AN ANALYSIS OF GENETIC VARIANTS ASSOCIATED WITH AUTISM SPECTRUM DISORDER

by

Daniel Benjamin Callaghan

B.Sc., Acadia University, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

December 2017

Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder affecting roughly 1% of the human population. Genomics research to date has discovered only a fraction of the variants causative for ASD. To this end, we whole-genome sequenced a cohort of 119 ASD individuals in order to find likely pathogenic variation. After quality and frequency filters, we prioritized variants as likely causal according to rarity and predicted damage scores (CADD and Snap2). Here, we report five *de novo* damaging variants and seven likely damaging variants of unknown inheritance. Since much of the variation reported in ASD cases is uncertain both in function and in significance in ASD, we aimed to functionally characterize missense variants from the ASD literature in PTEN and SYNGAP1, two well-characterized ASD genes. We curated missense variants of unknown significance from the ASD literature and assayed their functional effect in yeast using a Synthetic Genetic Array. We chose previously biochemically validated variants, population variants, and other variants in the genes of interest to gain insight into the functional diversity of PTEN and SYNGAP1 variation. We established functional effect of the ASD variants of unknown significance in PTEN and showed that computational predictors of damage are reasonable predictors of variants' functional effects in yeast. We found that agreement of computational metrics breaks down when predicting damage in certain genes, such as SYNGAP1. Functionalizing variants in this way contributes to our understanding of the range of functional effects of ASD variants.

Lay Summary

Autism Spectrum Disorder (ASD) is a genetic disorder causing social and communication deficits. While we have studied ASD for a long time, we have much to learn in terms of what genes are involved in ASD risk. We studied the genomes of a group of ASD individuals to discover mutations that are likely causing their ASD. We searched for rare mutations, and ones that they did not inherit from their parents, as these are more likely to be true ASD mutations. We found likely causal mutations in five individuals. In a separate project, we determined important ASD mutations to study the function of. Our collaborators studied these mutations by inserting them into yeast. We studied whether the effect of ASD mutations in yeast agreed with our computational predictions. While our predictions held up well, we found that for some genes, prediction was more difficult than for others.

Preface

Chapter 2 describes collaborative work on a whole-genome sequencing based autism variant discovery effort. Dr. Suzanne Lewis, Kristina Calli, Boris Kuzeljevic, Annie Yu, and Franz-Edward Kurtzke were responsible for cohort collection, deep phenotyping, and phenotype clustering. Chang Yu, Yanchen Li, and Yingrui Li from the Beijing Genomics Institute (BGI) performed the genome sequencing. Preliminary sequence analysis was performed by Dr. Guy Rouleau and Alexandre Dionne-Laporte. CNV calling and analysis was performed by Evica Rajcan-Separovic and Ying Qiao. Patrick Tan performed preliminary analysis on the cohort and was responsible for the original variant calling procedure, and the creation of ASPIREdb. Nathan Holmes, Matthew Jacobson, Manuel Belmadani, and Dr. Sanja Rogic are responsible for the creation of MARVdb. I was responsible for variant quality filtering steps, variant prioritization, and writing most of the manuscript, in addition to Dr. Paul Pavlidis and Dr. Sanja Rogic. Sanger sequencing was performed by Amy McNaughton, Melissa Hudson, and Xudong Liu. A portion of the discussion was originally written by Dr. Sanja Rogic and Dr. Paul Pavlidis. Ethics approval for research involving human subjects was obtained through the joint Research Ethics Board of the University of British Columbia and B.C. Children's Hospital Research Institute of British Columbia (Certificate # H01-70507)

Chapter 3 is based on work conducted as part of a UBC-based collaboration to functionally characterize autism missense variants of unknown significance. The gene prioritization, PTEN variant prioritization, and development of the computational pipeline was done primarily by myself. SYNGAP1 variant prioritization was done by myself, Manuel Belmadani and Eric Chu. Gene selection was done by Dr. Paul Pavlidis, Dr. Sanja Rogic, Dr. Kurt Haas, Dr. Shernaz Bamji, Dr. Chris Loewen, Dr. Tim O'Connor, Dr. Kathy Rankin and Dr. Doug Allen. Variant

functionalization was done by Kathryn Post, Barry Young, Troy McDiarmid, Riki Dingwall, Payel Ganguli, and Matt Edwards. Analysis of yeast functionalization results was done by Dr. Paul Pavlidis and Manuel Belmadani.

Table of Contents

List of Tables

List of Figures

List of Supplementary Materials

List of Abbreviations

ASD – Autism Spectrum Disorder

CADD – Combined Annotation Dependent Depletion

CNV – Copy Number Variation

GO – Gene Ontology

GOF – Gain of Function

GWAS – Genome-Wide Association Study

ID – Intellectual Disability

LGD – Likely Gene Disrupting

LOF – Loss of Function

MAF – Minor Allele Frequency

NDD – Neurodevelopmental Disorder

PCR – Polymerase Chain Reaction

SFARI – Simons Foundation Autism Research Initiative

SGA – Synthetic Genetic Array

SNP – Single Nucleotide Polymorphism

SNV – Single-Nucleotide Variation

SSC – Simons Simplex Collection

VCF – Variant Call Format

VUS – Variant of Unknown Significance

WES – Whole Exome Sequencing

WGS – Whole Genome Sequencing

Acknowledgements

I thank Dr. Paul Pavlidis for providing unwavering guidance throughout my Master's degree, and for the countless lessons in scientific research, analysis, and hard work that I now take onwards.

I thank Dr. Sanja Rogic for her enduring support in this endeavor, and for the many hours of instruction and encouragement throughout.

I thank my committee members: Dr. Kurt Haas, Dr. Dan Goldowitz, and Dr. Tara Klassen for prompting me to think deeply and critically about my research.

I thank the UBC GSAT program and NSERC for the financial support towards my degree.

I thank Shams Bhuiyan, Manuel Belmadani, Patrick Tan, Lilah Toker, and Shreejoy Tripathy for their expert counsel, and Matt Jacobson and Ogan Mancarci for their frequently sought and enthusiastically offered technical expertise. To all of Pavlab – I would be hard-pressed to find a more supportive and inspiring learning community. Thank you for challenging me and teaching me every day.

Thank you Ali, Veronique, Maddy, Min, Dan, Tal, Frank, Brent, Queenie, Kathryn, Troy, Riki, Payel, Matt, Warren, and Steph for (at different times) teaching and inspiring me, learning alongside me, and helping me enjoy the entire experience of my Master's.

To Matt and Danya: thank you for making Vancouver home.

I thank Margaux Ross for the constant and enduring love and support.

To all members, past, present, and future of AVFR: you kept me sane.

To Neil Spencer: thank you for encouraging me in all academic pursuits.

To Neal Callaghan: you are a constant source of inspiration.

To my parents: thank you for the support, the love, and the encouragement that I can always count on from you.

Dedication

To my family.

Chapter 1: Introduction

Modern genomics research has done much to improve our understanding of the genetic etiologies of Autism Spectrum Disorder (ASD), but a central question remains only partially solved: “What are the genetic variants that contribute to ASD incidence?”

ASD genetics research is steadily identifying candidate variants in ASD cases, but there are many cases for which no molecular diagnosis has been established. Furthermore, of the candidate variants reported in the genetic literature, many of them are variants whose true functional effect is uncertain. These variants evoke a second key question: “Of the genetic variants reported to contribute to ASD incidence, how many of them are truly causative for ASD, and what is their functional impact?”

In this thesis, I provide an introduction to these gaps in modern genetic knowledge, and describe my research done in pursuit of addressing these ASD research questions. Chapter 1 begins with an introduction to the key concepts that guided the goals of my thesis, including our current diagnostic and genetic understanding of ASD, and follows with an exploration of the current methodological approaches in ASD genomic research efforts. Chapter 2 describes my research to identify causative genetic variation in a cohort of ASD probands. Chapter 3 describes my research to prioritize ASD-associated variants (obtained from sequencing efforts such as the one described in Chapter 2) for functional characterization in an array of model organism assays. Chapter 4 presents conclusions from these two chapters and discusses implications for further research.

1.1 An evolution in our understanding of Autism Spectrum Disorder

What is now called Autism Spectrum Disorder was first described formally in 1943 by Leo Kanner. He commented on his patients' communication deficits and unique patterns of behaviour, summarizing his diagnosis as "children's inability to relate themselves in the normal way to people

and situations from the beginning of life" (Kanner & others, 1943, Robison, 2016). This foundational publication was followed soon after by a paper by Hans Asperger, a physician at the University Children's hospital in Vienna, who reported the symptoms of what he called "autistic psychopathy" in a cohort of children (Asperger, 1944, Silberman, 2015). Subsequent research in diverse fields such as neuroscience, psychology, and genetics would later inform new perspectives regarding ASD, and a gradual refining of our perception of the disorder. Indeed, since these early accounts, ASD has undergone many updates in definition and description. The latest change in diagnostic criteria was published in the fifth edition of The Diagnostic and Statistical Manual of Mental Disorders (DSM-5). This diversity of perspectives and ongoing evolution of our understanding reflects the etiological complexity of ASD.

We now define ASD to be a neurodevelopmental disorder characterized by deficits in social communication, repetitive behaviors, and restricted interests (Association & others, 2013). Recent estimates of ASD prevalence range from 1-2% of the population (Baird et al., 2006). ASD is estimated as being four times as prevalent in males compared to females (Chakrabarti & Fombonne, 2005; Kogan et al., 2009). ASD is typically diagnosed in early childhood by clinicians following diagnostic protocols outlined in two widely utilized diagnostic instruments. Clinicians use either the behavioural observation-based Autism Diagnostic Observation Schedule (ADOS) or parental interview Autism Diagnostic Interview-Revised ADI-R to assess levels of development of the referred individual in several key areas relating to language, social interaction, and repeated patterns of behaviour (Lord et al., 2012; Lord, Rutter, & Couteur, 1994). While our diagnostic techniques, availability of services, and public awareness of the disorder has only improved (Fombonne, 2003b; Schieve et al., 2011), the diagnostic journey for many children and their parents can often take several years. Although ASD can be diagnosed as early as 2 years, Crane

et. al (2015) reported a mean age of diagnosis of 7.5 years old, an estimate relatively consistent with that of studies now decades old (Howlin & Moore, 1997). Other estimates have been more moderate; Mandel *et al.* reported 3.1 years being the norm for age of diagnosis (Mandell, Novak, & Zubritsky, 2005). Delays in diagnosis can often bring about inadequate education and developmental interventions for these individuals in the interim. Studies have consistently indicated that early diagnosis and subsequent intervention in ASD cases are critical for many areas of development, including language ability, social behaviour, and IQ scores (Anderson & Romanczyk, 1999; Corsello, 2005; Fenske, Zalenski, Krantz, & McClannahan, 1985). One major obstacle towards early and accurate diagnosis lies in the extreme phenotypic heterogeneity of the disorder. Children with ASD differ widely in their presentation of the social and behavioural symptoms central to the disorder (Geschwind, 2009; Wiggins, Robins, Adamson, Bakeman, & Henrich, 2012). This marked heterogeneity can be considered to comprise the "spectrum" of "Autism Spectrum Disorder", and has ramifications both on proper diagnosis as well as adequate intervention, as there is no "one size fits all" approach to either in the case of ASD.

ASD diagnosis and intervention is further complicated by the co-occurrence of the many common comorbidities of ASD. Other neurodevelopmental disorders, such as intellectual disability and epilepsy, are more frequent in ASD patients than the general population (Comorbidity of ID: 70%, comorbidity of epilepsy: 30%) (Fombonne, 2003a; La Malfa, Lassi, Bertelli, Salvini, & Placidi, 2004; Olsson, Steffenburg, & Gillberg, 1988; Tuchman, Cuccaro, & Alessandri, 2010). Anxiety disorders, sleep problems, and gastrointestinal symptoms are also commonly reported alongside ASD. These comorbidities may disguise symptoms indicative of the condition, and can create extra challenges for successful management of the condition.

Far from Kanner's early suggestions that autism was spawned in children due to uncaring mothers, or Andrew Wakefield's fraudulent claims regarding the causative role of vaccines, we now know ASD is largely genetic in origin (Godlee, Smith, & Marcovitch, 2011; Kanner, 1949; Sandin et al., 2017; Wakefield et al., 1998). Twin studies have provided strong evidence for genetic etiologies of ASD, with reported concordance rates for monozygotic twins above 80%, and those of dizygotic twins at roughly 30% (Rosenberg et al., 2009). While there is conclusive evidence for the genetic roots of the condition, pinning down the exact genetic causes of ASD is much more difficult. Early genetic researchers of ASD discovered that it is not a simple Mendelian (single gene) disorder – ASD cohorts demonstrated weak, disparate genetic signals, making it difficult to identify important genes (Buxbaum et al., 2001; Ronald et al., 2006). As we will explore later in this chapter, in addition to being phenotypically heterogeneous, ASD is also extremely genetically heterogeneous (Bailey et al., 1995; Betancur, 2011).

Improving our understanding of the genetic etiology of ASD may lead to improved diagnostics of the disorder. In spite of the challenges that complicate the efficient and adequate identification of children with ASD, diagnoses - and thus estimates of prevalence - have only increased over the past several decades. This steady increase in apparent incidence has prompted substantial research into potential environmental causes. However, it is now broadly understood that the increase in reported prevalence is mainly due to expanded criteria in the DSM-5 and the refinement and improvement of the sensitivity of our diagnostic tools (Matson & Kozlowski, 2011; Wazana, Bresnahan, & Kline, 2007). Modern genetics techniques provide the solution to providing genetic explanations of ASD provenance for the ever-growing number of cases.

1.2 Characterizing and defining human genetic variation

Any research into human variation operates under the context of our modern knowledge of the human genome and variation therein. The human genome has some 6 billion base pairs of DNA. Every human genome contains on average 4-5 million variants in their DNA – deviations from the canonical reference genome (Consortium & others, 2015). These differences in the genetic code are a source of phenotypic diversity, including variation in disease risk. This diversity by way of effect in disease risk, and more broadly, the phenotypic effect, is often referred to as the range of *functional impact* or *effect size* of variants.

Several categories of variation are more likely than others to have functional (phenotypic) effects. The category of a variant as defined in this way is called its *functional category*. Large structural variants, such as chromosomal rearrangements or copy number variants (CNVs) spanning multiple genes, are often implicated in human disease. Insertions/deletions (indels) of one or more nucleotides and single nucleotide variants (SNVs) can have a variety of effects depending on the location and specific bases changed. Some indels can cause changes in the reading frame of DNA, causing so-called frameshifts and resulting in massive changes to a protein's code. SNVs can also cause large-scale impacts on a gene's resulting protein product. Stop-gain mutations are changes in DNA that create a premature stop codon – resulting in a truncated protein product. Splice site variants disrupt the splicing properties of genes, with often severe repercussions for the gene's resulting protein product. Frameshift-causing, stop-gain, and splice site variants are often described as a group as Likely Gene Disrupting (LGD) variants. Nonsynonymous or missense SNVs can cause changes in the protein sequence of a gene with a range of functional effects – from no effect at all to creating a completely nonfunctional protein product. Variants that are not in coding regions of genes (intergenic and intronic variants) are less

often implicated in human disease, as they are more difficult to study. Variant categories are often a useful preliminary consideration in predicting a variant's functional impact. However, as we will discuss, many other levels of information can be obtained about genomic variants. Variant annotation programs such as Annovar allow researchers to annotate variants with the results of multiple tools and databases (K. Wang, Li, & Hakonarson, 2010).

The functional impact of a variant is also intrinsically related to its allele frequency in the human population. Generally, deleterious variants are reduced in frequency in the human population through the process of natural selection. Therefore, inherited variants tend to be less deleterious, on average, than *de novo* variants; that is, genetic variants occurring for the first time in a family due to mutations in the germ cells of one of the parents. The most deleterious alleles typically only appear in humans by way of *de novo* variation (Veltman & Brunner, 2012). By considering variants in terms of both population frequency and functional impact, several important subsets of variants emerge. Different methods of studying variation tend to apply more readily – and yield more success – in studying certain subsets of variants, in regards to their position in this two-dimensional space (Figure 1-1). Rare, high impact variation has often been studied using family-based techniques, such as the transmission-disequilibrium test. Early genetic research of ASD syndromes often fell under this category of investigation (the results of which are discussed later in this chapter). Genome-wide association studies (GWAS) are used to map genomic loci to incidence of certain traits or disease via linkage analysis, and have typically been used to study common, low-impact variation. While the methods of GWAS have proven successful for detecting many trait-SNP associations (e.g. for height, Type 2 Diabetes, and Schizophrenia, to name a few) (Visscher et al., 2017), GWA studies have typically failed for

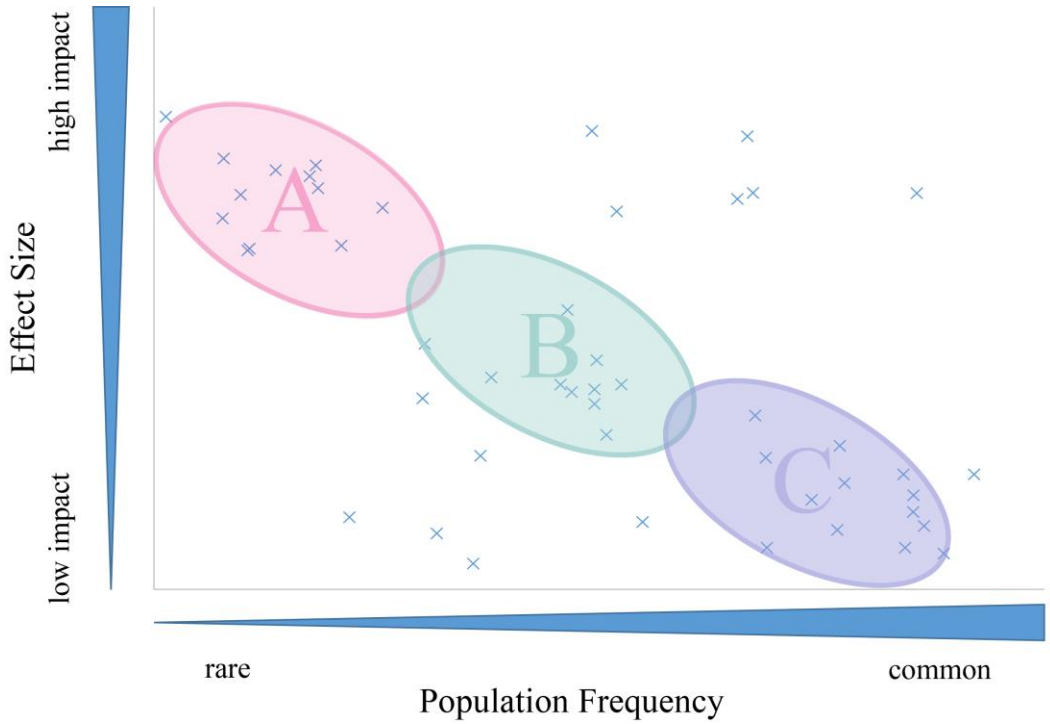

Figure 1-1: Simulated variants and their involvement in human disease.
Important categories of disease causing variants in humans (M. I. McCarthy et al., 2008). A) Rare, high-effect variants causing Mendelian disease. Often studied using family studies. B) Uncommon, moderate effect variants. Often implicated in sporadic diseases via WGS/WES. C) Common, low-impact variants. Often implicated in common diseases by GWAS. Variants below the diagonal line (rare, low-impact) are often neutral or benign in terms of phenotypic effect or disease risk – this is an especially large category of variation. Variants above the diagonal line (common, high-impact), are not often implicated in human disease – the high impact of these variants is often associated with non-disease associated phenotypic effects (eye colour, etc.)

characterizing more complex or less penetrant traits due to the extremely large cohort sizes that would be required to obtain statistical significance in these cases (Visscher, Brown, McCarthy, & Yang, 2012). Whole-exome sequencing (WES) and whole genome sequencing (WGS) studies searching for rare variants have provided many answers where GWAS has failed. Large cohort WGS or WES efforts such as 1000G have provided the basis for our understanding of human variation as a whole, whereas case-control cohort studies provide a means for us to characterize the genetic variation that causes disease. As sequencing costs have decreased, the number of variants of which we know both in the human population and as associated with diseases have increased drastically.

