Investigating the Role of Noncoding *De Novo* Single-Nucleotide Variants in Autism Spectrum Disorder

by

Christopher Castro

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2023

Doctoral Committee:

Associate Professor Alan Boyle, Chair
Associate Professor Stephanie Bielas
Professor Margit Burmeister
Associate Professor Ryan Mills
Associate Professor Xiang Zhou

Christopher Castro
castrocp@umich.edu
ORCID iD: 0000-0001-9727-1357

To my friends and family

ACKNOWLEDGEMENTS

There are not enough words to express my gratitude to everyone who helped make this work possible. I would not have been able to do this without the support of so many people.

Thank you to my thesis committee members for all your valuable feedback and guidance throughout my PhD. I appreciate all of the conversations I've had with all of you and your encouragement through the entire process. Thank you to Julia Eussen, who has helped me with too many things to list here. This whole thing would have been derailed from the very beginning without your help.

To the entire Boyle Lab, I couldn't imagine a better group of people to work alongside every day. Thank you for all the fun lab lunches, outings, and parties. Nobody does it like the Boyle Lab. Thank you, Adam, for your constant help and guidance. In particularly tough times, you were a major source of support who gave me the boost I needed. A special thank you to my PhD mentor, Alan. Thank you for everything you've done for me to help me succeed. Thank you for creating such a good environment for learning that people want to be a part of and can thrive in. Thank you for helping me grow as a researcher and in general as a person.

Thank you to all of my friends for keeping me sane. To my friends from Chicago, my friends across the country, and the friends I've made during my time in Michigan, thank you for being one of the best parts of my life. Thanks for all the fun times together, with many more to come.

Thank you to my sister, Veronica, for being such a constant source of positivity for me. To my mother and father, thank you for everything you've done for me throughout my life that has allowed me to reach this point. Your endless support has made all the difference and has kept me going.

Lastly, thank you Tanya, for being the most supportive, loving partner anybody could possibly have in their life. Thank you for always believing in me. Thank you for staying up late with me on tough work nights. Thanks for making sure I ate, and drank enough water, and for just helping keep me alive. I would not have been able to do this without you.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Autism spectrum disorder (ASD) is a complex and heterogeneous neurodevelopmental condition characterized by challenges with communication, social interaction, and behavior. With an estimated heritability between 50-90%, a strong genetic basis has been established ASD. While many ASD-associated genetic variants in coding regions have been identified, genome-wide association studies have shown that most trait-associated variants lie within noncoding regions. Therefore, here I have focused on characterizing the contribution of noncoding *de novo* SNVs (dnSNVs) to ASD risk.

To accomplish this, I leveraged whole-genome sequencing data from 1,917 families in the Simons Simplex Collection. In Chapter 2 describe the pipeline I have established to improve the accuracy of genotype and variant calls, in order to ensure a high quality list of dnSNVs. I then introduce the computational tools I used to prioritize variants and identify cis-regulatory elements. I show that there is a strong enrichment of high-impact coding dnSNVs in probands, but significance levels do not withstand multiple-testing correction in noncoding regions.

In Chapter 3 I present power analyses suggesting that a larger sample size may be necessary in order to detect association between ASD and noncoding dnSNVs in probands. I also show that certain annotation categories are better than others at capturing meaningful differences between probands and siblings. After discussing the challenges in screening for ASD-associated noncoding dnSNVs, I provide suggestions

as to how those challenges can be addressed for future studies.

I developed a web application database, *De Novo* Browser, which I introduce in Chapter 4. In this work I have curated an annotated list of 267,000 dnSNVs. I have made this list publicly available on the *De Novo* Browser, where the variants can be explored in table form and sorted by a variety of features and annotations. Together, my dissertation enhances our understanding of the role of noncoding variation in ASD, while also providing a tool and recommendations to benefit future studies.

CHAPTER I

Introduction

1.1 The Influence of Genetic Variation on Gene Expression

The central dogma of molecular biology was first proposed in the 1950's by Francis Crick, through which he outlined the theory that genetic information flows from DNA to RNA, to then create functional products, proteins [27]. The process by which information from DNA sequences leads to the functional product affects gene expression, which we see as the appearance of phenotypes and characteristics. Gene expression can be altered in a number of ways through genetic variation.

Genetic variation refers to alterations in DNA sequences, and these differences between our individual genomes are what make us unique from one another. Variation in our genome can be brought about in multiple ways. Insertions and deletions (INDELS) change sequences by adding or removing, respectively, one or more nucleotides in a sequence. Sequences may also be duplicated, when a stretch of nucleotides is copied and then placed back-to-back next to the original sequence. These copied sequences may repeat just a few times, or even hundreds of times. Repeating sequences are common, and indeed large portions of the human genome consist of repetitive DNA. When the number of times these sequences are repeated varies between individuals, they are known as copy-number variants (CNVs). In

some cases, CNVs are harmless. However, some CNVs have been discovered as the cause of certain diseases [126, 153]. Huntington's disease and Fragile X are two of the more well-known CNV-caused diseases, each caused by a trinucleotide sequence that is repeated in excess [149]. With translocations, segments of DNA are relocated to a different place in the genome. Translocations can also lead to inversions, which occur when the sequence reinserts itself in reverse orientation. Inversions could also occur from duplications.

The most common types of genetic variation in the human genome are single-nucleotide polymorphisms (SNPs), which occur when there is a substitution at a single base pair. In many cases, SNPs are biallelic, meaning that the observed allele may be that which is seen in the reference genome or it may be substituted for a variant allele. Generally, the allele which is more common across a population is considered to be the "major" allele, while a SNP that occurs less frequently across the population is considered to introduce a "minor" allele. The frequency at which the less-common allele is observed in a given population is known as the minor-allele frequency (MAF). Generally, this type of variation must have an allele frequency of 1% or greater to be considered a SNP [2]. In this dissertation I will also reference single-nucleotide variants (SNVs), the more commonly used term when referring to single-nucleotide polymorphisms present at a frequency of less than 1%.

As with the previously mentioned types of variation, in many cases SNPs may have no negative effect on health and will not contribute to the development of disease, depending on where they are located. For example, when speaking of coding regions, SNPs can be synonymous, meaning they do not alter the encoded amino acid. In many cases, this results in a mutation being silent and having no observable effect on phenotype. Nonsynonymous SNPs, however, can be further categorized

into two types of mutations that do affect the amino acid sequence: missense and nonsense mutations. Missense SNPs change a codon, resulting in the substitution for a different amino acid, which can have anywhere from neutral to highly negative effects on protein function [105]. A nonsense mutation, rather than substituting amino acids, introduces a stop codon prematurely. This can result in a shortened protein, which may affect its ability to function properly.

Genetic variation can influence gene expression in a number of ways, including through changes in the DNA sequence of the gene, modifications to the chromatin structure, or changes to the regulatory elements that control gene expression. Genetic variations that occur in coding regions can result in a change in the amino acid sequence of a particular protein, while noncoding genetic variants are genetic variations that do not directly affect the protein sequence but instead can still have an effect on the gene by influencing the gene's expression or by altering a regulatory element near the gene.

1.2 The Role of the Noncoding Genome

For some time, it was widely-believed that noncoding DNA was simply "junk", serving no biological purpose. With the advent of technologies that have allowed researchers to investigate the noncoding spaces of the human genome, it has become more clear that this line of thinking was incorrect. Chiefly among the most significant functions we now know noncoding DNA serves, is to control gene expression. This process of gene regulation controls the location, timing, and amount in which genes are expressed. More specifically, much of this regulation is driven by noncoding DNA sequences referred to as regulatory elements, which are essential for the process of transcription to occur. These regulatory elements provide binding sites for

proteins known as transcription factors, of which the right combination is necessary for transcription to be carried out properly (Figure 1.1).

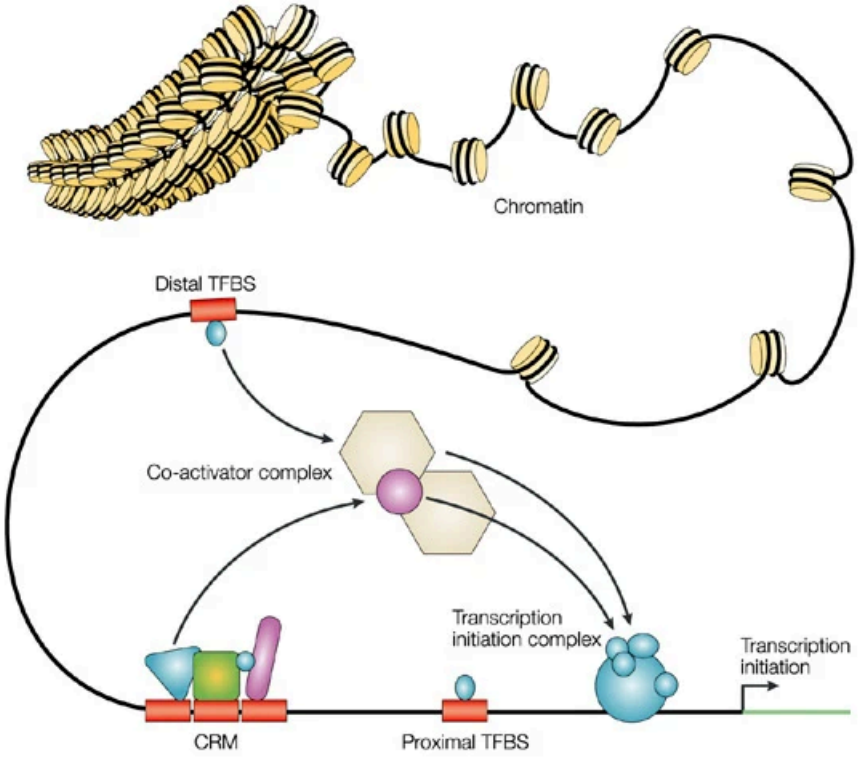

Figure 1.1: A simplified schematic of transcription factors binding at regulatory elements at proximal or distal binding sites in order to initiate transcription. Binding at cis-regulatory modules (CRM) allows for the regulation of gene expression. Figure adapted from Wasserman and Sandelin (2004) [145].

Among the main types of DNA regulatory elements are promoters, enhancers, silencers, and insulators. Promoters are regions upstream of where gene transcription begins, where transcription factors will bind along with RNA polymerase to initiate transcription. Promoters may be considered core promoters, usually within a couple hundred base pairs of the transcription start site, or proximal promoters, which are further but still within a few hundred base pairs of the TSS [86, 22].

Enhancers also allow for the binding of transcription factors to help activate transcription, often increasing the level of gene expression. Unlike promoters, enhancers

are generally much further from the TSS, sometimes thousands of base pairs away [16]. Enhancers generally are thought to function independent of distance and orientation to the target gene, although exceptions have been found [29]. The most popular theory to explain the mechanism by which enhancers can function from such a long distance is the looping theory. This theory explains that once the necessary transcription factors have bound to the enhancer, DNA will form a loop, bringing the enhancer region close to the promoter [29, 22].

Contrary to promoters and enhancers, silencers repress expression rather than activate or increase it. As with other regulatory elements, silencers will provide binding sites for transcription factors, but these transcription factors have repressive activity. Insulators can prevent activating or repressive effects by either acting as blockers of enhancer-promoter interactions, or preventing silencing effects by protecting against chromatin spreading.

Due to the fact that 98% of the human genome is made up of noncoding DNA, the vast majority of SNPs are not found in coding regions, but instead in noncoding regions [107]. Further, studies from GWAS have shown that 90% of phenotype-associated SNPs map to noncoding regions of the genome, suggesting that causal variants play a role in affecting gene expression [37, 53, 93, 20, 137, 87].

SNPs can play a large role in the regulation of gene expression by altering DNA sequences at, or near, the binding sites of transcription factors, particularly when these binding sites lie in regulatory elements [30]. Transcription factor binding sites are regions of particular interest for this reason as well as the finding that they make up 31% of GWAS SNPs [26]. Various studies have now provided evidence of SNPs affecting transcription factor binding in a variety of ways, and being associated with multiple diseases. SNPs can affect a transcription factor's binding affinity [97, 63,

103], by increasing [25] or decreasing [110] affinity (Figure 1.2). A SNP may also completely destroy a transcription factor binding site [103, 52] or change a binding site in a way that creates a binding site for a new transcription factor [110, 68, 7].

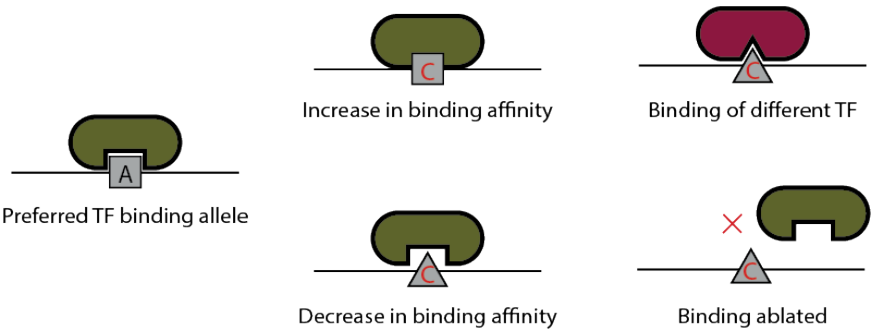

Figure 1.2: Different ways in which a SNP can affect transcription factor binding. Binding affinity may be increased or decreased. A change of allele may ablate binding entirely, or modify the site in a way that allows for a different transcription factor to bind.

1.3 Identifying Regulatory Elements

In order to identify SNPs in noncoding regions with the potential to affect gene expression, the important challenge of first annotating noncoding regions of the genome had to be addressed. To this end, the Encylopedia of DNA Elements (ENCODE) Consortium was established with the goal to create a comprehensive list of functional elements in the human genome [26]. The ENCODE project involves a large consortium of researchers from around the world who have used a variety of high-throughput techniques to map and annotate various genomic features such as DNA transcription, chromatin structure, and DNA methylation. Given that the binding of transcription factors is essential for proper gene regulation, it is necessary to have methods for identifying these binding sites. In this way, we can also define regulatory elements throughout the genome.

Identifying regulatory elements remains a challenge, but there are now several well-established techniques that are commonly used to do so. The ENCODE project

has assisted in improving our understanding of which regions of the genome are involved with regulating gene expression by providing computational and experimental data from these techniques, resulting in genome-wide maps on histone modification, chromatin accessibility, and transcription factor binding sites. Here I will describe four standard approaches for identifying transcription factor binding sites: using position weight matrices (PWMs), Chromatin immunoprecipitation followed by sequencing (ChIP-seq), DNase I hypersensitive site sequencing (DNase-seq), and DNase footprinting.

Position weight matrices are useful for describing the binding DNA sequence preferences of proteins. They represent the likelihood of proteins binding to a particular DNA sequence. The value in each cell of the matrix represents the frequency of that nucleotide at that position, normalized by the background frequency of that nucleotide in the genome. These PWMs can inform us on the preferential binding sequences for transcription factors, which can then in turn help us to identify regulatory regions.

ChIP-seq is an experimental technique used to measure the level of protein binding to DNA, which can provide data on transcription factor binding, histone modifications, methylated DNA regions, and nucleosome positioning [26, 46]. The technique begins with a crosslinking step in which proteins are bound to DNA. The chromatin containing the bound proteins is then isolated through fragmentation, creating fragments of approximately 300bp in length [139]. This is followed by the immunoprecipitation step, during which antibodies specific to the protein of interest bind to the proteins (a specific transcription factor, for example). Following a purification step to remove any unbound proteins, the remaining fragments containing the bound proteins are then sequenced. The reads can then be mapped back to the reference

genome to create a genome-wide map of locations where the proteins of interest are bound.

Another method used to locate functional elements relies on identifying genomic regions that are hypersensitive to cleavage by DNase I [88]. Each human cell contains about 2 meters of DNA, with estimates of at least 30 million human cells in our bodies [125]. Therefore, the human genome requires a method for packaging all of the DNA. Packaging DNA into the nucleus of human cells is accomplished by coiling and wrapping DNA around protein complexes called nucleosomes [39]. Regions of the genome that are tightly compacted by nucleosomes are less accessible for transcription factor binding. However, regions where nucleosomes have been displaced, and are accessible for binding by transcription factors at regulatory elements, are also more accessible for digestion by DNase I [17]. These DNase I hypersensitive sites (DHS) have been shown to overlap with genetic regulatory elements [50]. Locating these DHS can be accomplished in a high-throughput manner through a technique known as DNase-seq. With this technique, cells are treated with DNase I, an enzyme that will preferentially digest DNA in regions not bound by proteins, leaving behind regions of open chromatin. These regions are fragmented and undergo adapter ligation followed by high-throughput sequencing. Reads are then mapped to the reference genome, revealing areas of open chromatin.

Using a similar approach, DNase footprinting methods assist in identifying transcription factor binding sites by leaving patterns that leave a protein binding "footprint". This technique is based on the same principle that sites at which proteins are bound to DNA will be protected from digestion [47]. DNase I is introduced to a DNA sample and will cleave DNA molecules. DNA fragments are separated by gel electrophoresis and will reveal regions that are protected from cleavage as a

"footprint" on the gel [80].

1.4 The RegulomeDB framework

Each of the methods described in the previous section, on their own, are valuable techniques for which a wealth of data has been gathered and provided as public data by The ENCODE Project. These data provide researchers with information that can be used to glean insight on the noncoding portions of the human genome and annotate their own sequencing results or variants. For studies of genetic variation in regulatory regions, these data are an invaluable resource. Still, each method has its own limitations, and data for each must be accessed separately. A much more efficient approach for using these data is to apply annotations as a combination of these methods. To this end, RegulomeDB was established as a database and annotation tool, guiding variant interpretation by integrating data from ENCODE and other sources [18].

RegulomeDB allows users to prioritize SNPs using a heuristic scoring system based on overlap with known and predicted regulatory elements using data from ENCODE, The Roadmap Epigenome Project, GEO, and published literature. These data are leveraged jointly to score the likelihood that a variant of interest has functional effects on gene regulation. In the original implementation of RegulomeDB, variants are assigned a score ranging from 1 to 7, with lower numbers indicating a higher likelihood of functional significance. Scoring is based on the presence of experimental evidence overlapping the given location of a variant (Figure 1.3). The scoring scheme is based on a set of criteria including overlap with data evidence from ChiP-seq, DNase-seq, DNase footprinting, PWMs, and eQTLs (Table 1.1).

RegulomeDB originally included about 600 million annotations across over 100

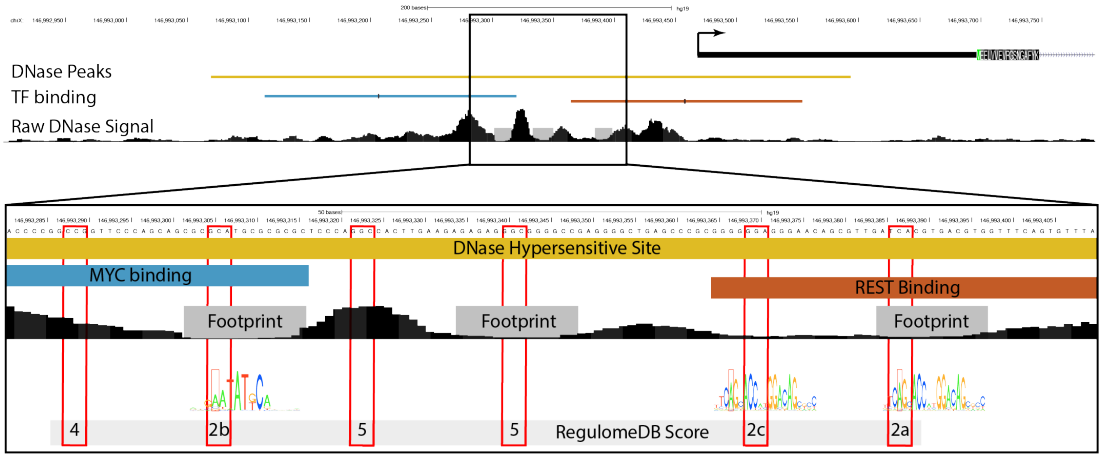

Figure 1.3: An example of RegulomeDB scoring in a zoomed promoter region of the FMR1 gene. All scored regions overlap with a DNase hypersensitive site, depicted by the yellow band. The regions that also overlap ChIP-Seq peaks (MYC and REST), along with DNase footprints and TF motifs are scored 2a and 2b. The regions overlapping a REST binding site and motif, but not a DNase footprint, is scored as 2c. Here, the regions with less evidence receive scores of 4 and 5, indicating a lower confidence of harboring regulatory elements.

tissue and cell lines, and the prioritization and scoring method was tissue-agnostic. While this tool is extremely useful for identifying functional regions in a general manner, a common challenge in studies focusing on specific phenotypes has been the inability to resolve differences in gene regulatory networks between different tissues [18, 51, 82]. Accordingly, the latest update to RegulomeDB builds upon the original framework to prioritize regulatory variants in non-coding regions of the human genome in a tissue-specific manner. This newer computational tool, Tissue-specific Unified Regulatory Features (TURF), has been integrated into RegulomeDB v2.0 [35]. In addition to the annotations previously available, TURF incorporates data from allele-specific transcription factor binding in six cell lines. Random forest models are trained on tissue-specific annotations to return both a generic score (tissue agnostic) and a tissue-specific score. The work done by Dong and Boyle showed that GWAS variants were enriched with regulatory variants predicted by TURF tissue-specific scores in trait-related organs, highlighting the usefulness of this tool in

prioritizing regulatory variants in studies of traits associated with specific tissues [35]. Importantly, this ability to make tissue-specific predictions has been evidenced to improve the prioritization of associated variants in complex brain disorders, including autism [35, 36, 81].

What does the RegulomeDB score represent?

The scoring scheme refers to the following available datatypes for a single coordinate.

Score	Supporting data
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding / DNase peak
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak
2b	TF binding + any motif + DNase Footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
4	TF binding + DNase peak
5	TF binding or DNase peak
6	other

Table 1.1: Scoring scheme for RegulomeDB. Variants receiving a score of 1 require overlap an eQTL and represent those most likely to reside within a functional region, therefore having the highest potential for having regulatory effects on gene expression. Higher scores indicate decreasing evidence for variants overlapping functional regions. Figure adapted from Boyle et al. [18].

1.5 Overview of Autism Spectrum Disorder

Autism was first described as "infantile autism" in 1943, by Austrian-American psychiatrist Leo Kanner when he described 11 children with "extreme autistic aloneness", "delayed echolalia", and an "anxiously obsessive desire for the maintenance of sameness." [60] The term "autism" was borrowed from Eugene Bleular, who had originally used the term to describe schizophrenic patients who withdrew into themselves. Kanner, however, established that the characteristics he viewed made autism unique from other existing conditions. After decades of establishing diagnostic concepts

and overcoming initial misunderstandings about the condition, when the Diagnostic and Statistical Manual of Mental Disorders-III (DSM-III) was published in 1980 it included autism for the first time[9].

Over the last decades, and through different editions of the DSM, the definition of autism has changed. In 2013, the term autism was changed to autism spectrum disorder in the DSM-IV, and became an umbrella term covering the previously-separate conditions of autistic disorder, pervasive development disorder – not otherwise specified (PDD-NOS), and Asperger syndrome. Today, the DSM-V, the latest edition of the DSM, defines autism as a neurodevelopmental disorder characterized by persistent difficulties in social communication and interaction, and restricted, repetitive patterns of behavior, interests, or activities [105, 8] (Fig 1.4).

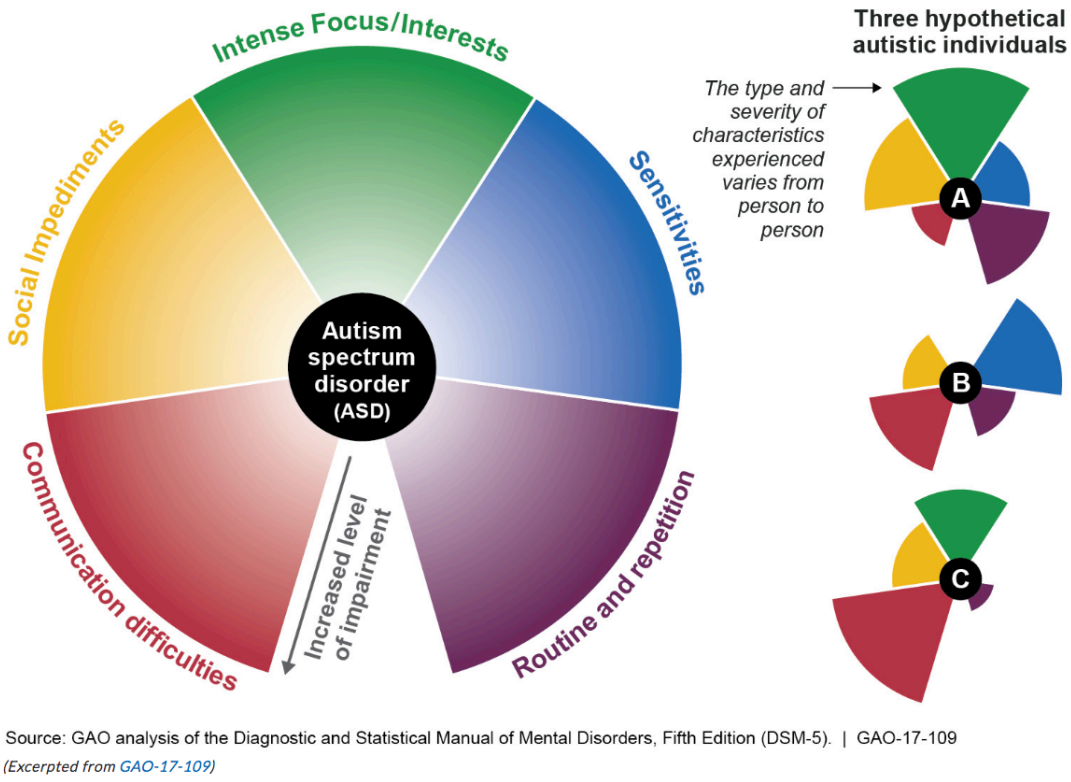

Figure 1.4: A simplified visualization of characteristics associated with autism spectrum disorder (ASD), grouped into five broad categories. Individuals diagnosed with ASD may experience combinations of different characteristics, each to varying degrees. This makes ASD a very individualized condition resulting in a wide range of characteristics.

The CDC established the Autism and Developmental Disabilities Monitoring (ADDM) Network in 2000 to monitor the prevalence of ASD. Since the year 2000, the estimated prevalence of ASD has increased from 1 in 150 children to the most recent estimate of 1 in 44 children [83]. It is thought that this increase in prevalence has more to do with increased global awareness than biological changes [152]. With an increased focus on ASD research, surveillance and diagnostic methods have improved, which has been reflected in the overall ability to properly identify cases. These changes in diagnostic accuracy can be seen in changes to the skewed assigned at birth sex ratio, which has consistently shown more boys than girls to be diagnosed. A Danish group in 1995 found the ratio ratio of diagnosed boys to girls to be 8:1. More recently though, this ratio has dropped to as low as 4.2:1 [152, 77]. It is thought that a big factor affecting this difference is diagnostic bias. Many studies have now suggested that girls are diagnosed later in life than boys, with one potential reason for this being that diagnostic assessments are skewed to identifying ASD characteristics in boys [38]. The apparent increase in ASD prevalence and drop in skewed sex ratio over the years indicates that, while research is still leading to diagnostic improvements, our understanding of how to characterize and diagnose ASD is an ongoing challenge.

