that BF is replaced by max(1, BF). We tested this in simulations and the size of the modified test was satisfactory (see the discussion in the next section).

TADA$^+$ incorporated all of the refinements delineated here. Using TADA$^+$, 102 genes with q-value less than 0.1 were identified, including three genes that have excessive PTVs in siblings (*EIF3G*, *KDM5B*, *RAI1*). By contrast, TADA^0 identified only 79 genes when applied to the same data. Clearly, the new relevant functional information embodied in the pLI and MPC scores improved the power of TADA by refining the model.

Simulations to evaluate TADA$^+$.

Simulations illustrated the performance of TADA$^+$ when applied to *de novo* mutations only. In this setting, we simulated three types of *de novo* variants: PTV, MisA, and MisB, using the mean risks and mutation rates from real data. Below is the detailed procedure.

1. Randomly select 5% of 17,484 genes as the signal genes; denote the set of signal genes as G_S and the null genes as G_N.
2. For each signal gene $g \in G_S$, generate risk γ_g for each three types of variants from a Gamma distribution, $\gamma_g^a \sim Gamma(\bar{\gamma}_g^a, \beta)$, $a = (PTV, MisA, MisB)$, where $\beta = 0.2$ and the hyper parameters $\overline{\gamma}_g^a$s are set to match the empirical counts. Note that $\overline{\gamma}_g^{MisA}$ and $\overline{\gamma}_g^{MisB}$ are the same across all genes, but $\overline{\gamma}_g^{PTV}$ are different.
3. For the null genes $g \in G_N$, set $\gamma_g^{PTV} = \gamma_g^{MisA} = \gamma_g^{MisB} = 1$
4. For each variant, generate the counts from a Multinomial distribution, where the total number is the expected total counts $2N \sum_g \mu_g^a \gamma_g^a$, $N = 6430$, and the probability is proportional to $\{\mu_g^a \gamma_g^a\}_{g=1}^M$. The mutation rates are taken from the real data.
5. Apply TADA$^+$ with Bayes Factors having a lower limit (floor) of 1, and calculate the empirical FDR.
6. Repeat steps 1-5 100 times.

Applying the floor principle increased the false discovery rate by a modest amount. In practice, we found there was considerable heterogeneity across genes and this adjustment was necessary.

TADA analyses

We explored the performance of TADA^0 and TADA$^+$ with three analyses:

A. TADA^0 applied to ASC2018, *de novo* only;
B. TADA$^+$ applied to ASC2018, *de novo* only; and
C. TADA$^+$ applied to ASC2018, *de novo* and case-control data.

By moving through the three analyses, we changed one variable at a time and analyzed the consequences. From A to B, we compared the improvements in the model by contrasting $TADA^0$ and $TADA^+$. From B to C, we assessed the impact of adding in the case-control data.

With additional data and a more powerful TADA model, we obtained substantial new discoveries. We identified 65 genes in A, 85 genes in B, and 102 genes in C. We visualized the q-values of the three analyses for the 114 genes with q-value less than 0.1 in at least one analysis—for most genes, the q-values decreased in sequence from analysis A to B to C, with the q-value of analysis C being the smallest. Twelve genes have a q-value greater than 0.1 in C, but less than 0.1 in at least one other analysis; of these 12 genes, most are downgraded in analysis C because of refinements in the new TADA model (with the genes or variants, having, for instance, low pLI score or low MPC score, particularly MPC less than 1 or missing and thus not categorized as MisA or MisB).

Comorbid phenotypes

Full-scale IQ scores were measured using several tests including, but not limited to, the Differential Ability Scales, Second Edition[49]; the Mullen Scales of Early Learning[50]; the Wechsler Intelligence Scale for Children[51]; and the Wechsler Abbreviated Scale of Intelligence[52]. The full-scale IQ estimates were taken from the full-scale deviation IQ variable when available and full-scale ratio IQ when it was not[53]. Full-scale IQ is normally distributed with a mean of 100 and a standard deviation of 15. We defined intellectual disability to be if a subject met one of the following conditions: a full-scale IQ (FSIQ) < 70 (i.e., two standard deviations below the mean), if the proband was administered but could not complete an IQ test, indicated by the subject having a date for their IQ test but no IQ score, or if the subject had an

HPO term or ICD code indicating intellectual disability or mental retardation. Age of walking unaided (in months) was taken from question 5A from the Autism Diagnostic Interview (ADI)[54]. We divided individuals into three possible categories for seizure status: yes, no, and unknown. A subject was put into the yes bin if he or she had a diagnosis of seizures or epilepsy, or a value of 2 on question 85 from the ADI (indicating a diagnosis of epilepsy). A subject was put into the no bin if no seizure/epilepsy diagnosis was indicated or if ADI question 85 had a value of 0. All remaining subjects were put into the unknown bin.

Burden of mutations in ASD as a function of IQ

We used full-scale IQ (FSIQ) to separate subjects into groups. Of the 5298 probands with any *de novo* mutation, 3010 have FSIQ information, 2055 (68.3%) with FSIQ > 70 and 1586 with FSIQ > 82. For a sample size N, the expected number of mutations within genes is computed as $E = 2N\phi$, where ϕ is the sum of the mutation rate, per variant type, over all relevant genes. (For example, to calculate ϕ for PTVs in genes with pLI > 0.995, we compute the sum of the PTV mutation rates for these genes.) We then compare *E* to the observed count for this mutation class, O, and evaluate the distribution of O/E as a chi-square statistic with 1 degree of freedom.

Burden of mutations over 102 TADA ASD genes

This analysis addressed the question of whether the signal found in the 102 genes with q<0.10 could have arisen solely from low IQ subjects, such that any mutations found in higher IQ subjects occurred by chance. To answer this question, we must address the bias inherent in choosing 102 genes because they have q<0.10. To do so, we performed model-based

simulations, similar to those used to evaluate the properties of the TADA model. We first selected 874 genes with the smallest q-values from the real data and labeled them "signal genes". Let $M = 0.306N$ be the number of subjects with IQ < 70, who accumulate mutations at rates greater than chance. We generated mutations for the signal genes using the TADA model and Poisson rate ($2M\gamma\mu$), where μ is the gene-specific mutation rate and γ is the increased rate of mutations due to this being a risk gene and the mutation of a particular type, and we generated additional mutations at a Poisson rate ($2[N - M]\mu$). We generated mutations in non-signal genes at a Poisson rate ($2N\mu$). We ran TADA to get the new top 102 genes and the new signal genes, and we recorded counts occurring in new signal genes by chance (i.e., for individuals with high IQ). We performed the simulation 500 times to obtain the distribution of counts in signal genes for individuals with high IQ and compared this to the observed data. For all four informative mutation types, the expected counts were consistently lower than the observed count; only for missense mutations with MPC between 1 and 2 does the expected count, 13.54 (± 4.2) approach the observed value, 23 (p=0.03). For all other mutation types, the empirical p-value was far smaller, based on 500 simulations (MPC >2: 13.9 ± 3.9 versus 28; PTV for pLI >0.995: 8.3 ± 3.0 versus 48; and PTV for 0.5 < pLI < 0.995: 3.0 ± 1.9 versus 15). We also performed these simulations for a split on IQ at 82 and reached the same conclusion, that the mutations in the higher IQ ASD subjects accumulate at a rate far greater than chance.

Genes in recurrent genomic disorders

We constructed a list of loci previously reported to be associated with ASD- or NDD-related phenotypes due to rare CNVs. We first collated coordinates of pathogenic genomic disorder (GD) regions as reported by nine previous studies[5,24,37,55-59] and converted all

coordinates to human reference genome build hg19 with UCSC liftOver tool as necessary. We next clustered the coordinates of all overlapping CNV regions using svtk bedcluster and a minimum 50% reciprocal overlap between segments, retaining the median clustered coordinates of all CNV regions appearing in at least two of the nine studies considered. After clustering, we excluded any CNV segments > 5Mb in size and all segments on sex chromosomes. Finally, we annotated each CNV segment passing all filters with all overlapping genes drawn from the list of autosomal genes considered during TADA analyses.

Assessment of overlap between ASD-associated genes and GD loci

We designed three permutation-based approaches to benchmark null expectations for the overlap of ASD-associated genes and GD loci. All approaches involved randomly drawing new sets of collinear genes for each GD locus from the list of all genes considered in TADA analyses, but differed in how these new genes were selected. These sampling approaches are summarized as follows:

1. Matched on number of genes: a new collinear list of genes was drawn for each GD locus, where the number of genes was matched to the number of genes in the original GD locus.
2. Matched on PTV mutation rates: a new collinear list of genes was drawn for each GD, where the number of genes was determined such that the sum of their estimated PTV mutation rates was at least as large as the sum of the estimated PTV mutation rates of the original list of genes in that GD locus.
3. Matched on brain expression, PTV mutation rates, and number of genes: prior to permutation, all genes were assigned a PTV mutation rate quintile and a brain expression quantile determined by the median brain expression value for that gene across all samples

and all brain regions present in GTEx release v7 calculated after excluding genes with non-zero median brain expression. During permutation, a new collinear list of genes was drawn for each GD such that the number of genes matched the original GD locus, with the additional requirements that the distribution of these genes across brain expression quintiles and PTV mutation rate quintiles were also preserved.

For each permutation, we performed one of the three above approaches for all 51 GD loci to obtain a new set of sampled genes, and we then counted the number of newly sampled genes that matched the TADA thresholds for ASD association in this study. We performed 1,000,000 permutations for each approach and computed p-values based on the fraction of all permutations where the number of GD loci with at least one randomly sampled ASD-associated gene matched or exceed the empirical observation in the original data. Fold-changes were determined as the observed number of GD loci with at least one ASD-associated gene divided by the mean number of GDs with at least one ASD-associated gene across all 1,000,000 permutations.

Finally, we titrated additional parameters to examine the variability of results from this permutation approach. For each of the three gene-sampling schemes above, we performed a separate 1,000,000 independent permutations for each combination of two additional factors, as follows:

1. ASD-associated gene list: we considered two different significance levels for ASD-associated genes, including (1) those determined as significant by the extended TADA model at FDR ≤ 0.1 (n=102 genes) and (2) those reaching Bonferroni-corrected significance (n=26).
2. Chromosome sampling weights: for each GD locus in each permutation, an autosomal chromosome was selected based on one of two weighting schemes prior to randomly sampling a new set of collinear genes. These weights were either (1) determined by the

fraction of all autosomal genes located on each chromosome, or (2) determined by the fraction of GD loci located on each chromosome.

All results were consistent across gene sampling strategies and the additional parameters had limited influence on our individual results or overall conclusions.

Author Contributions

Jack Kosmicki: method design, *de novo* QC, frequentist-based gene discovery using *de novo* variants, inherited variant QC, calling inherited variants, all inherited variant data analysis, collected and standardized all phenotype data, phenotype analyses (Figures 3.4B, C), ID/DD and ASD gene analyses (Figure 3.4A), writing

Mark Daly: method design, writing, overall guidance

F. Kyle Satterstrom: method design, writing, initial variant and sample QC, constructed pedigrees, performed sample overlap, initial *de novo* variant calling, case/control analyses

Kathryn Roeder, *Bernie Devlin*, *Jiebiao Wang:* extended TADA to incorporate MPC and pLI.

Elizabeth Guerrero, *Rachel Nguyen*, *Caroline Dias*, *Branko Aleksic*, *Hilary Coon*, *Andreas Chiocchetti*, *Irva Hertz-Picciotto*, *Edwin Cook*, *Louise Gallagher*: provided phenotypes for their respective samples

Kaitlin Samocha: created MPC, provided validation data for published *de novo* variants

Joon-Yong An: missense localization (Figure 3A-E)

Ryan Collins: CNV overlap (Figure 3F and 3G)

Shan Dong: performed validation of *de novo* variants

Mafalda Barbosa, *Alfredo Brusco*, *Christine Freitag*, *Jay Gargus*, *Irva Hertz-Picciotto*, *Christina Hultman*, *Dara Manoach*, *Nancy Minshew*, *Aarno Palotie*, *Mara Parellada*, *Maria Passos-Bueno*, *Margaret Pericak-Vance*, *Antonio Persico*, *Kaija Puura*, *Alessandra Renieri*: provided samples

Stephan Sanders: method design, writing, provided overall guidance

David Cutler: method design, writing, liability analyses

Joseph Buxbaum: method design, writing

References

1. Baio, J. *et al.* Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *MMWR. Surveillance Summaries* **67**, 1--23 (2018).

2. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* **46**, 881-885 (2014).

3. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).

4. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).

5. Sanders, S.J. *et al.* Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215-33 (2015).

6. He, X., Sanders, S.J. & Liu, L.a. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet* **9**, e1003671 (2013).

7. Ben-Shalom, R. *et al.* Opposing effects on NaV1.2 function underlie differences between SCN2A variants observed in individuals with autism spectrum disorder or infantile seizures. *Biological Psychiatry* **82**, 1-9 (2017).

8. Bernier, R. *et al.* Disruptive CHD8 mutations define a subtype of autism early in development. *Cell* **158**, 263-276 (2014).

9. Willsey, A.J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007 (2013).

10. Satterstrom, F.K. *et al.* ASD and ADHD have a similar burden of rare protein-truncating variants. *bioRxiv* (2018).

11. Kosmicki, J.A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat Genet* **49**, 504-510 (2017).

12. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).

13. Samocha, K.E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* (2017).

14. Power, R.A. *et al.* Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**, 22-30 (2013).

15. Christensen, D.L. *et al.* Prevalence and Characteristics of Autism Spectrum Disorder Among 4-Year-Old Children in the Autism and Developmental Disabilities Monitoring Network. *J Dev Behav Pediatr* **37**, 1-8 (2016).

16. Werling, D.M. The role of sex-differential biology in risk for autism spectrum disorder. *Biology of Sex Differences* **7**, 1-18 (2016).

17. Vulto-van Silfhout, Anneke T. *et al.* Mutations Affecting the SAND Domain of DEAF1 Cause Intellectual Disability with Severe Speech Impairment and Behavioral Problems. *The American Journal of Human Genetics* **96**, 178 (2015).

18. Chen, L. *et al.* Functional analysis of novel DEAF1 variants identified through clinical exome sequencing expands DEAF1-associated neurodevelopmental disorder (DAND) phenotype. *Hum Mutat* **38**, 1774-1785 (2017).

19. Heyne, H.O. *et al.* De novo variants in neurodevelopmental disorders with epilepsy. *Nature Genetics* **50**, 1048-1053 (2018).

20. Maljevic, S. *et al.* Novel KCNQ3 mutation in a large family with benign familial neonatal epilepsy: A rare cause of neonatal seizures. *Molecular Syndromology* **7**, 189-196 (2016).

21. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980-D985 (2014).

22. Johannesen, K.M. *et al.* Defining the phenotypic spectrum of SLC6A1 mutations. *Epilepsia* **59**, 389-402 (2018).

23. Leroy, C. *et al.* The 2q37-deletion syndrome: an update of the clinical spectrum including overweight, brachydactyly and behavioural features in 14 new patients. *Eur J Hum Genet* **21**, 602-12 (2013).

24. Coe, B.P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* **46**, 1063-71 (2014).

25. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat Genet* **43**, 838-846 (2011).

26. Bottomley, M.J. *et al.* The SAND domain structure defines a novel DNA-binding fold in transcriptional regulation. *Nat Struct Biol* **8**, 626-33 (2001).

27. Jensik, P.J., Huggenvik, J.I. & Collard, M.W. Identification of a nuclear export signal and protein interaction domains in deformed epidermal autoregulatory factor-1 (DEAF-1). *Journal of Biological Chemistry* **279**, 32692-32699 (2004).

28. Miceli, F. *et al.* A novel KCNQ3 mutation in familial epilepsy with focal seizures and intellectual disability. *Epilepsia* **56**, e15--e20 (2015).

29. Claes, L., Del-Favero, J., Ceulemans, B., Lagae, L. & Broeckhoven, C.V.a. De Novo Mutations in the Sodium-Channel Gene SCN1A Cause Severe Myoclonic Epilepsy of Infancy. *Am. J. Hum. Genet* **68**, 1327--1332 (2001).

30. Rosander, C. & Hallbook, T. Dravet syndrome in Sweden: a population-based study. *Dev Med Child Neurol* **57**, 628-633 (2015).

31. Deciphering Developmental Disorders, S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438 (2017).

32. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)* **316**, 445-449 (2007).

33. Soorya, L. *et al.* Prospective investigation of autism and genotype-phenotype correlations in 22q13 deletion syndrome and SHANK3 deficiency. *Mol Autism* **4**, 18 (2013).

34. Williams, S.R. *et al.* Haploinsufficiency of HDAC4 causes brachydactyly mental retardation syndrome, with brachydactyly type E, developmental delays, and behavioral problems. *Am J Hum Genet* **87**, 219-28 (2010).

35. Yip, B.H.K. *et al.* Heritable variation, with little or no maternal effect, accounts for recurrence risk to autism spectrum disorder in Sweden. *Biol Psychiatry* **83**, 589-597 (2018).

36. Reichenberg, A. *et al.* Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. *Proceedings of the National Academy of Sciences* **113**, 1098-1103 (2016).

37. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-372 (2010).

38. Buxbaum, J.D. *et al.* The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* **76**, 1052-6 (2012).

39. Lim, E.T. & Uddin, M.a. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nature Publishing Group* **20**(2017).

40. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11.10.1-33 (2013).

41. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).

42. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)* **26**, 2867-2873 (2010).

43. Werling, D.M. *et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* **50**, 727-736 (2018).

44. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-246 (2012).

45. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).

46. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950 (2014).

47. Efron, B. Bayesian inference and the parametric bootstrap. *Ann. Appl. Stat.* **6**, 1971-1997 (2012).

48. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).

49. Elliott, C.D. *Differential Ability Scales (2nd ed.)*, (Harcourt Assessment, San Antonio, TX, 2007).

50. Mullen, E.M. *Mullen Scales of Early Learning Manual*, (American Guidance Service, Circle Pines, MN, 1995).

51. Wechsler, D. *WISC III (Wechsler Intelligence Scale for Children)*, (Psychological Corporation, San Antonio, TX, 1992).

52. Wechsler, D. *Wechsler Abbreviated Scale of Intelligence WASI: Manual*, (The Psychological Corporation, 1999).

53. Chaste, P. *et al.* A Genome-wide Association Study of Autism Using the Simons Simplex Collection: Does Reducing Phenotypic Heterogeneity in Autism Increase Genetic Homogeneity? *Biological Psychiatry* **77**, 775-784 (2015).

54. Lord, C., Rutter, M. & Lecouteur, A.a. Autism Diagnostic Interview-Revised - a Revised Version of a Diagnostic Interview for Caregivers of Individuals With Possible Pervasive Developmental Disorders. *Journal of Autism and Developmental Disorders* **24**, 659--685 (1994).

55. Cooper, G.M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**, 628-640 (2011).

56. Dittwald, P. *et al.* NAHR-mediated copy-number variants in a clinical population: Mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Research* **23**, 1395--1409 (2013).

57. Schaefer, G.B. & Mendelsohn, N.J. Clinical genetics evaluation in identifying the etiology of autism spectrum disorders: 2013 guideline revisions. *Genet Med* **15**, 399-407 (2013).

58. Wright, C.F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet (London, England)* **385**, 1305-1314 (2015).

59. Wapner, R.J. *et al.* Chromosomal microarray versus karyotyping for prenatal diagnosis. *N Engl J Med* **367**, 2175-84 (2012).

Chapter 4

Influence of severe intellectual disability and developmental delay on *de novo* architecture and gene discovery in autism spectrum disorders and congenital heart disease

Abstract

De novo variants identified through exome sequencing have implicated numerous genes in both autism spectrum disorders (ASD) and congenital heart disease (CHD). In both traits, these variants are disproportionately observed in the subset of cases with comorbid intellectual disability / developmental delay (ID/DD). Using *de novo* and inherited variants from distinct ascertainments of 6430 ASD probands, 3683 CHD probands, and 5305 ID/DD probands coupled with extensive phenotype data, we examined the effect of ID/DD on *de novo* variant frequencies and gene discovery in ASD and CHD. We found the frequency of deleterious *de novo* coding variants was equivalent in both ASD and CHD with comorbid ID/DD as well as ID/DD with and without ASD and CHD. We identified 105 genes at genome-wide significance for at least one of the three disorders, but far more genes were identified in ID/DD (N=95) than ASD (N=18) and CHD (N=11). Using a frequentist-based approach, nine genes were significantly more strongly associated with ID/DD than ASD or CHD, and one gene – *KMT2D* – was more strongly associated with CHD than the either two traits after Bonferroni correction. Lastly, the excess of *de novo* and inherited variants reside in genes on opposite ends of the selection coefficient distribution.

Introduction

Autism spectrum disorder is a highly heritable, but phenotypically heterogeneous disorder, diagnosed by deficiencies in two core domains: (1) impairments in communication and social interaction and (2) restrictive interests and repetitive behaviors. Historically, ASD was most often diagnosed in individuals with comorbid intellectual disability / developmental delay (ID/DD)[1], and many early genetic associations to ASD were made in ASD cases with co-

occurring severe intellectual disability or syndromic features[2-6]. Thanks in large part through increased awareness of the disorder coupled with widening diagnostic criteria, ASD currently encompasses individuals that span the entire IQ spectrum, reaching those with average or above average IQ.

Congenital heart disease (CHD) has similarly often been described in the context of likely or established genetic syndromes that also have significant neurodevelopmental disabilities[7]; however, given the obvious lack of phenotypic and diagnostic overlap between the cardiac and neurodevelopmental aspects, such observations have simply suggested multiple or diverse sites of molecular action of specific genes, rather than generating contentious debate over diagnostic criteria and disease etiology. We sought here to explore the degree to which currently observed *de novo* variant excesses in both ASD and CHD might suggest general developmental abnormalities, and to what degree specific insights unique to each diagnosis may be available as studies expand.

Results

Frequency of *de novo* variants by ascertainment and comorbidity across studies

Here we analyze 21,288 published *de novo* variants: 12,166 *de novo* variants from 6430 individuals ascertained for ASD, 1404 *de novo* variants from 1012 individuals ascertained for intellectual disability (ID), 6791 *de novo* variants from 4293 individuals ascertained for developmental disorders (DD) as part of the Deciphering Developmental Disorders (DDD) project, 4585 *de novo* variants from 3683 individuals ascertained for congenital heart disease, and 3623 *de novo* variants from 2179 control individuals (**Table 4.1**; Materials and methods). We collected each individual's secondary diagnoses in each of the cohorts to identify the 2402

ascertained ASD and 1073 ascertained congenital heart disease individuals with comorbid ID/DD, and the 711 and 571 ascertained ID/DD individuals with comorbid ASD and CHD respectively (Materials and methods).

Table 4.1: Probands split by ascertainment and comorbidity status.

Ascertainment	Comorbidity	N	Male	Female
ASD	Total	6430	5333	1097
	without ID/DD	2895	2511	384
	comorbid ID/DD	2402	1924	478
	unknown	1133	898	235
CHD	Total	3683	2133	1550
	without ID/DD	1673	944	729
	comorbid ID/DD	1073	628	445
	unknown	937	561	376
ID/DD	Total	5305	2952	2353
	without ASD	4594	2457	2137
	comorbid ASD	711	495	216
	without CHD	4734	2638	2096
	comorbid CHD	571	314	257
Unaffected siblings	Total	2179	1031	1148

Because we meta-analyzed various datasets sequenced at different times on different platforms, we needed to correct for any technical discrepancies between studies. All *de novo* variants were reprocessed to ensure consistency across datasets (Materials and methods). As a control for the comparison of *de novo* missense and PTVs, we compared the frequencies of *de novo* synonymous variants across the separate ascertained cohorts because, as a class, true frequencies of *de novo* synonymous variants should be independent of phenotypic ascertainment. We observed no difference in the frequency of *de novo* synonymous variants (**Table 4.2**), nor did we observe a difference in the frequency of *de novo* PTVs in LoF-tolerant genes (pLI < 0.9) (**Table 4.3)** or benign missense variants (MPC < 1) (**Table 4.4**) between any of the ascertained cohorts.

Table 4.2: *De novo* synonymous variant frequencies. Testing for association between the *de novo* synonymous variant frequencies between the five different ascertained groups using a two-sided, two-sample Poisson exact test (also known as the C-test)[8]. We performed pairwise comparisons for a difference in the *de novo* synonymous variant frequency between ascertainments. The columns, Frequency 1 and 2, refer to the *de novo* synonymous variant frequencies for the two traits tested. Our significance threshold after Bonferroni correction for ten tests is $0.05 / 10 \sim 5 \times 10^{-3}$.