The sheer quantity of human variants (4 – 5 million genomic variants per individual, 10 – 12 thousand peptide sequence-altering variants, 150 - 200 protein-truncating variants, and 400 – 500 thousand variants in regulatory regions) precludes comprehensive functional testing (Consortium & others, 2015). As such many methods exist with which to estimate the functional effect or potential pathogenicity of individual variants. These methods are useful both for variant discovery efforts for which no functional validation will take place, as well as prioritizing variants for functional study. In addition to the evidence provided by a variant's functional category (such as nonsynonymous or LGD), many other sources of evidence exist. In this section I will explore the incorporation of variant data regarding frequency in the human population, inheritance, and computational predictors of damage into our estimates of variant pathogenicity.

One of the most straightforward criteria for judging a variant's potential for pathogenicity is found in its inheritance status. It is estimated that there are approximately 50 – 150 *de novo* SNVs per generation (Besenbacher et al., 2015; Kong et al., 2012; Turner, Coe, et al., 2017). Based on the proportion of exonic DNA in the human genome (roughly 1-2%) only one of these variants

is expected to be exonic, on average (Rands, Meader, Ponting, & Lunter, 2014). As most cases of ASD are sporadic (ASD-affected proband born to unaffected parents), previous studies have yielded reasonable success in identifying candidate variation by searching first for damaging *de novo* variants (Fischbach & Lord, 2010; Sanders et al., 2012). Focusing on *de novo* exonic variants yields a marked reduction in search space, and therefore typically higher success rates in determining candidate variation. Many trio WGS/WES studies have taken advantage of this line of reasoning to increase their yields of explanatory variation (Schaaf & Zoghbi, 2011). Such studies have reported molecular diagnosis success rates of between 8% - 16% (Iossifov et al., 2012; Sanders et al., 2015; Turner, Coe, et al., 2017; Yuen et al., 2017)

Population frequency databases have often been used as an important source of evidence for establishing potential variant pathogenicity. The 1000 Genomes project (2012) was the earliest attempt to quantify human variation on a relatively large scale, using low-coverage whole-genome sequencing data combined with exome sequencing to characterize genomic variation in a cohort of 1092 humans (Consortium, 2012). While the effort was successful in identifying over 95% of common (> 5% frequency) variants, a key insight of this study was that to create a more representative picture of human variation, particularly of rare variation, much larger, as well as more ethnically diverse samples were required(The 1000 Genomes Project Consortium, 2012). The release of the ExAC database of human population variation (whole-exome sequencing data from 60,706 individuals) followed by the gnomAD database in 2017 (exome sequencing data from 123,136 individuals and whole genome sequencing data from 15,496 individuals) represented significant steps towards increasing the number of human samples captured in such large-scale sequencing efforts(Lek et al., 2016).

The ExAC and gnomAD studies of human population variation provide us with extra information beyond individual variant frequencies. By comparing the expected amount of a particular class of variation (e.g. missense variants, truncating variants), to the actual number found in a particular gene in the human population, we can calculate gene-wise depletion of variation. This provides a proxy by which to estimate an individual gene's propensity to be depleted of damaging mutation. Lek *et al.* (2016) provide summary scores of this propensity in the form of missense-Z (MSZ) and probability of loss-of-function intolerance (pLI) scores. The MSZ score is a gene-specific score for the gene's number of standard deviations away from the mean of observed – expected missense variants. Similarly, the pLI score is a version of the same statistic for likely gene-disrupting variants, corrected for gene length, and ranging from 0-1. Genes with MSZ scores > 3 (n = 1727) are considered strongly depleted for missense variation, and those with pLI scores > 0.9 (n = 3230) are considered intolerant to loss-of-function variation (haploinsufficient) (Lek et al., 2016).

Many computational tools now exist to predict the effect of variants in the human genome. Prediction tools such as SIFT (Sorting Intolerant From Tolerant), PolyPhen-2, CADD (Combined Annotation-Dependent Depletion), and Snap2 employ a variety of classification methods, predictive features, and training data to predict a given variant's effect.

SIFT uses multiple sequences alignments of related sequences of a protein to calculate a normalized probability score for each possible amino acid substitution at each position in the alignment. From this probability score, SIFT infers a binary prediction for the amino acid substitution to be deleterious or benign (Ng & Henikoff, 2001). SIFT outputs the normalized probability score (ranging from 0-1) in addition to the binary prediction; annotating each amino

acid substitution in a protein as deleterious (SIFT score < 0.05) or tolerated (SIFT score ≥ 0.05) (Kumar, Henikoff, & Ng, 2009).

PolyPhen2 uses sequence-based and structure- based predictive features to predict the damaging effects of missense mutations via a Naïve Bayes classifier (Adzhubei et al., 2010). The classifier is trained on 3155 damaging variants associated with Mendelian diseases contained in the UniProt database, as well as 6321 (assumed benign) differences between human proteins and closely related mammalian homologues. PolyPhen-2 also predicts a probabilistic score from 0-1, classifying variants with a score greater than 0.85 as probably damaging, those greater than 0.15 as possibly damaging, and the remainder as benign variants.

CADD is a support vector machine trained on a variety of features, including other variant impact prediction scores (ex. SIFT, PolyPhen) and gene-level features. The algorithm is trained on 2 sets of variants: 14.7 M genetic differences from the human-chimpanzee ancestors, now fixed in the human genome, and 14.7 M simulated *de novo* events (Kircher, Witten, Jain, O'Roak, et al., 2014). CADD is typically interpreted via a Phred-scaled version of the resulting "C-score". All genomic variants are annotated with this score, where a variant that scores as the top 10% of all C-scores is given a score of 10, those in the top 1% are given a score of 20, and so on. CADD improves on previous methods by classifying all genomic variants (whereas many tools focus on coding variants), as well as by using an assortment of criteria by which to classify variants.

Snap2 is a neural network-based classifier trained on disease and neutral variants that uses a set of screened protein features (including physical amino acid properties, solvent accessibility, binding residues, etc.) to distinguish between neutral variants and those that cause effects (Hecht, Bromberg, & Rost, 2015). Nonsynonymous changes are given a score ranging from -100 (likely benign) to 100 (likely damaging). Snap2 improves on previous methods that rank nonsynonymous

variants via manual screening of the predictive features and incorporating cross-validation into training and evaluation steps.

In addition to the four common predictors discussed above, other methods abound in the literature. A key concern when using these tools is in how accurate the methods are, as well as the agreement of the predictors with each other. While moderate agreement has previously been reported between various predictive methods, this agreement is typically calculated on a genome-wide basis (Liu, Jian, & Boerwinkle, 2011). Since many analyses and functionalization efforts are focused on one, or only a few genes, it is a relevant concern whether this agreement holds true on the resolution of a single gene. To our knowledge no effort to date has attempted to quantify the correlation of scores of variant predictors on a gene-by-gene basis. We therefore present in Chapter 3 a novel investigation of gene-wise correlations between four popular variant prediction methods (SIFT, Polyphen2, CADD, and Snap2).

By combining diverse variant-level data, we can build a more comprehensive picture of the propensity for any particular variant to have an effect – or cause disease in the human population. The information discussed above, and much more, has been used extensively in untangling the genetic basis of ASD.

1.3 The complex genetic etiologies of Autism Spectrum Disorder: insights from variant discovery efforts

The advent of next-generation sequencing and related analysis techniques has provided researchers powerful new tools to identify genetic variation associated with ASD. Recent years have seen an explosion of GWAS and WGS/WES studies attempting to implicate various loci, genes, and variants within those genes in ASD causation.

Early attempts to study ASD under a genetic lens focused primarily on family-based association studies. These attempts proved successful in establishing the role of genes like FMR1 and TSC1 for syndromic forms of ASD. Other attempts to study ASD genetically used genome wide association to try and map genomic loci to the disorder. However, such efforts yielded limited results in terms of identifying associated loci, implying that the underlying genetic etiologies were either too low-impact or too rare to produce any discernable signals (Anney et al., 2010; Cantor, Lange, & Sinsheimer, 2010; Geschwind, 2011; K. Wang et al., 2009). GWAS are typically more effectively used to identify common variants associated with a disease; the vast majority of ASD-associated loci would require vast cohort sizes. A more recent GWAS effort reported greater success by leveraging meta-analysis techniques to assemble an extremely large cohort (16,000 cases). This large cohort size provided the researchers with the statistical power necessary to establish a locus (10q24) as significant, as well as to establish several other potentially interesting (albeit nonsignificant) regions previously implicated in schizophrenia (The Autism Spectrum Disorders Working Group of the Psychiatric Genomics Consortium, 2017). The results of recent sequencing studies have been harmonious with what earlier efforts suggested: rare, damaging variants are often the culprit for ASD cases, and furthermore are typically easier to identify than other categories of variants (Iossifov et al., 2015).

The lack of success of GWAS in ASD genetics research has been offset by the success gained by looking for rare *de novo* variants in ASD affected probands. Therefore, the field has largely settled on the paradigm of searching for rare, high-impact variants contributing to ASD risk, versus the combinatorial effects of multiple common, small impact variants. Several lines of evidence from genetic research support this paradigm. The first line of evidence is found in the various categories of variation that have been implicated in ASD. Studies exploring cohorts for *de*

novo CNVs have demonstrated high success rates, and have been able to identify many recurrently affected genes by way of large deletions or duplications (Glessner et al., 2009; Pinto et al., 2010; Sebat et al., 2007). It is estimated that 5-8% of simplex ASD cases are due to copy-number variation (Schaaf & Zoghbi, 2011). Cases with no explanatory CNV are typically investigated for damaging SNVs. In simplex cases, in which the parents of the affected proband are found to be unaffected, researchers typically look for rare, high effect (missense or likely gene-disrupting) de novo variation. Secondly, studies have shown that ASD probands have a modest increase in numbers of *de novo* CNVs and *de novo* LGD and missense SNVs over controls (Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2011).

Genetic research has identified several ASD syndromes – subgroups of Autism Spectrum Disorder characterized by a set of recurrent phenotypes and sharing common genetic causes. Damaging variation in FMR1 can cause Fragile X syndrome – an inherited condition causing intellectual disability and frequently ASD (Farzin et al., 2006). Similarly, tuberous sclerosis, which is a genetic disorder characterized by multiple benign tumours as well as frequent reports of autistic symptoms, is caused by damaging alleles in TSC1 or TSC2 (Bourgeron, 2009; Tavazoie, Alvarez, Ridenour, Kwiatkowski, & Sabatini, 2005; Williams, Dagli, & Battaglia, 2008). Eventually many forms of ASD, particularly severe ones with comorbid intellectual disability, will be characterized in this way, by leveraging the power of large cohort sizes to characterize the etiological subgroups of ASD. The less frequent subgroups (of which there are undoubtedly many) will take more time and effort to fully characterize.

Genetic heterogeneity notwithstanding, several biological pathways are enriched for candidate variation reported from WES and WGS efforts. Pathways involved in synaptic formation, transcriptional regulation, beta-catenin signaling, and chromatin remodeling have all

been repeatedly implicated by rare variants in ASD probands (De Rubeis et al., 2014a; O'Roak et al., 2012; B. J. O'Roak et al., 2014).

A further complication in assigning candidate genes for ASD via variant discovery is found in the overlap between candidate genes associated with different neurodevelopmental disorders. For example, Sodium Voltage-Gated Channel Alpha Subunit 2 (SCN2A) has been implicated in ASD and epilepsy, damaging *de novo* variants in Chromodomain Helicase DNA Binding Protein 8 (CHD8) and Autism Susceptibility Gene 2 (AUTS2) have been found in ASD and schizophrenia, and Forkhead Box P1 (FOXP1) is a candidate gene for ASD and ID (Ben-Shalom et al., 2017; Hamdan et al., 2010, p. 1; S. E. McCarthy et al., 2014). This overlap makes it at times difficult to disambiguate variants and their specific effects, particularly in cases where ASD is found comorbid with another neurodevelopmental disorder.

As I have described, while genomic research has made substantial progress towards uncovering the genetic etiologies of ASD, there is much left to uncover. Currently it is estimated that approximately 20% of ASD cases can be (reasonably conclusively) assigned a genetic cause, and this fraction is higher for cases with comorbid intellectual disability (Jeste & Geschwind, 2014; Neale et al., 2012; Tammimies et al., 2015). At the same time it is estimated that there are hundreds of ASD genes, while only approximately 50 genes are currently considered having strong evidence (Abrahams et al., 2013; Chahrour et al., 2016). More research is necessary to fully discover the breadth of ASD causal variation in the entire population, as well as to deliver satisfactory answers regarding the cause of any individual case of ASD. This motivates the first research question of this thesis: what genetic variation causes ASD?

In the case of ASD and many other genetic disorders, the goal of many genetics efforts is to identify potentially pathogenic variants in a cohort of probands. However, the sheer number of

variants occurring in each and every subject necessitates substantial filtering and / or ranking of variants according to discriminatory criteria in order to distinguish those variants more likely to be pathogenic from the large background of variants that are definitely benign. As discussed, many sources of information provide insight into a variant's potential to cause disease. Some of the most relevant criteria used to evaluate a variant's likelihood for ASD association are:

1. Mode of inheritance
2. Predicted coding consequence
3. Minor allele frequency in the human population
4. Gene function
5. Computational predictors of impact
6. Previous gene- or variant-level association with ASD or another NDD
7. Functional validation of variant effect

In this and subsequent chapters I will explain further the role each of the above criteria play in establishing variant candidacy. As previously discussed, ASD genomics researchers typically look for *de novo* rare, likely damaging variants in probands as potential candidates. Leveraging sophisticated variant prediction and population frequency evidence, in addition to the existing literature on implicated genes enables greater power with which to extract candidate variants from the huge number of variants unimportant to an individual's ASD. In Chapter 2, I report my work incorporating these sources of evidence to prioritize candidate variation in a cohort of whole-genome sequenced ASD probands.

With the growing amount of ASD genomic studies such as this, we are ever increasing our understanding of ASD candidate genes. We can use literature evidence of gene-level recurrence to prioritize those variants affecting genes that are more likely to be ASD-associated. Several

efforts have been made to categorically list reported variation in ASD probands, and disease variation in general. ClinVar, denovoDB, and NPdenovo are all web-accessible databases of disease-associated variation (Landrum et al., 2016; J. Li et al., 2015; Turner, Yi, et al., 2017). We built an in-house, harmonized database MARVdb (http://marvdb.msl.ubc.ca) to collect ASD variants reported in the literature. As of October 2017 we have collected over 12000 variants from 40 exome / WGS studies, and harmonized them using a computational pipeline. MARVdb is currently one of the most comprehensive databases of literature-derived ASD variation, and continues to grow in size (Rogic et al., Manuscript in preparation).

1.4 Functional characterization of candidate variants for Autism Spectrum Disorder

The expanding number of variants contained in databases such as MARVdb spans a variety of ASD candidate variation, in terms of genes as well as functional categories. This variety leads to a gap in our knowledge regarding the effects of variation: while it is often relatively straightforward to assign potential pathogenicity to LGD variants (LGD variants in likely haploinsufficient genes are likely to be deleterious), prediction of deleteriousness is not so clear in the case of missense variants. Computational predictors of impact generally correlate with functional effects of variants, but in some cases may give ambiguous, or even incorrect, assessments of pathogenicity. Furthermore, deleterious missense variants in a gene can be deleterious (and cause disease) via different mechanisms. An extreme example of this phenomenon was demonstrated by Ben-Shalom et al. (2017) while studying SCN2A variant effects on channel function. They found SCN2A missense variants implicated in epilepsy to be gain of function, whereas those implicated in ASD were found to be loss of function (Ben-Shalom et al., 2017). Similarly, population databases like ExAC demonstrate that most genes contain an abundance of common missense variants assumed mainly benign in population databases like ExAC – while rare

nonsynonymous variants in these genes can be extremely deleterious in disease cases. For these reasons, increasing amounts of research have gone into the functional characterization of missense variants for which the functional impact is unknown, also known as "Variants of Unknown Significance" (VUS). Figure 1-2 illustrates a hypothetical functional characterization assay, and the expectation of functional impact for several relevant categories of variants. More comprehensive functional knowledge regarding the range of effects of disease associated variants can increase our capacity to more accurately predict disease association, increase our knowledge regarding the function of a gene, and can provide insights into the underlying biological etiology of a disease.