Along with changes and improvements to classification and diagnostic methods for ASD over the years, the last several decades have brought a much greater understanding of the biology behind the disorder. In the 1950s and 1960s, when infantile autism was still being defined, many psychiatrists suggested aberrant parenting as a potential cause of the characteristics they were recognizing [38, 84]. At this time, it was also assumed that schizophrenia was caused by the “schizophrenogenic mother” [45]. This concept was popularized by Bruno Bettelheim, through his suggestion of

"refrigerator mothers" being a cause of autism, through lack of maternal (or paternal) warmth [45, 94]. Psychologists would claim that only through "correct" maternal behavior would children grow up to be hard-working, self-disciplined adults; deviation from this would result in weak-minded, badly behaved, aberrant adults [84]. Leading into the 1970s, the "refrigerator mother" concept had finally begun to be rejected as the body of research on autism grew.

Genetic studies of twins were among some of the more informative, providing further evidence that there was a strong biological basis to autism [84, 41]. These studies brought to light the importance of genetic variation in the etiology of autism. In the decades to follow, our understanding of the role of genetics in autism continued to evolve significantly. With the advent of next-generation sequencing came the ability to study autism genetics on the scale of whole exomes and genomes, which has led to the identification of hundreds of associated genes and variants and accelerated the path to our current understanding of the genetic basis of autism [124, 108, 100].

1.6 Summary of Dissertation

In this dissertation, I aim to contribute to the body of research in autism genetics by focusing on genetic variation in noncoding regions of the genome, an area for which there are significant gaps in our knowledge. I will outline my approach for identifying *de novo* single-nucleotide variants using whole-genome sequencing data from an autism cohort. Leveraging data from functional genomics experiments, I then annotate these variants in an effort to identify those that are more likely to impact gene regulatory function and test for their association with autism. I follow this up by describing some of the challenges encountered while performing a study of this type and provide suggestions as to how future studies might address these

challenges. Finally, I will introduce a web application that I developed with the goal of providing a straightforward way for other researchers to explore the data I have generated for this work.

CHAPTER II

Identifying and Profiling Noncoding *De Novo* Variants in an Autism Cohort using Whole-Genome Sequencing

2.1 Abstract

Autism spectrum disorder (ASD) refers to a broad range of neurodevelopmental conditions most often characterized by challenges with social interaction, communication, and behavior. While the genetic contribution to ASD has been extensively explored, most studies have restricted their focus to genetic variation in protein-coding DNA regions, which collectively account for only 2% of the human genome. This limitation is in sharp contrast to the results of many genome-wide association studies which have revealed that the vast majority of trait-associated variants lie within the noncoding regions. Therefore, investigation of noncoding variation in ASD individuals has the potential to uncover novel ASD-associated variants. In particular, *de novo* mutations have been implicated in the genetic etiology of ASD given that the rare nature of these uninherited variants makes them potentially more deleterious than common variants, as they are not subject to selective pressures. Here, I have identified and prioritized noncoding *de novo* single-nucleotide variants (SNVs) in regulatory regions of the genome in order to better understand their contributions to ASD. To identify predicted pathogenic *de novo* SNVs, I have analyzed whole genome sequence data from 1,917 families participating in the Simons Simplex Col-

lection autism cohort. I have used RegulomeDB, an established database containing nearly 60 million annotations of known and predicted regulatory elements. The sources of this data include public datasets from the Encyclopedia of DNA elements (ENCODE) Consortium, the Roadmap Epigenome Mapping Consortium, GEO, and published literature. In order to identify a set of SNVs that are more likley to impact function, SNVs were annotated using the RegulomeDB heuristic scoring, which is based on their overlap with predicted functional elements. Additionally, SNVs were annotated using TURF, a model built on the RegulomeDB framework that prioritizes which variants are likely to be functional in a tissue-specific manner. I hypothesized that a subset of these SNVs disrupt the function of regulatory elements that ultimately impair expression of ASD-associated genes. While I did detect a strong enrichment of proband *de novo* SNVs in high-impact coding regions, statistical tests performed on noncoding variants failed to reach significance. These results suggest that there may be challenges that need to be addressed when designing future autism studies. This work contributes to our understanding of the function of rare noncoding variants and their contributions to ASD genetics.

2.2 Introduction

2.2.1 The Genetic basis of ASD

Phenotypes that vary between people in a population generally do so because of differences in genotype and environment between those people. In an effort to tease apart and quantify to what extent the basis of those phenotypes lies mostly in genes or environment, researchers have used heritability estimates. Heritability is defined as the proportion of variation in a given trait or condition that can be attributed to genetic variation. This heritability estimate is generally based on a single population. It is a value that can range from zero to one, where a heritability

closer to zero would indicate that variability in a trait is less likely to be influenced by genetic variation rather than by environmental factors. For example, a characteristic such as the ability to speak another language has a heritability of zero because it is not controlled by genetics. On the other hand, disorders such as sickle-cell anemia or cystic fibrosis have heritability estimates closer to one because they are highly influenced by variants in single genes [109, 98]. Variation in traits with a heritability somewhere in between zero and one, such as variation in skin color, would be expected to be explained by a combination of genetic variation and environmental factors [104].

Often, particularly with complex phenotypes, there is a mixture of genetic and environmental factors at play. In many cases, it is an interaction between these factors that dictates a phenotype. Such is the case with autism. While environmental factors alone have not been shown to result in autism, there is evidence that environmental factors may be involved [61, 1]. The most recent studies, however, estimate the heritability of autism to be as high as 90%, suggesting a very strong genetic influence [120, 11].

Traditionally, heritability estimates for autism have been derived from twin studies. The earliest autism twin study was a study of 21 pairs of British twins in 1977 [41]. In the study, for 10 pairs of dizygotic twins in which one twin was diagnosed with autism, there was zero concordance. For the 11 pairs of monozygotic (genetically identical) twins, however, 36% were concordant. Although the sample size was small, it was very important at the time because it established clear evidence for a genetic basis of autism which would be built upon in the field. Many more twin studies have been performed since then, with increasingly large cohorts, and concordance for broader autism spectrum disorder diagnosis has been reported to be as high as 92% between monozygotic pairs [44, 13]. Twin studies led to the understanding that

there is a large genetic component to the characteristics that define autism spectrum disorder.

Prior to next-generation sequencing, one method for studying ASD was by identifying large chromosomal changes with karyotype analyses. These analyses revealed the presence of copy number variants (CNVs), translocations, and inversions among autism cases [146, 70, 147]. Many of these variants were found to be uninherited, *de novo*, mutations. As improvements were made to variant-detection methods using PCR-amplified DNA, candidate gene studies were performed in ASD, but lacked the ability to find associations.

Genotyping microarrays opened the door for large scale analyses of SNPs in the form of genome-wide association studies [115]. Microarrays also led to improved detection of CNVs, with greater accuracy than had previously been possible with karyotyping [54]. Using this technology, several studies were able to highlight the role of CNVs in ASD [108, 124, 119]. As before, many of the enriched CNVs were found to be *de novo* [49]. With the establishment of these new methods and the realization that *de novo* mutations were involved, it became more commonplace to perform studies in which *de novo* mutations in cases were compared to unaffected sibling controls [119].

Whole-exome sequencing (WES) later provided yet another improved method, with which researchers could investigate protein-coding regions of the genome. This advancement again proved to be an effective method for identifying ASD-associated variants, with *de novo* mutations still being implicated [100, 129]. Through WES, many studies have reported candidate genes which may be contributing to autism, with the number of genes implicated now being in the hundreds. Providing further evidence of the role of *de novo* mutations, using WES, several groups have published

their findings of *de novo* loss-of-function variants being associated with ASD [56, 95, 101].

In addition to improving the analysis of protein-coding regions [21, 150], when compared to WES, whole-genome sequencing has the benefit of offering more coverage across the genome. This opens up the possibility to study the role of variants in noncoding regions of the genome and improves the detection of rare and *de novo* variants.

2.2.2 The Significance of *de novo* single-nucleotide variants

Common variants have been suggested to contribute to up to 50% of ASD cases [48], but in those cases it is likely a case of combinations of many common variants of low effect size, and common risk loci haven't been found. Generally, the most detrimental variants in a population will be naturally selected against over time, resulting in them maintaining a lower allele frequency (Figure 2.1). When referring to rare genetic variation, *de novo* mutations are among the rarest. These are genetic changes that occur spontaneously in an individual and are not inherited from their parents.

The evolution of next-generation sequencing has allowed for the investigation of rare and *de novo* variants at an increasingly higher resolution. Structural variants (SVs) have been shown to be associated with ASD, by affecting larger segments of DNA. However, SNVs occur more frequently than SVs, and can sometimes be tied to specific genes or regulatory elements more easily, in cases when the size of an SV makes it more difficult to interpret its effect on gene function. Additionally, following up with experimental methods can be more difficult with larger portions of DNA, particularly in cases where the variant may affect multiple genes. Now we have the tools and technology to be able to investigate *de novo* variation at single

base resolution. Indeed, exome-wide studies have revealed rare and *de novo* variants to be associated with ASD, and have allowed for the discovery of hundreds of ASD-associated genes [48, 123]. Still, the coding regions in which these variants have been found account for only a small portion of the human genome, and the next logical step is to extend the research into the noncoding genome.

The work I have presented in this dissertation is based on the hypothesis that a subset of dnSNVs contribute to ASD by affecting the regulation of target genes, through interference of transcription factor binding at regulatory elements. The first step of this investigation relies on the ability to identify a high-quality set of dnSNVs from whole-genome sequence data.

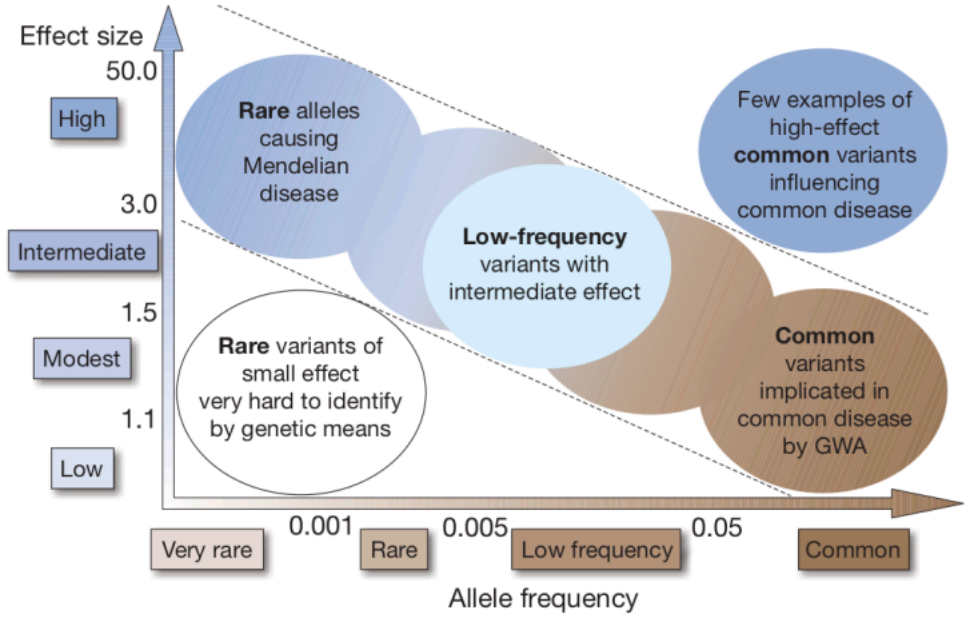

Figure 2.1: This figure illustrates the concept that we can expect variants with the strongest effects to have a low frequency. Adapted from Manolio et al. 2009 [85]

2.2.3 Cohort studies

An ideal way to identify *de novo* variants is to compare a child's sequencing data to that of its parents. When comparing the same genomic loci between child and

parents, any variants present in the child that are absent from both parents can be inferred to be uninherited. Then, we can begin to test for associations between the phenotype of interest and mutations that are found only in the person displaying said phenotype.

Although early twin studies in autism consisted of very few individuals, the ability of the studies to successfully establish a method for studying the genetic basis of autism prompted future studies to leverage increasingly larger cohorts. Today, genetic studies of autism require much larger population cohorts, with careful selection of participants [76]. One of the largest autism cohorts is the Autism Birth Cohort (ABC), which was established to study gene x environment interactions [136]. The ABC is an unselected birth cohort consisting of over 100k children who have been continuously screened in a longitudinal study started in 1999. This cohort has been useful for studying relationships between environmental exposures and genomic findings over developmental periods. However, it lacks the comprehensive whole-genome sequencing data necessary for a study of extremely rare, noncoding variants.

As next-generation sequencing has become more accessible and affordable, larger cohorts that include whole-genome sequencing data have been established. The Simons Simplex Collection was established in 2006 and provides access to WGS data for approximately 2,600 families (Table 2.1). Importantly, for the work in this dissertation, the SSC focuses on simplex families, in which only a single child of the family is diagnosed with ASD. This provides a family structure that allows for isolating variants only seen in affected children, in order to begin to identify *de novo* associated with ASD.

Many of the families enrolled in the SSC are referred to as being of quad structure; four family members consisting of two children, and two parents. In each family,

WGS phase	Funding (Project #)	# Families (# Samples)	Status	Data-sharing
Pilot	SFARI	40 quads (160)	Complete	Yes *(no embargo)*
	NIH (HHSN 27500503415C)	13 trios (39)	Complete	
Phase 1	SFARI	500 quads (2,000)	1,989 genomes (492 complete quads and 21 genomes belonging to partial families) are complete.	Yes *(no embargo)*
Phase 2	NIH (UM1 HG008901)	598 quads (2,392)	Complete	Yes *(A four-month publication embargo applies until **22 December 2017**)*
Phase 3 (Batch 1)	SFARI	591 quads (2,364)	540 quads (2,160 genomes) are complete	Yes *(no embargo)*
Phase 3 (Batch 2)	NIH (U01 MH105575)	228 quads (912)	226 families (904 genomes) are complete	Data is in transit from NYGC and is expected to be available by 15 September 2017 *(A four-month publication embargo applies until **15 January 2018**)*
Phase 4	SFARI	437 trios and 5 quads (1,331)	Complete	Yes *(no embargo)*

Table 2.1: Summary of available whole-genome sequencing data from the Simons Simplex Collection. For the work presented in this dissertation, I utilized data from 1,917 complete quads.[85]

one child is diagnosed with ASD, which I will often refer to as the proband going forward. Along with the proband, each family also includes an unaffected sibling and two unaffected parents. While the SSC also includes trio families (one child and two parents), the quad-structure provides the added benefit of including the unaffected sibling, which can be used as a control. The work that follows is derived from 1,917 quad-structure families from the SSC.

2.3 Methods

2.3.1 Data source

Investigating the noncoding region of the genome requires access to whole-genome sequence data. With the aim of this research being to focus on *de novo* variants, it was also necessary to have sequence data from a large number of individuals, as each

individual is only expected to possess a relatively small number of these variants. Additionally, this study required a carefully curated set of genomes from individuals diagnosed with ASD. The primary goal of the Simons Simplex Collection (SSC) was to establish a permanent repository of genetic samples and data from 2,600 simplex families, which is ideal for this study.

The SSC is run by the Simons Foundation Autism Research Initiative (SFARI) and began as a coalition of 12 university research clinics which would identify potential participants, from their own clinics already serving children with ASD. Participants were selected based on a "simplex" design, meaning that only one individual in the family had a confirmed diagnosis of ASD. This simplex family structure has the benefit of unaffected siblings providing an ideal control when identifying *de novo* variants. Stringent criteria was applied when validating probands, and all probands were evaluated using a number of diagnostic measures, such as the Autism Diagnostic Observation Schedule (ADOS) [78] and the Autism Diagnostic Interview - Revised (ADI-R) [79]. Comprehensive family medical history was collected and comorbidities including sleep irregularities, and gastrointestinal problems were noted. Probands were excluded for a number of reasons, the most common being: not meeting criteria for ASD, having relatives diagnosed with ASD, significant perinatal incidents, and low mental age [40]. As part of the Simons Simplex Collection inclusion criteria, all probands and unaffected siblings were at least four years old, with probands being no older than 18 at the time of diagnosis.

Blood samples were collected from each participant in the SSC and DNA was extracted from blood cells. Whole-genome sequencing data was then produced for approximately 9,200 individual genomes and data was processed by the Centers for Common Disease Genomics and the New York Genome Center (NYGC). Genomes

were sequenced at NYGC using 1 µg of DNA, an Illumina PCR-free library protocol, and sequencing on the Illumina X Ten platform with a mean coverage of 35.5. Post-sequencing, reads were aligned to hg38 using BWA-MEM (0.7.15) [75], duplicate reads marked (Picard version 1.83) and variant calling was done using GATK (v3.5) [31]. The resulting BAM and variant call format (VCF) files were transferred to Amazon Web Services (AWS) S3 storage system where they are accessible with approval from the Simons Foundation Autism Research Initiative. I accessed the VCF files containing the raw variant calls via AWS.

2.3.2 Genotype and Variant Quality Refinement

An important challenge in identifying genetic variants lies in the fact that raw sequencing data contain many errors. This leads to subsequent steps in the downstream analysis also being prone to containing errors, including the presence of artifacts from the alignment and variant calling processes [74]. This is particularly problematic when dealing with rare variants, where false positives and false negatives can have a larger impact due to the reduced subset of variants that are available to work with.

Compounding the challenge, the method by which I would be identifying *de novo* SNVs (dnSNVs) required comparing genotype/variant calls at the same position between four members of the same family, with putative *de novo* calls being instances of a child possessing an allele that was not present in any of the other three family members. Clearly, this requires that all four calls at a particular locus be accurate. Therefore, in this work, it was a priority for me to take steps to reduce error rates and improve the quality of the calls I would be working with, before careful filtering to identify the highest-quality set of *de novo* SNVs possible. To this end, I implemented an extensive quality refinement pipeline, using existing high-quality data to

improve the accuracy of the raw variant calls I started with.

2.3.3 Genotype Quality Score Refinement

For the first step of refining the quality of variant calls, I implemented GATK's Calculate Genotype Posteriors (CGP) tool to improve the quality of genotype calls themselves from the raw VCF files. The CGP tool takes a high-quality ("gold-standard") set of variant calls as input to use as priors for calculating posterior genotype likelihoods. For this I used the Phase 3 1000 Genomes set [3]. Because the data involved families, I also provided a pedigree file, to inform CGP on the relatedness of the individual calls, which CGP uses as family priors. After running CGP using the default parameters, an updated Phred-scaled genotype likelihood was available for each variant call.

2.3.4 Variant Quality Score Recalibration

As with sequencing and genotype calls, variant calls themselves contain artifacts [74]. In an effort to reduce erroneous calls, GATK developed the Variant Quality Score Recalibration process, in which a Gaussian mixture model is implemented to classify and filter variant callsets, taking advantage of highly validated known variant resources. Here, VQSR has been applied to the raw variant calls from the SSC, generating a quality score which can be used to balance sensitivity and specificity (Figure 2.2). This score is called the variant quality score log-odds (VQSLOD). During the recalibration process, a tranche sensitivity threshold is specified as your desired target sensitivity. For this set, variants were calibrated to filter out any that did not have a VQSLOD above which 99.8% of variants in the truth set were included. Variants with a score above the threshold were marked as "PASS" in the VCF file and I filtered out the remaining that did not pass the threshold filter.

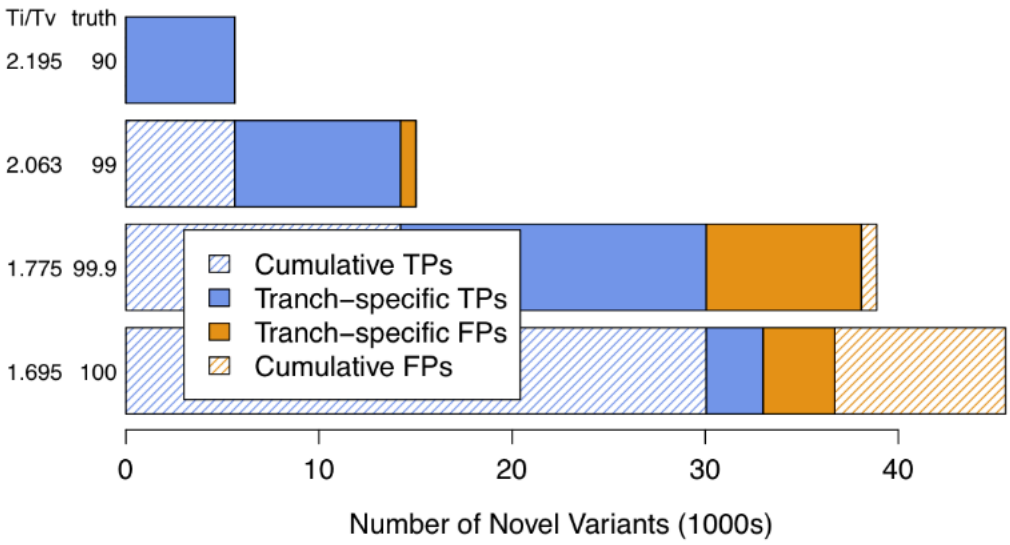

Figure 2.2: An example of how VQSR tranches affect variant filtering. Selecting 90% as a threshold has the lowest truth-sensitivity and returns fewer variants, but the transition/transversion ratio indicates the calls are of higher quality (based on an expected Ti/Tv around 2.0 or 3.0 for WGS or WES, respectively [14]) The specificity is very high, but many true variants would be missed. As the tranche threshold is increased, more true positive calls are gained, at the expense of introducing more false positives. [85]

2.3.5 Filtering out INDELS and low-complexity regions

In order to restrict my analyses to SNVs, I filtered out any remaining variants that were not bi-allelic, as there were many INDELS in the original VCF file. Even after an initial round of filtering to remove INDELS, however, it became apparent there were still unwanted variants that were missed. Additionally, because false-positive variant calls are introduced when including low-complexity regions, it was also important to exclude those regions [106]. To address both of these issues, I incorporated gold standard data sets of INDELS from the Mills and 1000 genomes INDELS set [92] and low-complexity regions from UCSC's RepeatMasker [130] and TRF [15] reference files. If regions from the reference files overlapped with variants from the VCF files, I excluded those variants from downstream analyses.

2.3.6 Lifting over from hg38 to hg19

Over the course of completing this work, computational tools have gradually been updated to include, or be compatible with, the hg38 genome build. However, at the time some of these analyses were done, there were still certain limitations as far as which tools or annotations were available for hg38. Therefore, in order to have access to as many resources as possible, I decided to lift the data over from hg38 to hg19. This allowed me to take full advantage of more well-established resources with a greater amount of data available. It should be noted that the negative side to this is that there are known problems with the conversion process, during which many variants can be lost or inaccurately converted [99], so there is a trade-off in the process of conversion for the sake of having more resources in downstream analyses.

2.3.7 Calling *de novo* SNVs

For the first step in identifying *de novo* SNVs, I applied GATK's VariantAnnotator tool to tag possible *de novo* variants. Using the calculated posterior Genotype Quality (GQ) score, I applied the threshold to filter out calls with GQ<20, a widely-accepted cutoff and recommendation from other studies that have found this cutoffs to result in a reasonable trade-off between precision and sensitivity [106], with a GQ20 indicating a 99% chance of the genotype call being correct. Any variant with a recorded depth (DP) of less than 10 was also not considered for the high quality set, with DP being the number of reads that support a particular variant call at a given position. The VariantAnnotator tool is designed to accept a pedigree file in order to also apply a "hiConfDenovo" to possible *de novo* variants for cases in which a variant is present in one individual but in neither of their parents. A limitation of the tool is that it is only designed to work with trios. Because the data I was

working with was from quad-structure families, VariantAnnotator didn't take the extra information from the unaffected child into account. To account for this, after processing variants with VariantAnnotator I applied a custom script to incorporate the additional data from the fourth family member, ensuring that putative *de novo* variants present in a child were absent from the other three family members. I ran this *de novo* variant identification process twice; once to identify dnSNVs in the proband group, and again to identify dnSNVs in the unaffected sibling group.

Due to the rare nature of *de novo* mutations, it was important to apply a filter based on population frequency. To reduce false-positive calls I required putative dnSNVs to either be completely absent from the gnomAD [62] population database, or be present at an allele frequency of less than 0.001, per recommendations from other studies [106].

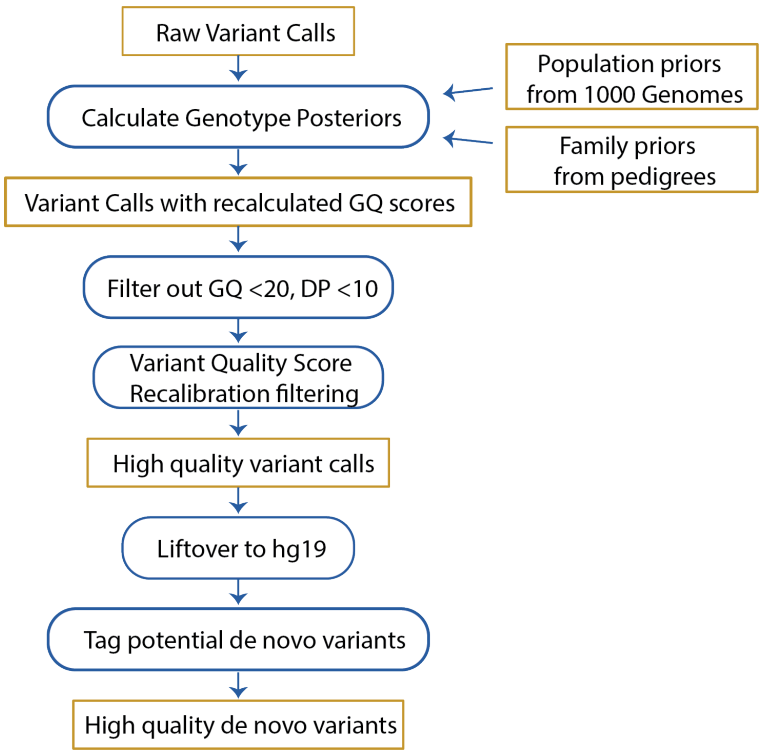

Figure 2.3: Overview of *de novo* variant detection pipeline with genotype and variant refinement steps

2.3.8 Identifying enhancers and promoter regions

For defining enhancer and promoter regions, I accessed DNase I and ChromHMM data derived from experiments from the Roadmap Epigenomics Project [117]. All data was derived from experiments in fetal brain tissue. To be considered an enhancer or promoter region, I identified regions of open chromatin (from DNase I data) overlapping with regions corresponding to enhancer or promoter histone marks, respectively, from ChromHMM data

2.4 Results

2.4.1 Raw count of identified *de novo* SNVs falls in line with expected human mutation rates

Establishing an accurate list of putative *de novo* variants was the first critical step toward investigating their association to ASD. Following the filtering and dnSNV identification steps resulted in a high quality list of variants, with a mean of 69 dnSNVs per proband (Figure 2.4). There was not a significant excess of dnSNVs in probands compared to siblings (134,969 vs 131,896 autosomal dnSNVs). This indicates that we should not expect ASD risk to be explained by an excess in raw dnSNV count, rather, more likely by a specific subset of deleterious mutations.

To confirm that this number fell within a reasonable range, I compared this count to those of nine other published studies that have identified dnSNVs [91, 10]. Of the studies I compared, the lowest mean count was 44 [42], with the highest being 90 [141] (Figure 2.5). Falling in line with these numbers, several other studies that have specifically focused on estimating the *de novo* mutation rate in humans place this number between 44 and 82 per generation [67, 4].