Comparison	Frequency 1	Frequency 2	*P*-value	Rate Ratio	95% CI
ASD vs. Control	1870 / 6430	640 / 2179	0.8363	0.9902	0.90 - 1.08
ASD vs. ID	1870 / 6430	252 / 1012	0.0207	1.1679	1.02 - 1.34
ASD vs. CHD	1870 / 6430	983 / 3683	0.0293	1.0896	1.01 - 1.18
ID vs. CHD	257 / 1012	983 / 3683	0.49	0.9515	0.83 - 1.09
ID vs. Control	257 / 1012	640 / 2179	0.0485	0.8646	0.75 - 1.00
CHD vs. Control	983 / 3683	640 / 2179	0.0608	0.9087	0.82 - 1.01
DDD vs. ASD	1276 / 4293	640 / 2179	0.8278	1.012	0.92 - 1.11
DDD vs. Control	1276 / 4293	1870 / 6430	0.5483	1.022	0.95 - 1.10
DDD vs. ID	1276 / 4293	257 / 1012	0.021	1.1704	1.02 - 1.34
DDD vs. CHD	1276 / 4293	983 / 3683	0.0113	1.1136	1.02 - 1.21

Table 4.3: *De novo* PTVs (pLI < 0.9) frequencies. Testing for association between the *de novo* protein truncating variant frequencies in LoF-tolerant genes (pLI < 0.9) between the five different ascertained groups using the Poisson exact test. We perform pairwise comparisons for a difference in the *de novo* PTV (pLI < 0.9) frequency between ascertainments. The columns, Frequency 1 and 2, refer to the *de novo* PTV (pLI < 0.9) frequencies for the two tested traits. Our significance threshold after Bonferroni correction for ten tests is $0.05 / 10 \sim 5 \times 10^{-3}$.

Comparison	Frequency 1	Frequency 2	*P*-value	Rate Ratio	95% CI
ASD vs. Control	423 / 6430	127 / 2179	0.2395	1.1287	0.92 - 1.39
ASD vs. ID	423 / 6430	60 / 1012	0.5066	1.1096	0.85 - 1.48
ASD vs. CHD	423 / 6430	201 / 3683	0.0305	1.2054	1.02 - 1.43
ID vs. CHD	60 / 1012	201 / 3683	0.5981	1.0864	0.80 - 1.46
ID vs. Control	60 / 1012	127 / 2179	0.9374	1.0172	0.74 - 1.39
CHD vs. Control	201 / 3683	127 / 2179	0.5681	0.9364	0.75 - 1.18
DDD vs. ASD	251 / 4293	127 / 2179	1	1.0032	0.81 - 1.25
DDD vs. Control	251 / 4293	423 / 6430	0.1458	0.8888	0.76 - 1.04
DDD vs. ID	251 / 4293	60 / 1012	0.9425	0.9861	0.74 - 1.33
DDD vs. CHD	251 / 4293	201 / 3683	0.4794	1.0713	0.89 - 1.30

Table 4.4: *De novo* missense variants (MPC < 1) frequencies. Testing for association between the d*e novo* missense variant (MPC < 1) frequencies between the five different ascertained groups using a two-sided, two-sample Poisson exact test. We perform pairwise comparisons for a difference in the *de novo* missense variants (MPC < 1) frequency between ascertainments. The columns, Frequency 1 and 2, refer to the *de novo* missense (MPC < 1) frequencies for the two tested traits. Our significance threshold after Bonferroni correction for ten tests is 0.05 / 10 ~ 8.33×10^{-3}.

Comparison	Frequency 1	Frequency 2	*P*-value	Rate Ratio	95% CI
ASD vs. Control	2117 / 6430	731 / 2179	0.6665	0.9814	0.90 - 1.07
ASD vs. ID	2117 / 6430	314 / 1012	0.3437	1.0611	0.94 - 1.20
ASD vs. CHD	2117 / 6430	1233 / 3683	0.6407	0.9834	0.92 - 1.06
ID vs. Control	314 / 1012	1233 / 3683	0.2401	0.9268	0.82 - 1.05
ID vs. CHD	314 / 1012	731 / 2179	0.2585	0.9249	0.81 - 1.06
CHD vs. Control	1233 / 3683	731 / 2179	0.9628	0.9979	0.91 - 1.10
DDD vs. ASD	1442 / 4293	731 / 2179	1	1.0013	0.92 - 1.10
DDD vs. Control	1442 / 4293	2117 / 6430	0.5609	1.0202	0.95 - 1.09
DDD vs. ID	1442 / 4293	314 / 1012	0.213	1.0826	0.96 - 1.23
DDD vs. CHD	1442 / 4293	1233 / 3683	0.9382	1.0033	0.93 - 1.08

While each study's ascertainment differs, both ASD and CHD studies include individuals with comorbid ID/DD and vice versa. We began by asking whether ascertainment affects the frequency of deleterious *de novo* missense ($MPC^9 \geq 2$) and PTVs in LoF-intolerant genes (pLI ≥ 0.9) absent from ExAC. We failed to observe a significant difference in the frequency of *de novo* missense (MPC ≥ 2) variants absent from ExAC between individuals ascertained for ID/DD with (0.12 per person) and without comorbid ASD (0.16 per person; $P = 0.04$; two-sided, two-sample Poisson exact test) and with (0.15 per person) and without comorbid congenital heart disease (0.15 per person; P=0.91; two-sided, two-sample Poisson exact test; **Figure 4.1**). Additionally, we failed to observe a significant difference in the frequency of *de novo* PTVs (pLI ≥ 0.9) absent from ExAC between individuals ascertained for ID/DD with (0.18 per person) and without comorbid ASD (0.16 per person; $P = 0.25$; two-sided, two-sample Poisson exact test) and with (0.15 per person) and without comorbid congenital heart disease (0.16 per person; P=0.62; two-sided, two-sample Poisson exact test; **Figure 4.1**).

By contrast, however, the presence of ID/DD in both ascertained ASD and ascertained congenital heart disease cases significantly increases the frequency of both *de novo* missense (MPC ≥ 2) and PTVs (pLI ≥ 0.9) absent from ExAC (**Figure 4.1**). Individuals ascertained for ASD with comorbid ID/DD have a 1.92-fold excess (95% CI: 1.50-2.46) of *de novo* missense (MPC ≥ 2) and a 1.49-fold excess (95% CI: 1.22-1.82) of *de novo* PTVs in LoF-intolerant genes (missense $P = 6.71 \times 10^{-8}$, PTVs $P = 6.27 \times 10^{-5}$; two-sided, two-sample Poisson exact test). Similarly, individuals ascertained for congenital heart disease with comorbid ID/DD have a 1.36-fold excess (95% CI: 1.05-2.29) of *de novo* missense (MPC ≥ 2) and a 2.33-fold excess (95% CI: 1.56-3.50) of *de novo* PTVs in LoF-intolerant genes (missense $P = 6.71 \times 10^{-4}$, PTVs $P = 1.62 \times 10^{-5}$; two-sided, two-sample Poisson exact test). These contrasting results indicates the presence of comorbid ASD or congenital heart disease within an ascertained severe ID/DD sample does not influence the frequency of *de novo* missense (MPC ≥ 2) or PTVs (pLI ≥ 0.9) absent from ExAC, but comorbid ID/DD strongly increases the frequency of both classes of *de novo* variation in ascertained ASD and congenital heart disease cases.

Beyond comparing comorbidities within ascertainment, we also compared the frequencies of *de novo* missense (MPC ≥ 2) and PTVs (pLI ≥ 0.9) absent from ExAC across ascertainments. After a Bonferroni correction for eight tests for each variant class (*P*-value threshold = 0.00625), we failed to detect a significant difference in the frequency of both classes of *de novo* variation when we compared individuals ascertained for ASD with comorbid ID/DD to individuals ascertained for congenital heart disease with comorbid ID/DD (missense $P = 0.03$; PTVs $P = 0.65$; two-sided, two-sample Poisson exact test); the same was true for individuals ascertained for ASD without comorbid ID/DD compared to individuals ascertained for congenital heart disease without comorbid ID/DD (missense $P = 0.59$; PTVs $P = 0.02$; two-

sided, two-sample Poisson exact test; **Figure 4.1**). These results are reminiscent of previous results demonstrating that ASD and ADHD with and without intellectual disability have similar rare PTV frequencies[10].

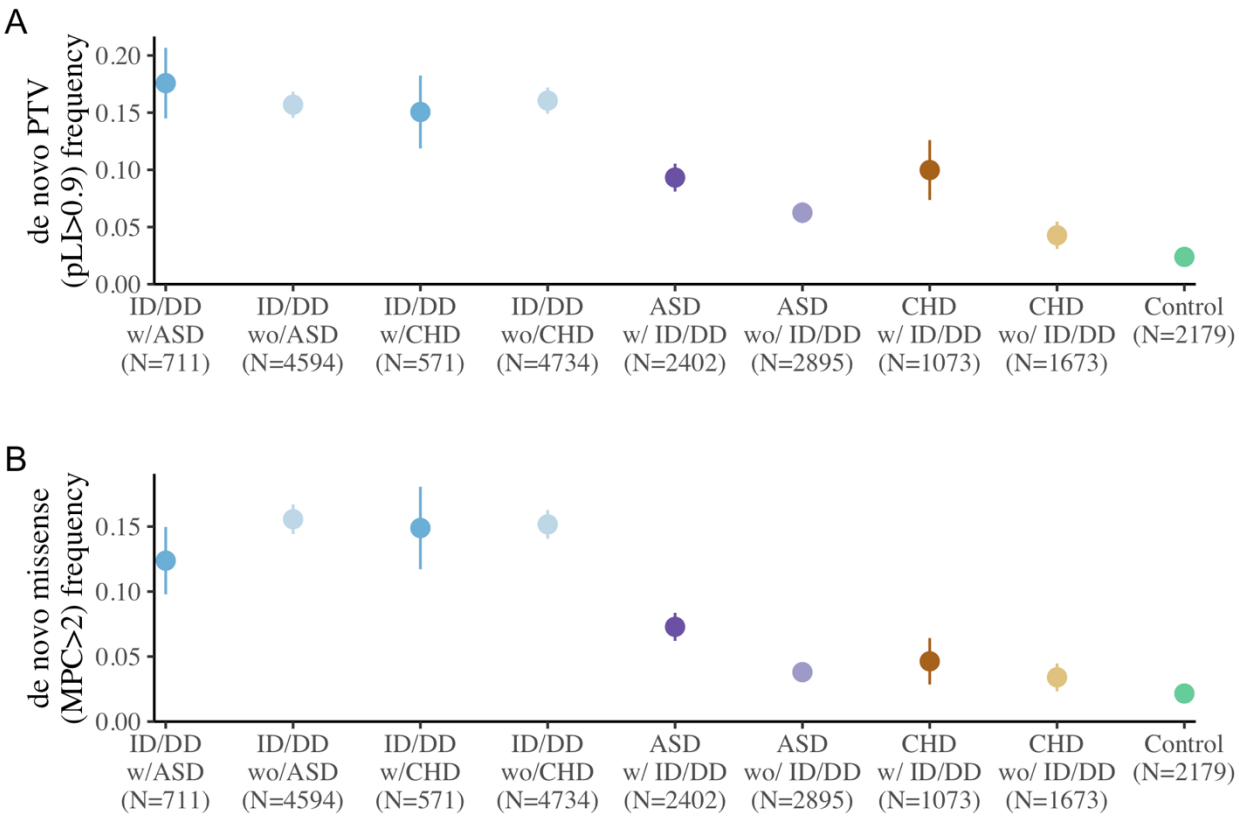


Figure 4.1: Frequency – variants per person – of *de novo* variants by ascertainment and comorbidity. In (**A**), the frequency of *de novo* PTVs (pLI ≥ 0.9) absent from ExAC. In (**B**), the frequency of *de novo* missense (MPC ≥ 2) variants absent from ExAC. Error bars represent 95% confidence intervals.