Certain research efforts study the effect of a gene's potential variants by saturation mutagenesis: creating cDNA libraries coding for all possible amino acid substitutions in a protein, introducing these constructs into a model, and subsequent functional characterization of all the resulting amino acid changes. Majithia et al. studied peroxisome proliferator-activated receptor γ, a protein encoded by PPARG and implicated in Type 2 Diabetes and familial partial lipodystrophy 3 (Majithia et al., 2016). Such efforts are valuable explorations of the range of functional impact that missense variation can have on a particular protein, and therefore the implications these variants can have for human disease. However, using such an approach to comprehensively study candidate genes for a genetically heterogeneous disorder such as ASD would necessarily only study one or very few ASD-associated genes. Therefore, prioritization of variants to identify those more likely to be informative is necessary in order to be able to practically and efficiently study the potential range of functional effect of missense variation in multiple ASD candidate genes.

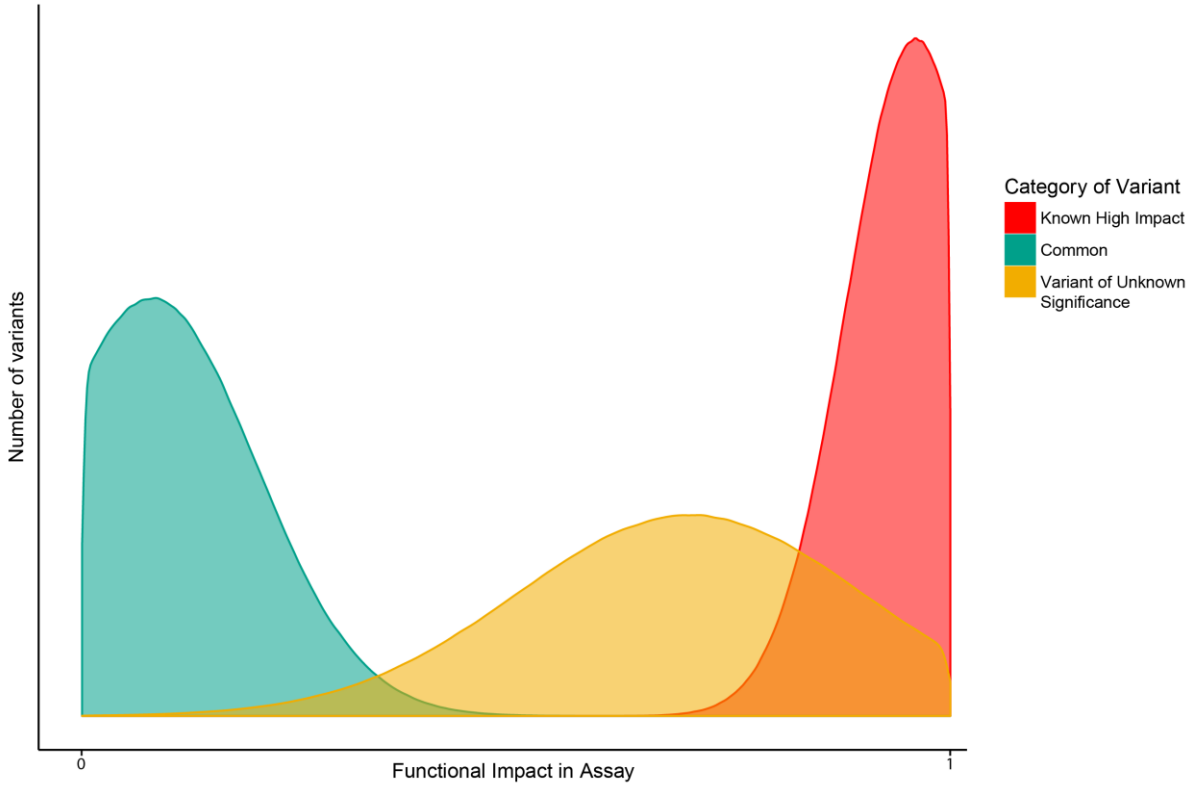

Figure 1-2: Functional impact of variants in a hypothetical functional characterization assay.
0 = Wildtype phenotype, 1 = Null phenotype. Many common variants derived from the human population would likely have very little (or zero) impact on phenotype in a functional assay, while known high impact variants (variants reported as high impact in previous functional characterization efforts or disease associated LGD variants) will likely have extremely high effects on function. We hypothesize variants of unknown significance will include high impact variants. A useful functional assay will provide appropriate discriminatory power between the distributions of effects of common variants and known high impact variants and will allow researchers to identify variants of unknown significance that have a significant functional effect.

Some genes are easier to study functionally than others, in regards to the functional characterization of missense VUSs. For this reason, gene selection is an appropriate first stage of functional characterization, and takes into consideration several factors. Genes with greater numbers of VUSs, greater depletion for missense variation in population studies and stronger prior evidence of ASD association clearly increase the expectations of informative results. A further factor for consideration is a gene's appropriateness for study in the functional assay used. Variants tested must display a significant range of functional effects by which to be able to discriminate high-effect variants from low-effect variants. Gene length, regulatory targets, and the requirement of other subunits for protein function are all additional considerations for designing a successful functional characterization assay.

Following gene selection, variants must be selected for functional characterization. In comprehensive saturation mutagenesis assays, every possible variant is tested. In other assays selecting for a subset of a gene's possible variants, literature and variant database review is undertaken to identify variants of interest. In order to calibrate interpretation of the functional characterization results, appropriate control variants must be identified, typically spanning a range of predicted effects. Variants are then functionally characterized in the model systems under study. Variants found to have significant functional effects (compared to wildtype controls) are thus validated as probable candidates, while nonsignificant effects may rule out a variant as a likely candidate for ASD.

1.5 Thesis outline

In this thesis I explore the two main research areas highlighted in the introduction. In Chapter 2, I describe my approach to identify possible ASD causing variants in a cohort of 119 ASD-diagnosed probands. To this end, I:

1. Identify variants in a cohort of ASD probands
2. Annotate and filter these variants on quality and MAF
3. Prioritize filtered variants on a per-subject basis according to a variety of computational predictors

Chapter 3 describes my approach to prioritize informative missense variation from the ASD literature and appropriate control variants for functional characterization in an array of model organism assays. In order to achieve this aim, I:

1. Identify genes of interest for functional characterization. Genes of interest have significant prior ASD association, as well as a substantial number of reported missense variants of unknown significance to test in biological assays.
2. Design a computational prioritization pipeline incorporating literature evidence of ASD association and computational prediction metrics of damage to annotate and prioritise variants of unknown significance and appropriate controls.
3. Computationally prioritize variants of interest in Phosphatase and Tensin Homolog gene (PTEN).
4. Computationally prioritize variants of interest in Synaptic Ras GTPase Activating Protein 1 gene (SYNGAP1).
5. Study the relationship between various computational prediction methods.

Chapter 4 explores insights from and limitations of these two research efforts, and proposes future research directions based on the conclusions of each.

Chapter 2: Whole genome sequencing and variant discovery in the ASPIRE Autism Spectrum Disorder cohort

2.1 Introduction

Autism spectrum disorder (ASD) is a largely genetic, albeit etiologically heterogeneous neurodevelopmental disorder (Krumm et al., 2015; Brian J. O'Roak et al., 2011). Rare, high-penetrance *de novo* variants have emerged as an important component of the genetic landscape of ASD. Whole-genome and -exome sequencing studies have identified candidate missense and Likely Gene Damaging (LGD) variants (including single nucleotide or indels resulting in nonsense mutations, frameshifts, and defects in canonical splice sites) occurring in genes involved with functions such as synaptic development and chromatin remodeling, among others (De Rubeis et al., 2014b; Krumm, O'Roak, Shendure, & Eichler, 2014; Pardo & Eberhart, 2007). However, the underlying genetic cause remains to be identified in many cases of ASD, and there is a continuing need for further discovery efforts.

In this study we prioritized candidate missense and LGD mutations on a per-subject basis in a cohort of whole-genome sequenced 77 ASD simplex and 42 ASD multiplex cases in order to discover LGD and missense ASD candidate variants.

2.2 Materials and Methods

Ethics approval for research involving human subjects was obtained through the joint Research Ethics Board of the University of British Columbia and Children's and Women's Health Centre of British Columbia. This study was conducted according to human subject research standards. Written informed consent for study was obtained from parents prior to study start.

2.2.1 Cohort Collection

This study is based on a cohort of 119 ASD subjects selected from a larger cohort of 318 ASD subjects recruited through the research registry of the Autism Spectrum Disorders – Canadian-American Research Consortium (ASD-CARC). The selection of the subset of 119 is described in the next section. All subjects were seen by a medical geneticist through the Provincial Medical Genetics Program located at the Children's and Women's Health Centre of British Columbia, where they underwent a comprehensive screening of medical systemic and morphological features. All subjects had normal gross karyotypes, and were further negative for targeted 22q11/22q13 and 15q11-q13 FISH, subtelomeric deletions, Fragile X and clinical chemistry screening (serum lactate, ammonia, creatine phosphokinase, lead, complete blood cell count and microscopy, uric acid, TSH, urine purine/pyrimidine and creatine metabolites). ASD diagnoses for all subjects were based on standardized Diagnostic and Statistical Manual, 4th ed (DSM-IV) criteria using Autism Diagnostic Interview-Revised (ADI-R) and/or Autism Diagnostic Observation Schedule-Generic (ADOS-G) standards (Lord et al., 2000, 1994). Each subject was evaluated by a clinical geneticist (Drs. Suzanne Lewis or Elena Lopez) blinded to the genetics findings.

2.2.2 Phenotyping

Our clinical collaborators selected 30 somatic and medical features (24 of which were craniofacial; see Supplementary Table 1 for collected phenotypes) with incidence between 5%-50% in the cohort and which had published evidence of increased frequency in ASD cases (Miles et al., 2008), and which were collected for each subject. K-means clustering and Discriminant (Canonical) Function analysis was used to identify clusters within the 246 subjects with complete feature data. Individuals from two major clusters (of 89 and 30 subjects) were selected from the

original cohort of 246, totaling a group of 119 ASD-diagnosed probands taken forward for sequencing. The two phenotype clusters significantly differed in their distributions of per-subject counts of observed phenotypes (Wilcoxon ranked-sum test, p-value = 0.00026, mean of number of positive phenotypes in cluster 1 = 5.9, mean in cluster 2 = 8.33). Other notable features not used in clustering are that (of the 75 subjects for which this information is available) 53 subjects were also diagnosed with a form of ID, ranging in severity from borderline to severe, 42 were diagnosed with mild to severe global developmental delay, and 17 subjects experience seizures. Subject demographics are presented in Supplementary Table 2.

2.2.3 DNA Collection

We collected blood samples from all 119 subjects, and from family members (parents and siblings) where available. Among the 119 subjects studied, 77 were from simplex families, of which we had full trio blood samples for 56. Forty-two subjects came from multiplex families. Of these, we were able to obtain blood samples from both parents and an affected sibling for 34 subjects. DNA was extracted from whole blood using the Puregene (Gaithersburg, MD, USA) DNA Isolation Kit.

2.2.4 Sequencing, Variant Calling, and CNV analysis

We sequenced the genomes of 119 probands with Illumina HiSeq 2000 paired-end (2x100 bp) sequencing. The average paired-end read count per sample was $1.12 * 10^9$. Sequencing depth ranged from 25.24 - 48.56 per base (mean 35.88; sequencing statistics are reported on a per-sample basis in Supplementary Table 3). We called variants using GATK 3.0 to obtain a variant call format (VCF) file containing single nucleotide variants (SNVs) and insertions / deletions (indels) for each proband (McKenna et al., 2010). Using multiple DNA microarray platforms, CNV analysis was performed on 79 cases (66% of subjects). We obtained CNV data for subjects using several CNV

platforms (Supplementary Table 4). The criteria for CNV classification was described previously (Qiao, 2013). Briefly, identified CNVs were screened for potential pathogenicity on the basis of size, gene content, and genome locus. CNVs were also screened for overlap with CNVs contained in the DGV and DECIPHER databases, with common CNVs with no known disease association ranking lowly for potential pathogenicity. Twelve subjects were found to have positive array results according to these criteria (Supplementary Table 4). The CNVs from these twelve subjects were then confirmed by fluorescent *in situ* hybridization (FISH). No subjects were excluded as a result of CNV analysis upon further review by our clinical collaborator.

2.2.5 Variant Filtering and Prioritization

Because our sequencing data set lacked parental DNA sequences, we could not directly identify *de novo* variation in our probands. Therefore, my prioritization approach was aimed at first identifying rare and predicted high-impact variants based on computational prioritization, and then screening these for those affecting genes with either existing evidence of association with ASD or another NDD or with plausible functions in neurodevelopment. For selected variants this was followed by determination of inheritance by targeted sequencing, when possible. The bioinformatics pipeline I developed is described in this section; an overview is given in Figure 2-1.

SNV and indel filtering: I used BCFtools (H. Li et al., 2009) to filter out SNVs and indels with low call quality or low read depth using the following thresholds: VQSR = PASS, QD > 10, Min(DP) ≥ 10, Avg(GQ) ≥ 40, and QUAL > 30. I then removed common variants according to population variation databases (MAF ≥ 0.1% in 1000 Genomes Project's "no known medical impact" variant list or in the ExAC database of 60 706 population samples)

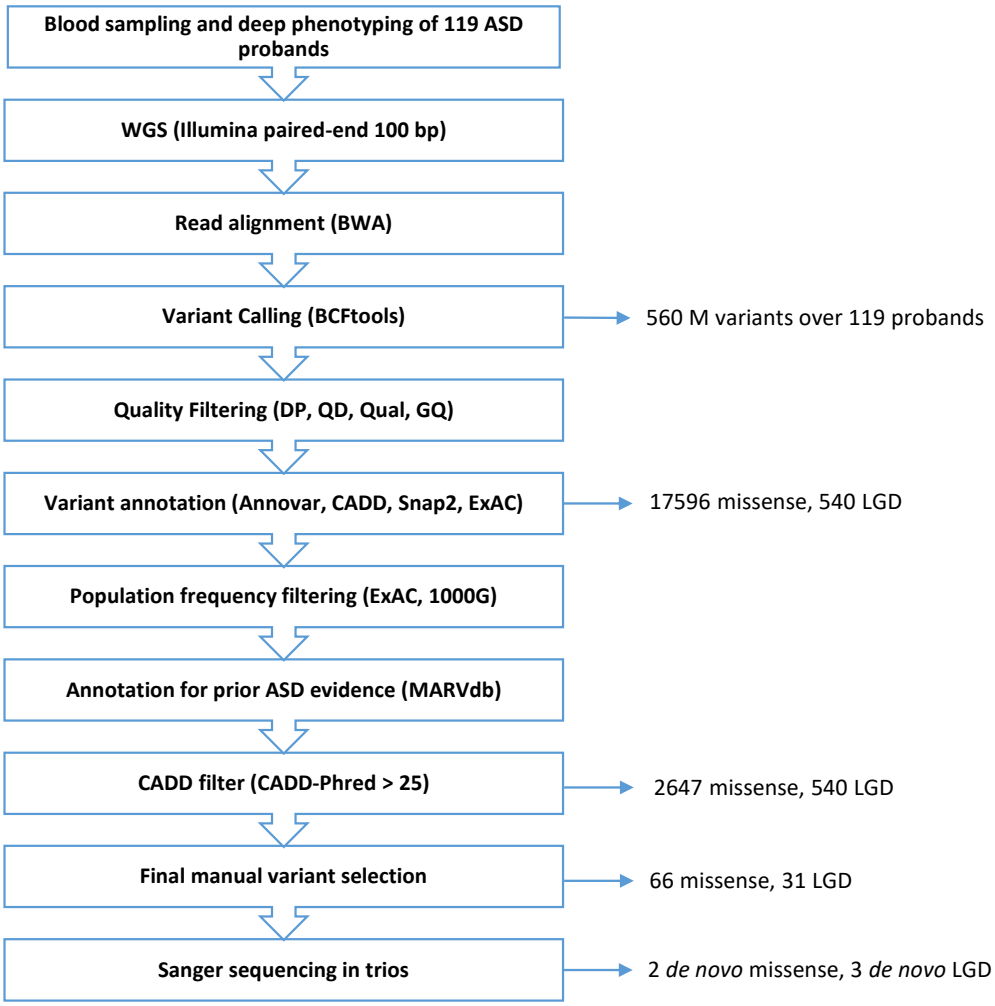

Figure 2-1: Workflow of computational variant prioritisation for likely causal ASD variation.
DP = read depth. QD = Quality score normalized by read depth. Qual = variant call quality. Avg(GQ) = average of genotype quality per base.

(Lek et al., 2016; The 1000 Genomes Project Consortium, 2012). Remaining variants were annotated with gene- and variant-level information (Gene name, transcript / amino acid change potential, CADD 1.2 score, ExAC frequency, occurrence in dbSNP) using Annovar (Kircher, Witten, Jain, O, et al., 2014; Lek et al., 2016; Sherry et al., 2001; K. Wang et al., 2010). I retained all indels predicted to be frameshift causing and all exonic SNVs and canonical splice-site affecting SNVs. I then filtered out any missense variants with CADD (version 1.2) Phred-scaled score < 25 (Kircher, Witten, Jain, O, et al., 2014).