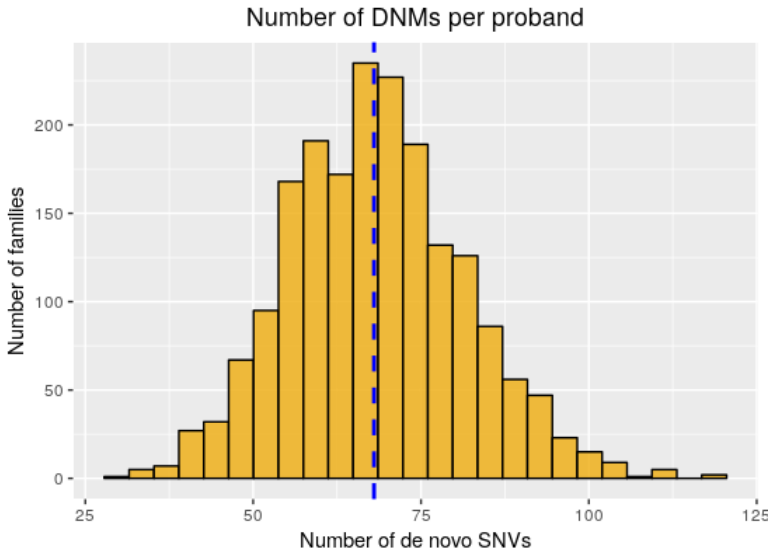

Figure 2.4: Counts of *de novo* SNVs in probands from 1,917 Simons Simplex Collection families. Blue dashed line indicates the observed median of 70 dnSNVs per proband.

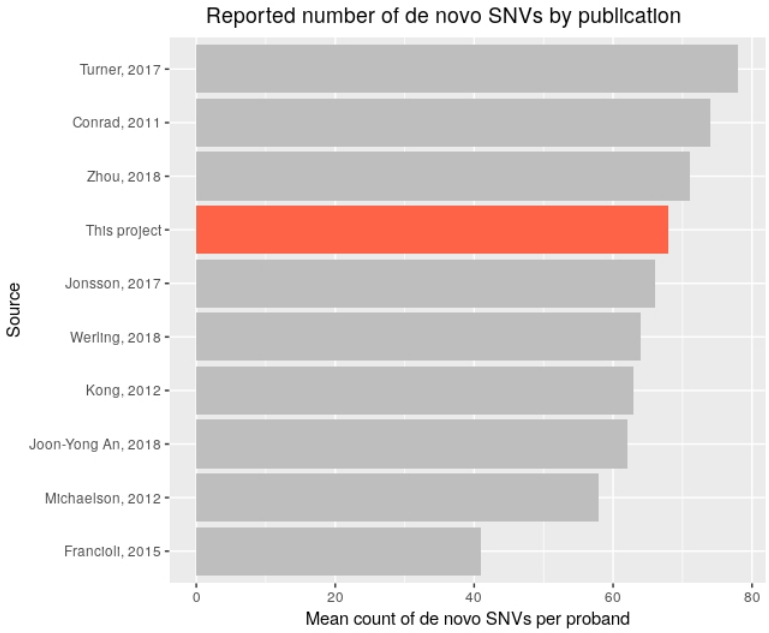

Figure 2.5: Comparison of number of *de novo* SNVs (dnSNVs) identified across several studies. Of the studies compared, the lowest observed mean count was 44 [42], while the highest was 90 [141]. The orange bar represents the mean of 69 dnSNVs per proband I identified from the Simons Simplex Collection for the work presented in this dissertation.

2.4.2 Advanced paternal age contributes to an increase in *de novo* variant count

Associations between paternal age and increased risk in neurodevelopmental disorders have been well established [138]. In particular, these associations have been

attributed to an increased number of *de novo* mutations [138, 67]. Using the ages of fathers from the SSC families, I confirmed that indeed there was a positive correlation between dnSNV count per proband and paternal age (r=0.51, p=2.6x10^{-126}) (Figure 2.6). These results estimate an increase of about 1.4 dnSNVs for each additional year of father's age at the time of conception of child. A published estimate of 1.5 additional dnSNVs per year of paternal age corroborates this finding [58]. The presence of this association serves as a validation of the *de novo* variant set.

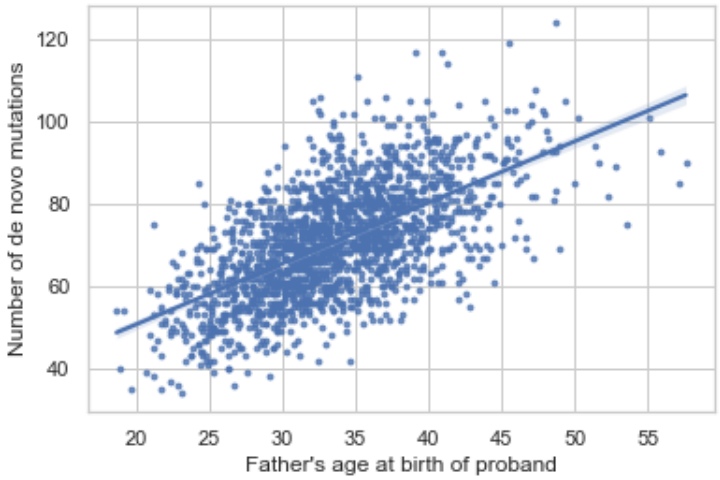

Figure 2.6: Scatterplot showing a positive correlation between paternal age and *de novo* SNV count in probands (r=0.51, p=2.6x10-126). There is an estimated increase of 1.4 dnSNVs for each additional year of father's age

2.4.3 Enrichment of dnSNVs in high-impact coding regions

Before turning my focus to noncoding dnSNVs, I wanted to first investigate the variants lying in coding regions. Given the reported associations between *de novo* variants and brain disorders from studies using whole-exome sequencing [144, 113], I decided to use this analysis as a way to further validate the set of dnSNVs I had curated by determining whether it did have the potential to reveal enrichment of certain classes of variants associated with the proband group.

For this, I annotated all dnSNVs using the Ensembl Variant Effect Predictor (VEP), an open-source tool for the annotation and prioritization of genomic variants [90]. For each variant that is annotated by VEP, a predicted consequence is calculated. These predicted consequences are then also categorized by severity and assigned an impact rating. Of the highest impact categories the set of dnSNVs were annotated as, they fell within three categories: premature stop codon gained, stop codon being lost, and a start codon being lost. These variants are all predicted to have large effects on transcription. Of the 266,865 total dnSNVs, the number of dnSNVs in probands and siblings predicted to lead to stop losses and start losses was small, with no significant difference between the groups (2 vs 4, and 4 vs 8, respectively). However, there was a surprisingly large number of variants predicted to lead to premature stop codons (120 in probands vs 62 in siblings) (Figure 2.7). This is a statistically significant enrichment in probands after multiple-testing correction (FDR-adjusted p=.001, Fisher's exact test).

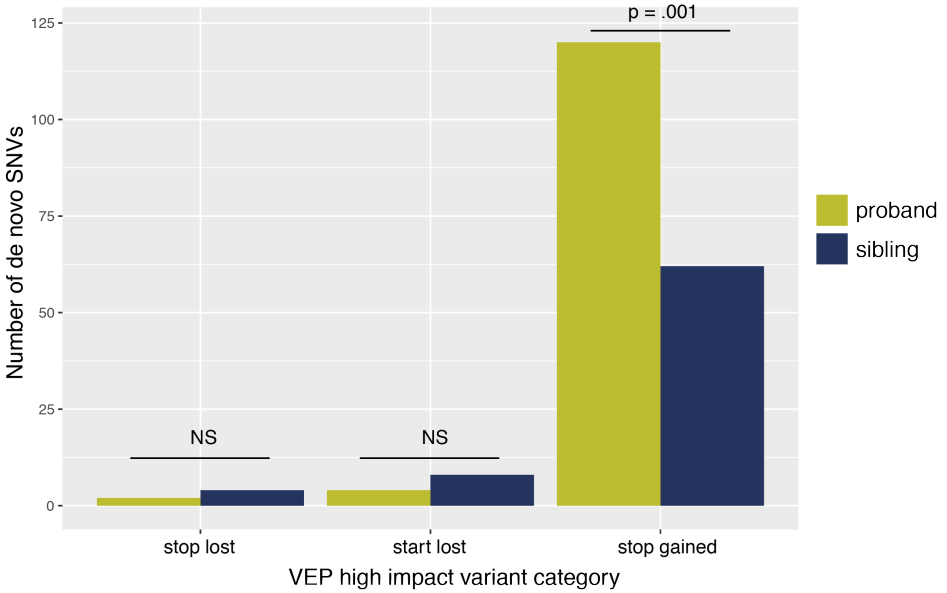

Figure 2.7: Counts of dnSNVs in high-impact coding regions. There is a significant enrichment of dnSNVs leading to premature stop codons in the proband group (120 vs 62 in siblings, FDR-adjusted p=.001, Fisher's exact test)

2.4.4 The vast majority of *de novo* SNVs are found within noncoding regions of the genome

To determine the distribution of dnSNVs by genomic region, I annotated all variants using the GENCODE v29 [43] reference file. Each variant was binned into one of five genomic regions: 3'UTR, 5'UTR, CDS, intergenic, or intronic (Figure 2.8). As expected, a large majority of variants (98%) were located in either intergenic or intronic regions. This falls in line with the approximation of 1.5% of the human genome consisting of protein-coding DNA [69]. Although I have provided evidence that there are dnSNVs with potentially large effects in coding sequences, this much larger proportion of dnSNVs in intronic and intergenic regions reinforces the importance of investigating the potential effects of dnSNVs in the noncoding regions. Proband and sibling dnSNVs were similarly distributed throughout the genome, with no enrichment of dnSNVs within any particular regions between the two groups.

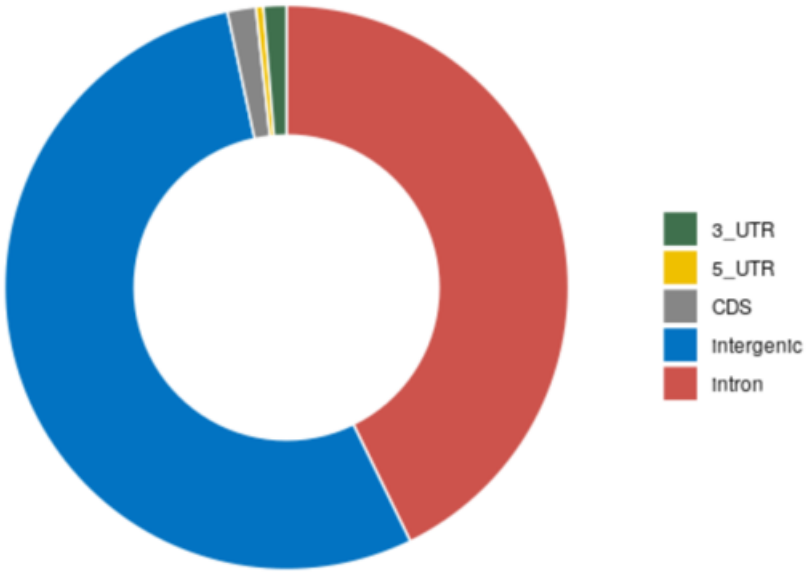

Figure 2.8: Distribution of dnSNVs in different genomic regions. Nearly 98% of dnSNVs (proband and sibling combined) are found in intergenic or intronic regions

2.4.5 Similar distribution of RegulomeDB prediction scores between probands and siblings

Because there is no significant difference in total raw count of dnSNVs between probands and siblings, it could be expected that it is a subset a functional dnSNVs which is associated with ASD risk. To identify this subset, I prioritized all dnSNVs using the RegulomeDB heuristic scoring system (described in Chapter 1), assigning scores to each variant based on evidence of having functional potential. Variants received scores between 2 and 7, with lower numbers indicating a higher probability of being a functional variant (Figure 2.9). Here, I prioritized variants with scores of 2 and 3 (and their respective sub-classes) as variants of interest. There was no significant difference of between the number of variants annotated as 2's (FET, p=.11) or 3's (FET, p=.58).

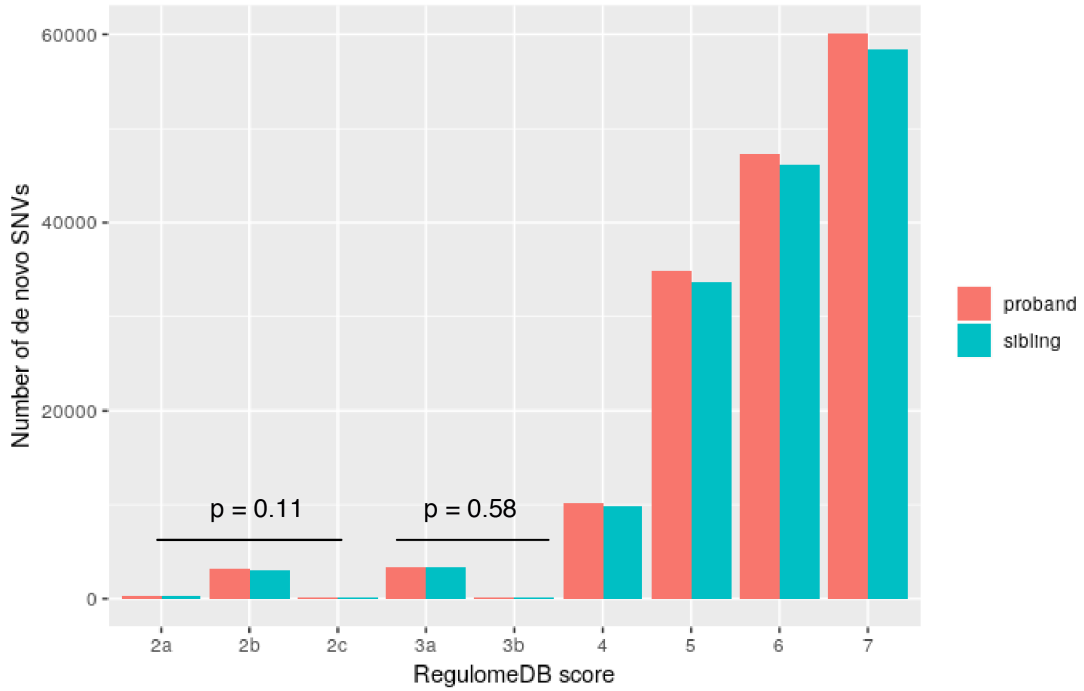

Figure 2.9: Variants prioritized using RegulomeDB. Each variant received a score between 2 and 7, with lower numbers indicating increasing evidence for overlap with regulatory elements. No significant difference was seen between probands and siblings for variants scored as 2's (p=.11, Fisher's exact test) or 3's (p=.58, FET)

2.4.6 Individual RegulomeDB annotations are not enriched after multiple-testing correction

A comparison of dnSNV RegulomeDB prediction scores between the proband and sibling groups revealed a very similar distribution of scores. This indicated to me that if there was an enrichment of regulatory variants to be detected, a more precise investigation of the functional annotations on which the prediction scores were calculated would be necessary. In this way, rather than looking for an enrichment of variants within general functional regions, I could examine individual annotations, in the event of truly enriched regions being lost in the process of combining them with other potentially irrelevant regions.

Consequently, using the RegulomeDB data, I identified each individual annotation that overlapped proband and sibling dnSNVs. This resulted in identifying 1,402 unique categories including annotations from ChromHMM, ChIP-seq, PWMs, DNase-seq and FAIRE-seq. ChromHMM annotations provided predictions on enhancer and promoter regions. ChIP-seq and PWM annotations identified predicted binding sites for specific transcription factors. DNase-seq and FAIRE-seq annotations identified regions of open chromatin. Because RegulomeDB includes annotations across diverse cell types, the cell type from which each annotation was generated was noted.

When testing for enrichment of dnSNVs within each of the 1,402 functional categories, excesses are seen in both cases and controls (Figure 2.10). Interestingly, only one functional category remains narrowly significant after multiple-testing correction, and it corresponds to an excess of dnSNVs (albeit a slight one) in ChromHMM-predicted enhancer regions in the sibling group (Fisher's exact test FDR-adjusted p=.045 , RR=.92). Additionally, the top three overall most significant results all

correspond to excesses in the sibling group. No functional category was significantly enriched in the proband group after multiple-testing correction, with the most significant proband category being ChromHMM-predictied enhancer regions from neural stem cell data (Fisher's exact test FDR-adjusted p=.63 , RR=1.26). The proband categories with the highest relative risk both came from predicted binding sites of ZNF274 and ESR1 transcription factors (RR=2.47 and RR=1.98, respectively). Although no conclusion can be made from these results, given their statistical insignificance, it is noteworthy that an association between SNPs in ESR1 and symptomatic severity of ASD has been reported before [32].

Notably, although some of the categories are associated with brain tissue, many are not. This brings into question the relevance of the non-brain tissue categories and whether even the slight excesses of dnSNVs in certain categories are truly associated with ASD or not.

2.4.7 Incorporating tissue-specific annotations is not sufficient to detect enrichment of functional dnSNVs in probands

Tissue-specific Unified Regulatory Features (TURF) was developed as an update to the original RegulomeDB scoring tool, with the goal being to provide a tool for prioritizing variants in a tissue-specific manner [35]. While the original implementation of RegulomeDB serves the important purpose of prioritizing variants according to general functional regions, it becomes increasingly important to prioritize in a tissue-specific manner when studying a phenotype with known associations to a particular organ. This has been illustrated to specifically impact studies of brain-related phenotypes [36].

Following the analysis of dnSNVs using the general RegulomeDB annotations, I proceeded to annotate all dnSNVs from probands and siblings using the TURF

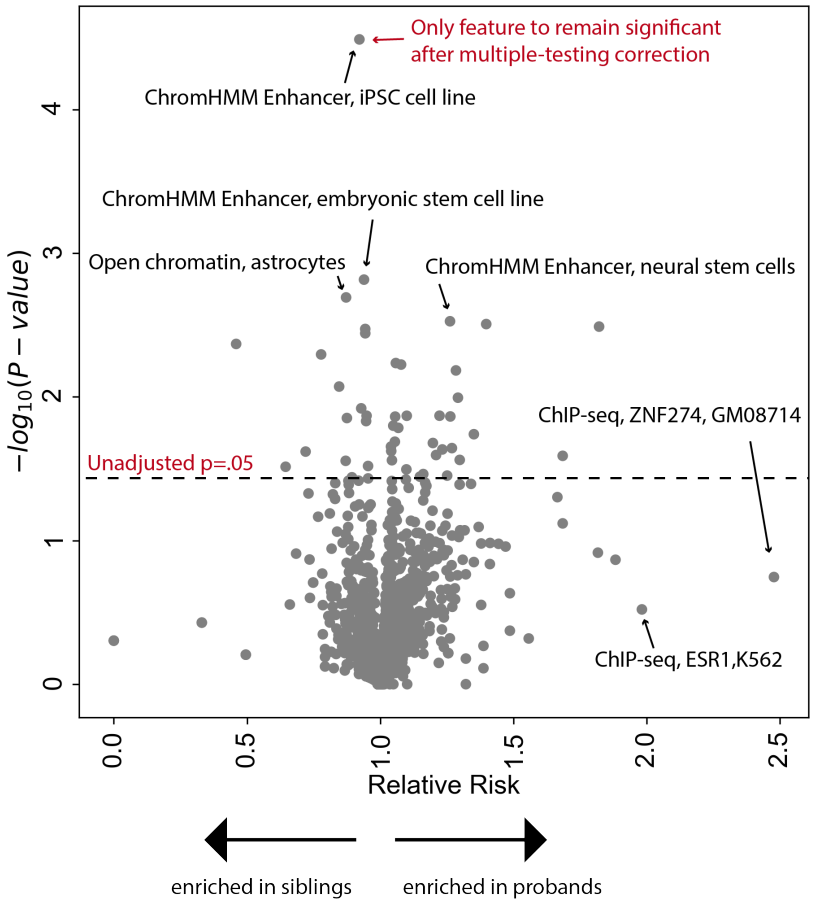

Figure 2.10: Enrichment of 1,402 individual RegulomeDB features. Points in the volcano plot represent the burden of mutations within each predicted regulatory annotation (enhancers, promoters, TF-binding sites, open chromatin regions). Excesses can be seen in both the proband and sibling groups. Only one category remains significant after multiple testing correction (ChromHMM-predicted enhancer in iPSC cell line) in the sibling group (Fisher's exact test FDR-adjusted p=.045 , RR=.92). No categories remain significant in the proband group after multiple-testing correction

algorithm, assigning a score to each variant using brain-specific annotations. When comparing TURF scores from all dnSNVs in the proband group to those from the sibling group, scores were not significantly higher in the proband group (p=0.61; Wilcoxon rank sum) (Figure 2.11). I performed the same statistical test to determine if TURF scores were higher in probands vs. sibling when restricting specifically to dnSNVs overlapping with enhancer and promoter regions, identified as described in the Methods section. As was the case when considering all dnSNVs, TURF scores

were not significantly higher in probands compared to siblings, although there was slight improvement when restricting to enhancer regions (p=0.16; Wilcoxon rank sum) and promoter regions (p=0.32; Wilcoxon rank sum).

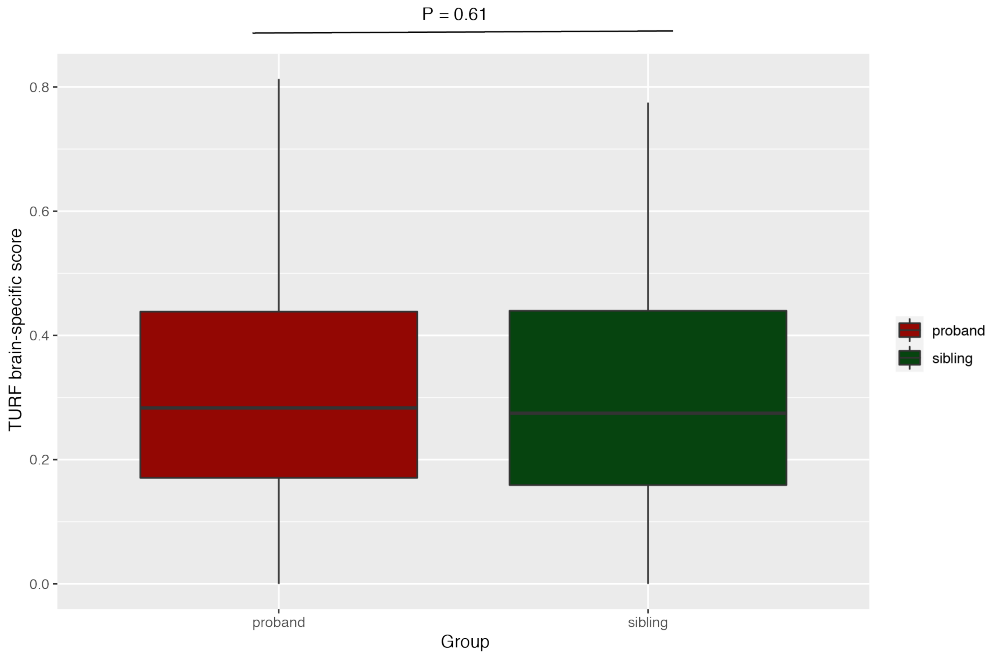

Figure 2.11: Comparison of Tissue-specific Unified Regulatory Features (TURF) scores between proband and sibling groups. All dnSNVs were scored using brain-specific annotations to calculate the TURF score. TURF scores in the proband group were not significantly higher than in the sibling group (p=0.61; Wilcoxon rank sum)

2.4.8 Multiple dnSNVs detected in the same enhancer regionss

Using DNase-Seq and ChromHMM data from the Roadmap Epigenomics Project I identified fetal brain enhancer and promoter regions (Methods). I found no significant enrichment of dnSNVs within those enhancer (3114 vs 3053; P=.7967) or promoter (1814 vs 1750; p=0.5) regions in probands compared to siblings. However, within these enhancer annotations I did identify three enhancer regions in probands which each contained three dnSNVs within 1,000bp of each other, and which I'll refer to as multi-hit enhancers. These multi-hit enhancers have been previously identified [102] and one of these enhancers has been shown to interact with the human gene

DYRK1A, for which an association with autism has been suggested [142, 59]. Importantly, I also saw two multi-hit enhancers in the sibling group, albeit not the same enhancers as in the proband group, indicating that while the multi-hit enhancers found in the probands may be biologically relevant, they were not significantly enriched in probands compared to the siblings. To confirm that multi-hit enhancers are a typical occurrence in dnSNV data and not, for example, a set of protective sibling mutations, I downloaded the full list of 184,379 dnSNVs identified in control samples from the Gene4Denovo database [154]. These are a collection of dnSNVs obtained from whole-genome sequencing data of individuals who participated as controls across various studies. Using the same filtering and annotation process as with the SSC cohort, I identified six multi-hit enhancers from the Gene4Denovo individuals, indicating no significant enrichment of multi-hit enhancers in the SSC cohort compared to an external control group.

2.4.9 Leveraging chromatin-interaction data to identify target genes

A common method for linking variants to target genes is to assign the nearest transcription start site as a variant's target. However, as few as 22% of looping interactions link elements to the nearest active TSS [122]. This suggests proximity is not the most accurate predictor of long-range interactions, such as those between enhancers and promoters. In order to address this, I have incorporated chromatin-interaction data into our analysis to more accurately link predicted enhancer regions to their target promoters. I have leveraged data from a previous study in which the authors linked risk variants to target genes using promoter-capture Hi-C (pcHiC) experiments in neural cells [131]. I applied those maps to the data in order to determine any potential contacts between enhancer dnSNVs and gene promoters. For each mutation I identified, I checked for contact with a promoter. If such a

contact was present, I assigned that mutation to the gene corresponding to that promoter. After assignment of target genes to all mutations for which contacts were available, each gene was tested for an enrichment of dnSNVs. No genes showed significant enrichment.

2.4.10 Prioritization of variants through manually selected combinations of relevant annotationss

To this point, after applying several methods to prioritize dnSNVs, no single method was successful in being able to identify a subset of functional noncoding variants enriched in probands compared to siblings. As a final approach to determine whether any of the annotations I had compiled could be used to discover ASD-associated variants, I decided to combine annotations. Much as with the concept of RegulomeDB being that regions overlapping multiple lines of evidence are more likely to have a functional impact, my rationale for combining annotations was that a single annotation may not be sufficient for filtering down to the variants with true impact.

In addition to the RegulomeDB and promoter-capture Hi-C data I have already described, here I have brought in additional relevant brain data. Whereas I previously had taken all protein-coding genes into account, I created a list of genes preferentially expressed in brain-tissue, based on data from A.B. Wells et al. [148]. I also referenced the SFARI Gene database for a list of genes with previous associations to ASD. Genes scored as a "SFARI 1" reflect the strongest evidence of true association. Other annotations I incorporated at this point were CADD scores, which are used to measure variant deleteriousness.

I proceeded to create filtering criteria based on several combinations of the available annotations. For example, instead of only focusing on variants scored as a 2

by RegulomeDB, here I tested for enrichment of variants that were not only scored as a 2, but also overlapped with the promoter region of brain-expressed genes. The tables below display different combinations of annotations I tested for enrichments of proband dnSNVs. The closest any category came to meeting significance was variants linked to the promoter of SFARI genes, based on promoter-capture Hi-C data (Table 2.2). However, as with previous cases, the significance did not survive multiple-testing correction (FDR-adjusted p=.07, Fisher's exact test).