Given the observed higher frequency of both *de novo* missense and PTVs in individuals ascertained for ID/DD with ASD (missense rate ratio = 1.70; missense $P = 8.99 \times 10^{-5}$; PTV rate ratio = 1.89; PTV $P = 4.51 \times 10^{-8}$) or CHD (missense rate ratio = 1.55; missense $P = 3.49 \times 10^{-3}$; PTV rate ratio = 1.51; PTV $P = 1.85 \times 10^{-3}$) than individuals ascertained for ASD or CHD with ID/DD suggests the current classification is not perfectly comparable. If we assume that the

relative frequency of these classes of *de novo* variation are indicative of severity, then the presence of ASD and CHD is secondary to that of ID/DD. By stratifying individuals in ascertained ASD and ascertained CHD based on the number of additional ID/DD phenotypes (i.e., ID, seizures, delayed walking, global developmental delay, cranio-facial abnormalities), the frequency of *de novo* missense and PTVs increases with increasing number of comorbid phenotypes (**Figure 4.2**). Once the ascertained ASD and congenital heart disease samples are limited to those with two or more ID/DD comorbidities, then the *de novo* variant frequencies are no longer significantly different after Bonferroni correction (**Figure 4.2**).

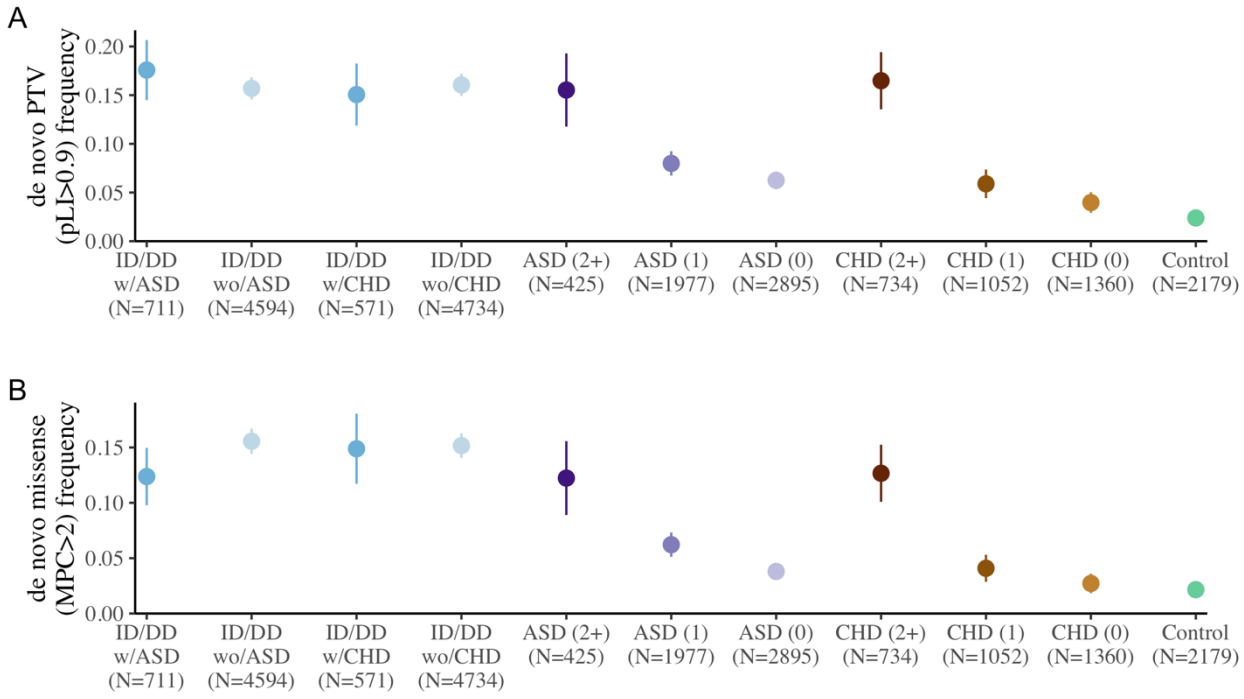


Figure 4.2: Frequency – variants per person – of *de novo* variants by ascertainment and number of comorbidities in ascertained ASD and congenital heart disease samples. In (**A**), the frequency of *de novo* PTVs (pLI≥0.9) absent from ExAC. In (**B**), the frequency of *de novo* missense (MPC≥2) variants absent from ExAC. Error bars represent 95% confidence intervals.

De novo variant genetic architecture

Despite prior evidence that ASD is a polygenic trait[11-13], a recent study suggested the contrary: having observed more individuals with ASD carrying multiple rare *de novo* variants

compared to their unaffected siblings indicated oligogenic inheritance[14]. For an oligogenic trait, one should observe more individuals carrying two or more trait-associated variants than expected as at least two trait-associated variants are necessary for the given trait to manifest. We tested this hypothesis in the three ascertained traits – ASD, ID/DD, and congenital heart disease – as well as unaffected ASD siblings by comparing the distribution of *de novo* synonymous, missense, and PTVs per person compared to what would be expected under a null Poisson distribution (Materials and methods). For ASD, congenital heart disease, and the unaffected ASD siblings, the observed distribution for all four sets of *de novo* variation did not differ from expectation (**Figure 4.3**). The distribution of 1) all *de novo* PTVs and 2) *de novo* missense and PTVs in ID/DD-significant genes were significantly different from expectation (**Figure 4.3**). There were significantly more individuals with one such *de novo* variant and a depletion at two that was responsible for the deviation from expectation in the chi-squared goodness of fit test (**Figure 4.3**). Based on this analysis, we do not find evidence that any of these traits are oligogenic.

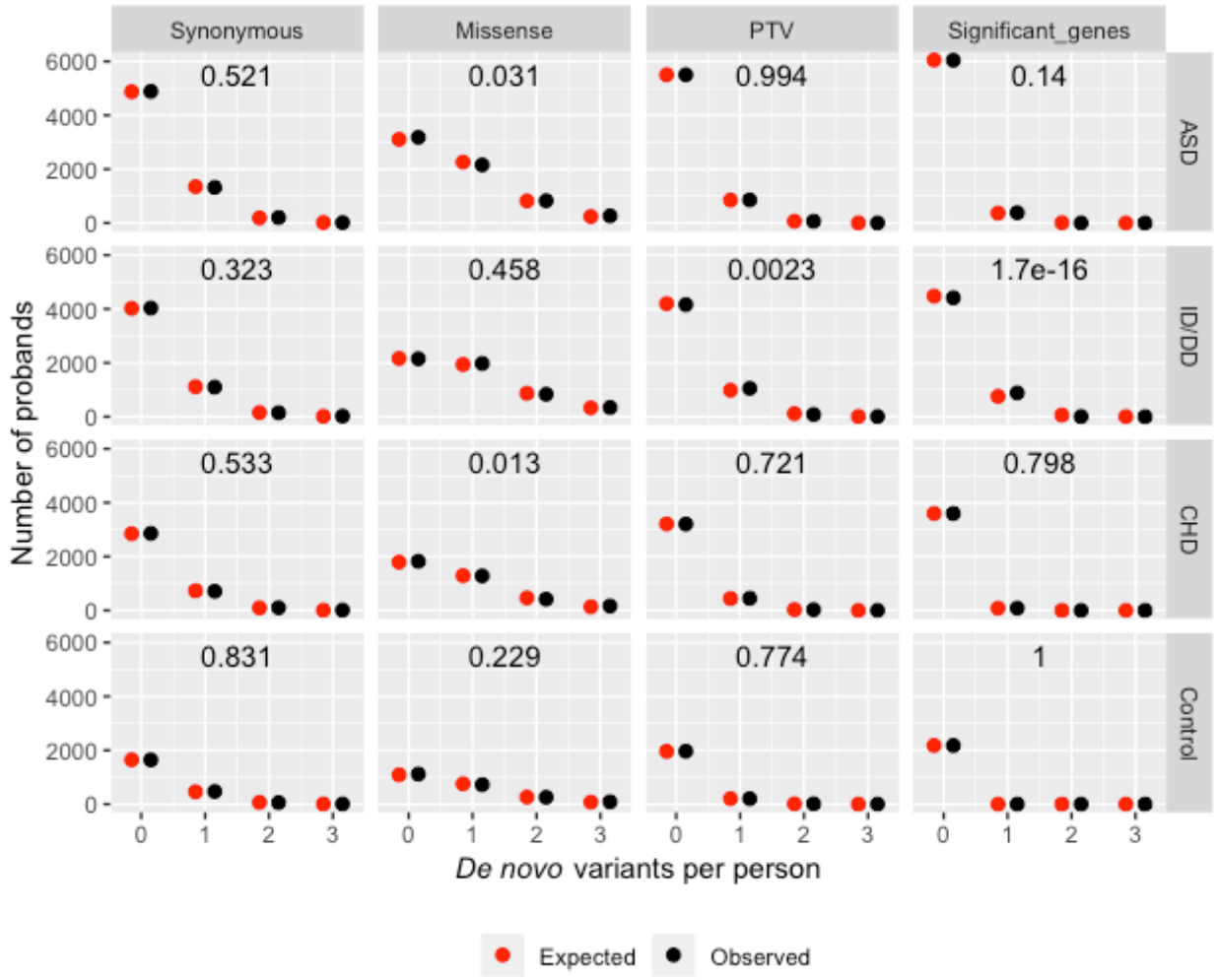


Figure 4.3: The observed (black) and expected (red) distribution of *de novo* synonymous, missense, PTVs, and lastly associated missense and PTVs in Bonferroni significant genes per person per ascertained trait. *P*-values come from the chi-square goodness of fit test with a Bonferroni significance threshold of 0.05 / 16 = 0.003125.

In the foundational paper by Jonathan Sebat and colleagues, they reported a 10-fold increase in the frequency of large CNVs in simplex ASD families (i.e., a family with a single child with ASD and no other affected family members out to first-degree relatives) and a 3-fold increase in multiplex families compared to unaffected siblings[15]. Such a striking observation motivated a subset of future family-based studies focusing on *de novo* variation to include only simplex ASD families[16]. While one of the earliest ASD *de novo* studies by Neale and colleagues[17] found no difference in the frequency of *de novo* variants between simplex and

multiplex families, their sample size was extremely small (N=175) and thus we revisited this analysis with a larger sample. Of the 6430 ASD probands, 3061 came from simplex families and 500 came from multiplex families. We lacked any information about the remaining 3839 probands. Consistent with the Neale et al. results, we failed to observe any difference in the frequency of five classes of *de novo* variation (**Table 4.5**) and thus, we suggest that at the very least for future family-based ASD studies examining the role of *de novo* single nucleotide variation, they not restrict samples based on family status.

Table 4.5: *De novo* variant frequencies – variants per ASD proband – between simplex and multiplex ASD families. We failed to observe a significant difference in the frequency of five classes of *de novo* variation between simplex and multiplex ASD families using a two-sided, two-sample Poisson exact test. Our significance threshold after Bonferroni correction for five tests is 0.01.

Variant class	Simplex frequency	Multiplex frequency	Rate ratio (P-value)
Synonymous	906 / 3061	155 / 500	0.96 (0.60)
Missense (All)	1582 / 3061	274 / 500	0.94 (0.38)
Missense (MPC ≥ 2)	158 / 3061	26 / 500	0.99 (0.92)
PTV (All)	460 / 3061	65 / 500	1.16 (0.32)
PTV (pLI ≥ 0.9)	231 / 3061	26 / 500	1.45 (0.07)

Besides *de novo* variants, we also had access to inherited variants from the entire ASD and congenital heart disease cohorts and a subset of the ID/DD cohorts (Materials and methods). We were specifically interested in whether rare inherited and *de novo* variants reside in the same or different genes. We counted the number of *de novo* and inherited PTVs absent from ExAC per gene and found they are inversely correlated ($r = -0.42$; $P = 3.74 \times 10^{-6}$; Pearson's product-moment correlation). Given that inherited variants have undergone at least one generation of selection, one could certainly imagine that genes under weak or no negative selection accumulate more *de novo* PTVs than genes under increasingly stronger degrees of negative selection. Using published selection coefficients for heterozygous protein truncating variants in 15,998 genes[18],

we examined the distribution of *de novo* and inherited PTVs per gene (**Figure 4.4**). Indeed, the increasing values of selection coefficients are associated with more *de novo* (β=0.41±0.02; $P < 2 \times 10^{-16}$; linear regression), but fewer inherited PTVs (β=-0.06±0.01; P=1.06 x 10^{-7}; linear regression).