These filtered variants were loaded into ASPIREdb (Tan et al., 2016; http://aspiredb.msl.ubc.ca/), our web-based interactive tool for analysis of genome/phenome datasets, to explore and evaluate candidate variants alongside proband phenotypes. I next ranked the retained variants for potential pathogenicity according to an array of established metrics. At this point, separate filtering criteria were used for subject's missense variants and for LGD variants. For each subject's remaining missense variants after quality and MAF filtering, I selected those with non-occurrence in the ExAC population database (Lek et al., 2016, p. 706), and in genes with evidence of constraint for missense mutation in ExAC (ExAC MSZ > 3). I then ranked the remaining variants according to their CADD Phred-scaled scores (Kircher, Witten, Jain, O, et al., 2014). I selected each subject's LGD SNVs and frameshift indels according to non-occurrence in ExAC, and in genes with ExAC pLI (probability of LGD intolerance) > 0.9, as such genes are likely to be intolerant to LGD mutation (Lek et al., 2016).

Variants were then further prioritized based on existing data on ASD genetics. I identified genes and variants previously reported to be associated with ASD. First we used MARVdb (http://marvdb.msl.ubc.ca/), our harmonised literature database of ASD-associated variants based

on 40 published sequencing studies (Rogic et al., in preparation). For the purposes of the current study, I filtered MARVdb variants according to the following criteria: *de novo* potentially damaging exonic variants (missense, nonsense or splice site) with ExAC frequency 0 and CADD score > 25. I then annotated our cohort variants with per-gene counts of reported variants. Similarly, we queried the MSSNG database of ASD WGS data (https://research.mss.ng) for *de novo* damaging variants using the same criteria and annotated our cohort variants with these per-gene variant counts. None of our priority variants were found to be previously reported in ASD probands.

The outcome of this prioritization is that for each subject, I selected up to two high priority candidate SNVs or indels for further consideration. I manually inspected the alignments of these candidate variants for artifacts using the Integrative Genomics Viewer (IGV) (J. T. Robinson et al., 2011). No variants were filtered out in this stage.

2.2.6 Sanger Resequencing

We validated prioritized variants via Sanger sequencing in the affected probands and all immediate family members for whom we had DNA. Primers flanking the variant site genomic location were designed using PrimerBLAST (Ye *et al.*, 2012) and ordered from Invitrogen (primer sequences given in Supplementary Table 5). Reactions contained 8μL Platinum PCR SuperMix High Fidelity (Invitrogen), 0.6μL each 5μM forward and reverse primers, and 12ng DNA. Cycling conditions were: 94°C for 2 min; 30 cycles of 94°C for 1 min, 58°C-64°C for 1 min, 68°C for 1 min; hold at 68°C for 5 min; hold at 4°C until taken out of machine. PCR products were visualized on agarose gel for correct size and quality. PCR products were sent to The Centre for Applied Genomics (SickKids, Toronto, ON) for Sanger Sequencing. Resulting chromatograms were

inspected visually using Finch TV v.1.4.0 (Geospiza Inc., Seattle, WA) to confirm the presence of the expected variant.

2.2.7 Burden analysis

In order to test for genetic disparities between the phenotype clusters initially used to select subjects for WGS, I tested for differences in the distribution of different classes of variants between subject clusters using Wilcoxon's rank sum test. I applied Fisher's exact test to 2x2 contingency tables of per-cluster counts of subjects presenting with different classes of variants. Correction for multiple comparisons was done using Benjamini-Hochberg control of the false discovery rate.

2.3 Results

2.3.1 Per-subject variant prioritization pipeline reveals likely causal variants

Because we did not have WGS data from the parents, our variant prioritization approach was designed to identify likely high-impact rare variants in protein-coding genes with at least some existing evidence of association with ASD or other neurodevelopmental disorders. Assessment of mode of inheritance was done *post hoc* with gene-targeted sequencing when parental DNA was available (Methods).

After selecting for rare (0% frequency in ExAC) LGD variation, and rare, likely damaging (CADD-Phred > 25) missense variation, 540 LGD SNVs, 1484 frameshift indels, and 2647 missense variants across the 119 probands remained. Prioritizing these variants on a per-subject basis for the top 1 or 2 variants per subject based on literature review and computational filtering identified a total of 31 LGD SNVs, 36 frameshift indels, and 68 missense variants with likely pathogenicity, distributed across 83 probands. Of these 83 probands we opted to test inheritance for the 42 probands containing SNVs of interest for which we had full trio DNA samples. We were therefore able to use targeted sequencing to test inheritance of 51 candidate SNVs identified in

these 42 families, comprising 15 LGD and 36 missense variants. Five were found to be *de novo*, all in probands belonging to simplex families (Table 2-1). No candidate variants were found to be *de novo* in the multiplex families we tested, nor did we find any candidate variants that segregated with the disorder.

While it was not possible to determine inheritance status of all prioritized potentially pathogenic variants due to lack of complete trio DNA, many prioritized variants occur in previously ASD- or other NDD-implicated genes (albeit, this is likely a consequence of our prioritization method). Table 2-2 lists other variants that rank highly according to pathogenicity metrics as well as literature evidence, but that we did not resequence (mostly due to lack of parental DNA). A complete list of prioritized variants is found in Supplementary Table 6.

2.3.2 Burden analysis

As our subjects were originally selected for sequencing on the basis of membership in the phenotype clusters (as described in Methods), it was of interest to examine whether any DNA variation characteristics correlated with cluster membership. I therefore applied burden analysis to several categories of variants (Table 2-3 and Table 2-4). Enrichment for indels in Cluster 2 reached a nominal p-value below 0.05 (Wilcoxon rank-sum test). No differences were considered significant at a false discovery rate (FDR) of 20%. While subjects in Cluster 1 were nominally enriched for hand-prioritized variants (Fisher's exact test: 66/89 subjects in Cluster 1, 16/30 subjects in Cluster 2, Wilcoxon rank-sum test p=0.03), this was not significant after correction for multiple testing (FDR 0.17) and the nominal significance did not hold when Sanger-verified inherited variants were dropped from analysis (Fisher's exact test: 38/89 in

Table 2-1: *De novo* candidate variants in the ASPIRE ASD cohort.
CADD 1.2 score is Phred-scaled. We prioritized rare (ExAC frequency = 0), predicted damaging (CADD > 25) missense SNVs and rare (ExAC frequency = 0) LGD SNVs for validation of inheritance. Five of these prioritized variants were found to be *de novo*.

SUBJECT ID	GENE	GENOMIC CHANGE (HG19)	VARIANT EFFECT	CADD 1.2
1543-22102	NIPBL	5:37059219-T/C	p.Leu2546Pro	29.8
21730-34469	SCN2A	2:166153564-G/A	p.Arg102Gln	34
21705-34281	SCN2A	2:166168534-G/A	splice acceptor variant	29
21974-35254	WDR45	X:48933022-C/T	splice donor variant	22.2
1885-23669	ARID2	12:46231491-G/A	splice donor variant	25.7

Table 2-2: Candidate variants of unknown inheritance in the ASPIRE ASD cohort.
CADD 1.2 score is Phred-scaled. The final column reports per-gene counts of de novo missense or LGD variants in ASD subjects as contained in MARVdb.

SUBJECT ID	GENE	GENOMIC CHANGE (HG19)	VARIANT EFFECT	CADD 1.2	# LGD VARIANTS
19275-30820	CUL3	2:225422467-T/C	p.Tyr64Cys	29	4
16260-27149	CACNA1B	9:140953048-G/A	p.Ala1446Thr	32	2
19283-30834	SPTBN2	11:66453910-T/C	p.Lys2237Arg	27.7	1
21758-34760	DGKD	2:234377170-C/T	p.Leu1176Phe	32	1
21758-34760	RAPGEF4	2:173787029-G/A	Splice donor variant	26	3
21879-36229	PLXNA2	1:208390189-A/G	p.Ile360Thr	25.5	1
1986-24087	ADNP	20:49508452-AC/A	p.Gly933Cysfs	-	17
20705-32533	ASXL3	18:31324700-CAAAA/C	p.Lys1631Argfs	-	3

Table 2-3: Cluster-specific differences in variants.

Student's T-Test (Welch Two-Sample t-test))	Cluster 1 (mean variants / subject)	Cluster 2 (mean variants / subject)	p-value	FDR
All Variants	4738128	4781556	0.1329	0.3363
Filtered Variants (exonic, MAF < 0.01)	142.98	158.93	0.2153	0.3747
Filtered Variants (exonic, NOT synonymous, MAF < 0.01, In MARV genes)	1.59	1.55	0.8548	0.9260
LGD SNVs	4.64	4.54	0.8141	0.9260
LGD SNVs (ExAC = 0)	1.70	2.21	0.08812	0.3341
Missense SNVs (exonic, MAF < 0.01)	138.65	153.80	0.2306	0.3747
Missense SNVs (ExAC = 0, CADD > 25)	4.67	5.67	0.1028	0.3341
Missense SNVs (ExAC = 0, CADD > 25, MSZ > 3)	0.61	0.40	0.1552	0.3363
Indels	236.33	243.07	**0.02621**	**0.1704**
Frameshift Indels	57.72	58.00	0.8075	0.9260

Bold numbers = significant at an FDR of 0.2.

Table 2-4: Differences in per-cluster individuals affected by damaging variation.

Fisher's Exact Test	Cluster 1 (N/89)	Cluster 2 (N/30)	p-value	FDR
Handpicked variants	67	16	**0.0227**	**0.1704**
Handpicked variants, NOT inherited	39	12	0.4415	0.6377
De novo variants	2	3	0.9859	0.9859

Bold numbers = significant at an FDR of 0.2.

Cluster 1, 12/30 in Cluster 2, p>0.05). Similarly, we did not find cluster-specific enrichment for the verified *de novo* variants (Fisher's exact test, 2 in cluster 1, 3 in cluster 2, p>0.05).

2.4 Discussion

Here we report the analysis of WGS of a cohort of 119 individuals with ASD, leading to the identification of variants potentially contributing to disease in several individuals. We were able to establish five predicted high-impact variants as *de novo* (Table 2-1). Other likely damaging, rare variants were established as inherited, and therefore of less certain significance. Our yield of 5 *de novo* predicted high-impact variants among 42 probands with parental DNA available (a yield of 9.5%) is a low estimate of the actual rate in the cohort, because variants were initially selected in the absence of parental DNA sequence. Despite this, the yield is comparable to that reported in a number of trio-based WES or WGS studies (Tammimies et al., 2015; Willsey et al., 2013). Robinson et al. (2014) observed a negative correlation between number of *de novo* LGD variants and IQ in the SSC (E. B. Robinson et al., 2014). As our cohort had a higher incidence of intellectual disability than the SSC, this may have increased our expectation for rate of *de novo* variants. Here, we discuss some of our specific findings in the context of the affected genes and the available phenotypic information. Genetic and phenotypic evidence in several subjects indicates the potential for additional clinical diagnoses, as discussed below. These subjects are currently being further evaluated clinically.

We found *de novo* damaging variants in Sodium Voltage-Gated Channel Alpha Subunit 2 (SCN2A) gene in two unrelated simplex cases: a predicted LGD (splice-site) variant in subject 21705-34281 and a missense variant in 21730-34469 (p.Arg102Gln) (Figure 2-2). SCN2A is one of the best-characterized ASD risk genes, with multiple LGD and missense *de novo* events

previously reported in ASD cases (De Rubeis et al., 2014b; Iossifov et al., 2014; Yuen et al., 2016). The variant in 21705-34281 is predicted to be a complete loss of function, as it disrupts the splice acceptor site between exons 6 and 7. It has been reported that some missense variants in SCN2A are gains of function (hypermorphs) associated with seizures rather than ASD (Ben-Shalom et al., 2017). Case 21730-34469 does not have seizures, and the location of the variant in the N-terminal domain is near the site of two ASD associated missense variants (D12N and D82G) shown by Ben-Shalom et al. to result in reduced channel function. Our findings are consistent with those of Ben-Shalom though functional testing will be necessary to establish the impact of the missense variant.

We found a missense *de novo* variant in simplex proband 1543-22102 in Nipped-B Homolog (NIPBL) (Figure 2-3), a gene associated with chromatin binding and implicated in 60% of Cornelia de Lange syndrome cases (Rohatgi et al., 2010). We noted that this proband has a number of clinical features characteristic of the syndrome: global developmental delay, short stature, intellectual disability, microcephaly, scoliosis, restricted movements of the joints, fifth finger clinodactyly, characteristic facial features (synophrys, long eyelashes, low-set ears and small upturned nose), and self-destructive behaviour. This subject is currently being further evaluated clinically for the possibility of Cornelia de Lange syndrome.

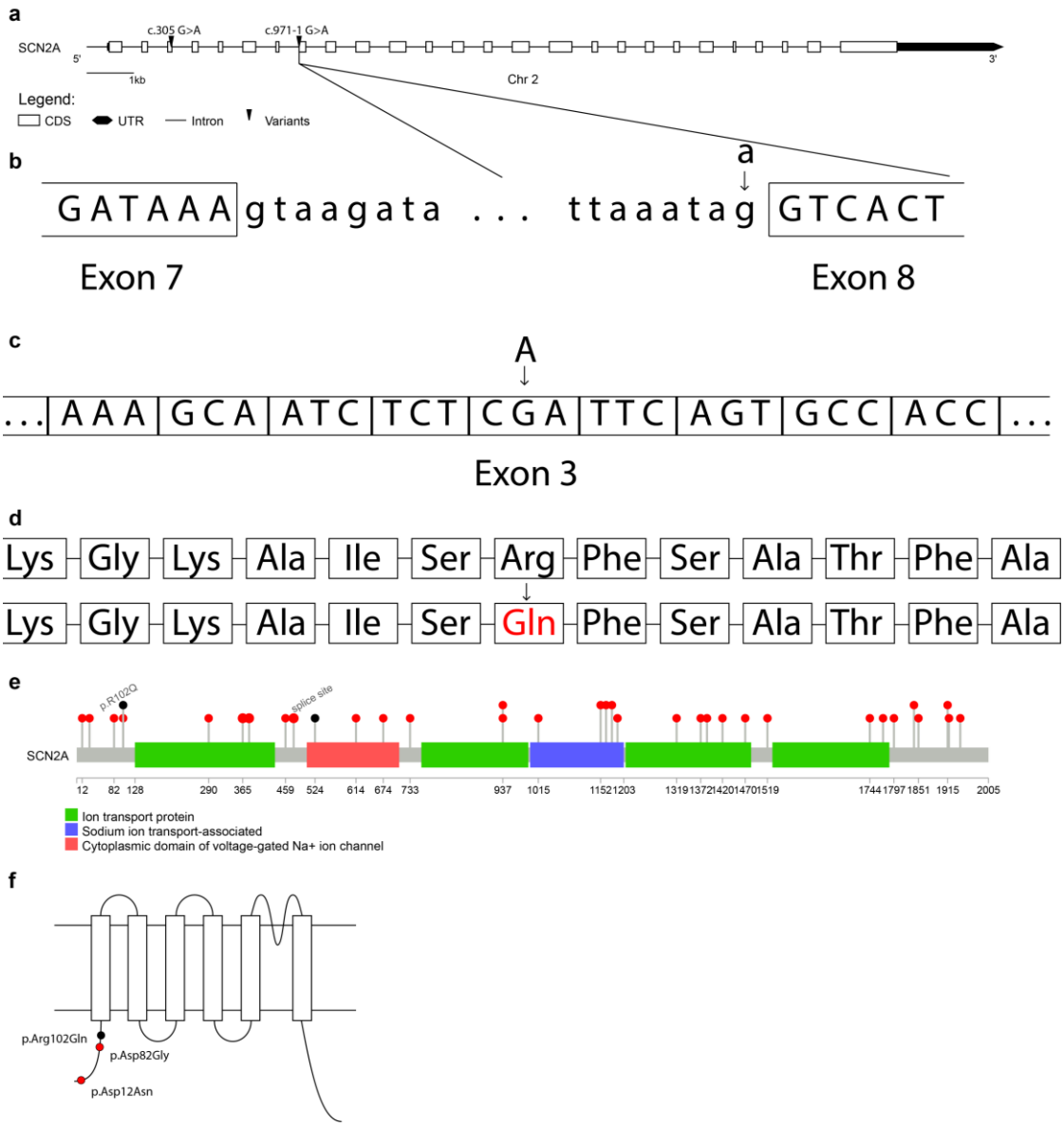

Figure 2-2: Two *de novo* variants in SCN2A.
a. Genomic locations of a predicted splice-site disrupting de novo variant (NM_021007.2:c.971-1G>A) in subject 21705-34281 and a de novo missense variant (NM_021007.2:c.305G>A) in subject 21730-34469. **b.** The splice site variant in 21730-34469 is predicted to disrupt a canonical splice acceptor site. **c., d.** The nonsynonymous variant in 21705-34281 causes an amino change (p.Arg102Gln). **e.** Reported damaging SCN2A variants in MARVdb. **f.** Location of our variant (black) in comparison to two *de novo* missense variants (red) previously reported in Domain I of SCN2A (Ben-Shalom et al., 2017). Figure based on Figure 1 of Ben-Shalom et al. (2017)

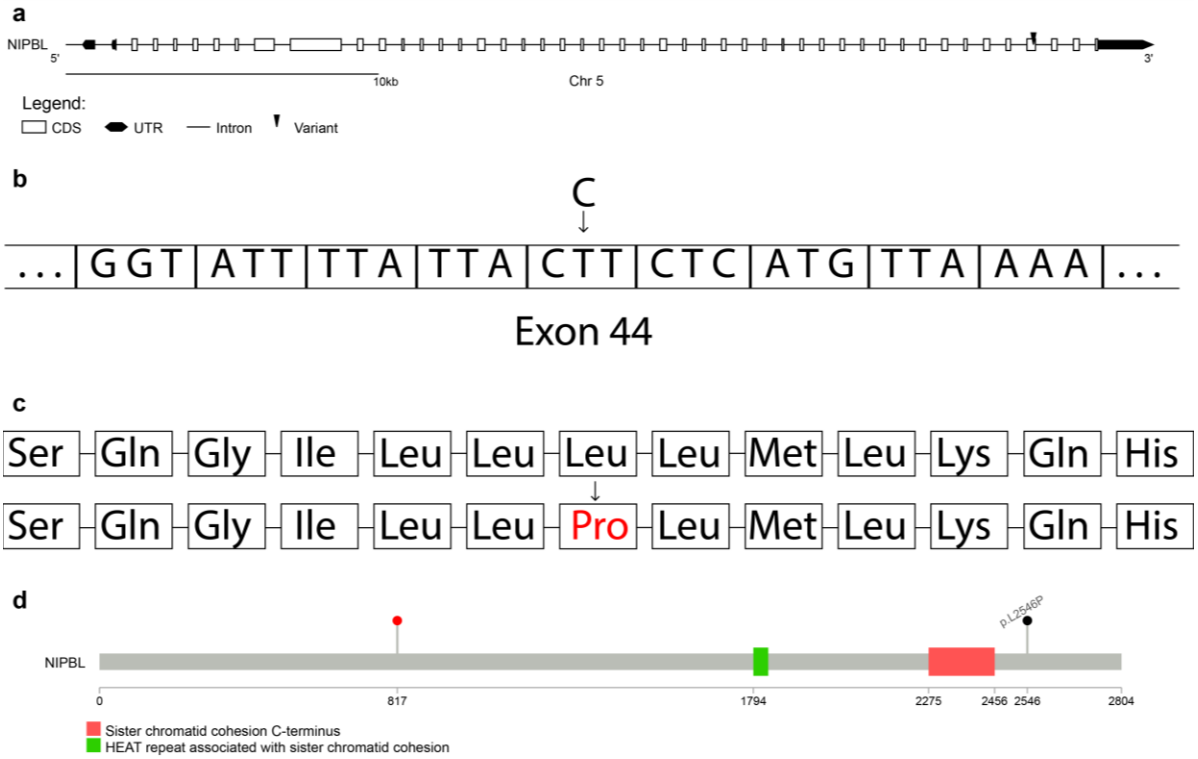

Figure 2-3: A *de novo* missense variant in NIPBL.

a.,b. Genomic and exonic locations of a *de novo* nonsynonymous NIPBL variant (NM_015384.4:c.7637T>C) in subject 1543-22102. **c.** The variant causes an amino change (p.Leu2546Pro.). **d.** Red: another reported NIPBL variant in MARVdb. Black: our *de novo* missense variant.