Annotation	Proband count	Sibling count	FET p-value	FDR-adjusted p-value
promoter-capture (pcHiC) SFARI gene	2942	2684	0.0044863	0.071780492
pcHiC SFARI gene and regdb top decile	355	299	0.0310018	0.364930588
CADD 15 and regdb 2	484	419	0.0362874	0.364930588
regDB 2a brain-expressed gene overlap	10	3	0.0502047	0.396390514
enhancer-linked pcHiC SFARI 1 gene	34	22	0.0825295	0.406299214
regDB 2s pLI.9 gene overlap	870	797	0.0953722	0.435987303
pcHiC SFARI gene and regdb top quartile	865	802	0.1440061	0.542140652
regDb 2 enhancer-linked pcHiC SFARI gene	23	15	0.1430757	0.542140652
regDB 2a pcHiC SFARI gene	15	9	0.1672525	0.56731949
CADD 20 and regdb 2	174	152	0.168423	0.56731949
pcHiC SFARI 1	441	411	0.2527344	0.613139465
regDB 2a overlap with any gene	265	244	0.2633837	0.613139465
regDB 2 pcHiC any gene	1288	1229	0.2762032	0.613139465
regDB 2 pcHiC SFARI geme	166	151	0.2789427	0.613139465
enhancer-linked pcHiC SFARI gene	286	260	0.2096564	0.613139465
regDb 2 enhancer-linked pcHiC SFARI gene	35	29	0.2966166	0.613139465
regDb 2 enhancer-linked pcHiC SFARI 1	5	2	0.2359013	0.613139465
top 5% generic TURF	7370	7146	0.316799	0.614398029
SFARI gene overlap	8181	7947	0.3376909	0.621211657
regDB 2a SFARI gene overlap	24	20	0.3508142	0.621211657

Table 2.2: Enrichment test results for several manually-selected combinations of annotations, sorted by FDR-adjusted p-values. No category remains significant after multiple-testing correction.

Annotation	Proband count	Sibling count	FET p-value	FDR-adjusted p-value
CADD 15, regdb2 pcHiC SFARI gene	16	13	0.3785509	0.621211657
enhancer-linked pcHiC SFARI gene	151	141	0.3691213	0.621211657
regDB2 enhancer-pcHiC brain-exp. gene	11	8	0.3418519	0.621211657
pcHiC brain-expressed gene	1685	1632	0.3989829	0.638372632
regDB 2a pcHiC any gene	113	107	0.4325701	0.675231307
regDB 2 brain-expressed gene overlap	108	103	0.4555958	0.687929626
phredCADD 15, regdb2s, p.o. any gene	106	104	0.5376301	0.748007058
regDb 2s enhancer-prom brain-expressed	19	18	0.527496	0.748007058
DHSenhancer-prom brain-expressed	68	67	0.552167	0.751886913
enhancer-prom SFARI 1 gene	49	49	0.5842393	0.778985674
pcHiC pLI.9 gene	8000	7882	0.6942943	0.871271217
pcHiC any gene	23948	23505	0.6821594	0.871271217
enhancer-linked pcHiC brain-exp. gene	148	157	0.7792977	0.9591356
regDb 2 enhancer-linked pcHiC any gene	143	154	0.8135355	0.98238246
regDb 2 enhancer-linked pcHiC pLI.9 gene	82	92	0.8371657	0.992196354
brain-expressed gene overlap	3334	3449	0.9911017	0.99428446
regDB any2s p.o SFARI 1	19	36	0.9942845	0.99428446
promoter phastCons top quartile	18	32	0.9865762	0.99428446
enhancer-pcHiC any gene	1782	1838	0.9492307	0.99428446
enhancer-pcHiC pLI.9 gene	704	746	0.9409622	0.99428446
regDb 2 enhancer-linked pcHiC any gene	220	243	0.9129528	0.99428446
regDb 2 enhancer-linked pcHiC SFARI 1 gene	4	8	0.9322993	0.99428446
enhancer-linked pcHiC any gene	942	994	0.9560953	0.99428446

Table 2.3: Enrichment test results for several manually-selected combinations of annotations, sorted by FDR-adjusted p-values. No category remains significant after multiple-testing correction.

2.5 Discussion

A genetic basis for autism spectrum disorder has been well established. Still, even as GWAS have revealed that the majority of trait-associated SNPs lie in noncoding regions, it has remained a challenge to find specific noncoding biomarkers or associations between noncoding SNVs and ASD. The Simons Foundation Autism Research Initiative has provided a valuable resource in the Simons Simplex Collection (SSC),

a collection of whole-genome sequence data from thousands of families, designed to assist in the discovery of *de novo* variants associated with ASD.

Here I have described the steps I have taken to identify a list of high-confidence noncoding *de novo* SNVs (dnSNVs). I have established a pipeline to ensure a high quality of genotype and variant calls. I have also implemented several filters in the pipeline to increase the sensitivity and specificity of dnSNV detection. Using the *de novo* discovery pipeline, I identified 266,865 autosomal dnSNVs in 1,917 SSC Families, with a mean of 69 per proband. The number of dnSNVs per proband was positively correlated with paternal age, with an estimated increase of 1.4 dnSNVs for each additional year of the father's age. Nearly 98% of all dnSNVs identified mapped to noncoding regions of the genome.

Supporting previous findings, I detected a strong enrichment of proband dnSNVs in high-impact coding regions. With nearly twice as many dnSNVs leading to premature stop codons in probands compared to siblings (120 vs 62, p=.001), the results support the validity of the dnSNV list and that the enrichment testing approach I have used can be applied to noncoding regions.

RegulomeDB was established as a database and annotation tool, providing a method for variant prioritization by integrating functional data from ENCODE and other sources. RegulomeDB scores variants based on several lines of evidence, indicating the likelihood of a variant residing within a regulatory element and having the potential to affect transcription factor binding. I have annotated the dnSNVs identified in the SSC with both the original RegulomeDB tool, and TURF, the more recent update to the RegulomeDB model, which calculates prediction scores in a tissue-specific manner. Using TURF, I was able to incorporate annotations specific to brain function.

Ultimately, there was no category or annotation that could successfully separate ASD-associated noncoding dnSNVs from all others. In some cases, annotations that were tested even revealed an excess of dnSNVs in the sibling group vs the proband. Following evidence from previous studies suggesting the importance of using annotations relevant to organs associated with the phenotype of interest, I incorporated brain-specific annotations. This addition of brain-related annotations was still insufficient to allow for the detection of any significant associations between ASD and noncoding dnSNVs in probands. These results indicate that there may be factors which continue to present obstacles when conducting association studies for brain disorders.

CHAPTER III

Challenges in Screening for *De Novo* Noncoding Variants Contributing to Genetically Complex Phenotypes

3.1 Abstract

Understanding the genetic basis for complex, heterogeneous disorders, such as autism spectrum disorder (ASD), is a persistent challenge in human medicine. Owing to their phenotypic complexity, the genetic mechanisms underlying these disorders may be highly variable across individual patients. Furthermore, much of their heritability is unexplained by known regulatory or coding variants. Indeed, there is evidence that much of the causal genetic variation stems from rare and *de novo* variants arising from ongoing mutation. These variants occur mostly in noncoding regions, likely affecting regulatory processes for genes linked to the phenotype of interest. However, because there is no uniform code for assessing regulatory function, it is difficult to separate these mutations into likely functional and nonfunctional subsets. This makes finding associations between complex diseases and potentially causal *de novo* single-nucleotide variants (dnSNVs) a difficult task. To date, most published studies have struggled to find any significant associations between dnSNVs from ASD patients and any class of known regulatory elements. We sought to identify the underlying reasons for this and present strategies for overcoming these challenges. We show that, contrary to previous claims, the main reason for failure to find ro-

bust statistical enrichments is not only the number of families sampled, but also the quality and relevance to ASD of the annotations used to prioritize dnSNVs, and the reliability of the set of dnSNVs itself. We present a list of recommendations for designing future studies of this sort that will help researchers avoid common pitfalls.

3.2 Introduction

In human medicine, heterogenous phenotypes are frequently grouped into single disorders based solely on similarities in clinical presentation. Diagnostic criteria for these conditions usually consist of lists of symptoms, of which individual patients typically exhibit only a subset. As such, it is possible for two patients to display completely disparate sets of symptoms while still meeting the diagnostic criteria for said disorder. Identifying the genetic basis of these disorders is necessary for understanding their underlying mechanisms and developing effective treatments. However, the breadth of phenotypes presented by individuals sharing the same diagnosis reflects an underlying genetic basis that is at least as complex. Indeed, the combination of subjective diagnostic criteria and the likely polygenic basis of most diagnostic symptoms, makes it possible for individuals with the same diagnosis to have completely distinct sets of underlying mutations affecting entirely different pathways. Dissecting this variation is necessary to identify common themes in the etiology and underlying mechanisms of these disorders.

Among these disorders, Autism Spectrum Disorder (ASD) stands out as one of the most complex, making it a good case study for developing robust statistical methods to identify novel variants contributing to complex disorders. Such mutations are difficult to identify using traditional statistical methods owing to difficulty objectively grouping individual patients into cohorts and the fact that even patients with similar

clinical presentation may not share the same underlying genetic and mechanistic basis. Here we use ASD to model complex, heterogeneous phenotypes, and test strategies for identifying *de novo* variants relevant to ASD given currently-available whole-genome-sequencing (WGS) datasets and functional genomic annotations.

ASD is a term used to describe a group of neurodevelopmental conditions often characterized by difficulties with social interaction, communication, and behavior. A genetic basis for ASD was first established through twin studies showing stronger concordance between monozygotic siblings compared to dizygotic siblings [116][71, 133]. Most estimates now place the heritability of ASD in the range of 50%-90% [48, 121, 140, 12]. With technological improvements, several classes of genetic variations have been revealed to contribute to ASD etiology. This includes point mutations and structural variants (such as copy-number variants), pointing to a genetically heterogeneous background [57][124]. It is clear that many different types of genetic factors play a role in ASD, and it would benefit the research community to begin bridging the gaps in our understanding of the underlying genetic heterogeneity. As a whole, common variants contribute strongly to ASD. Individually, however, each of these common variants are expected to have small effects. This can be explained, in part, by the fact that variants associated with large, harmful effects are less likely than neutral variants to be maintained in a population. Conversely, uninherited variants that arise spontaneously (*de novo* variants) may carry a higher risk than inherited variants because they have not yet been acted upon and removed by natural selection. While the combined effects of common variants may contribute to a large portion of the heritability of autism [48], *de novo* mutations may potentially have larger individual impacts.

A significant role for protein-coding variants in ASD has been established [55].

Still, the coding variants that have been identified account for only a fraction of the overall heritability of ASD. There is evidence that the genetic background of ASD likely involves a combination of both coding and noncoding variation. Although effect sizes of mutations in coding regions may, on average, be greater because of their ability to directly affect amino acid sequences, mutations in noncoding regions may contribute to autistic phenotypes in an alternate way: by disrupting regulatory sequences [55]. Regulatory elements, such as promoters and enhancers, are responsible for controlling the precise time, location, and level of gene expression. Mutations disrupting regulatory elements may interfere with proper expression of developmental genes in the brain, for example, leading to phenotypes that are characteristic of ASD. With more than 98% of the human genome composed of noncoding DNA, those noncoding regions present a logical place to potentially uncover some of the missing heritability of autism. Indeed, most genome-wide association study signals map to noncoding regions, highlighting their importance. Therefore, investigation of noncoding variation has the potential to uncover novel ASD-associated variants.

Beyond the identification of noncoding variants, the challenge of their functional interpretation remains. Previous studies have investigated possible roles of *de novo* single-nucleotide variants (dnSNVs) in ASD by testing for enrichment of functional evidence among dnSNVs in probands versus siblings used as matched controls. However, while these studies uncovered significant associations with several coding categories, ascertaining the functional impact of noncoding dnSNVs is more difficult: whereas known properties of open reading frames, splice sites, and the biochemical properties of amino acids facilitate coding dnSNV prioritization, no such code exists to prioritize noncoding dnSNVs. Despite this difficulty, previous work has implied a modest contribution of *de novo* noncoding variation in autism, although coding re-

gions exhibited the strongest associations and noncoding associations were not robust to multiple-testing correction [150]. The authors did not suggest that a noncoding association does not exist, rather, that the signal is not as strong as expected and that larger cohorts and careful attention to multiple-testing burden would be necessary to observe any true signal from *de novo* noncoding mutations. Indeed, other studies which have focused on noncoding mutations in ASD have also seen significant enrichment disappear once multiple-testing corrections have been applied [141]. Nonetheless, these studies establish that these rare mutations do play a role in ASD, underscoring the importance of improving methods with which to study them.

In recent years, we have learned certain features associated with active regulatory sequence, including epigenetic marks, sequence motifs for various TFs, chromatin accessibility, etc. Although we know that regulatory sequences are associated with characteristic combinations of functional annotations (H3K4me1 and open chromatin for enhancers, e.g.), our ability to predict the regulatory function of a locus based on overlapping annotations remains limited. Despite an abundance of publicly-available functional genomics datasets across many tissue and cell types, a uniform functional code for regulatory sequences remains elusive. In the absence of such a code, variants must be prioritized based solely on their overlap with various combinations of functional annotations thought to be important for regulatory function. Several approaches for screening and combining annotations have been used to prioritize variants in the ASD literature with varying degrees of success.

Most studies to date have leveraged the large number of public datasets by combining individual functional annotations into functional categories meant to reflect the regulatory potential of individual variants. A statistical test for enrichment must then be run on each of these functional categories to identify if an excess of dnSNVs

exists in probands, which may signal an association with development of ASD. Since it is not known in advance which categories will be informative, this strategy requires many individual tests to comprehensively screen for associations. Unfortunately, as the number of annotations increases, the number of individual tests increases exponentially, invoking a substantial multiple-testing burden and potentially reducing statistical power by orders of magnitude. As a result, very few studies have been able to identify robust statistical enrichment of dnSNVs within any functional category in ASD affected individuals. In nearly all of these studies, the solution presented for this problem is to increase the number of families until sufficient power is achieved.

We wanted to explore the implications and practicality of this solution in comparison to newer methods of variant prioritization, in hopes of identifying the optimal solution given current technology. To this end, we analyzed whole-genome sequencing data from 1,917 quad families in the Simons Foundation Autism Research Initiative (SFARI) Simons Simplex Collection (SSC) cohort. These families represent a total of 7,668 individuals: 1,917 affected children and 5,751 unaffected parents and siblings. This family structure is ideal for identifying *de novo* variants and provides us with a natural control group in the unaffected siblings. Using a set of newly-identified dnSNVs from this cohort, we quantified the ability of selected combinations of individual annotations, and scores from published variant-prioritization methods, to detect differential associations with dnSNVs. For each of these functional categories, we used 80% power curves to predict the optimal sample size, at which 80% of tests are expected to be significant. We further evaluated the relative effect size necessary to find a significant result given the current sample size in each functional category, expressed as the difference in dnSNV burden in probands versus siblings using 80% power curves. Finally, we used simulations to compare the actual effect

sizes observed in probands and siblings to random expectations to assess whether a significant test at any sample size is likely to reflect biologically-meaningful effects on ASD risk. Based on these experiments, we can draw several conclusions about current limitations in our ability to confidently identify dnSNVs contributing to ASD, and propose strategies to improve future studies on the impact of dnSNVs on ASD and other complex genetic disorders.

3.3 Methods

3.3.1 Identification and filtering of *de novo* single-nucleotide variants

In order to identify candidate *de novo* SNVs while minimizing false positives, we implemented a pipeline to improve the quality of genotype and variant calls before applying stringent filters. We first performed genotype refinement using GATK's [143] CalculateGenotypePosteriors tool and required all variants to pass the Variant Quality Score Recalibration (VQSR) filter, using a sensitivity threshold of 99.8%. We then filtered the variant set down to bi-allelic SNVs and tagged potential *de novo* mutations if a variant was present in a child and not any of the other family members, with the requirement that all four family members have GQ >= 20, DP >= 10, and the more stringent of AC < 4 or AF $< 0.1\%$. This same process was followed separately both for probands and unaffected siblings to identify *de novo* SNVs for each group. We further filtered mutations down to exclude those appearing in more than one individual. Because the CalculateGenotypePosteriors tool was only designed to handle trios, we created separate PED files for probands and unaffected siblings, based on the PLINK pedigree file format. We referenced a file mapping SSC individual IDs to SSC family IDs, provided as a resource by the SSC, to generate these pedigree files.

We annotated all SNVs with their minor allele frequencies (MAF) using the

Genome Aggregation Database (gnomAD) v2 [62]. Following annotation, we filtered the variant set down to those with a MAF less than .001. Variants for which a MAF was not available were also retained. We then used the Picard LiftoverVCF tool [19] to liftover variants from the hg38 build to hg19 since not all computational tools we used supported the hg38 build at the time of analysis.

Some families for which data was collected were initially enrolled in the SSC as simplex families but were later discovered to be multiplex and flagged as such by SFARI. These families were excluded from our analyses. Families were also excluded if they were part of the Simons Ancillary Collection or the Simons Twins Collection. To ensure that our identified mutations were true SNVs, variants overlapping with low-complexity regions were filtered out using UCSC's RepeatMasker [130] and TRF [15] reference files. Additionally, the Mills and 1000g gold standard set was used as a reference for filtering out mutations overlapping INDELS [92].

3.3.2 Annotation of coding dnSNVs

We defined coding mutations predicted to be damaging using Variant Effect Predictor (VEP) [90] annotations. For generating predictions, we used the "-most severe" tag in order to ensure that each variant was assigned just one "consequence" instead of taking every possible transcript into account for that variant. Variants were considered to be "high-impact" dependent on their predicted consequence, based on the VEP variant consequence table from Ensembl [28]. Additionally, we considered variants to be predicted loss-of-function if they were annotated by SIFT [96] as "deleterious" and by Polyphen [5] as "probably damaging". To identify genes that may be more susceptible to being affected by mutations, we annotated mutations with a score developed by the Exome Aggregation Consortium (ExAC) called pLI, which indicates a gene's probability of being intolerant to loss-of-function (LoF) mu-

tations [72]. A high pLI score implies a gene is LoF intolerant and we annotated genes with pLI>= 0.9 as extremely LoF intolerant. We downloaded pLI scores from https://gnomad.broadinstitute.org/downloads under "Gene Constraint Scores" [62].

3.3.3 Annotation of fetal brain enhancer and promoter regions

To identify enhancer regions, we used combined male and female fetal brain DNase-seq and ChromHMM data from the Roadmap Epigenomics Project [117]. Using core 15-state ChromHMM models we identified regions containing histone marks corresponding to the ChromHMM states "Genic Enhancers", "Enhancers", and "Bivalent Enhancer." We called those regions as enhancers if they also overlapped with DNase peaks. We used a similar process to identify promoter regions, focusing instead on the ChromHMM states "Active TSS", "Flanking Active TSS", and "Bivalent TSS", before determining which of those regions overlapped with DNase peaks. We also annotated promoters through an alternate method using the GENCODE [43] release 19 gene annotation file. Using all protein-coding genes, we defined promoters as the region within 1,500bp upstream of the TSS of the respective gene.

3.3.4 Annotation with functional scoring tools

We generated TURF generic and brain-specific scores as described in Dong and Boyle [35] with the tool available at https://github.com/Boyle-Lab/RegulomeDB-TURF. We generated Disease Impact Scores following the instructions for making predictions from the DNA model, provided at https://hb.flatironinstitute.org/asdbrowser/about.

3.3.5 Other annotations

We obtained a "rank" for dnSNVs using the original RegulomeDB scoring system, with RegulomeDB v2.0. Ranks can be obtained from https://regulomedb.org/regulome-search. We considered mutations to have potentially disruptive regulatory effects if

they received scores of 2 or 3 (note that a score of 1 is not possible in *de novo* mutations because it requires an eQTL annotation). For chromatin interaction analyses we used promoter-capture Hi-C data generated by Song et al. [131]. We applied those maps to our data to identify any potential contacts between dnSNVs and gene promoters. If such a contact was present, we assigned that mutation to the gene corresponding to that promoter. We annotated variants with CADD v1.4 [114] using offline scoring scripts, following instructions at https://github.com/kircherlab/CADD-scripts/ for the GRCh37 genome build. For identifying evolutionarily conserved elements we scored dnSNVs using the 46-way placental alignment phastCons [127] track from UCSC's Genome Browser [64]. We obtained a list of genes, along with rankings based on strength of evidence of their association with ASD, from the SFARI Gene database at https://gene.sfari.org/database/human-gene/. We generated a list of genes that were found to be preferentially-expressed in brain tissue using data from A.B. Wells et al. [148].

3.3.6 Enrichment testing proceduress

For each of the annotations tested, we compiled contingency tables for Fisher's exact test (FET) by counting the number of proband and sibling dnSNVs overlapping genomic regions falling within the category and those falling outside the category. Individual FETs were performed for each annotation category using the fisher.test method in R, with alternative="greater" [151]. When multiple tests were performed, multiple-testing adjustments were made using the FDR method in R [112, 118]. For TURF generic, TURF brain-specific, and DIS scores, FET contingency tables were constructed by counting proband and sibling dnSNVs scoring in the top 5% of scores for each category. These were subjected to FET as described above. Wilcoxon rank-sum tests were also performed on non thresholded data from these categories to

compare average score rankings between proband and sibling cohorts. For all tests, p <= 0.05 was used as the significance threshold.

3.3.7 Power analysis procedures

We performed the power analyses in R using the 'pwr' package. To calculate the power for each annotation category across varying sample sizes we implemented the two-proportion test within the 'pwr' package. Proband and sibling proportions were defined as the proportion of dnSNVs within a particular annotation category compared to the total number of dnSNVs in probands and siblings, respectively. We ran each calculation at a significance level of 0.05, and with the alternative hypothesis being "greater."

3.3.8 Reverse power analysis procedures

We produced "reverse" power curves, where we held the sample size fixed and plotted power over a range of differential overlapping proportions of proband and sibling dnSNVs for each annotation category in order to assess how much additional information each would need to convey in order to reach 80% power with 1,917 quad families. The same methods and thresholds described in the "Power analysis procedures" were then used to plot power curves with the 'pwr' R package.

3.3.9 Comparison to random permutations

In order to assess whether observed counts for each of our annotation categories significantly deviate from random expectations, we performed a permutation analysis using the same input data used for the FETs. Data were randomized by shuffling the "proband" and "sibling" labels across all dnSNVs, for 10,000 permutations. For each permutation, we stored the number of proband and sibling dnSNVs overlapping each annotation category in the permuted data. Counts were used to generate an

eCDF for each annotation category, and the mean of this distribution was used as the expected count for proband and sibling. We then calculated z-scores to quantify the deviation between the observed proband and sibling counts and the expected randomized mean based on the standard deviation of the eCDF. Z-scores of 2.0 or greater were considered significant evidence for departure from random expectation.

3.3.10 Comparison of dnSNV datasets across studies

For comparisons between studies, we obtained the publicly-available dnSNV datasets from each of the other groups we included in our analyses. As with our own set, we restricted the variants to autosomal dnSNVs. We generated the UpSet plot using Intervene [66] and UpSetR [73]. For direct comparison between our variant set and that of Zhou et al., we used BEDTools [73, 111] to obtain the intersection, union, and disjoint sets.

3.4 Results

3.4.1 *De novo* SNV calls show substantial overlap with previous studies

From the SSC, we identified a median count of 70 autosomal *de novo* single-nucleotide variants per proband (134,969 total autosomal proband dnSNVs). This is consistent with the estimated mutation rate in the general human population [65, 67]. Compared to the median of 68 autosomal *de novo* mutations we identified in the unaffected siblings (131,896 total autosomal sibling dnSNVs), we did not observe a statistically significant difference in dnSNV count between the proband and sibling groups (Fig. 3.1 A).

We annotated and categorized the dnSNVs into the following genomic regions: 3'UTR, 5'UTR, intergenic, intronic, and coding. Roughly 98% of dnSNVs overlap noncoding regions of the genome, which falls in line with coding regions compris-

ing 1.5% of the human genome. When taking all dnSNVs into account, within each genomic region, there is not a statistically significant difference in counts between probands and siblings (Fig 3.1 B).

We compared our list of identified dnSNVs to those from four other published studies which also used the SCC data (Fig. 3.1 C). Approximately 84% of the dnSNVs we identified have also been identified by at least one of these four groups [150, 141, 10, 155]. While all compared groups used data from the same cohort (SSC), the number of cohort families included in each group's respective analyses varied. The largest number of variants shared between studies came from the intersection of our study (1,917 families included in analyses) with that of An et al. (1,902 families). The next highest overlap in variants came from the intersection between three studies: our own, An et al. and Zhou et al. (1,790 families). As would be expected, there was a smaller amount of overlap with the studies which included fewer families on account of data availability at the time of their publication (Turner et al. – 516 families, Werling et al. – 519 families).