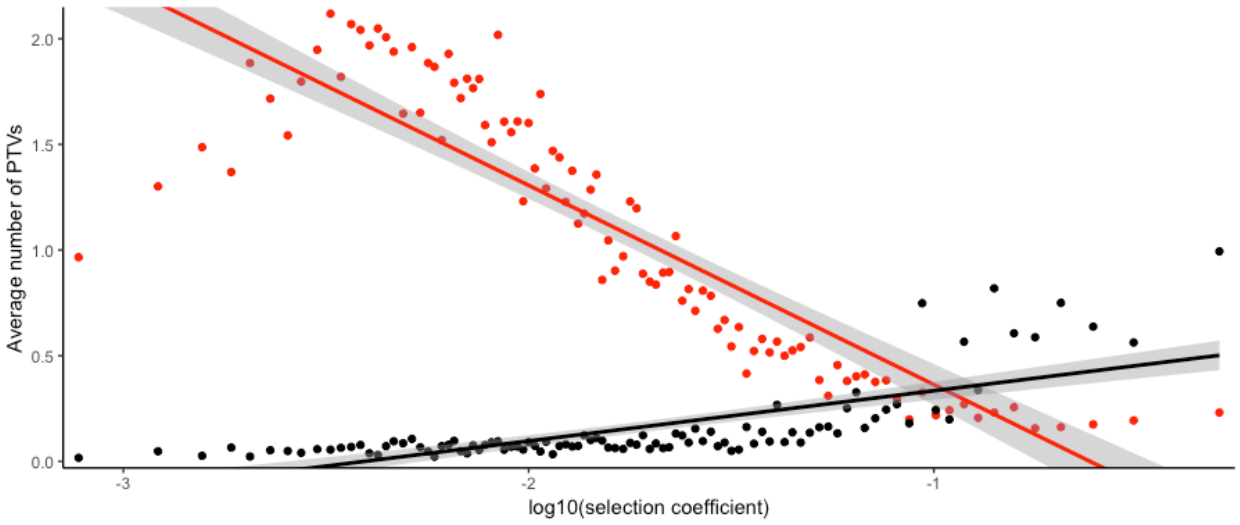


Figure 4.4: Distribution of *de novo* (black dots) and inherited protein truncating variants (PTVs; red dots) per bin of 160 genes (y-axis) across the negative $\log_{10}$ distribution of selection coefficients (x-axis). Linear regressions with 95% confidence intervals for both *de novo* (solid black line) and inherited (solid red line) PTVs.

Gene discovery across phenotypes and comorbidities

After examining how the frequency of *de novo* missense (MPC ≥ 2) and PTVs (pLI ≥ 0.9) absent from ExAC differs between and within ascertainments, we turned our attention to gene discovery. We identified genes with either 1) more *de novo* PTVs or 2) more *de novo* PTVs and *de novo* missense variants (MPC ≥ 2) than expected under a null mutation model[19]. Testing two sets of variant classes in 18,226 genes gives a Bonferroni significance threshold of ~1.37×10^{-6} (Materials and methods).

We first tested each ascertained trait separately identifying 95 genes in ascertained ID/DD, 18 genes in ascertained ASD, 11 genes in ascertained congenital heart disease, and no genes in unaffected ASD siblings (**Table 4.6**). Given the significantly different frequency of *de novo* missense and PTVs between ASD with and without ID/DD and CHD with and without ID/DD, we re-ran the gene discovery in these four sets and as well as the corresponding splits in ascertained ID/DD. Unsurprisingly, the number of Bonferroni significant genes differed drastically between the ascertained ASD with (N=15; **Table 4.6**; **Figure 4.5A**; $P < 0.001$; permutation; Materials and methods) and without ID/DD (N=4; **Table 4.6**; **Figure 4.5B**; $P = 0.06$; permutation) and ascertained congenital heart disease with (N=14; **Table 4.6**; **Figure 4.5C**; $P < 0.001$; permutation) and without ID/DD (N=0; **Table 4.6**; **Figure 4.5D**; $P < 0.001$; permutation). In particular, we discovered more Bonferroni significant genes in ascertained CHD with ID/DD than the entire set of ascertained CHD probands which is remarkable given that the former's sample size is 60.87% smaller which is hardly what one would expect given the conventional wisdom that larger sample sizes are (for the most part) always better[20].

Just as we evaluated how many Bonferroni significant genes could be discovered in ascertained ASD and ascertained CHD with and without ID/DD, we performed the same analysis with the ascertained 5305 ID/DD with and without ASD and similarly with and without congenital heart disease. Ten genes were Bonferroni significant in 711 ascertained ID/DD with ASD individuals (**Table 4.6**; **Figure 4.5E**; $P = 0.14$; permutation), 87 genes were Bonferroni significant in 4594 ascertained ID/DD without ASD individuals (**Table 4.6**; **Figure 4.5F**; $P = 0.15$; permutation), ten genes were Bonferroni significant in 571 ascertained ID/DD with congenital heart disease individuals (**Table 4.6**; **Figure 4.5G**; $P = 0.13$; permutation), and 88 genes were Bonferroni significant in 4734 ascertained ID/DD without congenital heart disease

individuals (**Table 4.6**; **Figure 4.5H**; $P = 0.16$; permutation). Unlike the gene discovery analysis in ascertained ASD and congenital heart disease, the number of Bonferroni significant genes did not differ from a random sample matched on sex and cohort (Materials and methods).

Lastly, if one assumes that ASD and ID/DD are simply arbitrary labels and differentiating between the two is semantic[21,22], then combining the data together to approximately double the sample size should drastically increase power for gene discovery. This is clearly not the case, given that only 111 genes surpass the Bonferroni significance threshold, of which 16 were novel to this analysis, but come at the cost of losing ten (**Table 4.6**). Furthermore, jamming all the data together agnostic to ascertainment produces 109 genes, thus losing power (**Table 4.6**).

Table 4.6: Number of Bonferroni significant genes discovered via *de novo* variation. ID/DD- and ASD-overlap columns represent how many of those genes were also Bonferroni significant in the full ascertained trait-only analysis.

Ascertained trait	Set	Sample size	N genes	ID/DD overlap	ASD overlap
ASD	ASD (All)	6430	18	13	-
ASD	ASD (with ID/DD)	2402	15	10	13
ASD	ASD (without ID/DD)	2895	4	0	2
ASD	meta-analysis (ASD + ID/DD [with ASD])	7141	25	19	16
ID/DD	ID/DD (All)	5305	95	-	13
ID/DD	ID/DD (with ASD)	711	10	10	6
ID/DD	ID/DD (without ASD)	4594	87	85	13
ID/DD	ID/DD (with CHD)	571	10	9	4
ID/DD	ID/DD (without CHD)	4734	88	85	13
ID/DD	meta-analysis (ID/DD + ASD [with ID/DD] + CHD [with ID/DD])	8279	112	92	17
CHD	CHD (All)	3683	11	7	3
CHD	CHD (with ID/DD)	1073	14	9	4
CHD	CHD (without ID/DD)	1673	0	0	0
Control	Control (All)	2179	0	0	0
NDD	ASD + ID/DD	11735	111	90	18
All	ID/DD + ASD + CHD	14917	109	84	18

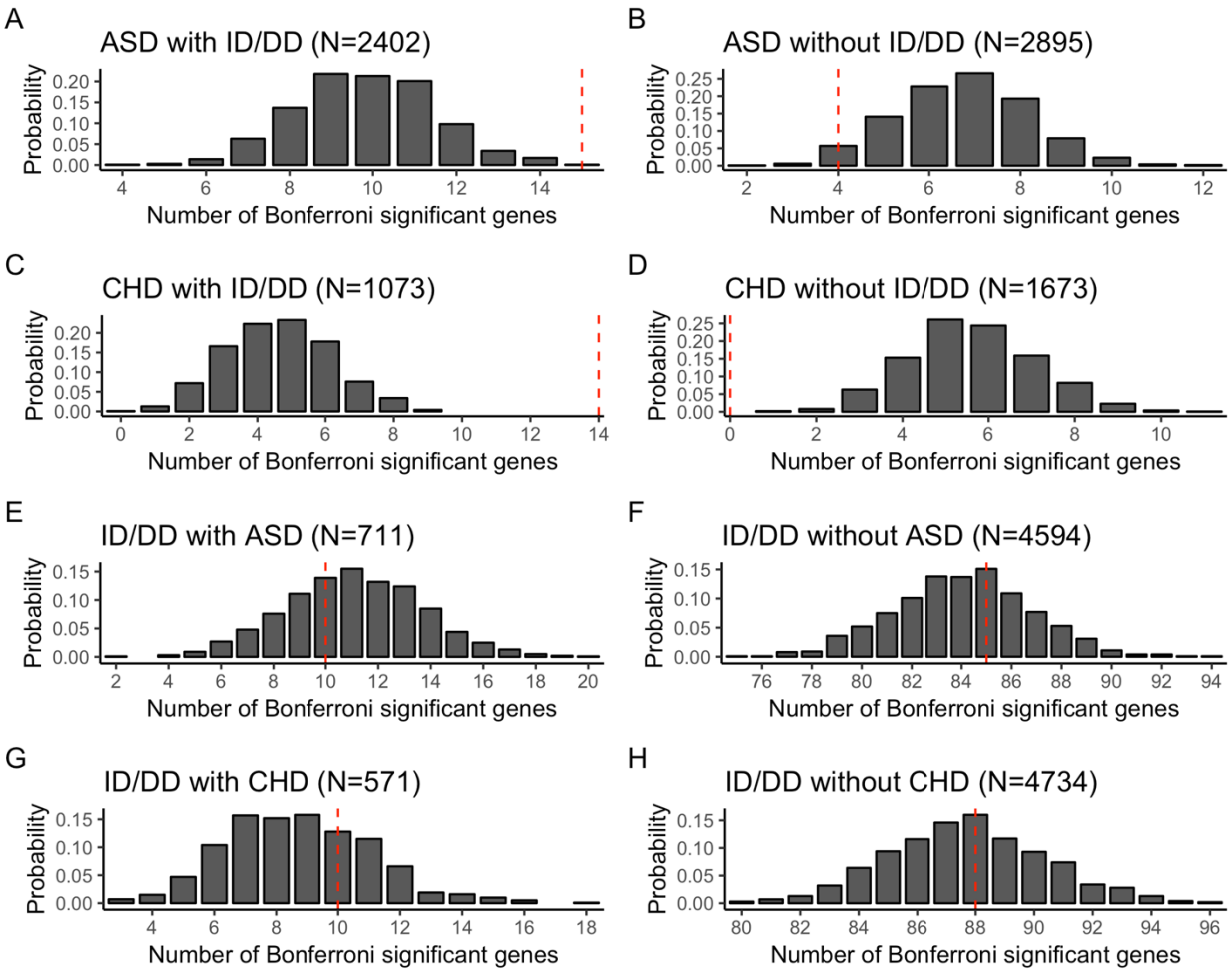


Figure 4.5: Number of Bonferroni significant genes from 1000 permutations of randomly sampled individuals controlling for sex and cohort (Materials and methods). The dashed, vertical red line marks the observed number of Bonferroni significant genes, and the *N* in parenthesis indicates the number of probands.

Evaluating disease specificity of associated genes

Despite the 18 genome-wide significant ASD genes discovered by ascertaining on ASD, we failed to observe a greater frequency of associated *de novo* variants among these eleven genes in ascertained ID/DD probands (0.042 per individual) than ascertained ASD probands (0.022 per individual; rate ratio=1.93; 95% CI: 1.55-2.39; $P = 7.63 \times 10^{-10}$; two-sided, two-sample Poisson

exact test). This observation extends to both a larger list of 80 manually curated ASD genes from SFARI (rate ratio = 2.63; 95% CI = 2.09-3.34; P=2.33x10^{-19}) and 102 genes at a false discovery rate < 0.1 from Satterstrom and colleagues[23] (rate ratio = 1.64; P = 9x10^{-17}; two-sided, two-sample Poisson exact test; Materials and methods). Both genome-wide significant genes in ASD and larger lists of ASD genes contain a mixture of genes that confer more risk to ASD or more risk to ID/DD in line with what others have observed[23]. Furthermore, this same pattern extends to congenital heart disease: the 11 congenital heart disease Bonferroni significant genes in are 1.33-fold enriched (95% CI: 1.02-1.75) in ascertained ID/DD probands (P = 0.03; two-sided, two-sample Poisson exact test).