We found a *de novo* splice site variant in simplex proband 21974-35254 in WD Repeat Domain 45 (WDR45) gene (Figure 2-4), which has been strongly associated with Beta-propeller protein-associated neurodegeneration, a disorder characterized by an early-onset global developmental delay and further neurological deterioration by early adulthood (Haack et al., 2012). The proband exhibits the following phenotypes that have been associated with WDR45 mutations: EEG abnormality, spasticity, ataxia, speech impairment, drooling and diminished pain sense. A majority of reported WDR45 mutations were *de novo* and many of them were splice site mutations (Haack et al., 2012; Hayflick et al., 2013; Hoffjan et al., 2016). Published reports suggest that WDR45 mutations are associated with a broader phenotypic spectrum, ranging from epileptic encephalopathy, Rett (-like) and West syndrome to only mild cognitive delay (Hoffjan et al., 2016; Long et al., 2015; Nakashima et al., 2016; Ohba et al., 2014; Redon et al., 2017). This patient is currently being further clinically investigated for any of the above symptoms associated with WDR45 variants.

We identified a *de novo* splice site variant in simplex proband 1885-23669 in AT-Rich Interaction Domain 2 (ARID2) gene (Figure 2-5), which is a component of the SWI/SNF chromatin-remodeling complex. Other subunits of this complex have already been implicated in intellectual disability and related neurobehavioral disorders and there are recent reports providing evidence for the similar role of ARID2. Shang et al. (2015) identified four individuals with predicted loss of function in ARID2. All four mutations result in frameshift/truncated proteins and cause the loss of the two conservative zinc finger motifs. Similar to these reports, the mutation in the proband falls proximal to the two zinc finger motifs. All four patients are reported to display global developmental delay and intellectual disability (Shang et al., 2015). While the proband in our study does not exhibit global developmental delay, she does have

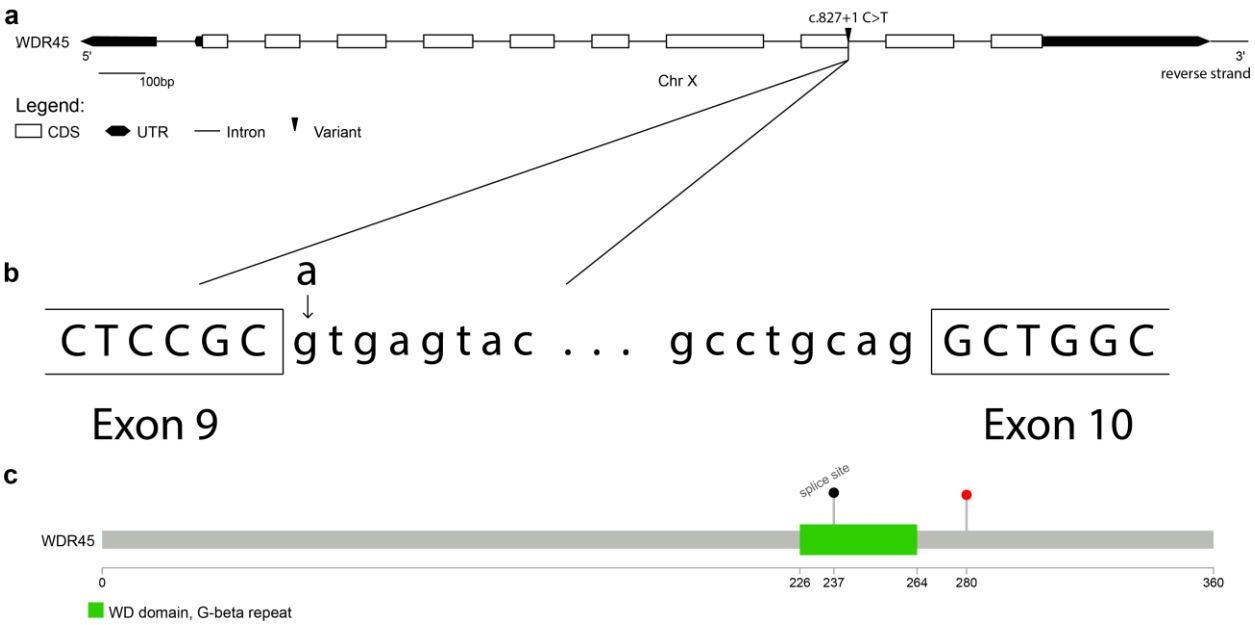

Figure 2-4: A *de novo* splice donor variant in WDR45.
a.,b. Genomic and exonic locations of a *de novo* splice donor WDR45 variant (NM_001029896.1:c.827+1G>A) in subject 21974-35254. **c.** Reported WDR45 variants in MARVdb.

Figure 2-5: A *de novo* splice donor variant in ARID2.
a.,b. Genomic and exonic locations of a *de novo* splice donor ARID2 variant (NM_152641.2:c.1330+1G>A) in subject 1885-23669. **c.** Reported ARID2 variants in MARVdb.

borderline intellectual disability and several other phenotypes reported in Shang et al: moderate communication impairment, hyperactivity and inability to concentrate, coarsening of facial features and hearing loss. Another study by Bramswig et al. (2017) identified de novo ARID2 frameshift mutations in two individuals with Coffin-Siris syndrome neurodevelopmental disorder. Both individuals present intellectual disability and coarsening of facial features.

Literature evidence supported the potential relevance of several of the high-ranking variants in probands for whom we did not have both parental DNA samples available (Table 2-2). I now discuss these variants. While inheritance cannot be determined in these subjects, these variants represent likely primary candidates to validate for these subjects, should their parental DNA be available in the future.

Several *de novo* LGD variants have been previously found in ASD subjects in Cullin 3 (CUL3), a gene involved in histone ubiquitination (Iossifov et al., 2014; Kong et al., 2012). Codina-Sola et al. (2012) found a likely damaging *de novo* missense CUL3 variant in an ASD case. These cases provide supporting evidence for the implication of the missense variant we find in 19275-30820. Conversely, the identification of this missense variant also further implicates CUL3 as an ASD candidate gene.

We found a rare missense variant in subject 16260-27149 in the Calcium Voltage-Gated Channel Subunit Alpha 1B (CACNA1B) gene, which codes for the pore-forming subunit of the N-type calcium channel (Cav2.2). ASD association of this gene is has been previously suggested by two *de novo* missense variants previously reported in ASD cases (De Rubeis et al., 2014b; Yuen et al., 2016). Yatsenko et al. (2012) reported on seven individuals found to carry monogenic duplications encompassing the CACNA1B gene. Four of these children with monogenic CACNA1B duplications were diagnosed with ASD.

A missense variant in subject 21758-34760 in Diacylglycerol Kinase Delta (DGKD) ranks highly in our analysis, DGKD has only been previously implicated in ASD as containing a *de novo* frameshift variant (De Rubeis et al., 2014b). Our patient presents with seizures, severe communication impairment, moderate ID, and moderate global developmental delay. Leach et al. (2007) found a patient with seizure and developmental delay phenotypes to have their DGKD gene disrupted by a balanced translocation. In the same subject, a rare splice donor variant was found in Rap Guanine Nucleotide Exchange Factor 4 (RAPGEF4). *De novo* variants have been previously reported in RAPGEF4 in ASD probands by Iossifov (2014) and De Rubeis (2014). RAPGEF4 codes for *Exchange Protein Directly Activated by cAMP 2* (EPAC2), and is linked to ASD and affects dendritic maturation (Martínez-Cerdeño, 2017; Srivastava et al., 2012). Bacchelli et al. (2003) identified four rare nonsynonymous variants in RAPGEF4. These variants were present in five families, where they segregated with the autistic phenotype, and were not observed in control individuals.

We found a rare missense variant in subject 19283-30834 in Spectrin Beta, Non-Erythrocytic 2 (SPTBN2) gene. De Rubeis et al. (2014) similarly reported a *de novo* missense variant in SPTBN2 in an ASD subject. Variants in SPTBN2, a gene coding for a subunit of a membrane-cytoskeleton component, have been previously implicated in spinocerebellar ataxia but mode of inheritance was found to be recessive (Elsayed et al., 2014). Our proband has spasticity, infantile muscular hypotonia, mild global developmental delay. Muscular phenotypes such as these, as well as global developmental delay, have all been previously observed in ataxia patients (Elsayed et al., 2014; Yıldız Bölükbaşı et al., 2017).

While *de novo* LGD variants have not been previously reported in the literature for Plexin A2 (PLXNA2), our prioritization of the variants found in the MSSNG database indicate a

candidate variant considered *de novo* by WGS. PLXNA2, coding for a semaphorin receptor, has been previously reported as a potential candidate by GWAS in schizophrenia (Mah et al., 2006).

I report the computational prioritization of variants in a cohort of 119 ASD probands for the purposes of ASD variant discovery. Five predicted damaging variants were found to *de novo*, all occurring in genes with prior ASD or other NDD association. Eight more damaging variants are considered as ASD candidate variants, occurring in probands for which the ascertainment of inheritance was not possible. While functional characterization of these variants could greatly add weight to our assessment of their ASD association, at present these variants add to the ASD literature base of damaging variation, and improve our understanding of the genetic provenance of ASD.

Chapter 3: Variant prioritization for functional characterization in model organisms in the SFARI Autism Spectrum Disorder collaboration

3.1 Introduction

While many variants have been identified in ASD probands by WGS / WES efforts similar to the one described in Chapter 2, many are missense variants of unknown significance (VUSs). Variants of unknown significance are those for which the functional impact is unknown and difficult to predict accurately by computational methods. The genes these VUSs are in are often already ASD candidates, although sometimes have limited (or zero) prior ASD association. Furthermore, the effects of missense variants within a gene can differ, depending on the location and precise nature of the amino acid change. For example, missense variants found to cause hypermorphic (increase in function) phenotypes in SCN2A have been implicated in epilepsy, whereas those found to have hypomorphic (decrease in function) phenotypes have been implicated in ASD (Ben-Shalom et al., 2017). Without knowledge of the effect of a particular VUS, to propose it as the genetic cause of a clinical diagnosis is to use inconclusive evidence. Therefore, characterizing the functional effect of VUSs is of key interest as it provides much more definite answers to a variant's potential for pathogenicity than computational predictions alone. In addition, the functionalization of multiple variants in a gene increases our understanding of the range of mechanisms by which a gene can be involved in human disease.

Some functional characterization efforts have attempted to functionally characterize the impact of every potential amino acid change of a protein. Majithia et al. reported such "saturation characterization" in the case of peroxisome proliferator-activated receptor γ, a protein encoded by PPARG and implicated in Type 2 Diabetes and familial partial lipodystrophy 3 (Majithia et al.,

2016). Such efforts are valuable explorations of the range of functional impact that missense variation can have on a protein, and the implications these variants can have for human disease. However, it is not always feasible to study all possible variants in all possible genes, particularly in a genetically heterogeneous disorder such as ASD. In these cases, variant prioritisation is necessary in order to focus on the most relevant ones. The number of genes implicated in ASD, combined with the number of potential variants to test necessitates a scalable, high-throughput, gene-independent approach to studying missense variants of unknown significance. To that end, we constructed a multi-level pipeline to functionally characterize ASD missense VUSs in an array of model organism systems (Figure 3-1). This project and resulting pipeline was developed in response to the Simons Foundation Autism Research Initiative's (SFARI's) Functional Screen of Autism-Associated Variants request for applications, announced in 2015 (SFARI, 2016). In this chapter, I describe my work in the context of this pipeline to provide computational prioritization of variants of unknown significance in two ASD candidate genes, PTEN and SYNGAP1, for functional characterization in model organism assays.

Phosphatase and Tensin Homolog (PTEN) is a tumor suppressor gene coding for Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase. PTEN is a negative regulator in the mTOR/PI3K pathway (Figure 3-2). Disruptions in this pathway have been shown to lead to alterations in synaptic growth (Barrows, McCabe, Chen, Swann, & Weston, 2017; Bourgeron, 2015). Germline variants in PTEN have been implicated in PTEN hamartoma tumour syndrome, Cowden Syndrome, and Bannayan-Riley-Ruvalcaba syndrome (Hansen-Kiss et al, 2016; Nelen MR et al, 1997; Butler et al, 2004). Patients with these symptoms present with a variety of tumour phenotypes, in addition to greater incidence of ASD and macrocephaly (Butler et al., 2005). Previous WGS/WES efforts have identified several

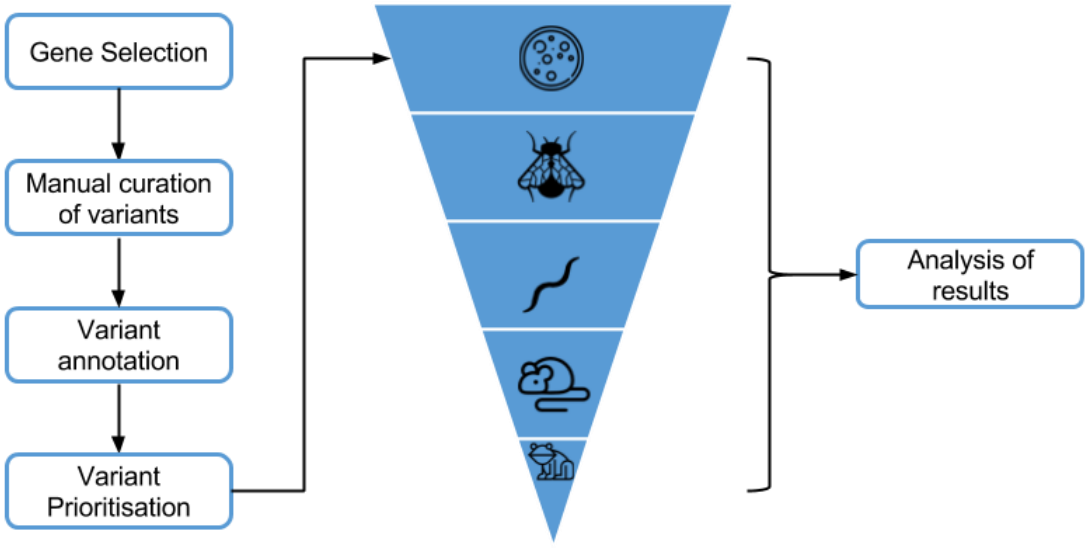

Figure 3-1: Our multi-platform pipeline for functionally characterizing ASD gene variants.
Through a combination of manual curation (of ASD variants of unknown significance and previously studied variants), computational variant annotation and computational variant prioritization steps we amass a list of variants of interest for functional characterization in genes with strong prior ASD evidence. We functionally characterize these variants in high-throughput model organism assays (Synthetic genetic lethal assays in Saccharomyces cerevisiae, development assays of Drosophila melanogaster, and behavioural assays in Caenorhabditis elegans). We then analyze these results to further prioritize variants for testing in our lower-throughput model systems (Neuron morphology and connectivity in Rattus rattus hippocampus cultures, neural growth assays in Xenopus laevis)

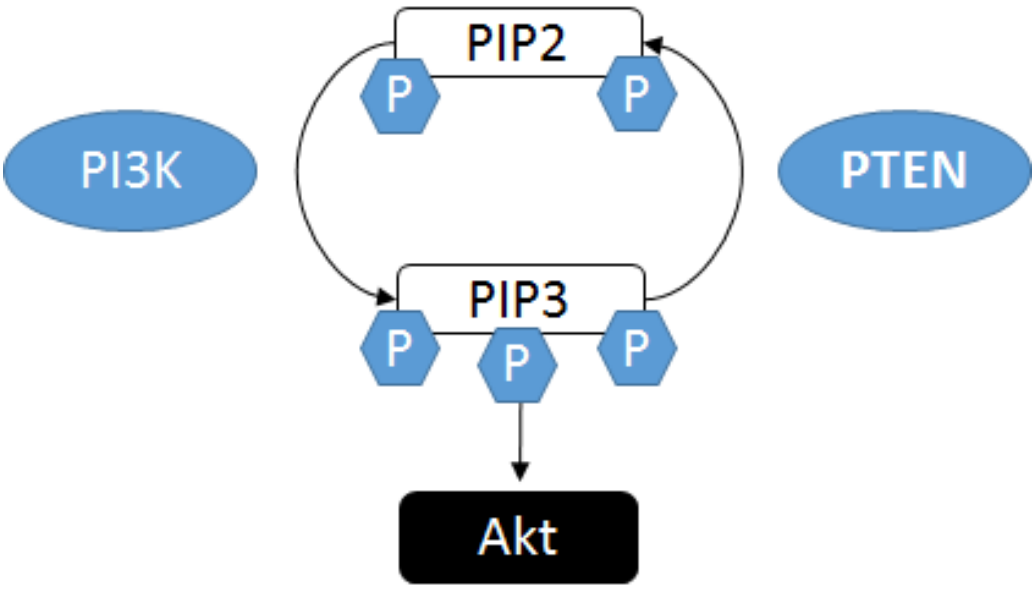

Figure 3-2 PTEN's role in the P13K / AKT pathway.
PTEN negatively regulates the PI3K / AKT pathway. PTEN works as a lipid phosphatase to directly oppose PI3K's phosphorylation of phosphatidylinositol (4,5)-bisphosphate (PIP2) to create Phosphatidylinositol (3,4,5)-trisphosphate (PIP3). Inactivation of PTEN causes activation of Akt by way of increased levels of PIP3. Activation of Akt has several downstream consequences, including involvement in axon growth and cell survival.

de novo LGD variants and *de novo* nonsynonymous variants in ASD probands (Iossifov et al., 2014; Kosmicki et al., 2017; B. J. O'Roak et al., 2014; Schaaf et al., 2011; Yuen et al., 2016). PTEN is considered a "High Confidence" ASD gene by SFARI which states that high confidence genes "…require evidence of recurrent and convincing mutations accompanied by a rigorous statistical comparison with the mutation frequency in controls, confirmed via independent replication" (Abrahams et al., 2013). Previous functional characterization has found that some ASD-associated PTEN variants affect the membrane binding and phosphatase activity of PTEN. (Redfern et al., 2008; Rodríguez-Escudero et al., 2011; Zhang, Piccini, Myers, Van Aelst, & Tonks, 2012).