3.4.2 *De novo* coding variants show significant association with ASD

In order to validate our dnSNV calling and enrichment testing strategies, we wanted to first show our ability to recover enrichments of high-impact coding mutations, which have been previously shown to be significantly associated with ASD [55], within proband dnSNVs. We prioritized dnSNVs by annotating them using Variant Effect Predictor (VEP) [90], SIFT [96], and PolyPhen [5]. Mutations were classified as predicted loss-of-function coding mutations (LOFCMs) if they were annotated as "high-impact" by VEP, or as both "deleterious" by SIFT and "probably damaging" by Polyphen. We used Fisher's Exact Test (FET) to weigh the evidence for a statistically significant excess of predicted LOFCMs in probands compared to

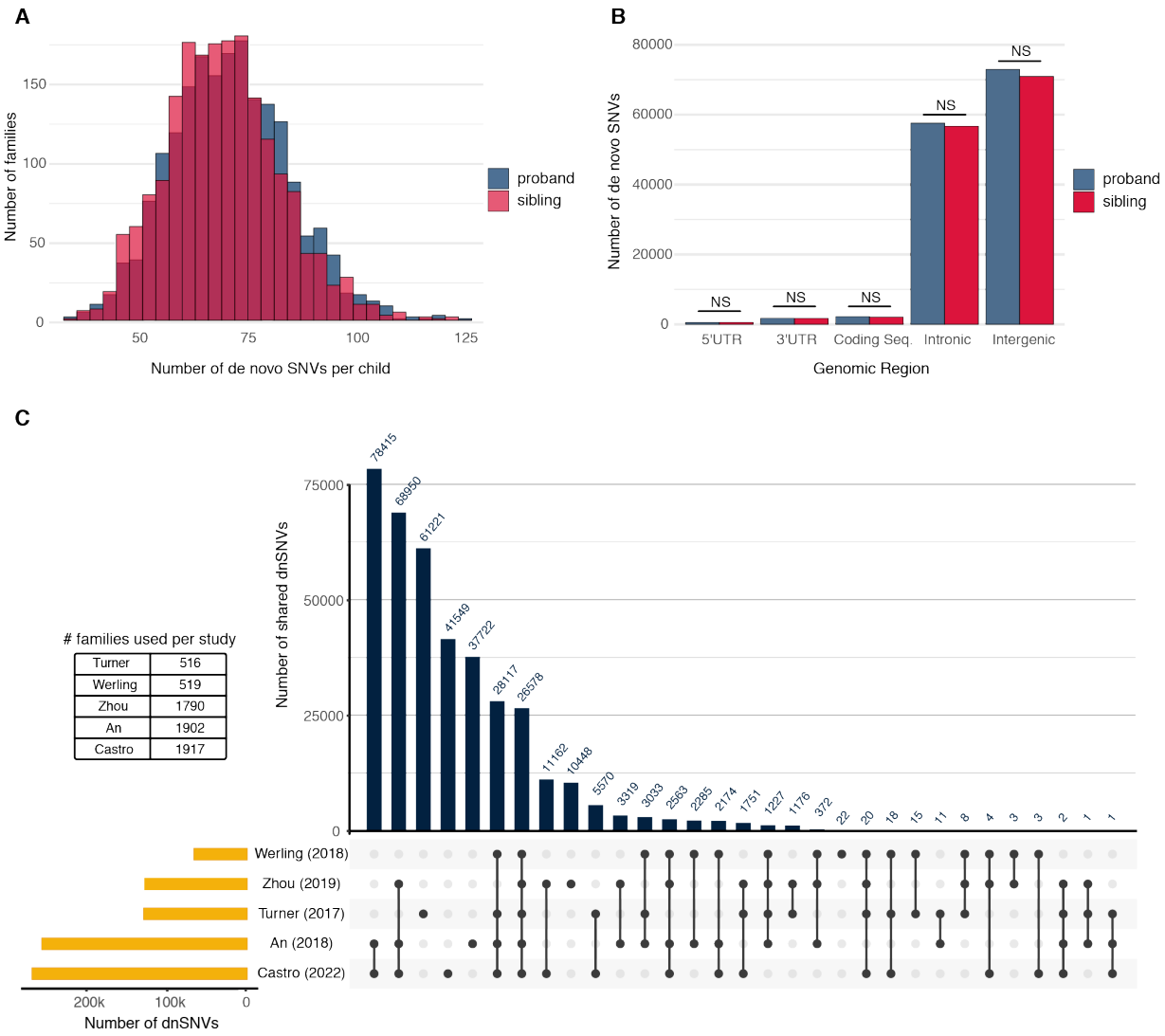

Figure 3.1: A) Distribution of *de novo* SNVs (dnSNVs) across 1,917 families from the SSC. Probands (blue) have a median of 70 dnSNVs per child while the median for unaffected siblings (light pink) is 68 dnSNVs. The darker pink bars indicate overlap in counts between probands and siblings. The difference in counts between the two groups is not statistically significant B) Distribution of dnSNVs by genomic region. Approximately 98% of all dnSNVs identified and used in this study land in intronic or intergenic regions. The number of dnSNVs in each category is not significantly different between probands and siblings. C) Total number of dnSNVs identified across different individual studies (gold bars), all using the Simons Simplex Collection cohort data. Blue vertical bars indicate the number of variants identified by more than one study (solid black points connected by black line) or variants only identified by a single group (solid black point). Although all families are part of the same cohort, the number of families utilized by each study varies, shown in the table.

siblings.

Consistent with our expectations, we observed 604 proband LOFCMs compared to

467 sibling LOFCMs, representing a statistically-significant enrichment (FET, FDR-adjusted p=0.002)(Figure 3.2). In all, over 90% of VEP high-impact variants were stop-gain mutations, which result in premature stop codons and, in turn, truncated transcripts. These stop-gain mutations were enriched even more strongly than all LOFCMs, with nearly twice as many found in probands compared to siblings (120 vs 62 mutations; FET, FDR-adjusted p=0.001). Taken together, these observations show we are able to recover known enrichments within our dnSNV dataset.

3.4.3 Proband dnSNVs are not enriched for predicted regulatory variants

We next wanted to extend our enrichment testing strategy to noncoding mutations. Specifically, we were interested in determining if we could detect an enrichment of proband mutations with the potential to affect regulatory function. We wanted to know whether a more comprehensive set of dnSNVs coupled with our FET screening strategy would yield any statistically significant enrichments in noncoding regulatory annotation categories.

This approach relies on our ability to predict how likely noncoding variants are to affect regulatory function. Most published studies have done so based on overlap with genomic annotations commonly associated with cis-regulatory regions. Annotations are derived from the Encyclopedia of DNA Elements (ENCODE) project and include, open chromatin (DNase-seq and DNase footprinting), and transcription factor (TF) binding sites (ChIP-seq), among others. These are commonly combined into annotation categories; either exhaustively, by selecting a subset manually, or by using computational methods. However, more recent studies have turned to machine learning to identify the relationships between functional annotations and ASD. The distinct advantage of this approach is that it may reduce or obviate the need for multiple-testing corrections. In order to compare these strategies, we selected a

combination of annotation categories and prioritization scores and proceeded with enrichment tests to determine if any significant associations could be recovered.

As representative manually-curated annotation categories, we combined annotations from the ENCODE database for ChromHMM and DNase-Seq, enhancer and promoter predictions from the Roadmap Epigenomics Project [117], chromatin interaction data from Song et al. [131], and brain-specific gene expression data from [148]. These were used to isolate sets of likely promoter and enhancer regions targeting genes expressed in the brain. Given known features of ASD etiology, these two genomic compartments seemed likely to harbor an enrichment of dnSNVs affecting relevant regulatory functions. However, when considering the number of mutations overlapping enhancers we saw no significant difference between the proband and sibling groups (3114 vs 3053, respectively; p=0.79, FET). Similarly, there was no enrichment of mutations within promoters (1814 vs 1750; p=0.5, FET).

As representative variant prioritization methods, we selected Tissue-specific Unified Regulatory Features (TURF), a probabilistic scoring model that prioritizes in a tissue-specific manner, which replaced the original categorical scores in the current release of RegulomeDB [18]. For noncoding variants, we scored each dnSNV with two different TURF prediction scores; one score was based on a previous implementation of TURF which scores variants in a generic context (independent of tissue) [33]. The second score was generated by the current implementation of TURF, in which functional variants are predicted in a tissue-specific context [35]. We calculated these scores based on functional evidence specifically from brain tissue, which we could reasonably expect to be more relevant to ASD. For both generic and brain-specific TURF scores, FET contingency tables were constructed based on overlap with positions scoring in the top 5% of annotated sites. In both cases, tests were not significant

after multiple-testing correction, similar to previous studies (generic TURF: FET, FDR-adjusted p=0.6; brain-specific TURF: FET, FDR-adjusted p=0.5)(Fig. 2). Since TURF scores are numeric and continuous, we retested for enrichment using Wilcoxon Rank-Sum tests, which also failed to reach significance (generic TURF: Wilcoxon Rank-Sum, p=0.83; brain-specific TURF: Wilcoxon Rank-Sum p=0.89).

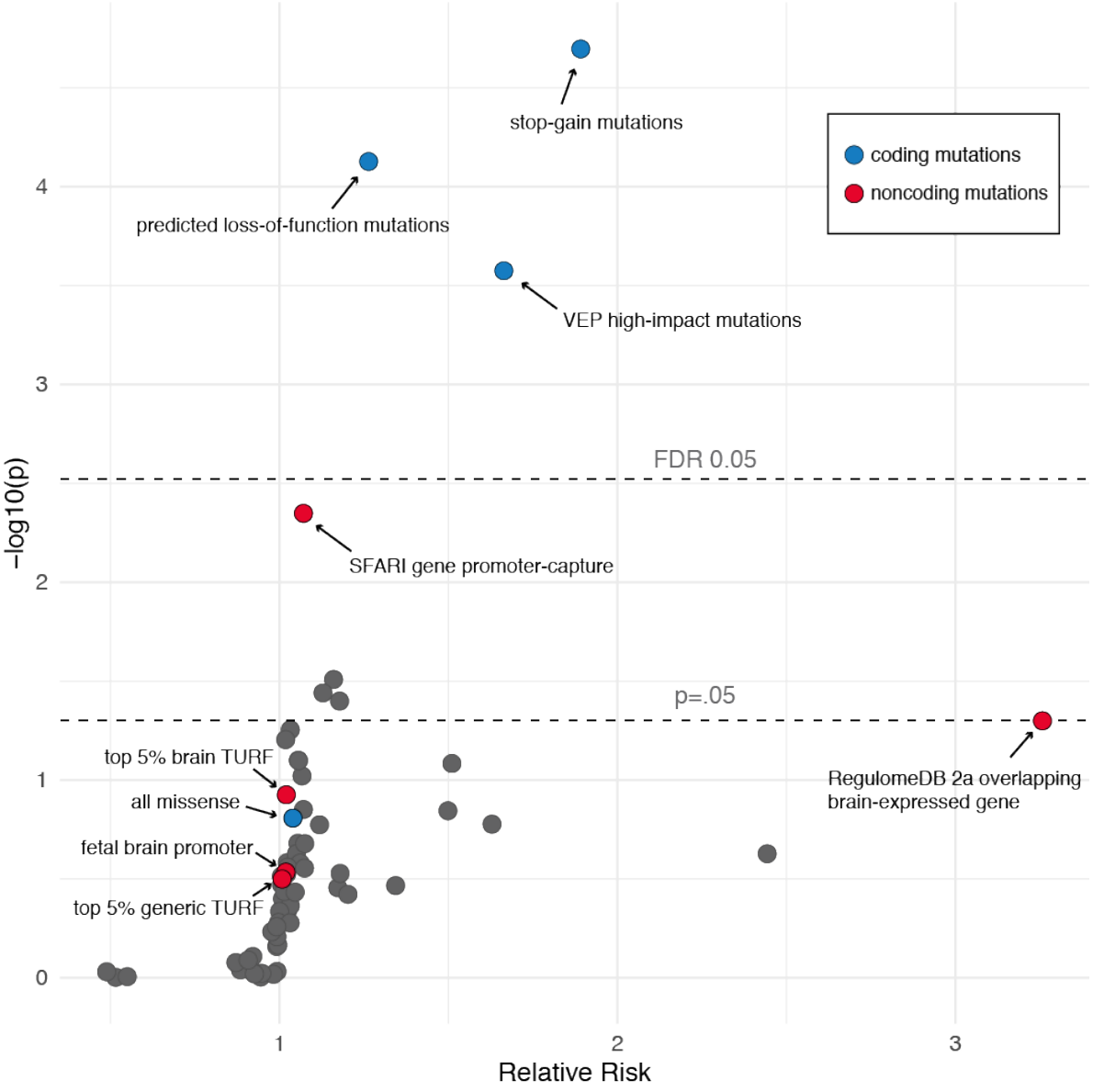

Figure 3.2: Relative risk of proband *de novo* single-nucleotide variants across 65 annotation categories, including combinations of different annotations. A relative risk >1 represents enrichment of dnSNVs in the proband group. The only categories that remain significant after multiple-testing correction are related to coding-region annotations.

3.4.4 Tissue- and disease-specific annotations are more informative than tissue-agnostic annotations

Given the lack of enrichments observed for any annotation categories or prioritization scores, we wanted to further explore how informative these annotations actually are. Since most previous studies failing to show significant enrichments have cited insufficient sample size as their primary limitation, we first asked what sample size would be necessary to achieve 80% power in our statistical tests. To do so, we plotted power curves for three noncoding annotation categories and prioritization scores: TURF generic, TURF brain-specific, and Fetal Brain Promoters. For comparison, we included two coding annotation categories: missense variants, regardless of severity of impact, and high-impact variants, which are likely to lead to loss-of-function.

Considering a desired power threshold of at least 80%, our analyses revealed that we are underpowered to detect enrichment of dnSNVs in probands for any of the noncoding categories given our sample size of 1,917 quad families (Fig. 3.3A). As a baseline, we estimated a power of 97% when testing for enrichment of high-impact coding dnSNVs using the same sample size. However, it is notable that the missense coding category yielded only 27% power at the current sample size, slightly below the highest-performing noncoding category: TURF brain-specific scores (32% power). TURF brain-specific scores were estimated to reach 80% power at a sample size of 10,000 families.

We note that if we were to instead use generic (non-brain-specific) TURF scores for prioritization, over 50,000 families would be necessary to achieve 80% power, highlighting the potentially profound effects of the choice of training strategy for variant prioritization. In particular, we see that including data directly relevant to the tissue and developmental stage under study offers a significant improvement for

this application. This is underscored by the observation that generic TURF scores (12% power) underperformed the manually-curated fetal brain promoter category (13% power) even though generic TURF scores incorporate more annotations to generate their prioritization scores. Even so, an estimated 37k families would be required to reach 80% power for detecting enrichment of fetal-brain promoter mutations, showing that neither of these annotation categories are particularly informative in terms of ASD risk.

3.4.5 Improving annotation quality has more impact on empirical power than increasing sample size

We wanted to resolve the question of whether a significant test result based on a larger sample size would actually reflect a meaningful association between any of these annotation categories and ASD. In order to further dissect the strength of associations between proband dnSNVs and ASD, we assessed statistical power for each annotation category at a fixed sample size over a range of effect sizes. We defined the effect size for an annotation category as the difference between the fractions of proband and sibling dnSNVs within a given annotation category. The magnitude of this difference is indicative of the strength of the association between an annotation and ASD. A highly informative annotation category would be expected to associate mostly with proband dnSNVs and only rarely with sibling dnSNVs, so "overlapping" counts in the FET contingency table would be highly skewed toward the proband column. The effect size, therefore, would be relatively large. By contrast, an uninformative classifier would associate randomly with proband and sibling dnSNVs; i.e. the observed proportions of proband and sibling dnSNVs overlapping the annotation category would be equal. Thus, the observed effect size would be very small. We wanted to assess how informative our annotation categories actually are in differ-

entiating between proband and sibling cohorts. More precisely, we wanted to ask, if we could improve an annotation to make it more informative, how much more information must it convey to achieve 80% power at the current sample size? We interpreted this as an ersatz quality metric for each annotation category.

We evaluated how much we would need to inflate the effect size of each annotation category by plotting curves relating statistical power to the proportion of excess information in probands relative to siblings in FETs. In this framework, zero effect size is observed when both proband and sibling have equal proportions of dnSNVs overlapping a given annotation category. We then emulate the consequences of increasing the effect size by increasing the proband overlapping proportion while holding the sibling proportion fixed, thus artificially increasing the effect size of the annotation category. By recalculating empirical power thusly over a range of effect sizes, we can plot curves showing the necessary effect size increase to reach 80% power at the current sample size (Figure 3.3B).

Consistent with our power analysis results, brain-specific TURF scores required the smallest increase in effect size to reach 80% power, approximately 2.2%, or roughly 150 more than the 6,828 we actually observed. This makes them somewhat more informative than generic TURF scores, which would require a 3.2% increase of dnSNVs overlapping the top 5% of scores, or 240 more dnSNVs than the 7,370 actually observed. However, both variant prioritization methods performed substantially better than the manually-curated fetal brain promoter category, which would require a 6.5% increase in information to achieve 80% power at the present sample size, or 117 variants in addition to the 1,806 observed.

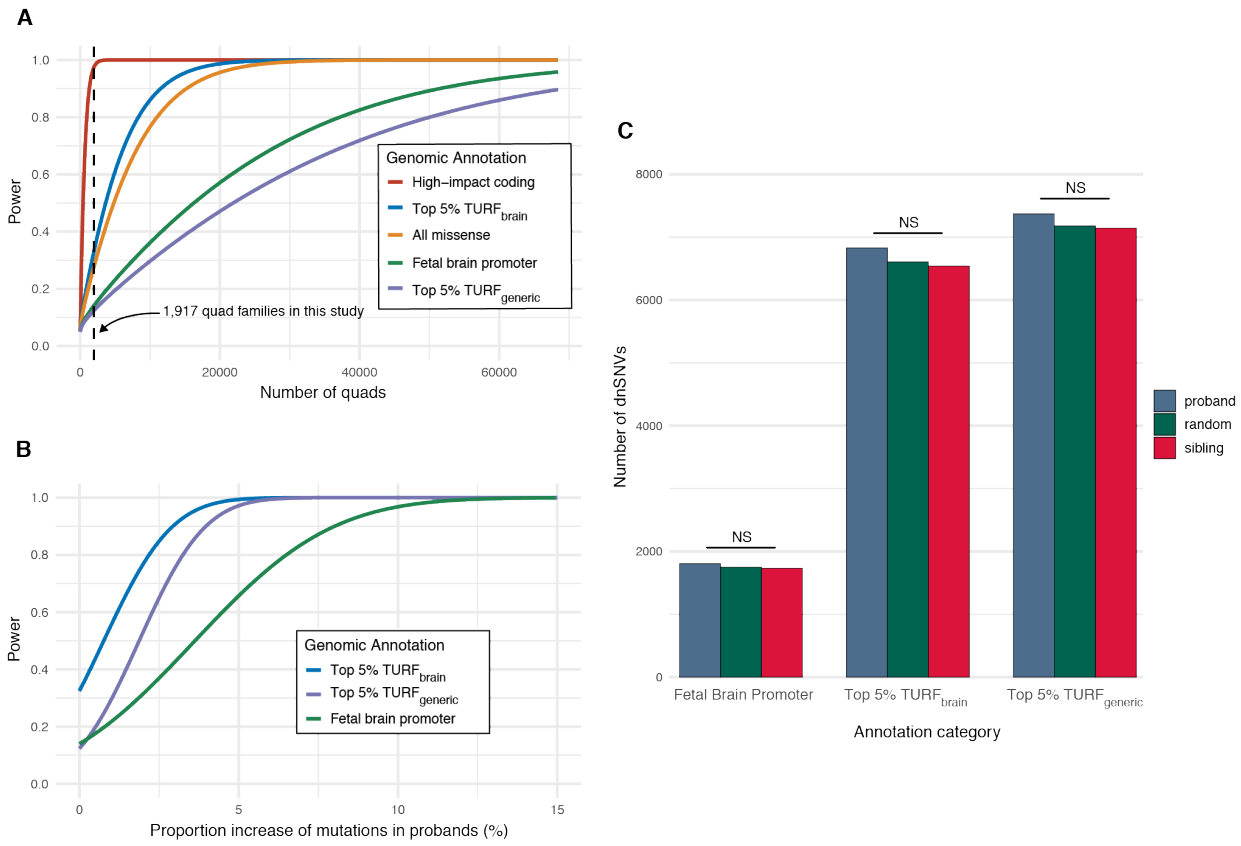

Figure 3.3: A) Power analysis for detecting proband enrichment of different categories of *de novo* SNVs. The black dashed line indicates our current sample size (1,917 quad families). We have estimated the sample sizes necessary in order to detect association of dnSNVs with ASD. We estimate a power of 97% when testing for enrichment of high-impact coding dnSNVs in probands at our current sample size. The missense coding category yields a power of 27%. Brain-specific TURF scores (32% power) would require 10,000 more families to achieve 80% power. Over 50,000 families would be necessary for generic TURF scores (12%) to reach that same 80% threshold. The fetal brain promoter category slightly outperforms the generic TURF scores at 13%. B) Generic TURF starting power = 0.12, achieves 80% at 3.2% increase (240 additional variants, 7,370 observed). Brain TURF starting power = 0.32, achieves 80% at 2.2% increase (150 additional variants, 6,828 observed). Fetal brain starting power = 0.14, achieves 80% at 6.5% increase (117 additional variants, 1,806 observed). C) Observed counts of proband (blue bars) and sibling (red bars) *de novo* SNVs prioritized with three different noncoding annotations. We observed no significant difference between random counts (green bars) and counts in probands or siblings (Z-scores: fetal brain promoters = 0.53, TURF generic = 0.48, TURF brain = 1.16, permutation tests).

3.4.6 Comparison of current annotations to random permutations

Given the results of both the power and effect size analyses, we can conclude that the most-informative noncoding annotation category among those tested were

TURF brain-specific scores. However, even though only a modest increase in either sample size or effect size was necessary to get to 80% power, the actual enrichment test still failed to reach significance. All others performed significantly worse: extreme increases in either sample or effect sizes would be necessary to reliably detect enrichments. This led us to question whether these annotation categories actually conveyed significantly more information than random expectations.

For each of the three noncoding annotation categories, we modeled the expected random overlap of proband and sibling dnSNVs with the given annotation category by randomly shuffling "proband" and "sibling" labels across all dnSNVs for 10,000 permutations. The mean number of overlapping proband and sibling dsSNVs across permutations were then compared to the applicable observed counts (Fig 3.3C). Z-scores were calculated to quantify the degree of departure between the observed count and its random expectation. Z-scores exceeding 2.0 were considered significant. However, the observed Z-scores for all categories were well under this threshold. Notably, TURF brain-specific scores, which performed the best in our other tests, was the only annotation category with a Z-score exceeding 1 (1.16). By comparison, the high-impact coding category produced a Z-score of 3.57 using these methods. Therefore, we can conclude that even the best of our annotation categories are relatively uninformative in regards to ASD risk or etiology.

3.4.7 dnSNV calls show variable quality across studies

We have seen the limitations of sample size and variant effect size across several studies conducted by other research groups who have used the same raw data from the Simons Simplex Collection to study noncoding mutations in ASD. Interestingly, while all these studies are subject to the same limitations, their results appear to vary substantially in terms of the specific associations found, with little reproducibility

between groups, even when testing methods were similar.

To date, only one study has reported a robust statistical enrichment in a noncoding category [155]. The authors use a disease impact score (DIS) to prioritize dnSNVs relative to their impact on brain disease. This method is an extension of DeepSea [156], a machine-learning model that assigns functional scores to individual variants based on overlap with functional annotations from ENCODE and other sources. DIS extends this model by way of training on a curated set of known features related to brain disease phenotypes. The authors found significant evidence for higher DIS scores in probands compared to siblings (p=.009, one-sided Wilcoxon rank sum test).

We wanted to know whether DIS scores would yield a significant enrichment within our data as well, so we tested our own set of dnSNVs but found no enrichment using either FET (p=0.214, one-sided FET) or Wilcoxon Rank-sum tests (p=.071, one-sided Wilcoxon rank-sum test). This prompted us to investigate the effect of the specific set of dnSNVs on test results. We investigated this by testing for association of DIS scores within the union, intersection, and disjoint fractions of our dnSNVs and those from Zhou et al. For each of these fractions, we repeated the Wilcoxon rank-sum test as before, making note of whether a significant difference was apparent.

The only significant result we observed was for the intersection of both datasets ($p < 0.0026$, one-sided Wilcoxon rank-sum test), which actually showed stronger evidence for enrichment than Zhou et al. originally reported (p=0.009, one-sided Wilcoxon rank-sum test). By contrast, neither the union (p=.083, one-sided Wilcoxon rank-sum) nor disjoint datasets produced a significant result, with the lowest performance observed for the disjoint sets (This study: p=0.99; Zhou et al.: p=0.73). This suggests that the quality of dnSNV calls varies, with the highest-quality calls also being the most reproducible.

We further tested this hypothesis by intersecting our variants with those from the previously-mentioned four other groups who have used their own methods to identify dnSNVs in the SSC cohort (Fig 3.1c). We filtered down our set of dnSNVs by keeping only those which appeared in at least two of these other groups' published sets. Once again, performing the Wilcoxon rank sum test on the intersection set of dnSNVs revealed a significant difference in DIS between probands and siblings ($p = .02837$), albeit at lower confidence compared to the intersection between our dnSNVs and those from Zhou et al. We speculate that this decrease results from the decrease in overall sample size incurred due to the smaller absolute size of the dnSNV datasets in other studies.

Altogether, these results suggest that intersecting variants across sets of calls leads to an overall increase in quality among the dnSNV calls. We postulate that the disjoint sets of variants from across studies are enriched for false-positive variant calls, which may arise due to sequencing errors, genotyping errors, or other unknown sources. If not filtered out, these false-positive dnSNVs may dilute the signal from true dnSNVs sufficiently to prevent a significant test result even when an annotation is genuinely associated with ASD.

3.5 Discussion

Prior to this analysis, several groups have used data from the SSC to seek associations between noncoding dnSNVs and ASD, with all but one yielding no statistically significant enrichments. Our results on three different noncoding annotation categories were consistent with their results in that we found no significant associations. Our testing methods reproduced previously-demonstrated enrichments of proband dnSNVs within high-impact coding annotation categories. However, it was not im-

mediately clear if we observed no noncoding enrichments due to insufficient sample size, as suggested by the authors of most previous studies, or the inability of the annotations we chose to reliably differentiate between regulatory and neutral noncoding variants.

In order to systematically investigate these possibilities, we set out to objectively evaluate how informative current annotations are in regards to differences in functional effects of dnSNVs in probands vs siblings. This allowed us to compare different strategies in terms of their effects on our ability to find potentially-revealing genomic associations. Specifically, we explored the impact of sample size, choice of annotations/variant prioritization methods, and choice of variant sets, on our power to detect associations between *de novo* noncoding genetic variation and ASD.

Based on observed dnSNV counts in probands and siblings, our power analysis suggests at least 10,000 quad families would be necessary to achieve a power of at least 80% to detect an association between our best-performing noncoding annotations category and ASD. Although autism cohorts are constantly growing, this is approximately five times as many quad families than are currently available in the SSC. More importantly, though, the necessity of such a large number of families suggests a very small effect size, begging the question of whether such effects are meaningful. We show evidence that, in fact, current annotations are only slightly more informative than random expectations. Therefore, the strategy of increasing sample size alone is likely to lead to erroneous conclusions.

What we see from this is that it is not only the sample size that is limiting our ability to detect the effects of the dnSNVs, but also their effect sizes. This suggests that certain annotations categories may not sufficiently capture meaningful differences between probands and siblings: i.e., the annotations used are unable to reli-

ably differentiate between noncoding dnSNVs that are neutral (without regulatory effects) and those capable of disrupting regulatory function. This may be amplified particularly in phenotypically heterogeneous disorders such as ASD. Therefore, an insignificant test result may not actually reflect lack of signal, but that the signal has been attenuated by high-scoring dnSNVs that actually lack functional significance. This effect may be particularly problematic if we rely solely on our intuition when choosing annotation categories. This is demonstrated by the poor performance of the fetal-brain-promoter category in this study; even though it is based on relevant, tissue-specific annotations, it still fails to produce a significant association. Thus, the most important choice when designing an experiment is likely which annotation category/ies and/or prioritization score(s) to use, keeping in mind the need to minimize multiple-testing burden. Werling et al. illustrated the importance of multiple-testing burden in a study that included a comprehensive set of >13k annotation categories, among which no significant associations were found after correcting for multiple-tests [150].

Accordingly, it is likely that improving annotations and prioritization scores, particularly in relation to their relevance to the specific tissue/disorder under study, is more likely to yield meaningful performance gains than increasing the number of available families for study. For example, we note that brain-specific TURF scores performed significantly better than generic TURF scores, highlighting the importance of using tissue-specific annotations when possible. We estimate nearly five times as many families would be necessary to achieve a power of 80% when using the generic TURF scores compared to tissue-specific TURF scores. Furthermore, DIS scores, which are specifically trained on disease-related features, outperformed TURF brain-specific scores even though the underlying training feature sets share

substantial overlaps. A clear limitation when investigating the impact of noncoding mutations in ASD is there are a great deal of ways with which we can choose to annotate and prioritize variants. In theory, we isolate the variants with evidence of being functionally relevant so that we can then use that subset of variants to test for genotype-phenotype associations. However, depending on the choice of annotations, the sets of variants being tested can be very different from each other, which would have downstream consequences on the observed results.