For each of the 105 Bonferroni significant genes in the three ascertained traits, we compared the combined frequency of *de novo* missense (MPC ≥ 2) and PTVs in ascertained ID/DD to both ascertained ASD and ascertained CHD using a two-sided, two-sample Poisson exact test under the assumption that a higher frequency of *de novo* would be indicative of the gene conferring more risk (and therefore more preferential) towards one trait than the other. After Bonferroni correction for 105 genes and two sets of comparisons, nine genes (*ARID1B* [P = 2.91 x 10^{-6}], *ANKRD11* [P = 6.65 x 10^{-8}], *ASXL3* [P = 4.46 x 10^{-5}], *GATAD2B* [P = 1.61 x 10^{-4}], *DDX3X* [P = 3.39 x 10^{-7}], *KAT6B* [P = 3.29 x 10^{-5}], and *KMT2A* [P = 1.38 x 10^{-6}]) had a higher combined frequency of *de novo* missense (MPC ≥ 2) and PTVs in ID/DD than ASD and congenital heart disease. Only one gene, *KMT2D*, had a significantly higher combined frequency of *de novo* missense (MPC ≥ 2) and PTVs in congenital heart disease than ID/DD (P = 1.73 x 10^{-4}).

Discussion

An oft-posed question within the field of psychiatric genetics is whether we are discovering genes not because they confer specific risk to the ascertained trait (i.e., ASD), but because they generally impair for cognition - which might either generally place individuals at higher baseline risk for diagnosis or generate associations specifically driven by the cognitively impaired subset of the sample[10,24-33]. Following this logic suggests whole exome and genome sequencing for gene discovery via *de novo* variation may be primarily useful for understanding ASD in the background of other severe neurodevelopmental disorders.

Here, we aggregated *de novo* and inherited coding variants across 6430 ascertained ASD, 5305 ascertained ID/DD, and 3683 ascertained congenital heart disease samples along with extensive phenotyping data and examined the influence of ID/DD on the frequency of *de novo* variation and gene discovery efforts in both ASD and congenital heart disease. We found that ASD and congenital heart disease did not affect the frequency of deleterious *de novo* variants within ascertained ID/DD samples, but the converse was not true. Furthermore, we observed no difference in the frequency of deleterious *de novo* variants between ascertained ASD and ascertained congenital heart disease samples with comorbid ID/DD. Using the same statistical method and significance threshold, we discovered 95 Bonferroni significant genes in ascertained ID/DD, 18 in ascertained ASD, and 11 in ascertained congenital heart disease. Removing the ascertained ASD and congenital heart disease samples with comorbid ID/DD eliminated all the genes in congenital heart disease and left only four in ASD, two of which, *ANK2* and *GIGYF1*, were novel to this analysis. With the individuals without comorbid ID/DD removed, both traits had improved power for gene discovery.

Lastly, we found that the bulk of rare *de novo* and inherited PTVs reside in different sets of genes along the distribution of selection. The greatest excesses of *de novo* PTVs reside in the genes under the strongest degree of negative selection where such variation may not survive beyond a couple generations. In contrast, the majority of inherited PTVs reside on the opposite end of the distribution. This suggests that *de novo* variation will be most useful (and most powered) to discover trait-associated genes under the strongest selection, inherited genetic variation will be most useful to identify trait-associated genes under the weakest selection, and genetic variation in case-control study designs will cover all genes up to the last tails of the distribution. As such, studies focusing on *de novo* variation will continue to identify genes under the strongest negative selection, and those genes are most often associated with severe ID/DD.

Materials and methods

Published *de novo* variants

We downloaded 12,166 *de novo* variants from 6430 individuals ascertained for ASD and 2623 *de novo* variants from 2179 unaffected siblings from Satterstrom, Kosmicki, Wang and colleagues[23]. We downloaded 8232 published *de novo* variants in 5305 individuals ascertained for ID/DD from five separate publications[20,34-38]. For congenital heart disease, we downloaded 5004 *de novo* variants from two publications of congenital heart disease[29,39] and removed 326 overlapping samples, and one sample that failed a sex-check post-publication (individual GT04014641; Steven DePalma, personal correspondence) and their *de novo* variants, bringing the total to 4594 *de novo* variants in 3683 individuals ascertained for congenital heart disease. To ensure uniformity in variant representation and annotation across datasets, we followed the same procedure described in Kosmicki et al[40]. Briefly, we standardized variant representation

through a Python implementation of vt normalize[41] and re-annotated all variants using Variant Effect Predictor (VEP)[42] version 81 with Gencode v19 on GRCh37. VEP provided the Ensembl Gene IDs, gene symbol, and the Ensembl Transcript ID for use in determining canonical transcripts. When a variant fell across multiple transcripts, we used the canonical transcript when possible and the most deleterious annotation in cases of multiple canonical transcripts. If no canonical transcript was available, the most deleterious annotation was used. Following the protocol from DDD[20], we restricted the number of *de novo* variants to one variant per person per gene prioritizing the *de novo* variant with the most severe consequence which removed 37 ID/DD and nine congenital heart disease *de novo* variants.

As the *de novo* variant data from congenital heart disease and intellectual disability came from multiple studies, we tested whether the frequency of *de novo* synonymous variants in each study was the same. If a study had too few *de novo* synonymous variants, then we would exclude it from the analysis. When we tested all of these cohorts, we failed to observe a significant difference in the frequency of *de novo* synonymous variants (**Table 4.7**) and proceeded to combine all of the intellectual disability and congenital heart disease data as a single group.

Table 4.7: *De novo* synonymous variant frequencies. Checking for differences in the *de novo* synonymous variant frequency between three cohorts ascertained for intellectual disability (ID) and two cohorts ascertained for congenital heart disease (CHD) using the Poisson exact test (also known as the *C*-test)[8]. We perform pairwise comparisons for a difference in the frequency of *de novo* synonymous variants between studies with a Bonferroni significant threshold of 0.05 / 7 = 0.0071. We observe no such difference in the frequency of *de novo* synonymous variants, and thus grouped the three ascertained ID cohorts together and two ascertained CHD cohorts into single cohorts of ascertained ID and ascertained CHD respectively.

Comparison	Frequency 1	Frequency 2	*P*-value	Rate Ratio	95% CI
Lelieveld vs. Hamdam	207 / 820	10 / 41	1.0000	1.0100	0.54 - 2.14
Lelieveld vs. Rauch	207 / 820	10 / 51	0.5602	1.2563	0.67 - 2.66
Lelieveld vs. de Ligt	207 / 820	30 / 100	0.2926	0.8211	0.56 - 1.25
Hamdam vs. Rauch	10 / 41	10 / 51	0.6580	1.2439	0.46 - 3.33
Hamdam vs. de Ligt	10 / 41	30 / 100	0.7278	0.8130	0.35 - 1.71
Rauch vs. de Ligt	10 / 51	30 / 100	0.3156	0.6536	0.29 - 1.37
Sifrim vs. Jin	364 / 1365	701 / 2645	0.9227	1.0062	0.88 - 1.14

Defining phenotypes

The ASD, ID/DD, and congenital heart disease cohorts comprise a heterogeneous collection of individuals. As such, one would certainly like to strive toward more homogenous groups through secondary diagnoses. To ensure a fair comparison, we defined intellectual disability and developmental delay within the ascertained ASD cohort in the same manner as was defined in the ascertained intellectual disability and ascertained developmental delay cohorts. For ascertained probands with ASD, a child was considered to have intellectual disability if they met one of the following conditions: a full-scale IQ (FSIQ) < 70 (N=1546), if the proband was administered but could not complete an IQ test, indicated by the child having a date for their IQ test but no IQ score (N=85), or if the child had an HPO term or ICD code indicating intellectual disability (N=591) or mental retardation (N=8). The primary phenotypes included in the developmental disorders study included intellectual disability, developmental delay, motor delay, and seizures, of which intellectual disability was noted in >90% of the cohort[20]. Any proband ascertained for ASD who had experienced seizures (N=496), had a previous diagnosis of

developmental delay or motor delay (N=31), or had begun walking after 18 months of age (measured in item 5A on the autism diagnostic interview[43]) (N=366) was considered to have a comorbid developmental disorder. Together, 2402 of the 6430 probands ascertained for ASD met at least one of these conditions and as such had comorbid ID/DD. Another 2895 probands ascertained for ASD had neither intellectual disability, developmental delay, seizures, nor motor delay and as such comprised a subset of ASD probands without ID/DD. The remaining 1133 ascertained ASD probands lacked information on all or a subset of intellectual disability, seizures, and motor delay status and were categorized as having an unknown ID/DD status. Similarly, among ascertained ID or ascertained developmental delay probands, those with a diagnosis of ASD were considered to have comorbid ASD (N=711) and those with either of the following conditions: ventricular septal defect, abnormality of the aortic valve, defect in the atrial septum, coarctation of aorta, tetralogy of Fallot, atrioventricular canal defect, abnormality of the pulmonary valve, hypoplastic left heart, patent ductus arteriosus, transposition of the great arteries with ventricular septal defect, abnormal branching pattern of the aortic arch, double outlet right ventricle, situs inversus totalis, abnormality of the mitral valve, abnormality of the vena cava, hypoplastic aortic arch, transposition of the great arteries with intact ventricular septum, pulmonary valve atresia, abnormality of cardiac morphology, cardiomyopathy, abnormality of the coronary arteries, abnormality of the left ventricular outflow tract, tricuspid atresia, abnormality of the tricuspid valve, abnormality of the pulmonary artery, total anomalous pulmonary venous return, mitral atresia, hypoplastic right heart, double inlet left ventricle, partial anomalous pulmonary venous return, pulmonary artery atresia, Ebstein's anomaly of the tricuspid valve, left atrial isomerism, congenitally corrected transposition of the great arteries, truncus arteriosus, right atrial isomerism, abnormality of the left ventricle, hypoplasia of right ventricle,

arrhythmia, abnormality of cardiac atrium, interrupted aortic arch, abnormality of the right ventricle, secundum atrial septal defect, pulmonic stenosis, were considered to have congenital heart disease (N=572). The remaining 4594 and 4734 ID/DD probands do not have ASD and congenital heart disease, respectively (**Table 4.1**). For the congenital heart disease cohorts,

Defining intellectual disability within the confines of ASD

Full-scale IQ scores were measured using several tests. These tests include, but are not limited to, the Differential Ability Scales, Second Edition[44]; the Mullen Scales of Early Learning[45]; the Wechsler Intelligence Scale for Children[46]; and the Wechsler Abbreviated Scale of Intelligence[47]. The full-scale IQ estimates were taken from the full-scale deviation IQ variable when available and full-scale ratio IQ when it was not[48]. Full-scale IQ is normally distributed with a mean of 100 and a standard deviation of 15.

The expected number of *de novo* variants per person under a null Poisson distribution

For a class of *de novo* variation, *c*, in trait *t*, one can estimate the expected number of individuals, *x*, carrying $d \in \{0, 1, 2, 3, 4, 5, \dots\}$ *de novo* variants under a null Poisson distribution from the probability mass function:

$$x_{t,c,d} = \frac{\lambda_{t,c}^{d} e^{-\lambda_{t,c}}}{d!} n_t$$

where $\lambda_{t,c}$ is the frequency of *de novo* variants of variant class *c* in trait *t*. We considered four separate variant classes for *c*: synonymous, missense, PTV, and combined missense (MPC ≥ 2) and PTVs in Bonferroni significant genes. We calculated $\lambda_{t,c}$ from the total number of observed *de novo* variants of class *c* in trait *t* such that each trait and each variant class will receive their own $\lambda_{t,c}$. This is scaled by the total number of individuals in trait *t* so that we can directly

compare the observed number of individuals carrying *d de novo* variants in variant class *c* in trait *t* to how many would be expected, assuming each *de novo* variant was randomly distributed (and with no association on the phenotype). We used a chi-square goodness of fit test in R with 4 degrees of freedom to compare the observed to expected counts at value of *d*. Because the value of *d* ranges from 0 to +∞, we capped the distribution to 3 by subtracting the total number of individuals n_t from the sum of $x_{t,c,d}$ where $d \in \{0, 1, 2, 3\}$ such that the sum of $x_{t,c,d}$ for values of $d \in \{0, 1, 2, 3\} = n_t$.