SYNGAP1 is a gene coding for synaptic Ras GTPase activating protein 1, involved in postsynaptic signaling. It has been shown that SYNGAP1 variants in mice affect cognitive development by interfering with dendritic spine synapse maturation; causing increased hippocampal neuron excitability (Clement et al., 2012). Variants in SYNGAP1 have been previously implicated in cases of intellectual disability, ASD and epilepsy (Carvill et al., 2013; Hamdan et al., 2011). *De novo* LGD and nonsynonymous variants have been identified in ASD probands in multiple WGS/WES studies (De Rubeis et al., 2014b; Iossifov et al., 2014; Kosmicki et al., 2017; B. J. O'Roak et al., 2014; T. Wang et al., 2016). SYNGAP1, like PTEN, is also listed by SFARI as a "High Confidence" ASD gene (Abrahams et al., 2013).

Prioritization of variants to identify those more likely to be informative is necessary in order to be able to practically and efficiently study the effective range of functional effects of ASD missense variation in genes like PTEN and SYNGAP1. Characterizing high-priority variants gives insight into biological mechanisms of ASD and can provide information to enhance the performance of computational predictors of ASD-implicated mutation. While there are numerous

ways to collect and categorize variants for functional characterization, it can be useful to think of them in three main groups: high impact variants, low impact variants, and missense variants of unknown significance.

High impact variants are any variants for which we hypothesize having a large impact on the phenotype under study. These variants include stop-gain or frameshift variants previously reported in ASD patients or in another disorder. As these variants cause early truncation of a protein, we can generally be reasonably sure of an effect on function from these variants. Another source of high impact variants comes from previous functional characterization efforts. Variants previously found to have a functional impact in a biochemical, biological, or other experimental assay are useful additions to our assays, as they have prior experimental evidence of functional impact. The functional characterization of these variants offers context into the validity of the assays themselves, and can serve to calibrate our expectations of functional effect of the other variants under study. Finally, computational prediction methods allow the selection of artificial variants (variants that have not been previously observed in human genomes) that are predicted to be higher effect by computational tools such as CADD. Such variants will likely result in higher functional effects in our model organism assays. By increasing the range and resolution of functional effects characterized in our assays, we increase our power to discriminate the effect of our missense variants of unknown significance.

Low impact variants are those that we hypothesize will behave similarly to the wild-type gene in the model organism assay. These variants include common variants (MAF > 0.1%) from population databases – these variants ones that are unlikely to be highly deleterious due to their frequency in the human population. We expect functional characterization will reveal these variants to have a weaker effect on the phenotypes examined in our assays. Functional

characterization efforts can also be a source of variants previously identified to have little or no impact on function in a certain assay. As in high impact variants, we can prioritize artificial variants that are not expected to have significant effect on phenotype.

The final broad category of variation we test is of missense variants of unknown significance. Characterizing these variants in the context of the previously mentioned categories of variants affords us better resolution to distinguish truly deleterious from truly benign variants. As the primary goal of this project was to characterize missense VUSs found in ASD probands, the majority of these variants are found in databases such as MARVdb, or in the broader WGS/WES ASD literature. The characterization of these variants provides us with insights into which reported variants are likely truly pathogenic, and may provide information on the biological mechanisms of pathogenic variation. Missense VUSs from other disorders can also be found, as many genes have more than one associated disorder. Functionally characterizing variants reported from NDD or other disease contexts allows more comprehensive study of the effects of pathogenic variation in the gene of interest.

We functionally characterized variants from each of these categories in yeast (Saccharomyces cerevisiae). In yeast, only a small proportion of human genes can be studied by simple knockout-rescue assays, where the human orthologue is knocked out. The human copy is then introduced into the knockout strain, which is functionally analyzed. Most human genes do not have an orthologue in yeast - a new strategy must be adopted in order to study the remainder of human genes. Therefore, we adopted a Synthetic Genetic Array (SGA) assay in yeast (Saccharomyces cerevisiae) to functionalize prioritized variants. SGA analysis explores the effects of genetic interactions using double mutant strains (Boone, Bussey, & Andrews, 2007). Our strategy looked for growth phenotypes stemming from gene interactions between a deletion strain

and our modified gene of interest (in this case, mutated human PTEN) (Nijman, 2011). While in our complete functionalization pipeline we incorporate additional model organism assays (in *Caenorhabditis elegans, Drosophila melanogaster, Xenopus laevius,* and *Rattus rattus* hippocampus neuronal cultures), at the time of this writing data are not available for these. Therefore, for the purposes off this thesis I discuss mainly the results of our computational prioritization of variants in the context of functionalization in *S. cerevisiae.*

An obvious concern regarding using computational predictors of deleteriousness to prioritize variants for functional characterization concerns the accuracy of these computational metrics. Related to the accuracy is in the agreement between various tools. Previous research into the pathogenic / benign variant discriminatory power of variant prediction methods, as well as their agreement with each other, has typically yielded favorable results (Dong et al., 2015; Kircher, Witten, Jain, O'Roak, et al., 2014; Kumar et al., 2009; Liu, Wu, Li, & Boerwinkle, 2016). However, both of these arcs of research have tended to be on a genome-wide basis. In contrast, we demonstrate later in this chapter that the usual correlation of variant predictors fails in certain cases, when comparing them gene-by-gene. As many variant prioritisation and functionalization efforts operate under the context of studying a particular gene, it is important to know for which genes which computational predictors will behave consistently with each other, and when they will not. To our knowledge, this relationship has not been previously studied on a per-gene basis. To this end in this chapter we present a novel gene-wise analysis of the agreement of four popular computational metrics tools (SIFT, Polyphen2, CADD, and Snap2).

3.2 Methods

3.2.1 Informing gene selection based on comprehensive gene annotation

I obtained per-gene counts of ASD missense variants found in MARVdb, our in-house database of variation from the ASD literature (marvdb.msl.ubc.ca). I annotated all genes with ExAC MSZ and pLI, which assign each gene a score indicating its depletion in the human population for missense variants and LGD variants, as introduced earlier. I scaled the MSZ scores between 0-1 to make them directly comparable to pLI. I then ranked each gene represented in MARVdb by the following formula:

(# of *de novo* missense variants in MARVdb) * scaled MSZ + (# of *de novo* LGD variants in MARVdb) * pLI

Top genes on this list therefore ranked highly both on number of missense variants and/or LGD variants to test as well as on general depletion for the relevant categories of damaging variation in the human population (Table 3-1). Final gene selection took into consideration this ranking as well as other criteria, including but not limited to appropriateness for study in model systems, prior research volume, and gene-specific functional knowledge. We selected PTEN for the preliminary round of functional characterization, followed by SYNGAP1.

3.2.2 Development of a computational variant annotation and prioritization pipeline

In order to ensure consistency in subsequent variant prioritization efforts of our collaboration, I created a computational pipeline in R to enable rapid and consistent variant annotation and prioritization and inform variant selection for genes of interest (Figure 3-3). The pipeline works according to several options, which can be changed in the *config.R* script: gene name, chromosome number, transcript identifier from RefSeq (NM_ accession), and the location of a query file of manually curated variants. The query file is a tab-delimited file identifying

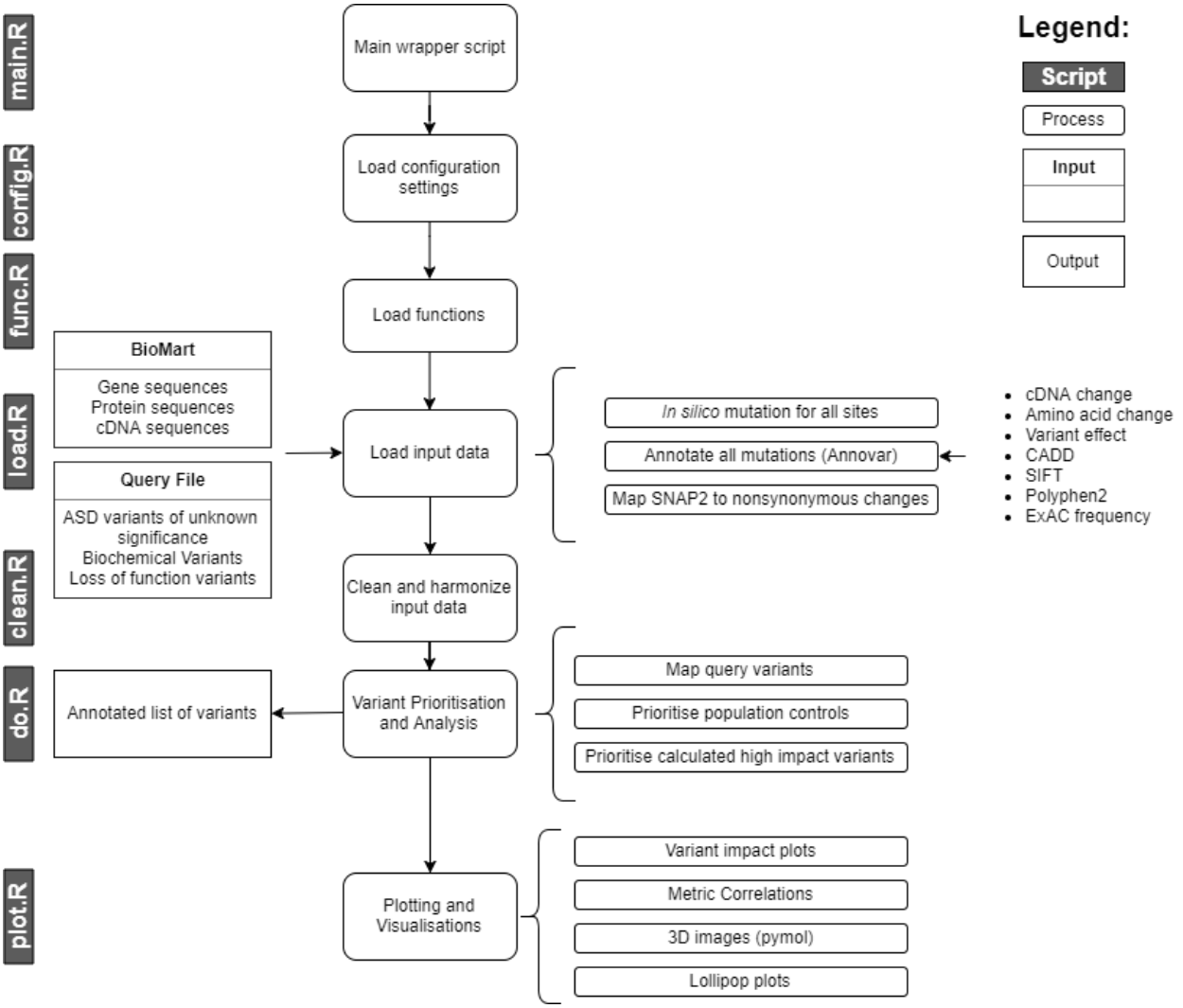

Figure 3-3: The computational workflow of variant prioritization for functional characterization.

variants selected during the manual curation process (this process is described below in the context of the manual curation stage for PTEN). Running *main.R* loads dependencies and runs the rest of the scripts.

As part of the pipeline, I annotate all exonic variants of the gene of interest with CADD, SIFT, Polyphen2, and ExAC population frequency using Annovar (November 14, 2012 release). I annotated variants with SNAP2 scores manually as Annovar's included databases do not include SNAP2. Further variants were selected according to several categories of interest, using filtering steps developed in R:

3.2.3 Functional Characterization of PTEN missense variants

3.2.3.1 Collection of ASD variants of interest and previously characterized variants

I collected ASD missense variants from published WGS / WES studies. The majority of these were found in MARVdb. I found additional ASD missense variants of unknown significance from a general literature search and in ClinVar's list of pathogenic variants (in addition to several cancer syndrome associated variants). Several ASD-associated stop-gain variants were selected to test the effect of early truncation of the protein. Additional NDD-associated variants were found in various web-based databases.

I additionally selected previously biochemically tested PTEN variants to inform interpretation of functional characterization results. All manually curated variants were computationally annotated as previously described. Additional variants were selected for functionalization as described below.

3.2.3.2 Computational prioritization of additional genomic variants

The computational pipeline I developed allows the annotation and subsequent prioritization of additional categories of variants:

Population-derived variants: I filtered all exonic PTEN variants to select common (ExAC MAF > 0.1%), predicted low impact (CADD < 15) variants.

Computationally-predicted high-impact SNVs: I filtered all exonic PTEN variants for rare (ExAC MAF = 0), predicted high damaging (CADD > 25 & SNAP2 > 75) variants to select predicted high-impact SNVs.

A complete list of PTEN variants used in this study is displayed in Appendix A.1. We later used the same criteria for SYNGAP1 variant selection as we did for PTEN variant selection.

3.2.3.3 Testing synthetic lethality of PTEN ASD variants in yeast

Our collaborators used a Synthetic Genetic Array (SGA) screen to assess functional effects of prioritized variants. They first transformed 5000 yeast mutant strains with human PTEN and measured the difference in colony size compared to an empty vector. Of the 5000 strains, 8 "sentinel" strains were identified - FIG4, VAC14, VAC7, VAM3, VAM7, YPT7, VPS38, and VPS30, as the deletion strains that displayed the highest sensitivity (in terms of growth reduction phenotype) to human PTEN overexpression. These sentinels were then used to design a "miniarray" that allowed higher-throughput screening, obviating the need to screen all 5000 strains. Our collaborators then engineered our prioritized variants into a human PTEN expression construct and transfected the previously identified sentinel strains with this mutated copy of PTEN. We measured the change in growth of PTEN variant double mutants compared to the wild-type PTEN double mutants. The sensitized yeast strains incorporating wildtype PTEN displayed a decrease in growth as compared to wildtype yeast. PTEN variants having low or no effect on the function of PTEN caused growth phenotypes similar to the PTEN wildtype strain. Variants having a higher effect disrupted the activity of the human PTEN, thus causing an increase in growth. Variants having a gain-of-function phenotype resulted in an additional decrease in growth as

compared to the wildtype PTEN strains. We obtained raw colony growth data from Balony, a software package for analysis of data from SGA experiments (Young & Loewen, 2013). We quantified and normalized these colony growth differences in R to obtain normalized yeast activity scores, where lower values (towards 0) indicate loss-of-function (hypomorphic) phenotypes, values closer to 1 indicate phenotypes similar to wildtype PTEN, and values greater than one indicate gain-of-function (hypermorphic) phenotypes. We characterized PTEN protein levels in each double mutant via western blot.

3.2.4 Per-gene correlation of computational predictors of damage

For each exonic base in each human gene, I calculated all possible nucleotide changes to obtain a list of all possible exonic variants in the genome. I then annotated each of these variants with four variant prediction scores (CADD 1.3, PolyPhen2, SIFT, and Snap2) using Annovar. I used an inverted version of SIFT (1 = likely damaging, 0 = likely tolerated), so that for any pair of variant scores, agreement would be indicated by a positive correlation. I calculated pair-wise Spearman's rank correlation coefficients for each pair of scores on a gene-wise basis for a subset of genes in the genome. As not all genes annotated by one tool were represented in others, in total I tested 32 559 652 variants across the 11554 genes for which comprehensive scores were available for all four variant impact prediction methods. I created a control set of 2.5 M random genomic variants to test the same correlations on a genome-wide basis. I similarly annotated these variants with the four variant prediction scores, and then calculated pairwise Spearman correlations.