Of the variant prioritization methods we implemented in this study, we achieved the greatest power when using tissue-specific TURF scores. Using a combined-annotation scoring system, such as TURF, comes with advantages compared to using individual annotations. For one, it minimizes the multiple-testing burden because multiple annotations are already built into the scoring system without having to test them individually. Additionally, combined-annotation scoring systems also help eliminate any bias introduced by the manual selection of annotations. Manual selection of annotations relies on the investigator's pre-conceived notions of which genomic regions may or may not be relevant or functional, and therefore has the potential to introduce irrelevant data or miss sources of true signal. Previous investigators who have focused on a few specific genomic regions (e.g. promoters, UTRs) have themselves pointed out that not all possible classes of noncoding regulatory elements were considered in their study [141], which could allow for other important regions to be missed. Zhou et al. have provided further evidence of the utility of combined-annotation scoring systems in their work in which they detected a significant burden of mutations affecting transcriptional and post-transcriptional regulation in ASD probands as compared to their unaffected siblings, using their Disease Impact Score [155]. We note that some differences in results between our group's work and that of

Zhou's could be attributed to the fact that their prioritization scoring method was trained using a set of mutations specifically associated with disease from the Human Gene Mutation Database, while TURF was trained on SNVs associated with general regulatory function.

Different methods for identifying dnSNVs will yield different lists of mutations, even when starting with the same sequencing data or variant calls. These differences in lists between research groups can negatively affect reproducibility, and in fact we provide evidence that when working with the intersection of dnSNVs from other groups we can improve the overall ability to detect associations. When comparing the sets of mutations identified by different groups, including our own, it is encouraging to see that many of the same dnSNVs can be reproduced across groups (Fig 1c). We can reasonably expect to have higher confidence that the intersections of the sets represent true dnSNVs. Indeed, the fact that the intersection of our dataset and Zhou et al. yielded a stronger enrichment for DIS scores than either dataset alone suggests that the intersection is itself enriched for regulatory variants as compared to variants in the disjoint set. This is consistent with the possibility that variants discovered by only one group may be more likely to be false positives. We do note that some differences in dnSNV sets may be attributed to the fact that some studies included families that others did not, and there was little consistency in dnSNV identification methods across studies. Taking this into account, we suggest that improved methods of dnSNV identification and validation are likely to generate substantial improvements.

We were surprised to note that the use of Wilcoxon rank-sum tests instead of FET had disparate effects when using TURF brain vs DIS scores. Specifically, TURF brain scores performed better with FET (FET p=0.5; Wilcoxon rank-sum p=0.89) whereas

DIS performed better with Wilcoxon rank-sum (Wilcoxon rank-sum p=0.071; FET p=0.214). This suggests that different prioritization scores may include biases that affect our ability to detect associations with a phenotype of interest. For example, while dnSNVs scoring within the top-5% of TURF brain-specific scores are modestly (but not significantly) enriched in probands, proband dnSNVs do not systematically rank higher than sibling dnSNVs based on their Wilcoxon rank-sum test results. By contrast, the reverse is true for DIS. While it is not immediately clear what may be responsible for the difference, it raises the question of whether a single significant test result can be considered definitive evidence of correlation or whether hidden structure may sway test results if only a single testing method is used, or whether disparate test results reflect shortcomings in the quality of a given annotation. This suggests that it may behoove researchers to compare results across different testing methods, giving preference to annotations that show consistent performance regardless of method.

Taking all these findings into account, we can make several recommendations for testing associations between genetic disorders and rare *de novo* variants:

1. Start with a high-confidence set of dnSNV calls, possibly leveraging intersections with other published datasets.

2. Select annotations and/or training features relevant to the tissue and/or phenotype of interest. Our results showed that brain-specific TURF scores outperformed generic TURF by a wide margin. Likewise, DIS outperformed brain-specific TURF in average score rank in probands and vs siblings, the difference being that the DIS model was trained on disease-specific regulatory variants, not general regulatory variants.

3. Do not rely on intuition alone in selecting annotations. Currently available

machine-learning models do a better job of isolating signal from noise among a large and varied set of individual annotations.

4. Improving prioritization scores rather than increasing sample size is more likely to yield positive results. In particular, choosing training data that is relevant to the tissue or phenotype under study is of critical importance.

3.6 Publication

The manuscript of the work presented in this chapter has been submitted and is accessible on bioRxiv [23]: Christopher P. Castro, Adam G. Diehl, Alan P. Boyle. Challenges in screening for *de novo* noncoding variants contributing to genetically complex phenotypes

CHAPTER IV

De Novo Browser

4.1 Abstract

This dissertation presents an annotated list of 267,000 *de novo* single-nucleotide variants from 1,917 quad-structure families in the Simons Simplex Collection (SSC). I have developed a web application database called the *De Novo* Browser to allow for a straightforward way to explore the data. The variants can be explored in table form and are sortable by a variety of features and annotations. Additionally, one may filter the variants to be displayed by combinations of those features. Functional prediction scores, evolutionary conservation scores, and overlap with genomic features are among the annotations included. This database can help researchers identify genetic markers associated with ASD, as well as providing insight into the genetic basis of ASD. The database is a valuable resource for the research community, as it can save time and effort from repeating the same work, or from searching through multiple different sources for the data they need.

4.2 Introduction

The work I have presented in this dissertation has resulted in identifying, annotating, and classifying approximately 267,000 *de novo* single-nucleotide variants (dnSNVs) from just over 3,800 individuals participating in the Simons Simplex Col-

lection (SSC) [40]. Many of these variants have also been identified across other autism genomics studies, highlighting their likely importance. While the focus of my work has been on autism spectrum disorder (ASD), genes harboring *de novo* variants have been shown to overlap between several neurodevelopmental disorders [135]. Therefore, much of the data on genes and mutations I have curated could also be relevant to related studies and either serve as a supplement to other existing data, or save researchers the time and effort of repeating the work I have already done here. To that end, in an effort to make the work I have done more accessible to the rest of the scientific community, I have developed a web application database which can be used to explore the comprehensive list of all annotated *de novo* SNVs from my work, called the *De Novo* Browser.

Through the *De Novo* Browser, researchers may make comparisons between individuals or families, cases vs controls, and apply a variety of filters to displaying the variants. The aim of the database is to allow for a straightforward way to explore the large amount of data, in whichever way is of the most interest to the researcher. Whether the focus is on ASD, or *de novo* mutations in general, *De Novo* Browser provides the research community with an accessible tool which can be used to provide insights on patterns of *de novo* mutations in humans

4.3 Methods

4.3.1 Data collection and processing

The dnSNVs listed in the database are derived from whole-genomes of 3,834 individuals participating in the SSC , of whom 1,917 have an autism diagnosis. Variant calling was originally performed by the Centers for Common Disease Genomics and the New York Genome Center. I processed VCF files containing the raw variant calls through pipelines for genotype and quality score recalibration using the Genome

Analysis Toolkit (GATK) in order to improve the quality and only retain those of the highest quality. I filtered variants down to high-quality autosomal SNVs and lifted them all over to the hg19 genome build. Finally, I used a custom script to identify *de novo* variants in each family by identifying variants present in the probands, but absent from the unaffected sibling and both parents. I repeated this process to also identify *de novo* SNVs in the unaffected siblings.

4.3.2 Annotations

Three of the functional prediction scores provided in the *De Novo* Browser are based on the RegulomeDB framework: the RegulomeDB v2.0 rank score[18], Score of Unified Regulatory Features (SURF) [34], and Tissue-specific Unified Regulatory Features (TURF)[35]. The rank score is calculated using the original RegulomeDB algorithm, based on overlap with DNase hypersensitive regions and predicted transcription factor binding sites. Scores will range between 2 and 7, with lower-number scores representing a larger predicted regulatory variant effect. Scores of "1" will not be present in this set of dnSNVs because such a score requires the presence of an eQTL. SURF builds upon the existing annotations used for calculating rank scores by implementing a machine-learning based framework which combines features from RegulomeDB and DeepSEA [18, 156], incorporating data from massively parallel reporter assays to predict the effect of variants on expression in promoters and enhancers. SURF scores range between 0 and 1.0, with scores closest to 1.0 representing variants with the largest predicted effects. Lastly, TURF extends on SURF by training the model on tissue-specific functional genomic annotations. This provides a score ranging from 0 to 1.0 for a specific tissue of interest. For the web application, I have provided brain-specific TURF scores.

Other tools that I used for variant annotation include Combined Annotation De-

pendent Depletion (CADD) [114], Variant Effect Predictor (VEP) [89], phastCons [89, 127], SIFT [128], and PolyPhen [6] . CADD scores predict the deleteriousness of variants by integrating multiple annotations and conservation information. In the web application, I have provided the raw CADD scores as well as Phred-scaled scores, with higher scoring variants more likely to be deleterious. SIFT and PolyPhen were developed primarily for predicting the effects of amino acid substitutions, and therefore are more appropriate for analyzing coding variants.

To determine overlap of dnSNVs with enhancer and promoter regions, I used a combination of ChromHMM and DNase-seq data from the Roadmap Epigenomics Project [117]. All data used was derived from fetal brain tissue. Once fetal brain enhancers and promoters were identified, I intersected their loci with those of the dnSNVs. Instances of overlap are noted in the browser.

To identify overlap of dnSNVs with protein-coding gene promoters I used GENCODE release 19 gene annotations [43]. I defined promoters as the regions within 1,500bp upstream of gene TSS's. In cases where dnSNVs overlapped with those regions, the gene name of the respective TSS is listed in the browser table.

4.3.3 Web application

The *De Novo* Browser can be accessed at

`https://boylelab.shinyapps.io/denovo_app/`

4.4 Results

4.4.1 Browser interface

The *De Novo* browser contains information for 266,865 *de novo* SNVs from 1,917 quad-structure families. Each family consists of one individual diagnosed with ASD, and one unaffected sibling. Variants in the database are listed for the 3,834 probands

and neurotypical siblings, and are restricted to autosomal variants. The variants can be explored in table form and are sortable by a variety of features/annotations (Fig 4.1). Additionally, one may filter the variants to be displayed by combinations of those features.

For each variant entry, the first seven columns of the table contain information on the variant's coordinates (chrom, start end), alleles (ref, alt), and the individual from which the variant was derived (SSC family ID, proband/sibling). The remaining columns for each entry contain a variety of annotations, including functional prediction scores (as described in Methods), evolutionary conservation scores, and overlap with genomic features.

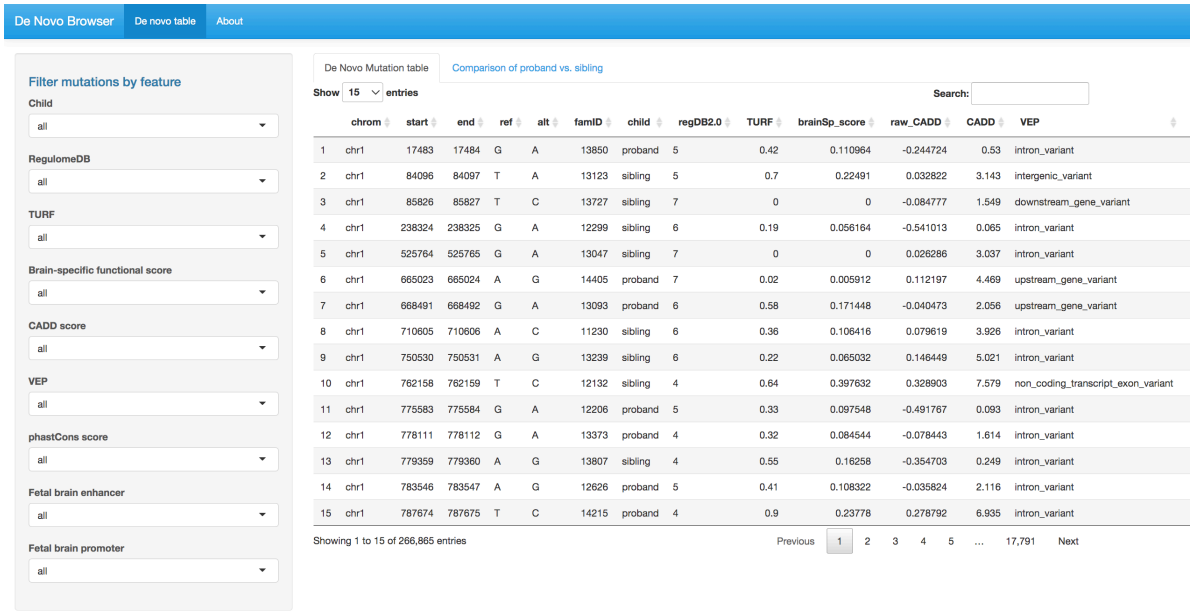

Figure 4.1: Sortable table of all annotated dnSNVs from the Simons Simplex Collection

The browser table can be sorted by each feature column, and there is a search bar provided which can be used to search for specific feature entries, including gene names. By default all 266,865 dnSNVs are included in the table. However, users may choose to filter the table down according to feature options using drop-down

menus on the left side of the table. For example, one may choose to only view dnSNVs pertaining to probands, and choose the "proband" option from the "Child" drop-down menu. Five menus show categorical variables: "Child", "RegulomeDB" and "VEP" (Variant Effect Predictor), "Fetal brain enhancer", and "Fetal brain promoter". All available options for those features can be viewed by clicking on the menu. For the enhancer and promoter menus, the user should select "yes" if they wish to display dnSNVs overlapping with those regions. The remaining options are numerical annotations: "TURF", "Brain-specific functional score", "CADD score", and "phastCons score". For these numerical annotations, different thresholds are provided and users may select to only view dnSNVs scoring in the top quantile of their choosing for the respective annotation. In the case of wishing to apply multiple filters at the same time, options from different drop-down menus can be selected together (Figure 4.2

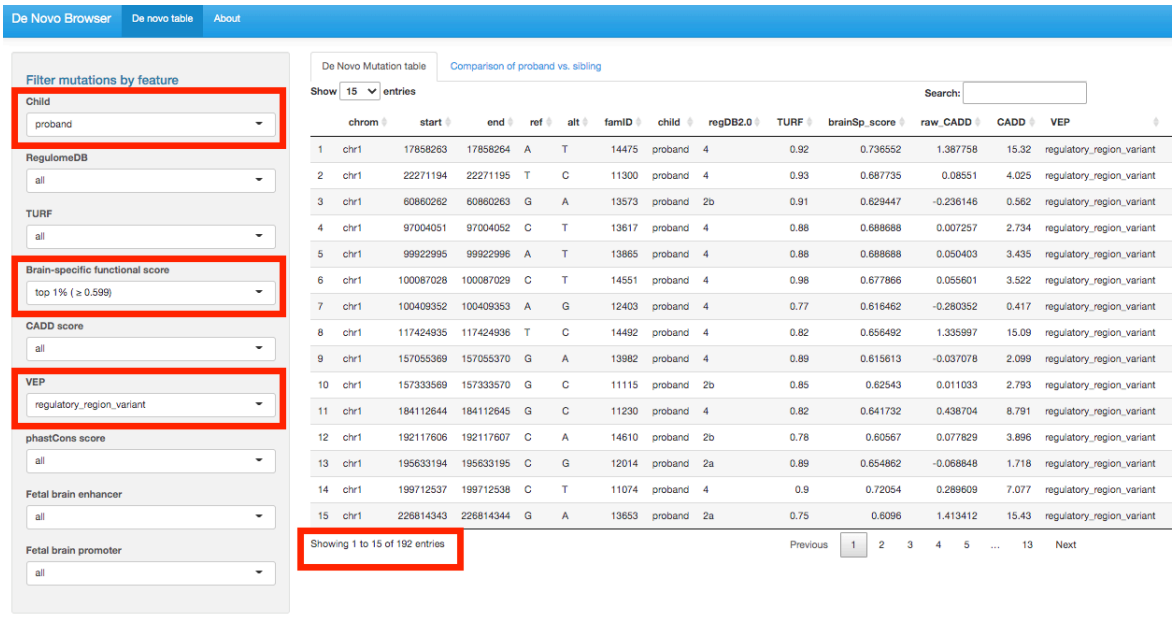

Figure 4.2: Page displaying dnSNVs meeting specific criteria based on the drop-down menu selections. Here we can see we've selected to only view "proband" dnSNVs, scoring in the top 1% of the "Brain-specific functional score", and classified as a "regulatory region variant" by VEP. This reduces the list of dnSNVs to 192 variants of interest.

Another feature of the *De Novo* Browser is the ability to compare counts of dnSNVs in probands vs. unaffected siblings. Clicking on the tab above the table labeled "Comparison of proband vs. sibling" displays a 2x2 matrix of counts based on the filters selected from the drop-down menus (Figure 4.3). In addition to displaying the counts of dnSNVs meeting the selected criteria, the browser also displays the count as a proportion of the total number of dnSNVs in probands and siblings respectively. A Fisher's exact test is also performed to test for enrichment of the selected subset of variants in probands, returning a p-value. When menu selections are changed, counts will automatically be updated and the enrichment test will be performed using the updated counts.

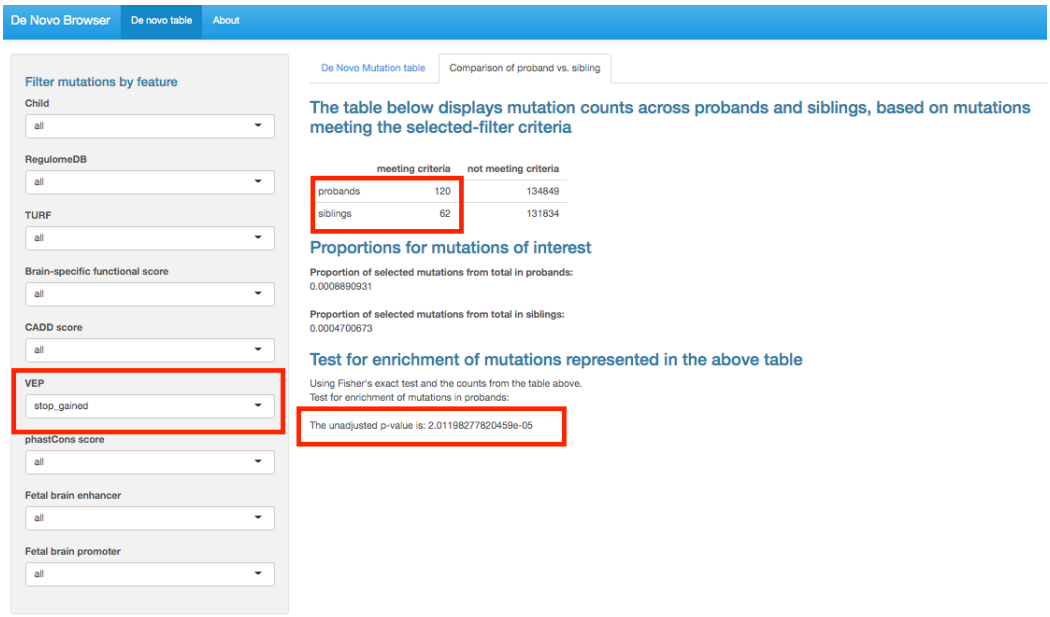

Figure 4.3: Page displaying dnSNV count comparison between probands and sibling based on user-selected filters from the left panel. Results from a Fisher's exact test are also displayed, providing a p-value for enrichment of dnSNVs in probands compared to siblings. In this example, we see a comparison of counts for dnSNVs classified as stop-gains. The matrix tells us there are nearly twice as many of these mutations present in probands compared to siblings (120 vs 62), and the result of the Fisher's exact test tells us this is a statistically significant enrichment in probands.

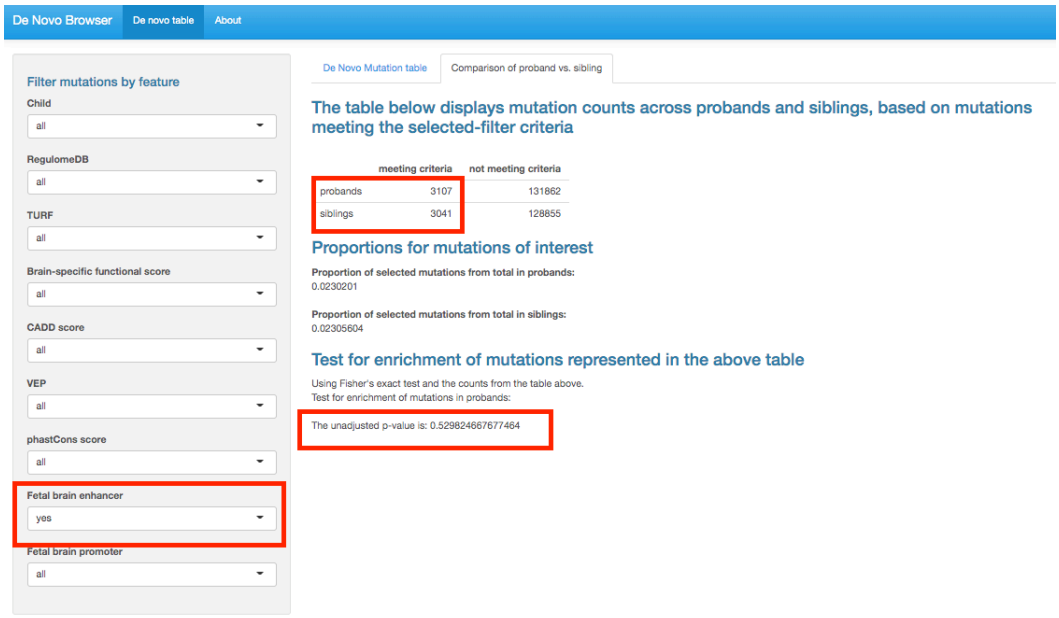

Figure 4.4: Example enrichment test result from *De Novo* Browser. Choosing to view only dnSNVs overlapping with fetal brain enhancer regions, the counts matrix shows us the distribution is fairly even between probands and siblings. The result of the enrichment test confirm there is not a statistically significant enrichment of dnSNVs overlapping fetal brain promoters in probands.

4.5 Discussion

The *De Novo* Browser provides a simple way to explore annotated *de novo* single-nucleotide variants identified in neurotypical individuals as well as individuals diagnosed with ASD. The web application allows users to filter variants by a number of different annotations, including functional scores from the RegulomeDB framework used for identifying variants with gene-regulatory potential. Although the original motivation for curating this list of dnSNVs was to identify variants in noncoding regions, the browser allows for researchers to to view variants that may have importance in other regions of the genome.

My hope is that this resource can save other researchers time and effort by providing a centralized location to access valuable variant data from one of the largest whole-genome sequencing ASD cohorts currently available. This not only can save

others the time of repeating the same work, but also saving the effort of searching through multiple different sources to access data they need. For researchers who have their own data, the *De Novo* Browser can provide additional data with which to make comparisons.

This database can facilitate the discovery of new genetic markers and hopefully help provide insights into the genetic basis of ASD. In addition to helping identify variants potentially associated with ASD, the *De Novo* Browser can help in discovering associated genes. Through this, researchers may gain a better understanding of reasons for the development of ASD, which can potentially improve the ability to diagnose it earlier in a child's life in order to provide them with the appropriate resources

CHAPTER V

Conclusion and Future Direction

5.1 Summary

The focus of this dissertation was to investigate the contribution of genetic variants to ASD. Most estimates for the heritability of ASD fall in the range of 50-90%, supporting a strong genetic basis [48, 121, 140, 12]. Past studies have shown a substantial contribution of common genetic variants, often with additive effects, but studies have failed to show strong evidence of associations from individual variants. Further, most associations from genetic variants have been from coding regions of the genome, despite nearly 90% of phenotype-associated SNPs identified by GWAS being found within noncoding regions [87]. This represents a large portion of the genome which remains to be explored for *de novo* SNVs (dnSNVs) associated with ASD. The work here presents my approach for identifying dnSNVs in an autism cohort and applying several methods to prioritize those with the potential for noncoding regulatory activity. I have tested for associations between the noncoding dnSNVs identified here and ASD-risk. I have also addressed existing challenges in screening for *de novo* noncoding variants contributing to ASD and how they might be overcome in the future.

5.2 Future directions

In Chapter 2, I demonstrated that an enrichment of functional noncoding variants was not detectable in probands compared to siblings. I took several approaches to prioritizing variants in order to isolate those most likely to affect gene regulation, including the incorporation of brain-specific annotations. Despite testing a variety of classification methods, any excess of categories of dnSNVs never maintained statistical significance following multiple-testing correction. In Chapter 3, I explored factors that could be limiting our ability to detect ASD-associated noncoding dnSNVs. Below I outline future changes to the field of ASD genetics that could strengthen future studies and improve our chances of successfully discovering noncoding biomarkers.

5.2.1 Larger autism cohorts

Power analyses revealed that when focusing on this rare subset of noncoding variants, this type of study is unpowered at the sample size of 1,917 families I included here. Considering a desired power threshold of 80%, this study achieved little more than 30% power in the best scenario. To contrast, when testing for enrichment of high-impact coding dnSNVs, the power was estimated to be 97%. Although the necessary sample size to achieve 80% for noncoding dnSNV tests varied, depending on prioritization method, power analyses indicated at least 10,000 families would be required if one were to conduct this study in the same way.

Over the years, ASD cohorts have continued to grow larger. At the time of initiating this work, the Simons Simplex Collection provided the largest collection of whole-genome sequencing for an ASD cohort. Since then, the Simons Powering Autism Research (SPARK) has been established, with the goal of providing genomic data for nearly 72,000 individuals with ASD and their families [132]. While this is

a large number of individuals, WGS data is only available for 3,227 autistic individuals, of the 11,545 total whole-genomes available in SPARK, with the remaining individuals only having WES available. Still, the existence of the SPARK cohort provides encouragement that cohorts will continue to grow in size. As sequencing costs continue to go down and technology continues advancing, it is only a matter of time before cohorts include WGS data for tens of thousands of autistic individuals. This will be an important step toward establishing sufficiently powered studies of rare noncoding variants.

5.2.2 Improvements to noncoding annotations

I showed in Chapter 3 that, although having a large enough sample size is important, improving annotation quality is also necessary for detecting associations between proband dnSNVs and ASD. The difference that annotations can make was evident in the comparison of tissue-agnostic prioritization scores to brain-specific TURF scores. Given the calculated effect sizes of dnSNVs annotated with each method, an estimated 50,000 additional families would be necessary to achieve 80% power when using tissue-agnostic prioritization scores for noncoding dnSNVs, compared to 10,000 families that would be necessary when using brain-specific annotations. This highlights the importance of selecting annotations derived from relevant tissues.

While brain-specific TURF scores were still unable to achieve statistical significance in terms of enrichment of dnSNVs in probands, one study has been successful in detecting significant enrichment of noncoding dnSNVs using a similar scoring model [155]. In their study, Zhou et al. applied their deep-learning-based framework to generate Disease Impact Scores (DIS) to prioritize dnSNVs. Much like the TURF model, their model relies on identifying overlap with functional annotations based on

ENCODE. Using their model, the authors found higher DIS in probands compared to siblings at a statistically significant level (p=.009).

When seeking for the difference between TURF and DIS that could potentially be making the difference between enrichment tests being significant or not significant, one thing that stands out is that DIS is trained on SNPs specifically associated with disease from the Human Gene Mutation Database [134]. This is an indication that while TURF is successful at prioritizing regulatory variants in a more general manner, perhaps the additional data from disease-associated SNPs that is being incorporated into DIS is important when seeking associations to ASD.

Altogether, it is evident that as more relevant data is included in future prioritization methods, be it tissue-specific annotations or known disease-related variants, annotations will become increasingly effective at separating out truly associated variants.

5.2.3 Combining data sets and studying multiple classes of variants together

In Chapter 2 I gave an overview of the discovery of different classes of variants associated with ASD. With a better understanding of common variants, and variants in coding regions, an existing gap in our knowledge remains in understanding the role of rare noncoding variants in ASD. This was the motivation for restricting the work done here to noncoding dnSNVs. However, while it is the case that there are some rare mutations of large effect driving ASD phenotypes, it has also been shown that a large majority of genetic variants associated with ASD are common variants with additive effects. It has been estimated that the combined effects of many common variants working together could account for almost 50% of the genetic basis of ASD [48]. Therefore, restricting analyses to one specific type of rare variation, likely leaves a study susceptible to missing many other driving genetic factors.