Gene discovery using *de novo* variants

Following the framework proposed by Samocha et al.,[19] we evaluated 18,226 genes for enrichment of *de novo* variants across a variety of different traits and trait subsets (**Table 4.6**). Statistical significance was calculated under the null hypothesis that the observed and expected number of *de novo* variants given the gene-specific mutation rate and the number of chromosomes were equal. For all 18,226 genes with measured mutation rates, the number of expected *de novo* variants in gene *i* of consequence class *j* follows a Poisson distribution

$$Poisson(\lambda_{i,j})$$

$$\lambda_{i,j} = \mu_{i,j} c$$

where $\mu_{i,j} c$ is the mutation rate for gene *i*, consequence class *j* multiplied by the number of chromosomes (*c*). For autosomal genes and pseudo-autosomal genes on X, $c = 2n_f + n_m$ where n_f and n_m are the number of females and males, respectively. For non-pseudo-autosomal genes on X, $c = 2n_f + \phi_f n_m$, and lastly for genes on the Y chromosome, $c = \phi_m n_m$, where $\phi_m = \frac{2}{1+\frac{1}{\alpha}}$ and $\phi_f = \frac{2}{1+\alpha}$ are correction factors for the different mutation rates in males and females,

respectively (as males have a higher mutation rate than females). The scaling factor, α, was obtained from the phased *de novo* variants in the Scottish Family Health Study and defined over the values $\{\alpha \mid \alpha \in \mathbb{R}, \alpha \neq \{-1,0\}\}$ which in our case is 3.4. Under our null model, we calculated the probability of finding an equal or more extreme number of *de novo* variants in gene *i* of consequence class *j*, compared to the observed number of *de novo* variants in the given cohort. We used the ppois function in R to calculate the *P*-value. We considered two sets of consequence classes: 1) PTVs and 2) PTVs + missense (MPC ≥ 2). The resulting Bonferroni significance threshold for 18,226 genes and 2 variant consequence classes gives a *P*-value threshold = 0.05 / (2*18226) = 1.37 x 10^{-6}.

We calculated the expected number of *de novo* missense (MPC ≥ 2) variants in the same manner as Samocha et al[19]. Briefly, for all 66,939,307 possible single nucleotide variants in 17,915 genes with MPC scores[9], we used the expected mutation rate for trinucleotide -> trinucleotide *i* as the expected mutation rate for the given variants, such that each variant now has a value of μ. $\mu_{MPC \geq 2}$ was then calculated by removing all SNVs with a MPC < 2, grouping by gene, and summing the remaining the variants to get the expected $\mu_{MPC \geq 2}$ for each gene. Due to the nature of MPC, only 5,146 genes have at least one possible SNV with a MPC ≥ 2. As such, the remaining 12,769 genes have $\mu_{MPC \geq 2} = 0$.

Bonferroni significant gene permutations

When we subsetted the ascertained samples of ASD, congenital heart disease, and ID/DD to specific comorbidities, we observed a number of Bonferroni significant genes. However, because for almost every single one of these analyses, we observed fewer Bonferroni significant genes (as one would expect given a smaller sample size), we inquired whether these observations

were unexpected from a random subset. To evaluate the significance of these observations, we performed 1000 permutations, randomly sampling the same number of individuals controlling for sex and cohort. With a random sample of individuals, we extracted their *de novo* variants and tested each gene for significance in the same procedure as laid out the previous section.

Estimating negative selection from *de novo* and inherited variation

Following the logic described by Zuk and colleagues[49], one can estimate the negative selection for a given class of variation in a gene based on the fraction of variation that is *de novo* (i.e., s = *de novo* / total number of variants). We used the 15,998 empirical estimates of s for heterozygous protein truncating variants from Cassa and colleagues[18], and examined the relationship between *de novo* and inherited PTVs with respect to s. Due to the sparse nature of the data (11,857 genes with at least one inherited PTV and 1639 genes with at least *de novo* PTV), we created 160 bins of ~100 genes in ascending values of s. In each of the 160 bins, we calculated the average number of *de novo* and inherited PTVs across all genes in the bin from the observed *de novo* and inherited data. We ran two linear regressions (one for *de novo* and the other for inherited PTVs) regressing s on the average number of PTVs.

Testing for trait specificity

Given the large overlap between genes discovered across the ascertained traits, it is natural to inquire whether any genes are preferential towards one trait or the other. For the 105 significantly associated genes across the three traits, we compared the combined frequency of *de novo* missense (MPC $\geq$ 2) and PTVs in ascertained ID/DD to both ascertained ASD and ascertained CHD with a two-sided, two-sample Poisson exact test. Our Bonferroni correction for

105 genes and two sets of comparisons was 2.38×10^{-4}. That being said, this approach is very conservative, and we are underpowered at current sample sizes to differentiate between genes whose differential risk conferred to each trait is small. To illustrate, *CHD7* has 13 *de novo* PTVs and one *de novo* missense (MPC ≥ 2) in congenital heart disease (combined frequency 0.0038 per congenital heart disease proband) compared to three *de novo* PTVs and 0 *de novo* missense (MPC ≥ 2) variants in ascertained ID/DD (combined frequency = 5.66×10^{-4}) – so it is roughly an order of magnitude greater in congenital heart disease than ID/DD. But yet the frequency difference is not significant after Bonferroni correction ($P = 1.45 \times 10^{-3}$). Furthermore, a gene such as *GIGYF1* that was only Bonferroni significant in ascertained ASD without ID/DD with all four *de novo* PTVs present in individuals without intellectual disability, developmental delay, seizures, and delayed walking has no chance of being significant even though there are zero *de novo* PTVs in the ascertained ID/DD cohorts. As such, there are very likely more trait-preferential genes, but larger sample sizes are necessary using this approach or a better approach (perhaps a permutation-based approach) that accounts for comorbidities could be developed to resolve these issues.

Author Contributions

Jack Kosmicki: method design, data analyses (exceptions below), writing

Mark Daly: method design, writing, overall guidance

Liu He: provided phenotypes for the DDD trios.

Jeffrey Barrett: provided data for DDD, method design, overall guidance,

Matthew Hurles: provided insight into DDD and provided helpful comments

Christian Gilissen: provided phenotypes for the RUMC cohort and helpful comments

References

1. Baio, J. *et al.* Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *MMWR. Surveillance Summaries* **67**, 1--23 (2018).

2. Amir, R.E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185-8 (1999).

3. Brown, W.T. *et al.* Association of fragile X syndrome with autism. *Lancet* **1**, 100 (1982).

4. Hunt, A. & Shepherd, C. A prevalence study of autism in tuberous sclerosis. *J Autism Dev Disord* **23**, 323-39 (1993).

5. Sanders, Stephan J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams Syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011).

6. Weiss, L.A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**, 667-675 (2008).

7. Zaidi, S. & Brueckner, M. Genetics and Genomics of Congenital Heart Disease. *Circ Res* **120**, 923-940 (2017).

8. Przyborowski, J. & Wilenski, H. Homogeneity of Results in Testing Samples from Poisson Series: With an Application to Testing Clover Seed for Dodder. *Biometrika* **31**, 313-323 (1940).

9. Samocha, K.E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* (2017).

10. Satterstrom, F.K. *et al.* ASD and ADHD have a similar burden of rare protein-truncating variants. *bioRxiv* (2018).

11. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* **46**, 881-885 (2014).

12. Weiner, D.J. *et al.* Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nature Genetics* **49**, 978--985 (2017).

13. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics* **51**, 431-444 (2019).

14. Turner, T.N. *et al.* Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710--722.e12 (2017).

15. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)* **316**, 445-449 (2007).

16. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192-195 (2010).

17. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-246 (2012).

18. Cassa, C.A. *et al.* Estimating the Selective Effects of Heterozygous Protein Truncating Variants from Human Exome Data. *Nature genetics* **49**, 806-810 (2017).

19. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950 (2014).

20. Deciphering Developmental Disorders, S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438 (2017).

21. Coe, B.P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat Genet* **51**, 106-116 (2019).

22. Nguyen, H.T. *et al.* mTADA: a framework for analyzing de novo mutations in multiple traits. *bioRxiv*, 406868 (2018).

23. Satterstrom, F.K. *et al.* Novel genes for autism implicate both excitatory and inhibitory cell lineages in risk. *bioRxiv*, 484113 (2018).

24. Chiurazzi, P. & Pirozzi, F. Advances in understanding – genetic basis of intellectual disability. *F1000Research* **5**, F1000 Faculty Rev-599 (2016).

25. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-84 (2014).

26. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-1266 (2015).

27. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-372 (2010).

28. Robinson, E.B. *et al.* Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proc Natl Acad Sci U S A* **111**, 15161-15165 (2014).

29. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat Genet* **48**, 1060-1065 (2016).

30. Singh, T. *et al.* Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci* **19**, 571-7 (2016).

31. Singh, T. *et al.* The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nat Genet* **49**, 1167-1173 (2017).

32. Skuse, D.H. Rethinking the nature of genetic vulnerability to autistic spectrum disorders. *Trends Genet* **23**, 387-395 (2007).

33. Wilfert, A.B., Sulovari, A., Turner, T.N., Coe, B.P. & Eichler, E.E. Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. *Genome Med* **9**, 101 (2017).

34. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* **367**, 1921-9 (2012).

35. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344-7 (2014).

36. Hamdan, F.F. *et al.* De novo mutations in moderate or severe intellectual disability. *PLoS Genet* **10**, e1004772 (2014).

37. Lelieveld, S.H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nature Neuroscience* **19**, 1194-1196 (2016).

38. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674-82 (2012).

39. Jin, S.C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet* **49**, 1593-1601 (2017).

40. Kosmicki, J.A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat Genet* **49**, 504-510 (2017).

41. Tan, A., Abecasis, G.R. & Kang, H.M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202-4 (2015).

42. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).

43. Lord, C., Rutter, M. & Lecouteur, A.a. Autism Diagnostic Interview-Revised - a Revised Version of a Diagnostic Interview for Caregivers of Individuals With Possible Pervasive Developmental Disorders. *Journal of Autism and Developmental Disorders* **24**, 659--685 (1994).

44. Elliott, C.D. *Differential Ability Scales (2nd ed.)*, (Harcourt Assessment, San Antonio, TX, 2007).

45. Mullen, E.M. *Mullen Scales of Early Learning Manual*, (American Guidance Service, Circle Pines, MN, 1995).

46. Wechsler, D. *WISC III (Wechsler Intelligence Scale for Children)*, (Psychological Corporation, San Antonio, TX, 1992).

47. Wechsler, D. *Wechsler Abbreviated Scale of Intelligence WASI: Manual*, (The Psychological Corporation, 1999).

48. Chaste, P. *et al.* A Genome-wide Association Study of Autism Using the Simons Simplex Collection: Does Reducing Phenotypic Heterogeneity in Autism Increase Genetic Homogeneity? *Biological Psychiatry* **77**, 775-784 (2015).

49. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-64 (2014).

Chapter 5

Discussion

The main goal of this dissertation was to investigate the role of rare *de novo* and inherited coding variation in neurodevelopmental disorders and use these sources of rare coding variation for gene discovery. To that end, we explored the properties of recurrently mutated sites in the genome and their contribution to trait etiology, performed the largest exome sequencing study of ASD to date identifying 26 Bonferroni significant associated genes and upwards of 102 genes (FDR<0.1), and the influence of intellectual disability on the *de novo* variant architecture and gene discovery in autism spectrum disorder and congenital heart disease. Furthermore, we were able to better investigate older hypotheses (mutational recurrence, oligogenic architecture, *de novo* architecture of simplex and multiplex families) illustrating how prior assumptions and conclusions can be re-evaluated with significantly larger sample sizes.

Summary of results

Recurrently mutated sites in the genome

Many studies examining *de novo* variation[1-4] either explicitly or implicitly made the assumption that each *de novo* variant site was novel, in line with Kimura's infinite sites model[5]. We found that ~1 in 3 observed *de novo* variants in intellectual disability / developmental delay (ID/DD), autism spectrum disorder (ASD) and unaffected ASD siblings were also observed as standing variation in the 60,706 reference exomes from the exome aggregation consortium (ExAC)[6] (**Figure 2.2A**). We ran five secondary analyses to confirm the validity of these findings, ranging from evaluating the frequency of CpG variants to testing validation rates and the allele frequency distribution (**Figure 2.2C**).