For the purposes of correlating gene-level characteristics (gene length, protein length, gene expression, and number of isoforms) with pairwise correlations of variant prediction scores, I obtained gene-level characteristics from a variety of sources. I calculated gene length and protein length from each gene's cDNA sequence and protein sequence, both obtained from BioMart

(Durinck et al., 2005). For gene expression, I calculated mean reads per gene for 11688 transcriptomes contained in the GTEx dataset (accession: phs000424.v7.p2; eGTEx, 2017). I calculated the number of unique Ensembl transcripts using data from BioMart. I calculated Spearman correlation between each of these scores with the pairwise metric correlations (Table 3.2). I calculated the mean rank of correlations of all six comparisons for each gene studied. Using ErmineJ, I then analyzed this ranking of the genes using receiver-operator curve (ROC) analysis (Parameters: *Maximum gene set size* = 100, *Minimum gene set size* = 5, *Gene replicate treatment* = "Use best scoring replicate", *Test effects of multifunctional genes* = True) to search for patterns of enrichment in either the genes showing highest pairwise score correlations, or the genes showing the least pairwise score correlations (Ballouz, Pavlidis, & Gillis, 2017; Gillis, Mistry, & Pavlidis, 2010). I also used the multifunctionality score output of ErmineJ to test for enrichment of genes scoring highly in multifunctionality according to their ranking based on mean correlation across computational scores (Gillis & Pavlidis, 2011).

3.3 Results

3.3.1 Gene selection

Many candidate ASD genes contain substantial numbers of missense variants of unknown significance. Many of these genes also rank highly in terms of population-level depletion of missense variation (Table 3-1). A complete list of genes is found in Supplementary Table 7. We selected PTEN for the initial round of variant prioritization and functional characterization, followed by SYNGAP1.

Table 3-1: Top ASD-associated genes in MARVdb.
Number of variants columns are derived from MARVdb. ExAC MSZ and ExAC pLI scores are from ExAC's per-gene constraint metrics. Genes selected for functional characterization are bolded.

Gene Name	***# Variants***	***# LOF Variants***	***# MS variants***	***# de novo LOF variants***	***# de novo MS variants***	***ExAC MSZ***	***ExAC pLI***
CHD8	63	27	31	27	10	5.54	1.00
SCN2A	62	23	34	19	14	6.58	1.00
ADNP	41	25	15	22	1	1.02	1.00
DYRK1A	39	22	10	19	1	3.37	1.00
SYNGAP1	**22**	**14**	**7**	**12**	**4**	**7.15**	**1.00**
TBR1	19	9	10	9	10	5.57	0.99
CHD2	30	9	21	9	7	5.09	1.00
ARID1B	15	11	4	10	2	3.39	1.00
POGZ	31	11	20	9	3	3.36	1.00
GRIN2B	29	8	17	7	5	6.74	1.00
ANK2	20	11	9	9	4	1.06	1.00
WDFY3	39	9	30	7	5	5.72	1.00
TRIP12	27	7	14	6	4	4.59	1.00
DYNC1H1	8	0	7	0	7	13.88	1.00
NCKAP1	12	8	3	7	0	3.98	1.00
NRXN1	16	8	8	6	4	3.02	1.00
ASH1L	37	6	31	6	3	3.05	1.00
DSCAM	34	7	25	6	1	4.35	1.00
CNOT3	6	6	0	6	0	3.89	1.00
CSDE1	6	6	0	6	0	3.22	1.00
VCP	7	6	1	5	1	6.47	1.00
NFIA	11	7	4	5	2	3.06	1.00
EPHB2	11	6	5	5	1	3.45	1.00
PTEN	**18**	**4**	**13**	**2**	**12**	**3.71**	**0.98**
PAX5	8	7	1	5	1	3.26	0.98
ATP1B1	7	6	1	5	1	2.31	1.00
WAC	8	5	3	5	0	1.57	1.00
PHF2	10	7	2	5	0	3.26	0.99
CTNNB1	7	4	3	4	3	4.44	1.00
CDC42BPB	7	4	3	4	3	4.30	1.00
RAB2A	6	5	1	5	0	3.16	0.97
MPHOSPH8	5	5	0	5	0	0.66	0.93
LRP1	6	0	6	0	6	10.62	1.00
SETD2	9	4	5	4	2	3.48	1.00
DIP2A	10	8	2	6	0	2.37	0.73
KDM6B	9	6	2	4	2	2.44	1.00
KMT2C	8	5	3	4	3	1.53	1.00
CUL3	6	4	1	4	1	4.95	0.97

3.3.2 Functional characterization of prioritized PTEN variants

We manually selected 60 PTEN variants (ASD variants, biochemically-validated variants, cancer and other NDD variants) from MARVdb, ClinVar, and the broader literature to test in model organisms. Using the computational prioritization procedures described above provided 43 further variants (common variants and predicted high impact variants) to test in model organisms. These 103 variants are listed in Appendix A.1, and shown in context of all genomic variants in Figure 3-4. Our collaborators tested the activity of these variants by way of measuring the differences in colony growth of yeast in an SGA experiment as detailed in the methods. Normalizing the differences of colony growth provided us with yeast activity scores, where 1 is like wildtype, < 1 indicates a hypomorphic phenotype, and > 1 indicates a hypermorphic phenotype (Figure 3-5). CADD and Snap2 scores were negatively correlated with yeast activity scores ($r = -0.521$ and $r = -0.646$, respectively). This agrees with our expectations, as higher scores in both CADD and SNAP2 imply higher deleteriousness, whereas lower scores in our yeast assay imply lack of a function phenotypes. Protein quantification results were positively correlated with normalized yeast activity ($r = 0.31$). This also agrees with our expectations; some changes in yeast activity are undoubtedly driven by changes in protein expression and/or stability. However, CADD and SNAP2 scores were not strongly correlated with protein quantification results ($r = -0.127$ and $r = -0.023$). While we would expect the prediction of deleteriousness by variant prediction methods to provide some level of predictive power for variant effects on protein expression / stability, there are many ways in which variants can have effects on function, some in which protein stability is a factor, and others in which it is not. If variant impact prediction tools are focusing on the factors unrelated to protein stability, this may explain the low correlation we found.

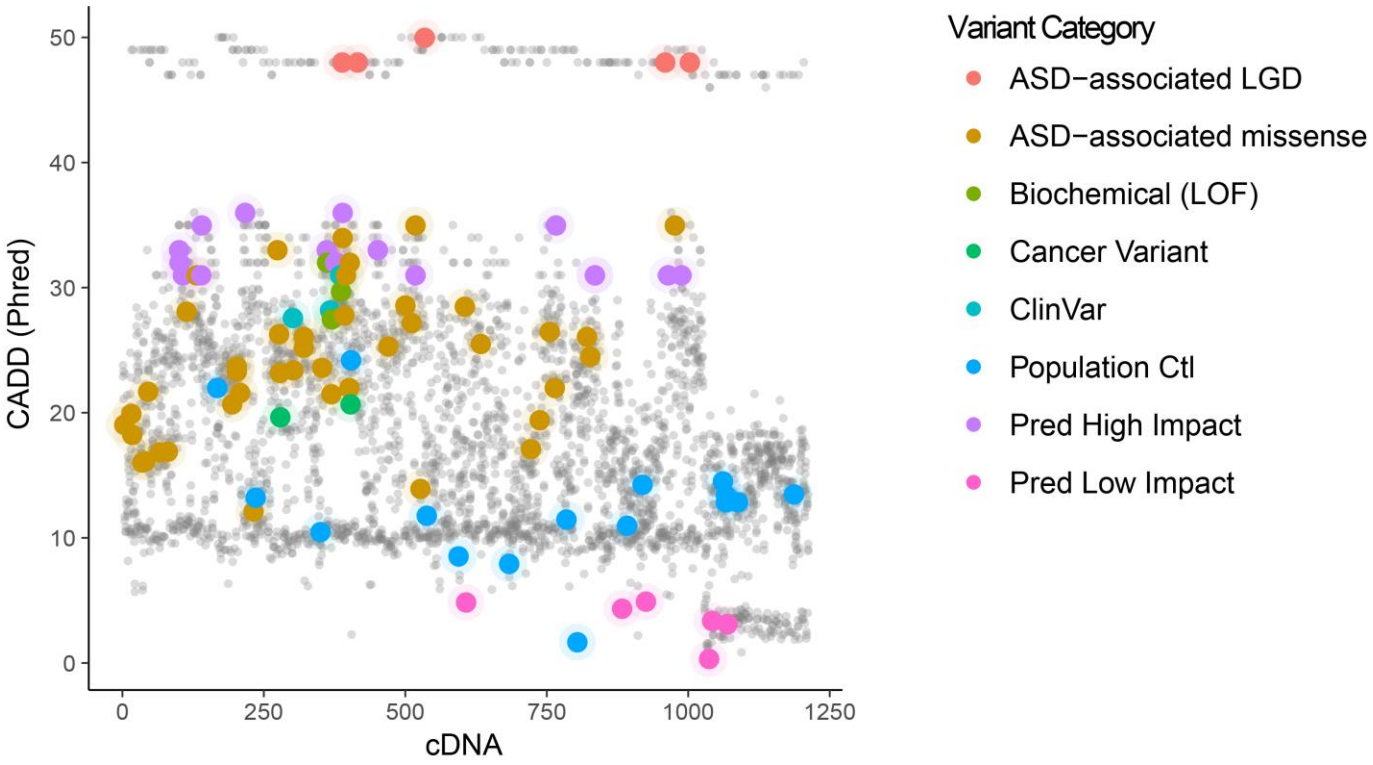

Figure 3-4: CADD score by cDNA position of all genomic PTEN variants.
Variants of interest are highlighted. All other genomic variants in grey.

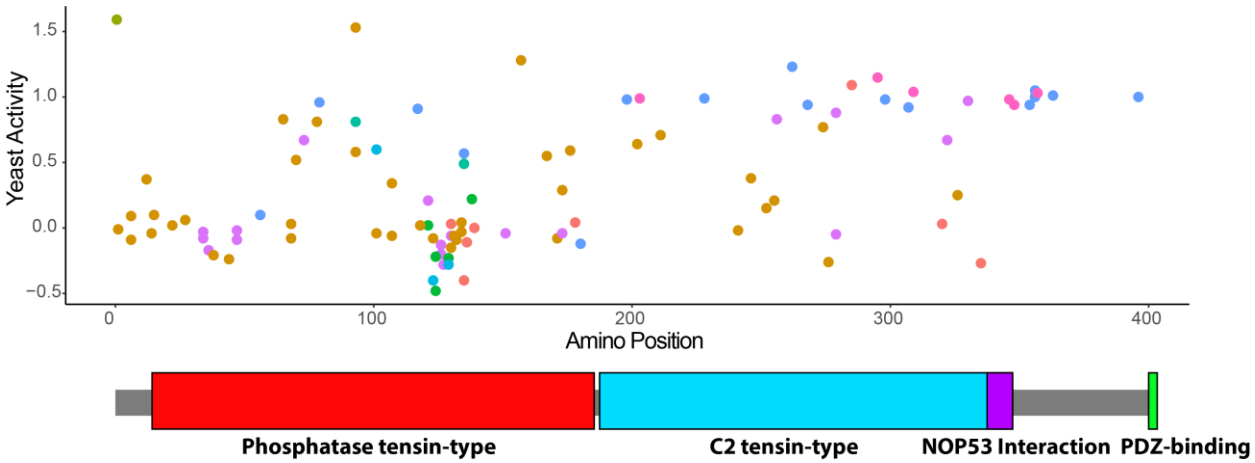

Figure 3-5: Yeast activity of prioritized PTEN variants according to amino position and UniProt domains
Variants assayed spanned multiple protein domains of PTEN. A cluster of deleterious variants span PTEN's catalytic core domain (residues 124-130).

Yeast assay results generally correlated with computational predictions (Figure 3-6). For example. P357S, P246L, and Q171E were predicted by CADD and SNAP2 to be variants of low, intermediate, and high effect, respectively, and their corresponding yeast activity scores confirm these predictions. The majority of the ASD missense variants we tested show substantial agreement between our computational predictions and their resulting yeast activities. However, several variants were shown to be outliers in terms of yeast activity results. Y180H, I135T, and F56C, all population variants from ExAC, displayed loss-of-function phenotype in yeast, although we had expected these variants to perform similarly to wild-type. H93R and E157G, two ASD missense variants, displayed gain-of-function phenotypes (yeast activity 1.53 and 1.28 times that of wildtype). Three ASD missense variants - T78A, Y65C, and W274L, and one cancer variant – H93Q, displayed functional effects similar to wildtype. These low functional changes could imply that these variants are in fact not true pathogenic variants, or that our assay is not adequately testing their function.

3.3.3 Prioritization of SYNGAP1 variants

We manually selected 62 SYNGAP1 variants (ASD variants, biochemically-validated variants, likely benign variants and other NDD variants) to functionally characterize in model organisms. We computationally selected 38 further variants - rare, high-predicted damage variants and common, low-predicted damage variants. These 110 variants are listed in Appendix A.2 and shown in context of all genomic SYNGAP1 variants in Figure 3-7. At this writing the functionalization of these variants is in progress.

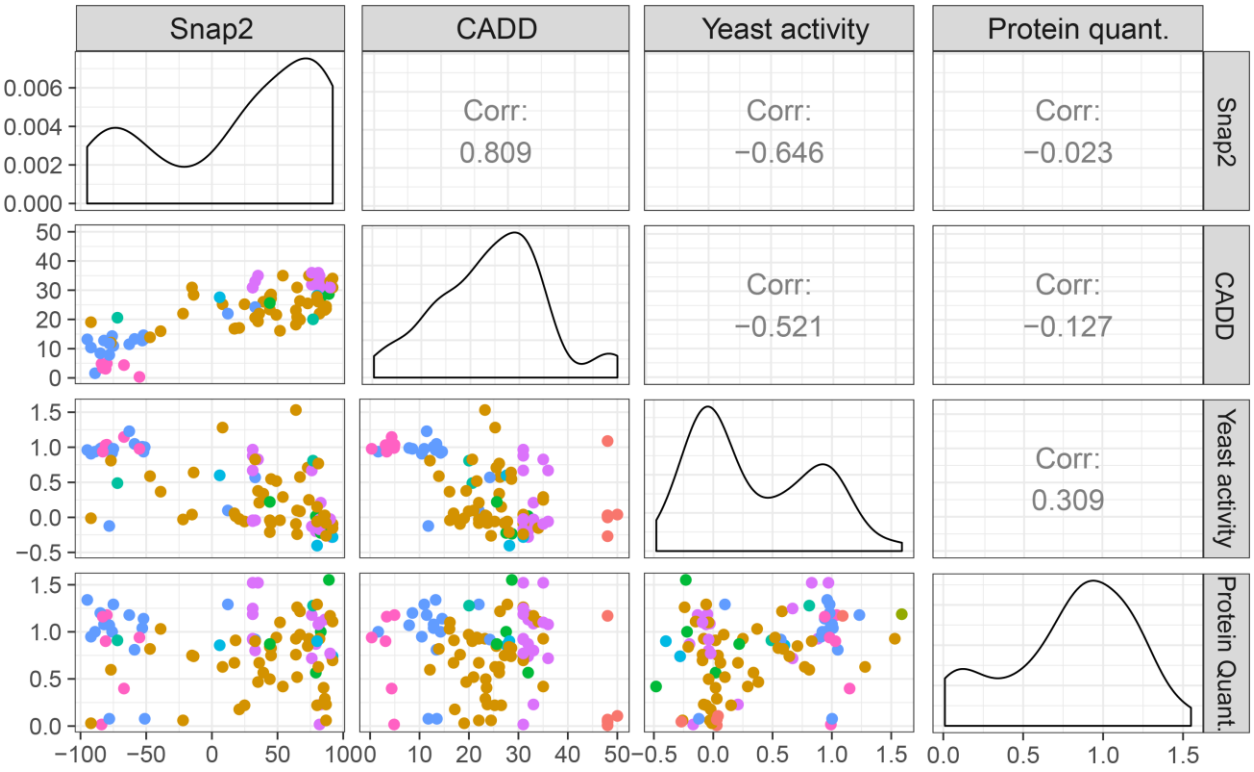

Figure 3-6: Pairwise comparisons of SNAP2, CADD, yeast activity in SGA, and protein quantification of prioritized PTEN variants

Yeast activity: 1 = like wildtype, >1 = gain-of-function, < 1 = loss-of-function. Protein quantification: 1 = like wildtype, >1 = increase in protein quantity, < 1 = decrease in protein quantity.