Following this logic, in 2022, researchers used data from the SPARK cohort to identify *de novo* coding variants as well as rare inherited loss of function variants [157]. By integrating both classes of variants together, the authors were able to identify 60 ASD-associated genes, five of which had not been previously reported. In the same vein, another study conducted during the same year identified 74 risk genes by combining *de novo* and inherited variants in addition to incorporating variants from previous studies [24]. Seven of the genes they identified had not been previously reported in any study.

Studies of several individual classes of genetic variation have bolstered our understanding of the genetic heterogeneity of ASD. Although we cannot yet fully understand even the individual contribution of each class of variant, it is clear that many different types of variation are simultaneously involved. As more variant lists are published and made publicly available, it will allow researchers to study their effects in a combined manner and detect enrichments that may have been otherwise missed.

5.3 Concluding Remarks

In this dissertation I have established a pipeline for the detection of *de novo* SNVs using variant calls derived from whole-genome sequence data from the Simons Simplex Collection. The set of variants I curated revealed an enrichment of protein-coding dnSNVs predicted to lead to loss-of-function. I annotated and classified these dnSNVs with the goal of identifying a subset of noncoding dnSNVs associated with ASD. Ultimately, no enrichments of dnSNVs in probands within noncoding categories were detectable following multiple-testing corrections. My examination into what factors may be preventing the detection of noncoding enrichment indicate that, in addition to larger sample sizes, more precise tissue-specific annotations would be

necessary. I have developed a web application, the *De Novo* Browser, to share some of the work I have done with the research community in an effort to facilitate access to relevant data for future studies. As ASD cohorts continue to grow in size and autism classification methods and noncoding functional annotations improve, further progress in our understanding of the true effects of noncoding *de novo* SNVs can be made. It is my hope that the work I have presented in this dissertation can shed some light on the current state of research in the genetic basis of ASD and contribute to the path forward.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Perinatal factors and the development of autism: a population study, 2004.

[2] 1000 Genomes Project Consortium, Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010.

[3] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[4] Rocio Acuna-Hidalgo, Joris A Veltman, and Alexander Hoischen. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.*, 17(1):241, November 2016.

[5] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, Chapter 7:Unit7.20, January 2013.

[6] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2, 2013.

[7] Marina A Afanasyeva, Lidia V Putlyaeva, Denis E Demin, Ivan V Kulakovskiy, Ilya E Vorontsov, Marina V Fridman, Vsevolod J Makeev, Dmitry V Kuprash, and Anton M Schwartz. The single nucleotide variant rs12722489 determines differential estrogen receptor binding and enhancer properties of an IL2RA intronic region. *PLoS One*, 12(2):e0172681, February 2017.

[8] American Psychiatric Association. *DSM-5 Classification*. American Psychiatric Publishing, August 2015.

[9] American Psychiatric Association Staff. *Diagnostic and Statistical Manual of Mental Disorders (DSM-III)*. 1980.

[10] Joon-Yong An, Kevin Lin, Lingxue Zhu, Donna M Werling, Shan Dong, Harrison Brand, Harold Z Wang, Xuefang Zhao, Grace B Schwartz, Ryan L Collins, Benjamin B Currall, Claudia Dastmalchi, Jeanselle Dea, Clif Duhn, Michael C Gilson, Lambertus Klei, Lindsay Liang, Eirene Markenscoff-Papadimitriou, Sirisha Pochareddy, Nadav Ahituv, Joseph D Buxbaum, Hilary Coon, Mark J Daly, Young Shin Kim, Gabor T Marth, Benjamin M Neale, Aaron R Quinlan, John L Rubenstein, Nenad Sestan, Matthew W State, A Jeremy Willsey, Michael E Talkowski, Bernie Devlin, Kathryn Roeder, and Stephan J Sanders. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science*, 362(6420), December 2018.

[11] Dan Bai, Benjamin Hon Kei Yip, Gayle C Windham, Andre Sourander, Richard Francis, Rinat Yoffe, Emma Glasson, Behrang Mahjani, Auli Suominen, Helen Leonard, Mika Gissler, Joseph D Buxbaum, Kingsley Wong, Diana Schendel, Arad Kodesh, Michaeline Breshnahan, Stephen Z Levine, Erik T Parner, Stefan N Hansen, Christina Hultman, Abraham Reichenberg, and Sven Sandin. Association of genetic and environmental factors with autism in a 5-country cohort, 2019.

[12] Dan Bai, Benjamin Hon Kei Yip, Gayle C Windham, Andre Sourander, Richard Francis, Rinat Yoffe, Emma Glasson, Behrang Mahjani, Auli Suominen, Helen Leonard, Mika Gissler, Joseph D Buxbaum, Kingsley Wong, Diana Schendel, Arad Kodesh, Michaeline Breshnahan, Stephen Z Levine, Erik T Parner, Stefan N Hansen, Christina Hultman, Abraham Reichenberg, and Sven Sandin. Association of genetic and environmental factors with autism in a 5-country cohort. *JAMA Psychiatry*, 76(10):1035–1043, October 2019.

[13] A Bailey, A Le Couteur, I Gottesman, P Bolton, E Simonoff, E Yuzda, and M Rutter. Autism as a strongly genetic disorder: evidence from a british twin study. *Psychol. Med.*, 25(1):63–77, January 1995.

[14] Matthew N Bainbridge, Min Wang, Yuanqing Wu, Irene Newsham, Donna M Muzny, John L Jefferies, Thomas J Albert, Daniel L Burgess, and Richard A Gibbs. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol.*, 12(7):R68, July 2011.

[15] G Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, 27(2):573–580, January 1999.

[16] E M Blackwood and J T Kadonaga. Going the distance: a current view of enhancer action. *Science*, 281(5373):60–63, July 1998.

[17] Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, January 2008.

[18] Alan P Boyle, Eurie L Hong, Manoj Hariharan, Yong Cheng, Marc A Schaub, Maya Kasowski, Konrad J Karczewski, Julie Park, Benjamin C Hitz, Shuai Weng, J Michael Cherry, and Michael Snyder. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, 22(9):1790–1797, September 2012.

[19] Broad Institute. Picard toolkit, 2019.

[20] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L Whetzel, Ridwan Amode, Jose A Guillen, Harpreet S Riat, Stephen J Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, 47(D1):D1005–D1012, January 2019.

[21] Ryan K C Yuen, Daniele Merico, Matt Bookman, Jennifer L Howe, Bhooma Thiruvahindrapuram, Rohan V Patel, Joe Whitney, Nicole Deflaux, Jonathan Bingham, Zhuozhi Wang, Giovanna Pellecchia, Janet A Buchanan, Susan Walker, Christian R Marshall, Mohammed Uddin, Mehdi Zarrei, Eric Deneault, Lia D'Abate, Ada J S Chan, Stephanie Koyanagi, Tara Paton, Sergio L Pereira, Ny Hoang, Worrawat Engchuan, Edward J Higginbotham, Karen Ho, Sylvia Lamoureux, Weili Li, Jeffrey R MacDonald, Thomas Nalpathamkalam, Wilson W L Sung, Fiona J Tsoi, John Wei, Lizhen Xu, Anne-Marie Tasse, Emily Kirby, William Van Etten, Simon Twigger, Wendy Roberts, Irene Drmic, Sanne Jilderda, Bonnie Mackinnon Modi, Barbara Kellam, Michael Szego, Cheryl Cytrynbaum, Rosanna Weksberg, Lonnie Zwaigenbaum,

Marc Woodbury-Smith, Jessica Brian, Lili Senman, Alana Iaboni, Krissy Doyle-Thomas, Ann Thompson, Christina Chrysler, Jonathan Leef, Tal Savion-Lemieux, Isabel M Smith, Xudong Liu, Rob Nicolson, Vicki Seifer, Angie Fedele, Edwin H Cook, Stephen Dager, Annette Estes, Louise Gallagher, Beth A Malow, Jeremy R Parr, Sarah J Spence, Jacob Vorstman, Brendan J Frey, James T Robinson, Lisa J Strug, Bridget A Fernandez, Mayada Elsabbagh, Melissa T Carter, Joachim Hallmayer, Bartha M Knoppers, Evdokia Anagnostou, Peter Szatmari, Robert H Ring, David Glazer, Mathew T Pletcher, and Stephen W Scherer. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.*, 20(4):602–611, April 2017.

[22] Michael Carey, Michael F Carey, and Stephen T Smale. *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques*. CSHL Press, 2000.

[23] Christopher P Castro, Adam G Diehl, and Alan P Boyle. Challenges in screening for*de novo*noncoding variants contributing to genetically complex phenotypes.

[24] Timothy S Chang, Matilde Cirnigliaro, Stephanie A Arteaga, Laura Pérez-Cano, Elizabeth K Ruzzo, Aaron Gordon, Lucy Bicks, Jae-Yoon Jung, Jennifer K Lowe, Dennis P Wall, and Daniel H Geschwind. The contributions of rare inherited and polygenic risk to ASD in multiplex families.

[25] Jiyeon Choi, Tongwu Zhang, Andrew Vu, Julien Ablain, Matthew M Makowski, Leandro M Colli, Mai Xu, Rebecca C Hennessey, Jinhu Yin, Harriet Rothschild, Cathrin Gräwe, Michael A Kovacs, Karen M Funderburk, Myriam Brossard, John Taylor, Bogdan Pasaniuc, Raj Chari, Stephen J Chanock, Clive J Hoggart, Florence Demenais, Jennifer H Barrett, Matthew H Law, Mark M Iles, Kai Yu, Michiel Vermeulen, Leonard I Zon, and Kevin M Brown. Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat. Commun.*, 11(1):2718, June 2020.

[26] The Encode Project Consortium and The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome, 2012.

[27] F H Crick. On protein synthesis. *Symp. Soc. Exp. Biol.*, 12:138–163, 1958.

[28] Fiona Cunningham, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, Andrew Berry, Jyothish Bhai, Alexandra Bignell, Konstantinos Billis, Sanjay Boddu, Lucy Brooks, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Jose Gonzalez Martinez, Cristina Guijarro-Clarke, Arthur Gymer, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, José Carlos Marugán, Shamika Mohanan, Aleena Mushtaq, Marc Naven, Denye N Ogeh, Anne Parker, Andrew Parton, Malcolm Perry, Ivana Piližota, Irina Prosovetskaia, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, José G Pérez-Silva, William Stark, Emily Steed, Kyösti Sutinen, Ranjit Sukumaran, Dulika Sumathipala, Marie-Marthe Suner, Michal Szpak, Anja Thormann, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Brandon Walts, Natalie Willhoft, Andrea Winterbottom, Elizabeth Wass, Marc Chakiachvili, Bethany Flint, Adam Frankish, Stefano Giorgetti, Leanne Haggerty, Sarah E Hunt, Garth R IIsley, Jane E Loveland, Fergal J Martin, Benjamin Moore, Jonathan M Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J Trevanion, Sarah Dyer, Peter W Harrison, Kevin L Howe, Andrew D Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2022. *Nucleic Acids Res.*, 50(D1):D988–D995, January 2022.

[29] G Das, D Henning, D Wright, and R Reddy. Upstream regulatory elements are necessary and sufficient for transcription of a U6 RNA gene by RNA polymerase III. *EMBO J.*, 7(2):503–512, February 1988.

[30] Arina O Degtyareva, Elena V Antontseva, and Tatiana I Merkulova. Regulatory SNPs: Altered transcription factor binding sites implicated in complex traits and diseases. *Int. J. Mol. Sci.*, 22(12), June 2021.

[31] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43(5):491–498, May 2011.

[32] Hirokazu Doi, Takashi X Fujisawa, Ryoichiro Iwanaga, Junko Matsuzaki, Chisato Kawasaki, Mamoru Tochigi, Tsukasa Sasaki, Nobumasa Kato, and Kazuyuki Shinohara. Association between single nucleotide polymorphisms in estrogen receptor 1/2 genes and symptomatic severity of autism spectrum disorder. *Res. Dev. Disabil.*, 82:20–26, November 2018.

[33] Shengcheng Dong and Alan P Boyle. Predicting functional variants in enhancer and promoter elements using RegulomeDB. *Hum. Mutat.*, 40(9):1292–1298, September 2019.

[34] Shengcheng Dong and Alan P Boyle. Predicting functional variants in enhancer and promoter elements using RegulomeDB, 2019.

[35] Shengcheng Dong and Alan P Boyle. Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome. *Nucleic Acids Res.*, 50(1):e6, January 2022.

[36] Abolfazl Doostparast Torshizi, Iuliana Ionita-Laza, and Kai Wang. Cell Type-Specific annotation and fine mapping of variants associated with brain disorders. *Front. Genet.*, 11:575928, December 2020.

[37] Stacey L Edwards, Jonathan Beesley, Juliet D French, and Alison M Dunning. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.*, 93(5):779–797, November 2013.

[38] Spencer C Evans, Andrea D Boan, Catherine Bradley, and Laura A Carpenter. Sex/Gender differences in screening for autism spectrum disorder: Implications for Evidence-Based assessment. *J. Clin. Child Adolesc. Psychol.*, 48(6):840–854, 2019.

[39] Gary Felsenfeld and Mark Groudine. Controlling the double helix, 2003.

[40] Gerald D Fischbach and Catherine Lord. The simons simplex collection: a resource for identification of autism genetic risk factors. *Neuron*, 68(2):192–195, October 2010.

[41] S Folstein and M Rutter. Infantile autism: a genetic study of 21 twin pairs. *J. Child Psychol. Psychiatry*, 18(4):297–321, September 1977.

[42] Laurent C Francioli, Paz P Polak, Amnon Koren, Androniki Menelaou, Sung Chun, Ivo Renkens, Genome of the Netherlands Consortium, Cornelia M van Duijn, Morris Swertz, Cisca Wijmenga, Gertjan van Ommen, P Eline Slagboom, Dorret I Boomsma, Kai Ye, Victor Guryev, Peter F Arndt, Wigard P Kloosterman, Paul I W de Bakker, and Shamil R Sunyaev. Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.*, 47(7):822–826, July 2015.

[43] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu,

Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, 47(D1):D766–D773, January 2019.

[44] Thomas W Frazier, Lee Thompson, Eric A Youngstrom, Paul Law, Antonio Y Hardan, Charis Eng, and Nathan Morris. A twin study of heritable and shared environmental contributions to autism. *J. Autism Dev. Disord.*, 44(8):2013–2025, August 2014.

[45] F Fromm-Reichmann. Notes on the development of treatment of schizophrenics by psychoanalytic psychotherapy. *Psychiatry*, 11(3):263–273, August 1948.

[46] Terrence S Furey. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions, 2012.

[47] D J Galas and A Schmitz. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, 5(9):3157–3170, September 1978.

[48] Trent Gaugler, Lambertus Klei, Stephan J Sanders, Corneliu A Bodea, Arthur P Goldberg, Ann B Lee, Milind Mahajan, Dina Manaa, Yudi Pawitan, Jennifer Reichert, Stephan Ripke, Sven Sandin, Pamela Sklar, Oscar Svantesson, Abraham Reichenberg, Christina M Hultman, Bernie Devlin, Kathryn Roeder, and Joseph D Buxbaum. Most genetic risk for autism resides with common variation. *Nat. Genet.*, 46(8):881–885, August 2014.

[49] Sarah R Gilman, Ivan Iossifov, Dan Levy, Michael Ronemus, Michael Wigler, and Dennis Vitkup. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron*, 70(5):898–907, June 2011.

[50] David S Gross and William T Garrard. NUCLEASE HYPERSENSITIVE SITES IN CHROMATIN, 1988.

[51] Yuanfang Guan, Dmitriy Gorenshteyn, Margit Burmeister, Aaron K Wong, John C Schimenti, Mary Ann Handel, Carol J Bult, Matthew A Hibbs, and Olga G Troyanskaya. Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Comput. Biol.*, 8(9):e1002694, September 2012.

[52] Dennis J Hazelett, Suhn Kyong Rhie, Malaina Gaddis, Chunli Yan, Daniel L Lakeland, Simon G Coetzee, Ellipse/GAME-ON consortium, Practical consortium, Brian E Henderson, Houtan Noushmehr, Wendy Cozen, Zsofia Kote-Jarai, Rosalind A Eeles, Douglas F Easton, Christopher A Haiman, Wange Lu, Peggy J Farnham, and Gerhard A Coetzee. Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet.*, 10(1):e1004102, January 2014.

[53] Barbara Hrdlickova, Rodrigo Coutinho de Almeida, Zuzanna Borek, and Sebo Withoff. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim. Biophys. Acta*, 1842(10):1910–1922, October 2014.

[54] A John Iafrate, A John Iafrate, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer, and Charles Lee. Detection of large-scale variation in the human genome, 2004.

[55] Ivan Iossifov, Brian J O'Roak, Stephan J Sanders, Michael Ronemus, Niklas Krumm, Dan Levy, Holly A Stessman, Kali T Witherspoon, Laura Vives, Karynne E Patterson, Joshua D Smith, Bryan Paeper, Deborah A Nickerson, Jeanselle Dea, Shan Dong, Luis E Gonzalez, Jeffrey D Mandell, Shrikant M Mane, Michael T Murtha, Catherine A Sullivan, Michael F

Walker, Zainulabedin Waqar, Liping Wei, A Jeremy Willsey, Boris Yamrom, Yoon-Ha Lee, Ewa Grabowska, Ertugrul Dalkic, Zihua Wang, Steven Marks, Peter Andrews, Anthony Leotta, Jude Kendall, Inessa Hakker, Julie Rosenbaum, Beicong Ma, Linda Rodgers, Jennifer Troge, Giuseppe Narzisi, Seungtai Yoon, Michael C Schatz, Kenny Ye, W Richard McCombie, Jay Shendure, Evan E Eichler, Matthew W State, and Michael Wigler. The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515(7526):216–221, November 2014.

[56] Ivan Iossifov, Michael Ronemus, Dan Levy, Zihua Wang, Inessa Hakker, Julie Rosenbaum, Boris Yamrom, Yoon-Ha Lee, Giuseppe Narzisi, Anthony Leotta, Jude Kendall, Ewa Grabowska, Beicong Ma, Steven Marks, Linda Rodgers, Asya Stepansky, Jennifer Troge, Peter Andrews, Mitchell Bekritsky, Kith Pradhan, Elena Ghiban, Melissa Kramer, Jennifer Parla, Ryan Demeter, Lucinda L Fulton, Robert S Fulton, Vincent J Magrini, Kenny Ye, Jennifer C Darnell, Robert B Darnell, Elaine R Mardis, Richard K Wilson, Michael C Schatz, W Richard McCombie, and Michael Wigler. De novo gene disruptions in children on the autistic spectrum, 2012.

[57] M-L Jacquemont, D Sanlaville, R Redon, O Raoul, V Cormier-Daire, S Lyonnet, J Amiel, M Le Merrer, D Heron, M-C de Blois, M Prieur, M Vekemans, N P Carter, A Munnich, L Colleaux, and A Philippe. Array-based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. *J. Med. Genet.*, 43(11):843–849, November 2006.

[58] Hákon Jónsson, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdottir, Florian Zink, Eirikur Hjartarson, Marteinn T Hardarson, Kristjan E Hjorleifsson, Hannes P Eggertsson, Sigurjon Axel Gudjonsson, Lucas D Ward, Gudny A Arnadottir, Einar A Helgason, Hannes Helgason, Arnaldur Gylfason, Adalbjorg Jonasdottir, Aslaug Jonasdottir, Thorunn Rafnar, Mike Frigge, Simon N Stacey, Olafur Th Magnusson, Unnur Thorsteinsdottir, Gisli Masson, Augustine Kong, Bjarni V Halldorsson, Agnar Helgason, Daniel F Gudbjartsson, and Kari Stefansson. Parental influence on human germline de novo mutations in 1,548 trios from iceland. *Nature*, 549(7673):519–522, September 2017.

[59] Inkyung Jung, Anthony Schmitt, Yarui Diao, Andrew J Lee, Tristin Liu, Dongchan Yang, Catherine Tan, Junghyun Eom, Marilynn Chan, Sora Chee, Zachary Chiang, Changyoun Kim, Eliezer Masliah, Cathy L Barr, Bin Li, Samantha Kuan, Dongsup Kim, and Bing Ren. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.*, 51(10):1442–1449, October 2019.

[60] L Kanner. Autistic disturbances of affective contact. *Acta Paedopsychiatr.*, 35(4):100–136, 1968.

[61] Mojgan Karahmadi, Padideh Karimi, Elahe Kamali, and Seyyedmohammad Mousavi. Environmental factors influencing the risk of autism, 2017.

[62] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, Laura D Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M England, Eleanor G Seaby, Jack A Kosmicki, Raymond K Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X Chong, Kaitlin E Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S Ware, Christopher Vittal, Irina M Armean, Louis Bergelson, Kristian Cibulskis, Kristen M Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade,

Michael E Talkowski, Genome Aggregation Database Consortium, Benjamin M Neale, Mark J Daly, and Daniel G MacArthur. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, May 2020.

[63] M Kasowski, F Grubert, C Heffelfinger, M Hariharan, A Asabere, and S M Waszak. Variation in transcription factor binding among humans, 2010.

[64] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, June 2002.

[65] Michael D Kessler, Douglas P Loesch, James A Perry, Nancy L Heard-Costa, Daniel Taliun, Brian E Cade, Heming Wang, Michelle Daya, John Ziniti, Soma Datta, Juan C Celedón, Manuel E Soto-Quiros, Lydiana Avila, Scott T Weiss, Kathleen Barnes, Susan S Redline, Ramachandran S Vasan, Andrew D Johnson, Rasika A Mathias, Ryan Hernandez, James G Wilson, Deborah A Nickerson, Goncalo Abecasis, Sharon R Browning, Sebastian Zöllner, Jeffrey R O'Connell, Braxton D Mitchell, National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Population Genetics Working Group, and Timothy D O'Connor. De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the amish founder population. *Proc. Natl. Acad. Sci. U. S. A.*, 117(5):2560–2569, February 2020.

[66] Aziz Khan and Anthony Mathelier. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets, 2017.

[67] Augustine Kong, Michael L Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Wendy S W Wong, Gunnar Sigurdsson, G Bragi Walters, Stacy Steinberg, Hannes Helgason, Gudmar Thorleifsson, Daniel F Gudbjartsson, Agnar Helgason, Olafur Th Magnusson, Unnur Thorsteinsdottir, and Kari Stefansson. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–475, August 2012.

[68] Kirill V Korneev, Ekaterina N Sviriaeva, Nikita A Mitkin, Alisa M Gorbacheva, Aksinya N Uvarova, Alina S Ustiugova, Oleg L Polanovsky, Ivan V Kulakovskiy, Marina A Afanasyeva, Anton M Schwartz, and Dmitry V Kuprash. Minor C allele of the SNP rs7873784 associated with rheumatoid arthritis and type-2 diabetes mellitus binds PU.1 and enhances TLR4 expression, 2020.

[69] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock,

H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[70] M Lauritsen, O Mors, P B Mortensen, and H Ewald. Infantile autism and associated autosomal chromosome abnormalities: a register-based study and a literature survey. *J. Child Psychol. Psychiatry*, 40(3):335–345, March 1999.

[71] A Le Couteur, A Bailey, S Goode, A Pickles, S Robertson, I Gottesman, and M Rutter. A broader phenotype of autism: the clinical spectrum in twins. *J. Child Psychol. Psychiatry*, 37(7):785–801, October 1996.

[72] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, Taru Tukiainen, Daniel P Birnbaum, Jack A Kosmicki, Laramie E Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M Peloso, Ryan Poplin, Manuel A Rivas, Valentin Ruano-Rubio, Samuel A Rose, Douglas M Ruderfer, Khalid Shakir, Peter D Stenson, Christine Stevens, Brett P Thomas, Grace Tiao, Maria T Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C Florez, Stacey B Gabriel, Gad Getz, Stephen J Glatt, Christina M Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M Neale, Aarno Palotie, Shaun M Purcell, Danish Saleheen, Jeremiah M Scharf, Pamela Sklar, Patrick F Sullivan, Jaakko Tuomilehto, Ming T Tsuang, Hugh C Watkins, James G Wilson, Mark J Daly, Daniel G MacArthur, and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, August 2016.

[73] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, 20(12):1983–1992, December 2014.

[74] Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, October 2014.

[75] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, March 2010.

[76] W Ian Lipkin, W Ian Lipkin, Michaeline Bresnahan, and Ezra Susser. Cohort-guided insights into gene–environment interactions in autism spectrum disorders, 2023.

[77] Rachel Loomes, Laura Hull, and William Polmear Locke Mandy. What is the Male-to-Female ratio in autism spectrum disorder? a systematic review and Meta-Analysis. *J. Am. Acad. Child Adolesc. Psychiatry*, 56(6):466–474, June 2017.

[78] C Lord, S Risi, L Lambrecht, E H Cook, Jr, B L Leventhal, P C DiLavore, A Pickles, and M Rutter. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.*, 30(3):205–223, June 2000.

[79] C Lord, M Rutter, and A Le Couteur. Autism diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.*, 24(5):659–685, October 1994.

[80] C M Low, R K Olsen, and M J Waring. Sequence preferences in the binding to DNA of triostin a and TANDEM as reported by DNase I footprinting. *FEBS Lett.*, 176(2):414–420, October 1984.

[81] Qiongshi Lu, Ryan L Powles, Sarah Abdallah, Derek Ou, Qian Wang, Yiming Hu, Yisi Lu, Wei Liu, Boyang Li, Shubhabrata Mukherjee, Paul K Crane, and Hongyu Zhao. Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset alzheimer's disease, 2017.

[82] Qiongshi Lu, Ryan Lee Powles, Qian Wang, Beixin Julie He, and Hongyu Zhao. Integrative Tissue-Specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS Genet.*, 12(4):e1005947, April 2016.

[83] Matthew J Maenner, Kelly A Shaw, Amanda V Bakian, Deborah A Bilder, Maureen S Durkin, Amy Esler, Sarah M Furnier, Libby Hallas, Jennifer Hall-Lande, Allison Hudson, Michelle M Hughes, Mary Patrick, Karen Pierce, Jenny N Poynter, Angelica Salinas, Josephine Shenouda, Alison Vehorn, Zachary Warren, John N Constantino, Monica DiRienzo, Robert T Fitzgerald, Andrea Grzybowski, Margaret H Spivey, Sydney Pettygrove, Walter Zahorodny, Akilah Ali, Jennifer G Andrews, Thaer Baroud, Johanna Gutierrez, Amy Hewitt, Li-Ching Lee, Maya Lopez, Kristen Clancy Mancilla, Dedria McArthur, Yvette D Schwenk, Anita Washington, Susan Williams, and Mary E Cogswell. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, united states, 2018. *MMWR Surveill. Summ.*, 70(11):1–16, December 2021.

[84] Paul Maloret. Autism: A social and medical history waltz mitzi autism: A social and medical history200pp £50 palgrave macmillan 9780230527508 0230527507, 2014.

[85] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[86] Glenn A Maston, Sara K Evans, and Michael R Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.

[87] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph,

Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, September 2012.

[88] J D McGhee, W I Wood, M Dolan, J D Engel, and G Felsenfeld. A 200 base pair region at the 5' end of the chicken adult beta-globin gene is accessible to nuclease digestion. *Cell*, 27(1 Pt 2):45–55, November 1981.