We then sought to determine whether both *de novo* protein truncating variants (PTVs) present and absent from ExAC contributed equally to ASD and ID/DD risk. We found no

difference in the frequency of *de novo* PTVs present in ExAC between ASD, ID/DD, and unaffected ASD siblings (**Figure 2.3B**). By contrast, *de novo* PTVs absent from ExAC were significantly more enriched in individuals with ASD and ID/DD as opposed to unaffected ASD siblings (**Figure 2.3B**). The lack of excess case burden in *de novo* PTVs present in ExAC was consistent with what would be expected for a neutral class of variation, similar to *de novo* synonymous variants.

Moving from the variant level to the gene level, we evaluated whether the overall frequency of PTVs per gene in ExAC provided a similar guide to which ASD and ID/DD variants were relevant. We used the gene-level constraint metric, pLI[16] (probability of loss-of-function intolerance), to evaluate whether *de novo* PTVs absent from ExAC conferred more risk in LoF-intolerant (pLI $\geq$0.9) than LoF-tolerant genes (pLI <0.9). We found a stronger enrichment of *de novo* PTVs absent from ExAC in LoF-intolerant gene in both ASD (rate ratio = 3.24; $P = 3.14 \times 10^{-16}$) and ID/DD (rate ratio = 6.7; $P = 6.34 \times 10^{-38}$), and no observed excess in LoF-tolerant genes (**Figure 2.3D**). Hence, all detectable *de novo* PTV signal in these phenotypes can be localized to 18% of genes with clear intolerance to PTVs in ExAC, with, consequently, substantially amplified rate ratios in this gene set. Furthermore, all of the previous observed phenotypic associations with *de novo* variants were not only preserved, but also enhanced, when removing *de novo* variants present in ExAC (**Figure 2.4**).

We also investigated whether inherited and case-control variation could be similarly enhanced using ExAC as a variant-level filter and pLI as a gene-level filter. Comparing transmitted to untransmitted PTVs absent from ExAC and in LoF-intolerant genes, we found a modest excess of transmitted PTVs in ASD cases (rate ratio = 1.16; $P = 9.85 \times 10^{-3}$), which was minute in comparison to that of the *de novo* signal. Lastly, using a Swedish cohort of 404 ASD

cases and 3654 controls, we found yet another significant association of singleton PTVs absent from ExAC in LoF-intolerant genes (2.63 OR; $P = 1.37 \times 10^{-18}$), which was the first instance of an exome-wide excess of PTVs in an ASD case-control analysis.

Gene discovery via exome sequencing in ASD

Here we carried out the largest exome sequencing study of ASD to-date with 35,584 samples. This sample comprised both family-based data (6,430 ASD cases, 2,179 unaffected controls, and both parents) and case-control data (5,556 ASD cases, 8,809 controls). We found a significant enrichment in *de novo*, inherited, and case-control missense (MPC ≥2) and PTVs in genes with a pLI ≥ 0.995 (**Figure 3.1B**). Using an enhanced Bayesian gene discovery method that incorporated both gene- and variant-level constraint (**Figure 3.2A**), we found 26 Bonferroni significant genes and upwards of 102 genes (FDR < 0.1; **Figure 3.2B**) associated with ASD, 31 of which represent novel associations.

We next examined what biological and phenotypic insights this larger list of genes could provide us about the underlying biology of ASD. First, we examined the distribution of missense variants and found missense clustering in the functional domains of *DEAF1*, *SCN1A*, and *SLC6A1* and gain-of-function variants contributing to ASD in *KCNQ3* (**Figure 3.3A-E**). Given that large CNVs confer risk to ASD, we examined whether any of the 102 genes could provide insight into which, if any, gene(s) were driver in 51 genomic disorder loci. We found 13 ASD-associated genes resided in 12 genomic disorder loci, which was unexpected by chance (2.3-fold increase; $P = 2.3 \times 10^{-3}$) and nominated *HDLBP* in 2q37.3 and both *SHANK2* and in 11q13.2-q13.4 (**Figure 3.3F** and **3.3G**).

Lastly, given that many individuals with ASD also have comorbid ID/DD, we attempted to determine whether any of the 102 (FDR < 0.1) genes conferred more risk to ID/DD or conferred more risk to ASD. By comparing the frequency of *de novo* missense (MPC > 1) and PTVs in 5264 ascertained ID/DD individuals. We found 53 genes conferred more risk to ASD (ASD-preferential) and 49 genes conferred more risk to ID/DD (NDD-preferential; **Figure 3.4A**). We then tested to see whether individuals with ASD who carried associated variants in the NDD-preferential genes were phenotypically different from ASD individuals carrying associated variants ASD-preferential variants. We found that ASD individuals with associated variants in NDD-preferential genes walked 2.6 ± 1.2 months later in life (**Figure 3.4B**) and had a full-scale IQ 11.9 ± 6.0 points lower (**Figure 3.4C**) compared to ASD individuals carrying associated variants in ASD-preferential genes. Furthermore, we found more inherited PTVs in the ASD-preferential genes than the NDD-preferential genes potentially suggesting that ASD-preferential genes are under less negative selection than NDD-preferential genes.

<u>Influence of intellectual disability / developmental delay on the genetic architecture and gene discovery in autism spectrum disorders and congenital heart disease</u>

Even after ascertaining on ASD, roughly half of the genes discovered in Chapter 3 were more often observed in individuals ascertained not for ASD but for ID/DD. This finding indicated that perhaps some of the genes might not contribute to ASD at all. In this chapter, we sought to examine how the effect of ID/DD influences the frequency of *de novo* variants across and within ascertainments as well as their effect on gene discovery. We took three traits with the largest trio sample sizes, ID/DD (N=5305), ASD (N=6430), and congenital heart disease (N=3683) and aggregated all the genetic and phenotypic data. We found that within ascertained

ID/DD samples, comorbid congenital heart disease and ASD did not influence the frequency of *de novo* missense (MPC ≥ 2) and PTVs (pLI ≥ 0.9). In contrast, both of the above classes of *de novo* variation were elevated within ascertained ASD (missense rate ratio 1.92; PTV rate ratio 1.49) and ascertained congenital heart disease individuals with comorbid ID/DD (missense rate ratio 1.36; PTV rate ratio 2.33; **Figure 4.1**). We also found that the frequency of both *de novo* missense (MPC ≥ 2) and PTVs (pLI ≥ 0.9) were not significantly different in ascertained ASD with ID/DD and ascertained congenital heart disease with ID/DD, suggesting that comorbid ID/DD contributes equally in both disorders for these two classes of *de novo* variation. Lastly by collecting phenotype data, we did not find any difference in the frequency of any class of *de novo* variation between simplex and multiple ASD families, in contrast to the results from CNV studies (**Table 4.5**).

We tested 18,226 genes for excess *de novo* missense (MPC ≥ 2) and PTVs and found 95 genes in ID/DD, 18 genes in ASD, and 11 genes in congenital heart disease. As expected, we found more genes in ASD with than without ID/DD (15 vs. 4) and similarly in congenital heart disease (14 vs. 0) as all of the individuals carrying the *de novo* variants driving these gene associations had comorbid ID/DD. Two genes, *ANK2* and *GIGYF1*, rose above the Bonferroni significance threshold for the first time in ASD without ID/DD because all of the individuals carrying *de novo* variants in these genes did not have intellectual disability, developmental delay, seizures, or delayed walking. Splitting ID/DD by ASD or congenital heart disease comorbidity status did not significantly affect the amount of associated genes discovered (**Figure 4.4E-H**).

Future directions

<u>Correcting for differences in exome capture</u>

Exome sequencing has come a long way since its initial use in 2009[7]. In order to sequence solely the exome, exome sequencing uses exome capture kits with baits designed to target just the exons that are then sequenced. Over time, the exome capture kits have improved originally covering roughly 80% of exons to covering nearly 95% of exons. Differences in exon capture can create issues with meta-analyses, as older studies will contribute fewer variants because fewer exons were sequenced. As such, differences in both the total number of variants and individual variant classes (e.g., synonymous, missense, PTV) between studies and cohorts may be partially due to differences in capture and not QC or other technical artifacts that are currently not accounted for. In particular, there exists a large number of schizophrenia trios with published de novo variants (2544 trios), most of which was sequenced using older exome capture kits. When we compare the global frequency of *de novo* synonymous variants between the aggregated schizophrenia data and ASD, ID/DD, and congenital heart disease, we find that the schizophrenia data is significantly depleted of *de novo* synonymous variants as such we have not explored there data further but this depletion could very well be accounted for by the older sequencing.

Insertions / deletions

All of our analyses discussed throughout this dissertation depend on accurately calling variants with high sensitivity and specificity. While variant calling is quite accurate for single nucleotide variants, calling insertions and deletions (indels) still remains a challenge. As frameshift variants can sometimes comprise upwards of 50% of the PTVs in a study, there is a strong incentive to improve our indel calling. A consequence of poor indel calling is two-fold: the mutational model currently lacks the ability to predict the expected number of indels per

gene[8] and pLI does not account for the observed frameshift variants in ExAC[6]. To include frameshift variants in the mutational model, we estimated the per-gene frameshift mutation rate to equal the per-gene nonsense mutation rate multiplied by the ratio of frameshifts to nonsense mutations – a clear, and definite, hack. We foresee both of these issues to be resolved with larger sample sizes, improved variant calling technology, longer reads, and better prediction algorithms such as convolutional neural networks.

Updating TADA for the sex chromosomes

Currently, the transmission and *de novo* association (TADA) method only operates on the autosomes and thus ignores the sex chromosomes[9]. Given that the X-chromosome has the most haploinsufficient genes in the genome (31.63%) far above expectation (observed 230; expected 128.88; $\chi^2 = 79.34$), there are most likely a non-zero number of ASD associated genes residing on chromosome X that have yet to be discovered. Some of the challenges for the sex chromosomes include different number of copies between women and men, potential for different prior effect sizes for women and men, and sex-specific germline mutation rates. While we already incorporated the sex-specific germline mutation rates into our gene discovery models (Chapter 4), the remaining issues have yet to be resolved. We know there is a large potential for ASD-associated genes given the large imbalance in males and females (other traits, such as ID/DD with much smaller male:female imbalances currently have many associated genes on chromosome X as we discovered in Chapter 4) and we already identified one gene on chromosome X in ASD with comorbid ID/DD, *IQSEQ2*, using a frequentist-based approach. Therefore, many more genes on X will be discovered via TADA if this were resolved, and associated genes on X is practically guaranteed with a liberal FDR-based approach.

Final thoughts

Throughout this dissertation, we sought to investigate the role of rare *de novo* and inherited coding variants in neurodevelopmental disorders and identify trait-associated genes. We identified the subset of PTVs most likely to confer risk, provided an analysis strategy for PTV prioritization and association that has been used in studies of ASD[9,10], schizophrenia[11], and epilepsy[12], and found that rare *de novo* and inherited PTVs reside in different sets of genes. With increasingly larger exome sample sizes on the horizon, such as the ASD SPARK cohort of 150,000 exomes, and the exome sequencing of UK Biobank (500,000 exomes), it will fascinating to see if the results presented here hold in the face of massive sample sizes.

References

1. Picoraro, J. & Chung, W. Delineation of New Disorders and Phenotypic Expansion of Known Disorders Through Whole Exome Sequencing. *Current Genetic Medicine Reports* **3**, 209-218 (2015).

2. Ku, C.S., Vasiliou, V. & Cooper, D.N. A new era in the discovery of de novo mutations underlying human genetic disease. *Hum Genomics* **6**, 27 (2012).

3. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).

4. Hamdan, F.F. *et al.* De novo mutations in moderate or severe intellectual disability. *PLoS Genet* **10**, e1004772 (2014).

5. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893-903 (1969).

6. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).

7. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272--276 (2009).

8. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950 (2014).

9. Satterstrom, F.K. *et al.* Novel genes for autism implicate both excitatory and inhibitory cell lineages in risk. *bioRxiv*, 484113 (2018).

10. Satterstrom, F.K. *et al.* ASD and ADHD have a similar burden of rare protein-truncating variants. *bioRxiv* (2018).

11. Howrigan, D. *et al.* Schizophrenia risk conferred by protein-coding de novo mutations. *bioRxiv*, 495036 (2018).

12. Heyne, H.O. *et al.* De novo variants in neurodevelopmental disorders with epilepsy. *Nature Genetics* **50**, 1048-1053 (2018).