[89] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor.

[90] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biol.*, 17(1):122, June 2016.

[91] Jacob J Michaelson, Yujian Shi, Madhusudan Gujral, Hancheng Zheng, Dheeraj Malhotra, Xin Jin, Minghan Jian, Guangming Liu, Douglas Greer, Abhishek Bhandari, Wenting Wu, Roser Corominas, Aine Peoples, Amnon Koren, Athurva Gore, Shuli Kang, Guan Ning Lin, Jasper Estabillo, Therese Gadomski, Balvindar Singh, Kun Zhang, Natacha Akshoomoff, Christina Corsello, Steven McCarroll, Lilia M Iakoucheva, Yingrui Li, Jun Wang, and Jonathan Sebat. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, 151(7):1431–1442, December 2012.

[92] Ryan E Mills, Klaudia Walter, Chip Stewart, Robert E Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtai Chris Yoon, Kai Ye, R Keira Cheetham, Asif Chinwalla, Donald F Conrad, Yutao Fu, Fabian Grubert, Iman Hajirasouliha, Fereydoun Hormozdiari, Lilia M Iakoucheva, Zamin Iqbal, Shuli Kang, Jeffrey M Kidd, Miriam K Konkel, Joshua Korn, Ekta Khurana, Deniz Kural, Hugo Y K Lam, Jing Leng, Ruiqiang Li, Yingrui Li, Chang-Yun Lin, Ruibang Luo, Xinmeng Jasmine Mu, James Nemesh, Heather E Peckham, Tobias Rausch, Aylwyn Scally, Xinghua Shi, Michael P Stromberg, Adrian M Stütz, Alexander Eckehart Urban, Jerilyn A Walker, Jiantao Wu, Yujun Zhang, Zhengdong D Zhang, Mark A Batzer, Li Ding, Gabor T Marth, Gil McVean, Jonathan Sebat, Michael Snyder, Jun Wang, Kenny Ye, Evan E Eichler, Mark B Gerstein, Matthew E Hurles, Charles Lee, Steven A McCarroll, Jan O Korbel, and 1000 Genomes Project. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, February 2011.

[93] Aashiq H Mirza, Simranjeet Kaur, Caroline A Brorsson, and Flemming Pociot. Effects of GWAS-associated genetic variants on lncRNAs within IBD and T1D candidate loci. *PLoS One*, 9(8):e105723, August 2014.

[94] Sam B Morgan. The autistic child and family functioning: A developmental-family systems perspective, 1988.

[95] Benjamin M Neale, Yan Kou, Li Liu, Avi Ma'ayan, Kaitlin E Samocha, Aniko Sabo, Chiao-Feng Lin, Christine Stevens, Li-San Wang, Vladimir Makarov, Paz Polak, Seungtai Yoon, Jared Maguire, Emily L Crawford, Nicholas G Campbell, Evan T Geller, Otto Valladares, Chad Schafer, Han Liu, Tuo Zhao, Guiqing Cai, Jayon Lihm, Ruth Dannenfelser, Omar Jabado, Zuleyma Peralta, Uma Nagaswamy, Donna Muzny, Jeffrey G Reid, Irene Newsham, Yuanqing Wu, Lora Lewis, Yi Han, Benjamin F Voight, Elaine Lim, Elizabeth Rossin, Andrew Kirby, Jason Flannick, Menachem Fromer, Khalid Shakir, Tim Fennell, Kiran Garimella, Eric Banks, Ryan Poplin, Stacey Gabriel, Mark DePristo, Jack R Wimbish, Braden E Boone, Shawn E Levy, Catalina Betancur, Shamil Sunyaev, Eric Boerwinkle, Joseph D Buxbaum, Edwin H Cook, Jr, Bernie Devlin, Richard A Gibbs, Kathryn Roeder, Gerard D Schellenberg, James S Sutcliffe, and Mark J Daly. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–245, April 2012.

[96] Pauline C Ng and Steven Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, 31(13):3812–3814, July 2003.

[97] Sierra S Nishizaki, Natalie Ng, Shengcheng Dong, Robert S Porter, Cody Morterud, Colten Williams, Courtney Asman, Jessica A Switzenberg, and Alan P Boyle. Predicting the effects of SNPs on transcription factor binding affinity. *Bioinformatics*, 36(2):364–372, January 2020.

[98] Wanda K O'Neal and Michael R Knowles. Cystic fibrosis disease modifiers: Complex genetics defines the phenotypic diversity in a monogenic disease, 2018.

[99] Cathal Ormond, Niamh M Ryan, Aiden Corvin, and Elizabeth A Heron. Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Brief. Bioinform.*, 22(5), September 2021.

[100] Brian J O'Roak, Pelagia Deriziotis, Choli Lee, Laura Vives, Jerrod J Schwartz, Santhosh Girirajan, Emre Karakoc, Alexandra P Mackenzie, Sarah B Ng, Carl Baker, Mark J Rieder, Deborah A Nickerson, Raphael Bernier, Simon E Fisher, Jay Shendure, and Evan E Eichler. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.*, 43(6):585–589, June 2011.

[101] Brian J O'Roak, Laura Vives, Santhosh Girirajan, Emre Karakoc, Niklas Krumm, Bradley P Coe, Roie Levy, Arthur Ko, Choli Lee, Joshua D Smith, Emily H Turner, Ian B Stanaway, Benjamin Vernot, Maika Malig, Carl Baker, Beau Reilly, Joshua M Akey, Elhanan Borenstein, Mark J Rieder, Deborah A Nickerson, Raphael Bernier, Jay Shendure, and Evan E Eichler. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246–250, April 2012.

[102] Evin M Padhi, Tristan J Hayeck, Brandon Mannion, Sumantra Chatterjee, Marta Byrska-Bishop, Rajeeva Musunuri, Giuseppe Narzisi, Avinash Abhyankar, Zhang Cheng, Riana D Hunter, Jennifer Akiyama, Lauren E Fries, Jeffrey Ng, Nick Stong, Andrew S Allen, Diane E Dickel, Raphael A Bernier, David U Gorkin, Len A Pennacchio, Michael C Zody, and Tychele N Turner. De novo mutation in an enhancer of EBF3 in simplex autism.

[103] Biswajit Padhy, Bushra Hayat, Gargi Gouranga Nanda, Pranjya Paramita Mohanty, and Debasmita Pankaj Alone. Pseudoexfoliation and alzheimer's associated CLU risk variant, rs2279590, lies within an enhancer element and regulates CLU, EPHX2 and PTK2B gene expression, 2017.

[104] Seung Hwan Paik, Hyun-Jin Kim, Ho-Young Son, Seungbok Lee, Sun-Wha Im, Young Seok Ju, Je Ho Yeon, Seong Jin Jo, Hee Chul Eun, Jeong-Sun Seo, Oh Sang Kwon, and Jong-Il Kim. Gene mapping study for constitutive skin color in an isolated mongolian population, 2012.

[105] Lipika R Pal and John Moult. Genetic basis of common human disease: Insight into the role of missense SNPs from Genome-Wide association studies. *J. Mol. Biol.*, 427(13):2271–2289, July 2015.

[106] Brent S Pedersen, Joe M Brown, Harriet Dashnow, Amelia D Wallace, Matt Velinder, Martin Tristani-Firouzi, Joshua D Schiffman, Tatiana Tvrdik, Rong Mao, D Hunter Best, Pinar Bayrak-Toydemir, and Aaron R Quinlan. Effective variant filtering and expected candidate variant yield in studies of rare human disease. *NPJ Genom Med*, 6(1):60, July 2021.

[107] Elena Perenthaler, Soheil Yousefi, Eva Niggl, and Tahsin Stefan Barakat. Beyond the exome: The non-coding genome and enhancers in neurodevelopmental disorders and malformations of cortical development. *Front. Cell. Neurosci.*, 13:352, July 2019.

[108] Dalila Pinto, Alistair T Pagnamenta, Lambertus Klei, Richard Anney, Daniele Merico, Regina Regan, Judith Conroy, Tiago R Magalhaes, Catarina Correia, Brett S Abrahams, Joana Almeida, Elena Bacchelli, Gary D Bader, Anthony J Bailey, Gillian Baird, Agatino Battaglia, Tom Berney, Nadia Bolshakova, Sven Bölte, Patrick F Bolton, Thomas Bourgeron, Sean Brennan, Jessica Brian, Susan E Bryson, Andrew R Carson, Guillermo Casallo, Jillian Casey, Brian H Y Chung, Lynne Cochrane, Christina Corsello, Emily L Crawford, Andrew Crossett, Cheryl Cytrynbaum, Geraldine Dawson, Maretha de Jonge, Richard Delorme, Irene Drmic, Eftichia Duketis, Frederico Duque, Annette Estes, Penny Farrar, Bridget A Fernandez, Susan E Folstein, Eric Fombonne, Christine M Freitag, John Gilbert, Christopher Gillberg, Joseph T Glessner, Jeremy Goldberg, Andrew Green, Jonathan Green, Stephen J Guter, Hakon Hakonarson, Elizabeth A Heron, Matthew Hill, Richard Holt, Jennifer L Howe, Gillian Hughes, Vanessa Hus, Roberta Igliozzi, Cecilia Kim, Sabine M Klauck, Alexander Kolevzon, Olena Korvatska, Vlad Kustanovich, Clara M Lajonchere, Janine A Lamb, Magdalena Laskawiec, Marion Leboyer, Ann Le Couteur, Bennett L Leventhal, Anath C Lionel, Xiao-Qing Liu, Catherine Lord, Linda Lotspeich, Sabata C Lund, Elena Maestrini, William Mahoney, Carine Mantoulan, Christian R Marshall, Helen McConachie, Christopher J McDougle, Jane McGrath, William M McMahon, Alison Merikangas, Ohsuke Migita, Nancy J Minshew, Ghazala K Mirza, Jeff Munson, Stanley F Nelson, Carolyn Noakes, Abdul Noor, Gudrun Nygren, Guiomar Oliveira, Katerina Papanikolaou, Jeremy R Parr, Barbara Parrini, Tara Paton, Andrew Pickles, Marion Pilorge, Joseph Piven, Chris P Ponting, David J Posey, Annemarie Poustka, Fritz Poustka, Aparna Prasad, Jiannis Ragoussis, Katy Renshaw, Jessica Rickaby, Wendy Roberts, Kathryn Roeder, Bernadette Roge, Michael L Rutter, Laura J Bierut, John P Rice, Jeff Salt, Katherine Sansom, Daisuke Sato, Ricardo Segurado, Ana F Sequeira, Lili Senman, Naisha Shah, Val C Sheffield, Latha Soorya, Inês Sousa, Olaf Stein, Nuala Sykes, Vera Stoppioni, Christina Strawbridge, Raffaella Tancredi, Katherine Tansey, Bhooma Thiruvahindrapduram, Ann P Thompson, Susanne Thomson, Ana Tryfon, John Tsiantis, Herman Van Engeland, John B Vincent, Fred Volkmar, Simon Wallace, Kai Wang, Zhouzhi Wang, Thomas H Wassink, Caleb Webber, Rosanna Weksberg, Kirsty Wing, Kerstin Wittemeyer, Shawn Wood, Jing Wu, Brian L Yaspan, Danielle Zurawiecki, Lonnie Zwaigenbaum, Joseph D Buxbaum, Rita M Cantor, Edwin H Cook, Hilary Coon, Michael L Cuccaro, Bernie Devlin, Sean Ennis, Louise Gallagher, Daniel H Geschwind, Michael Gill, Jonathan L Haines, Joachim Hallmayer, Judith Miller, Anthony P Monaco, John I Nurnberger, Jr, Andrew D Paterson, Margaret A Pericak-Vance, Gerard D Schellenberg, Peter Szatmari, Astrid M Vicente, Veronica J Vieland, Ellen M Wijsman, Stephen W Scherer, James S Sutcliffe, and Catalina Betancur. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368–372, July 2010.

[109] Fernanda C G Polubriaginof, Rami Vanguri, Kayla Quinnies, Gillian M Belbin, Alexandre Yahi, Hojjat Salmasian, Tal Lorberbaum, Victor Nwankwo, Li Li, Mark M Shervey, Patricia Glowe, Iuliana Ionita-Laza, Mary Simmerling, George Hripcsak, Suzanne Bakken, David Goldstein, Krzysztof Kiryluk, Eimear E Kenny, Joel Dudley, David K Vawdrey, and Nicholas P Tatonetti. Disease heritability inferred from familial relationships reported in medical records. *Cell*, 173(7):1692–1704.e11, June 2018.

[110] Matthias Prestel, Caroline Prell-Schicker, Tom Webb, Rainer Malik, Barbara Lindner, Natalie Ziesch, Monika Rex-Haffner, Simone Röh, Thanatip Viturawong, Manuel Lehm, Michal Mokry, Hester den Ruijter, Saskia Haitjema, Yaw Asare, Flavia Söllner, Maryam Ghaderi Najafabadi, Rédouane Aherrahrou, Mete Civelek, Nilesh J Samani, Matthias Mann, Christof Haffner, and Martin Dichgans. The atherosclerosis risk variant rs2107595 mediates Allele-Specific transcriptional regulation of via E2F3 and rb1. *Stroke*, 50(10):2651–2660, October 2019.

[111] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features, 2010.

[112] R Core Team, R Foundation for Statistical Computing, Vienna, Austria. R: A language and environment for statistical computing, 2020.

[113] Elliott Rees, Jun Han, Joanne Morgan, Noa Carrera, Valentina Escott-Price, Andrew J Pocklington, Madeleine Duffield, Lynsey S Hall, Sophie E Legge, Antonio F Pardiñas, Alexander L Richards, Julian Roth, Tatyana Lezheiko, Nikolay Kondratyev, Vasilii Kaleda, Vera Golimbet, Mara Parellada, Javier González-Peñas, Celso Arango, GROUP Investigators, Micha Gawlik, George Kirov, James T R Walters, Peter Holmans, Michael C O'Donovan, and Michael J Owen. De novo mutations identified by exome sequencing implicate rare missense variants in SLC6A1 in schizophrenia. *Nat. Neurosci.*, 23(2):179–184, February 2020.

[114] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, 47(D1):D886–D894, January 2019.

[115] Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases, 1996.

[116] E R Ritvo, B J Freeman, A Mason-Brothers, A Mo, and A M Ritvo. Concordance for the syndrome of autism in 40 pairs of afflicted twins. *Am. J. Psychiatry*, 142(1):74–77, January 1985.

[117] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, Ginell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015.

[118] RStudio Team. RStudio: Integrated development for R, 2019.

[119] Stephan J Sanders, A Gulhan Ercan-Sencicek, Vanessa Hus, Rui Luo, Michael T Murtha, Daniel Moreno-De-Luca, Su H Chu, Michael P Moreau, Abha R Gupta, Susanne A Thomson, Christopher E Mason, Kaya Bilguvar, Patricia B S Celestino-Soper, Murim Choi, Emily L Crawford, Lea Davis, Nicole R Davis Wright, Rahul M Dhodapkar, Michael DiCola, Nicholas M DiLullo, Thomas V Fernandez, Vikram Fielding-Singh, Daniel O Fishman, Stephanie Frahm, Rouben Garagaloyan, Gerald S Goh, Sindhuja Kammela, Lambertus Klei, Jennifer K Lowe, Sabata C Lund, Anna D McGrew, Kyle A Meyer, William J Moffat, John D Murdoch, Brian J O'Roak, Gordon T Ober, Rebecca S Pottenger, Melanie J Raubeson, Youeun Song, Qi Wang, Brian L Yaspan, Timothy W Yu, Ilana R Yurkiewicz, Arthur L Beaudet, Rita M Cantor, Martin Curland, Dorothy E Grice, Murat Günel, Richard P Lifton, Shrikant M Mane, Donna M Martin, Chad A Shaw, Michael Sheldon, Jay A Tischfield, Christopher A Walsh, Eric M Morrow, David H Ledbetter, Eric Fombonne, Catherine

Lord, Christa Lese Martin, Andrew I Brooks, James S Sutcliffe, Edwin H Cook, Jr, Daniel Geschwind, Kathryn Roeder, Bernie Devlin, and Matthew W State. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 williams syndrome region, are strongly associated with autism. *Neuron*, 70(5):863–885, June 2011.

[120] Sven Sandin, Paul Lichtenstein, Ralf Kuja-Halkola, Christina Hultman, Henrik Larsson, and Abraham Reichenberg. The heritability of autism spectrum disorder, 2017.

[121] Sven Sandin, Paul Lichtenstein, Ralf Kuja-Halkola, Christina Hultman, Henrik Larsson, and Abraham Reichenberg. The heritability of autism spectrum disorder, 2017.

[122] Amartya Sanyal, Bryan R Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, September 2012.

[123] F Kyle Satterstrom, Jack A Kosmicki, Jiebiao Wang, Michael S Breen, Silvia De Rubeis, Joon-Yong An, Minshi Peng, Ryan Collins, Jakob Grove, Lambertus Klei, Christine Stevens, Jennifer Reichert, Maureen S Mulhern, Mykyta Artomov, Sherif Gerges, Brooke Sheppard, Xinyi Xu, Aparna Bhaduri, Utku Norman, Harrison Brand, Grace Schwartz, Rachel Nguyen, Elizabeth E Guerrero, Caroline Dias, Autism Sequencing Consortium, iPSYCH-Broad Consortium, Catalina Betancur, Edwin H Cook, Louise Gallagher, Michael Gill, James S Sutcliffe, Audrey Thurm, Michael E Zwick, Anders D Børglum, Matthew W State, A Ercument Cicek, Michael E Talkowski, David J Cutler, Bernie Devlin, Stephan J Sanders, Kathryn Roeder, Mark J Daly, and Joseph D Buxbaum. Large-Scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*, 180(3):568–584.e23, February 2020.

[124] Jonathan Sebat, B Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, Anthony Leotta, Deepa Pai, Ray Zhang, Yoon-Ha Lee, James Hicks, Sarah J Spence, Annette T Lee, Kaija Puura, Terho Lehtimäki, David Ledbetter, Peter K Gregersen, Joel Bregman, James S Sutcliffe, Vaidehi Jobanputra, Wendy Chung, Dorothy Warburton, Mary-Claire King, David Skuse, Daniel H Geschwind, T Conrad Gilliam, Kenny Ye, and Michael Wigler. Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449, April 2007.

[125] Ron Sender, Shai Fuchs, and Ron Milo. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.*, 14(8):e1002533, August 2016.

[126] Tamim H Shaikh. Copy number variation disorders. *Curr. Genet. Med. Rep.*, 5(4):183–190, December 2017.

[127] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050, August 2005.

[128] Ngak-Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C Ng. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.*, 40(Web Server issue):W452–7, July 2012.

[129] Asli Sirmaci, Michail Spiliopoulos, Francesco Brancati, Eric Powell, Duygu Duman, Alex Abrams, Guney Bademci, Emanuele Agolini, Shengru Guo, Berrin Konuk, Asli Kavaz, Susan Blanton, Maria Christina Digilio, Bruno Dallapiccola, Juan Young, Stephan Zuchner, and Mustafa Tekin. Mutations in ANKRD11 cause KBG syndrome, characterized by intellectual disability, skeletal malformations, and macrodontia. *Am. J. Hum. Genet.*, 89(2):289–294, August 2011.

[130] Smit, AFA, Hubley, R & Green, P. RepeatMasker, 2013.

[131] Michael Song, Xiaoyu Yang, Xingjie Ren, Lenka Maliskova, Bingkun Li, Ian R Jones, Chao Wang, Fadi Jacob, Kenneth Wu, Michela Traglia, Tsz Wai Tam, Kirsty Jamieson, Si-Yao Lu, Guo-Li Ming, Yun Li, Jun Yao, Lauren A Weiss, Jesse R Dixon, Luke M Judge, Bruce R Conklin, Hongjun Song, Li Gan, and Yin Shen. Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat. Genet.*, 51(8):1252–1262, August 2019.

[132] SPARK Consortium. Electronic address: pfeliciano@simonsfoundation.org and SPARK Consortium. SPARK: A US cohort of 50,000 families to accelerate autism research. *Neuron*, 97(3):488–493, February 2018.

[133] S Steffenburg, C Gillberg, L Hellgren, L Andersson, I C Gillberg, G Jakobsson, and M Bohman. A twin study of autism in denmark, finland, iceland, norway and sweden. *J. Child Psychol. Psychiatry*, 30(3):405–416, May 1989.

[134] Peter D Stenson, Matthew Mort, Edward V Ball, Katy Howells, Andrew D Phillips, Nick St Thomas, and David N Cooper. The human gene mutation database: 2008 update. *Genome Med.*, 1(1):13, January 2009.

[135] Holly A F Stessman, Tychele N Turner, and Evan E Eichler. Molecular subtyping and improved treatment of neurodevelopmental disease. *Genome Med.*, 8(1):22, February 2016.

[136] C Stoltenberg, S Schjølberg, M Bresnahan, M Hornig, D Hirtz, C Dahl, K K Lie, T Reichborn-Kjennerud, P Schreuder, E Alsaker, A-S Øyen, P Magnus, P Surén, E Susser, W I Lipkin, and ABC Study Group. The autism birth cohort: a paradigm for gene-environment-timing research. *Mol. Psychiatry*, 15(7):676–680, July 2010.

[137] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies, 2019.

[138] Jacob L Taylor, Jean-Christophe P G Debost, Sarah U Morton, Emilie M Wigdor, Henrike O Heyne, Dennis Lal, Daniel P Howrigan, Alex Bloemendal, Janne T Larsen, Jack A Kosmicki, Daniel J Weiner, Jason Homsy, Jonathan G Seidman, Christine E Seidman, Esben Agerbo, John J McGrath, Preben Bo Mortensen, Liselotte Petersen, Mark J Daly, and Elise B Robinson. Paternal-age-related de novo mutations and risk for five disorders. *Nat. Commun.*, 10(1):3043, July 2019.

[139] Lorane Texari, Nathanael J Spann, Ty D Troutman, Mashito Sakai, Jason S Seidman, and Sven Heinz. An optimized protocol for rapid, sensitive and robust on-bead ChIP-seq from primary cells. *STAR Protoc*, 2(1):100358, March 2021.

[140] Beata Tick, Patrick Bolton, Francesca Happé, Michael Rutter, and Frühling Rijsdijk. Heritability of autism spectrum disorders: a meta-analysis of twin studies. *J. Child Psychol. Psychiatry*, 57(5):585–595, May 2016.

[141] Tychele N Turner, Bradley P Coe, Diane E Dickel, Kendra Hoekzema, Bradley J Nelson, Michael C Zody, Zev N Kronenberg, Fereydoun Hormozdiari, Archana Raja, Len A Pennacchio, Robert B Darnell, and Evan E Eichler. Genomic patterns of de novo mutation in simplex autism. *Cell*, 171(3):710–722.e12, October 2017.

[142] B W M van Bon, B P Coe, R Bernier, C Green, J Gerdts, K Witherspoon, T Kleefstra, M H Willemsen, R Kumar, P Bosco, M Fichera, D Li, D Amaral, F Cristofoli, H Peeters, E Haan, C Romano, H C Mefford, I Scheffer, J Gecz, B B A de Vries, and E E Eichler. Disruptive de novo mutations of DYRK1A lead to a syndromic form of autism and ID. *Mol. Psychiatry*, 21(1):126–132, January 2016.

[143] Geraldine A Van der Auwera and Brian D O'Connor. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, April 2020.

[144] Weidi Wang, Roser Corominas, and Guan Ning Lin. De novo mutations from whole exome sequencing in neurodevelopmental and psychiatric disorders: From discovery to application, 2019.

[145] Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements, 2004.

[146] Thomas H Wassink, Joseph Piven, and Shivanand R Patil. Chromosomal abnormalities in a clinic sample of individuals with autistic disorder, 2001.

[147] E Weidmer-Mikhail, S Sheldon, and M Ghaziuddin. Chromosomes in autism and related pervasive developmental disorders: a cytogenetic study, 1998.

[148] Alan B Wells, Nathan Kopp, Xiaoxiao Xu, David R O'Brien, Wei Yang, Arye Nehorai, Tracy L Adair-Kirk, Raphael Kopan, and Joseph D Dougherty. The anatomical distribution of genetic associations.

[149] Katharine D Wenstrom. Fragile X and other trinucleotide repeat diseases. *Obstet. Gynecol. Clin. North Am.*, 29(2):367–88, vii, June 2002.

[150] Donna M Werling, Harrison Brand, Joon-Yong An, Matthew R Stone, Lingxue Zhu, Joseph T Glessner, Ryan L Collins, Shan Dong, Ryan M Layer, Eirene Markenscoff-Papadimitriou, Andrew Farrell, Grace B Schwartz, Harold Z Wang, Benjamin B Currall, Xuefang Zhao, Jeanselle Dea, Clif Duhn, Carolyn A Erdman, Michael C Gilson, Rachita Yadav, Robert E Handsaker, Seva Kashin, Lambertus Klei, Jeffrey D Mandell, Tomasz J Nowakowski, Yuwen Liu, Sirisha Pochareddy, Louw Smith, Michael F Walker, Matthew J Waterman, Xin He, Arnold R Kriegstein, John L Rubenstein, Nenad Sestan, Steven A McCarroll, Benjamin M Neale, Hilary Coon, A Jeremy Willsey, Joseph D Buxbaum, Mark J Daly, Matthew W State, Aaron R Quinlan, Gabor T Marth, Kathryn Roeder, Bernie Devlin, Michael E Talkowski, and Stephan J Sanders. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.*, 50(5):727–736, April 2018.

[151] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, November 2019.

[152] Jinan Zeidan, Eric Fombonne, Julie Scorah, Alaa Ibrahim, Maureen S Durkin, Shekhar Saxena, Afiqah Yusuf, Andy Shih, and Mayada Elsabbagh. Global prevalence of autism: A systematic review update. *Autism Res.*, 15(5):778–790, May 2022.

[153] Feng Zhang, Wenli Gu, Matthew E Hurles, and James R Lupski. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.*, 10:451–481, 2009.

[154] Guihu Zhao, Kuokuo Li, Bin Li, Zheng Wang, Zhenghuan Fang, Xiaomeng Wang, Yi Zhang, Tengfei Luo, Qiao Zhou, Lin Wang, Yali Xie, Yijing Wang, Qian Chen, Lu Xia, Yu Tang, Beisha Tang, Kun Xia, and Jinchen Li. Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res.*, 48(D1):D913–D926, January 2020.

[155] Jian Zhou, Christopher Y Park, Chandra L Theesfeld, Aaron K Wong, Yuan Yuan, Claudia Scheckel, John J Fak, Julien Funk, Kevin Yao, Yoko Tajima, Alan Packer, Robert B Darnell, and Olga G Troyanskaya. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.*, 51(6):973–980, June 2019.

[156] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, 12(10):931–934, October 2015.

[157] Xueya Zhou, Pamela Feliciano, Chang Shu, Tianyun Wang, Irina Astrovskaya, Jacob B Hall, Joseph U Obiajulu, Jessica R Wright, Shwetha C Murali, Simon Xuming Xu, Leo Brueggeman, Taylor R Thomas, Olena Marchenko, Christopher Fleisch, Sarah D Barns, Leeanne Green Snyder, Bing Han, Timothy S Chang, Tychele N Turner, William T Harvey, Andrew Nishida, Brian J O'Roak, Daniel H Geschwind, SPARK Consortium, Jacob J Michaelson, Natalia Volfovsky, Evan E Eichler, Yufeng Shen, and Wendy K Chung. Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nat. Genet.*, 54(9):1305–1319, September 2022.