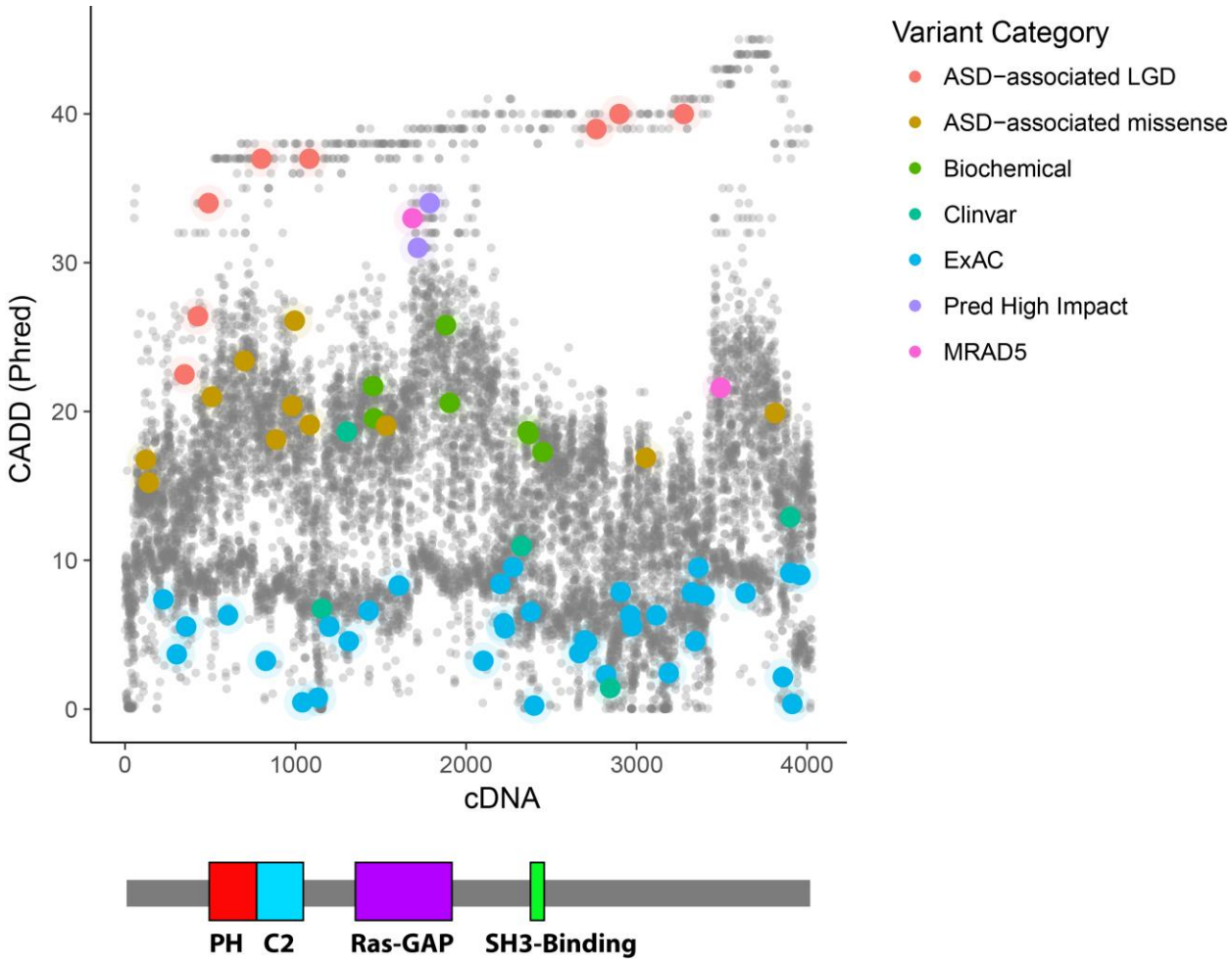

Figure 3-7: Prioritized SYNGAP1 variants.
CADD scores (Phred scaled) per cDNA position of all prioritized SYNGAP1 variants and corresponding Uniprot domains. MRAD5 – Mental Retardation, Autosomal Dominant 5.

3.3.4 Agreement of computational metrics across genes

During the prioritization process for PTEN and SYNGAP1, it became apparent that while CADD and Snap2 scores generally agreed for nonsynonymous variants in some genes (e.g. PTEN), correlation was extremely poor for other genes (e.g. SYNGAP1) (Figure 3-8). To investigate this phenomenon, I calculated pairwise correlations of four metrics of variant damage (CADD, Snap2, SIFT and Polyphen2) for each of the 11554 genes which were comprehensively annotated in terms of these metrics. These computational prediction metrics showed a range of per-gene agreement, as shown in Figure 3-9. A summary of these relationships is displayed in Table 3-2. CADD displayed lower correlations with all other variant scores, with bimodal distributions of correlations.

At first glance it was unclear why variant scoring methods would show this type of gene-specific behavior. As the disagreement appeared to occur on a per-gene basis, and variant prediction tools often use gene-level characteristics as features in their classification, I hypothesized some of these features may hold an explanation. In a preliminary attempt to seek an explanation, I considered whether the genes which have poor inter-score agreement share some biological characteristics. I calculated correlation between several gene-level characteristics (gene length, protein length, gene expression in GTEx transcriptomes, and number of isoforms) and each of the pairwise comparisons of computational metrics (Table 3-3: Correlations of four gene-level characteristics with pairwise metric correlations.Table 3-3). No gene characteristic tested displayed sufficient correlation to explain the pairwise computational metric comparisons (either singly or in combination with other gene characteristics).

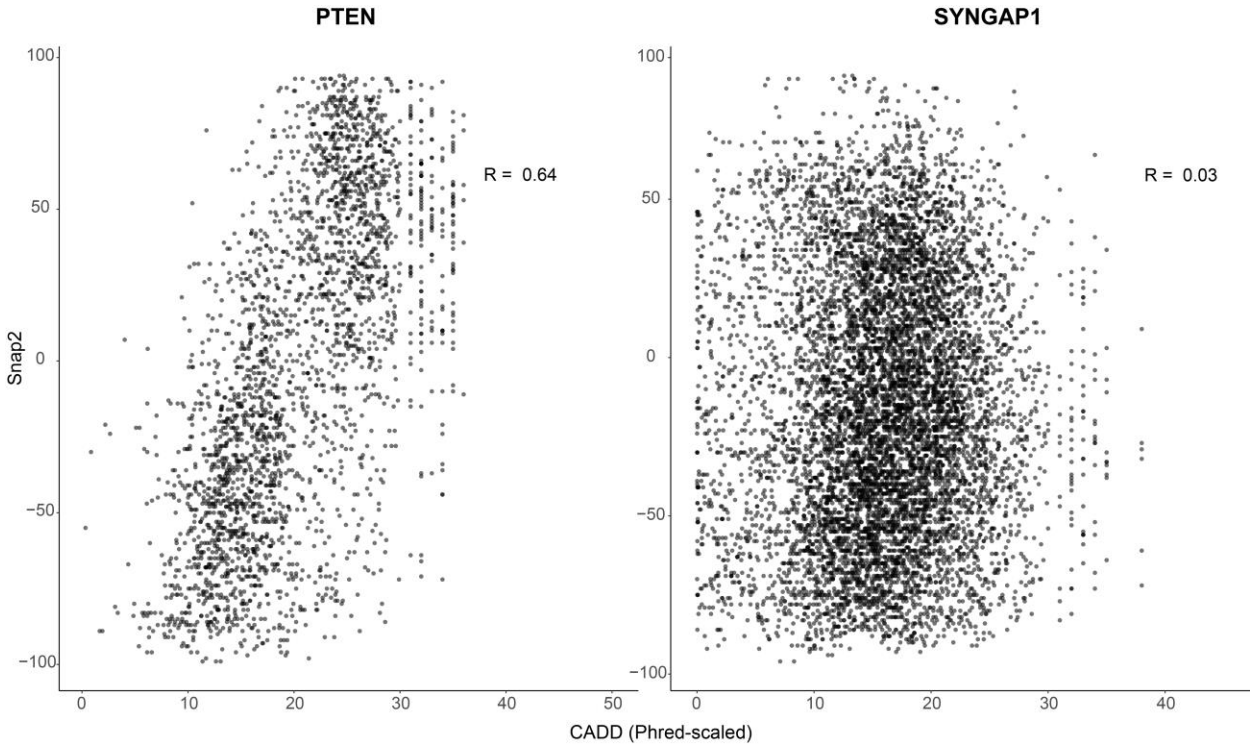

Figure 3-8. Correlation of CADD and Snap2 for all PTEN and SYNGAP1 variants.
Missense variants in PTEN display higher Spearman correlation between CADD and Snap2 than do those in SYNGAP1.

Table 3-2: Summary statistics for pairwise comparisons of variant impact prediction scores.
Variant prediction scores often display lower agreement when consider on a per-gene basis, compared to their genome-wide correlation with other scores. A large proportion of pairs of variant impact prediction scores (particularly pairs with CADD) display very low (< 0.1) correlation when considered in this manner.

	Genome-wide correlation	Median of per-gene correlation	% Genes with Per-Gene Correlation:			
			< 0.1	< 0.3	< 0.5	< 0.7
CADD vs SNAP2	0.46	0.34	16.3	45.1	64.5	98.5
CADD vs SIFT	0.67	0.35	16.2	44.6	65.3	94.4
CADD vs PolyPhen2	0.66	0.42	14.1	38.1	57.1	88.7
SNAP2 vs SIFT	0.57	0.62	3.0	10.5	27.0	74.5
SNAP2 vs PolyPhen2	0.59	0.61	1.9	7.2	25.0	80.4
SIFT vs PolyPhen2	0.68	0.57	2.1	10.3	35.3	80.4

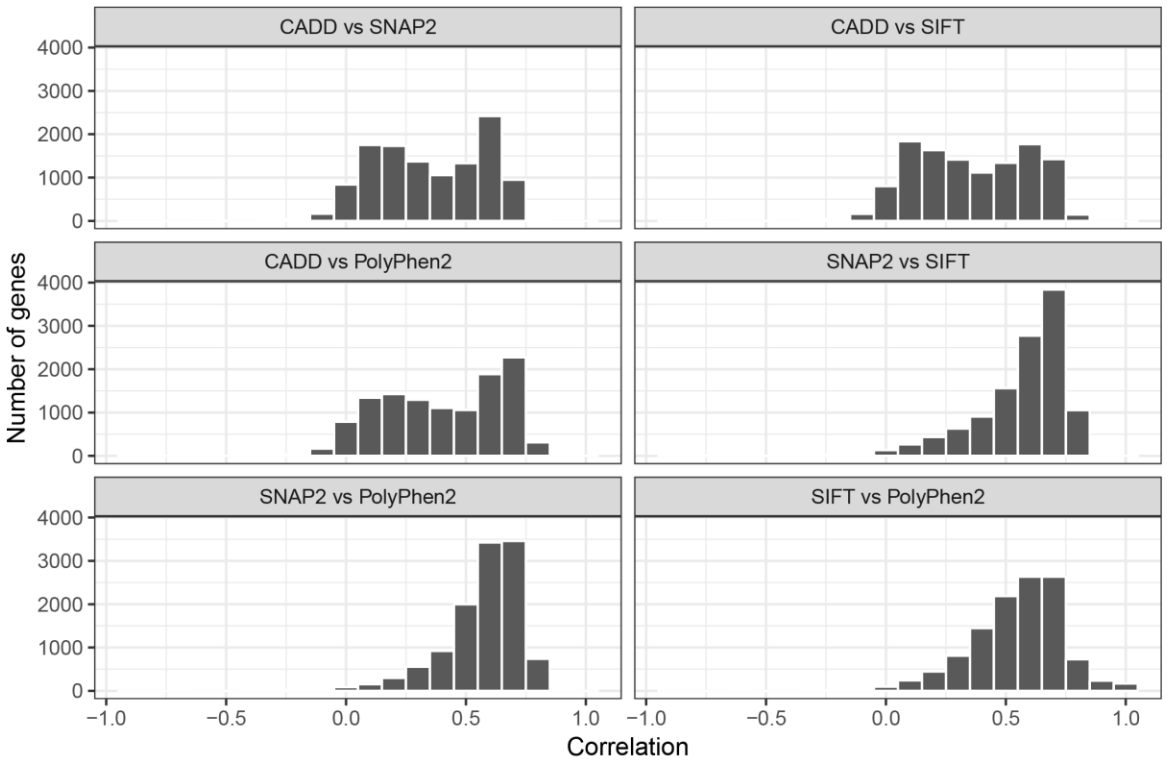

Figure 3-9: Gene metric correlations for 11554 genes.
Correlations were calculated on a gene-wise basis for all exonic variants. 11554 genes had comprehensive score annotations for all variants, and thus were included in the analysis.

Table 3-3: Correlations of four gene-level characteristics with pairwise metric correlations.
PP2 = PolyPhen2. MRC = Gene-by-gene mean rank of correlations for each pairwise comparison of variant impact predictors.

	CADD/ SNAP2	CADD/ SIFT	CADD/ PP2	Snap2/ SIFT	SNAP2/ PP2	SIFT/ PP2	MRC
Gene Length	0.082	0.061	0.113	0.001	0.060	0.021	0.08
Protein Length	0.050	-0.007	0.047	-0.043	0.038	-0.045	0.01
Gene expression	0.082	0.108	0.096	0.088	0.073	0.107	0.13
Number of isoforms	0.008	0.038	0.046	-0.027	0.014	0.047	0.03
Multifunct-ionality	0.09	0.12	0.10	0.19	0.19	0.20	0.20

As these results were inconclusive, I next explored whether genes which show poor inter-method agreement are concentrated in particular gene functions. I used ErmineJ to test for GO term enrichment in the ranking of genes by their pairwise correlations. First, both the top and bottom of my gene rankings were enriched for several GO terms (Supplementary file 8; ErmineJ project file). Interestingly, I observed a relationship between multifunctionality and the mean rank of pairwise correlations (0.2). Multifunctionality refers to the propensity of a gene or genes to hold multiple molecular functions, and has been shown to impact GO enrichment results (Gillis & Pavlidis, 2011). Both the GO enrichment and the multifunctionality results suggest at the very least that the propensity of a gene to show low or high correlations of computational metrics of deleteriousness is not completely random. However, I did not identify a definitive explanation for these patterns and leave it as a topic for future study.

3.4 Discussion

In this section, I discuss the creation and implementation of a computational pipeline for informing gene selection and subsequent variant prioritization for the functional characterization of ASD missense variants of unknown significance.

3.4.1 Gene selection

I developed a bioinformatics-informed method to provide a preliminary prioritization of genes for functional study, based on evidence from WGS/WES ASD literature as well as gene-level depletion of damaging variation in ExAC. This prioritization ranks higher genes containing more missense VUSs, as well as those likely less tolerant of damaging missense variation. This provides a useful starting point for determining genes of interest for functional characterization, however, as discussed previously, further information must be taken into account. One primary consideration for gene selection is a particular gene's appropriateness for study in our biological

assays. For example, while many ASD missense VUSs have been reported in Glutamate Ionotropic Receptor NMDA Type Subunit 2B (GRIN2B), functionally characterizing missense variants would be difficult without concurrent expression of genes coding for the other subunits of the N-methyl-D-aspartate channel of which it is a component. Other criteria that influence gene selection are gene length, and genomic and proteomic background in the model organisms under study.

3.4.2 PTEN variant prioritization and functional characterization

Manual collection efforts and computational annotation and variant prioritization procedures enabled us to amass a list of variants of interest to functionally characterize in model organism assays. Subsequent functionalization of these variants in S. cerevisiae revealed several informative results regarding missense variants of unknown significance in PTEN. Firstly, most missense VUSs observed in ASD probands exhibited a functional effect in our yeast assay. The functional effects of the variants ranged from gain-of-function phenotypes (i.e. H93R, E157G), to no effect (effects similar to wildtype - implying low deleteriousness in humans), to strong loss-of-function phenotypes (implying deleteriousness in humans). This diversity in functional effect may demonstrate a diversity in the mechanisms by which missense variants may disrupt a protein's function. On the other hand, this diversity may simply demonstrate a limitation in our ability to detect the effect of variants on gene function in our assay – this could indicate a lack of sensitivity in our assay, or perhaps that some variants are affecting a different function than we are able to test. Furthermore, the disconnect between variant prediction scores and functional assay results reveals another level of disagreement between expected and observed results. It is likely that functional characterization results from other assays will help to clear up discrepancies.

Many, but not all, variants performed as expected. The biochemically-validated variants previously reported as loss-of-function (A121E, C124S, C124S-4A, G129E, and Y138L)

displayed loss-of-function phenotypes in our yeast assay (all with normalized activity < 0.25). We found the predicted gain-of-function variant 4-Ala to indeed behave as gain-of-function in our assay. The agreement of these variants with our expectations serves to increase our confidence in the generalizability of our assay.

On the other hand, several variants behaved contrary to predictions. G285X, a stop-gain variant found in an ASD proband, demonstrated functional effect similar to wild-type (normalized activity = 1.09). Some NDD missense and cancer variants did not appear to display strong phenotypes in yeast (namely T78A, H93Q, Y65C, and W274L). These results could imply one of three things. First, the results of the stop-gain variant in particular may suggest some technical problems in our assays, as G285X is predicted to cause an earlier truncation of PTEN than L320X and R335X, both of which displayed strong functional effects in yeast (normalized activity (0.03 and -0.27, respectively). Secondly, these findings could indicate that such variants may represent lower pathogenicity subsets of candidate variants – or perhaps are not truly pathogenic variants at all. Indeed, T78A, while reported as *de novo* by Schaaf et al. (2001), is not predicted as strongly deleterious by either CADD or SNAP2, and was identified in a proband with "high-functioning autism spectrum disorder". This suggests that T78A may represent a milder ASD effect – or may be inconsequential in the etiology of ASD. W274L, an inherited variant found in a patient with Developmental Delay by McBride et al. (2010), was also found to exhibit a milder effect in yeast, although computational predictors indicated relatively high deleteriousness (SNAP2 = 81 and CADD = 26.1). Thirdly, due to the fact that we are only testing functional effects in yeast for one phenotype, perhaps our assay is not able to detect the full range of mechanisms by which a missense variant can disrupt PTEN's function. However, the fact that we detect functional effects for variants in multiple domains of PTEN suggests that our assay provides at least a moderate

degree of sensitivity for detecting functional effect. Given the significantly different genomic and proteomic backgrounds of yeast and humans, complementing yeast functional assays with those in other organisms enables us to resolve ambiguities between computational predictions and functional results in any one organism. Functional characterization of PTEN variants is still underway in our other model systems. Preliminary results suggest reasonable agreement between functional characterization results in yeast and in *D. melanogaster*. For the sake of discussion, if a predicted high-damage VUS were to show little functional effect in yeast, but high effect in *D. melanogaster*, this may suggest that the yeast assay was lacking sensitivity in establishing a functional effect for that variant. Testing the variant in additional assays would provide further discriminatory power towards identifying the variant's likely role in ASD.

The computational methods I used can now provide a consistent platform for variant annotation and prioritization in ASD genes of interest. The method as implemented in R allows versatile adjustment of parameters to adjust the resulting categorical composition of variants. Although we have to this point only completed functional characterization of prioritized variants in PTEN, initial findings are favourable - functional characterization results of variants in yeast generally agree with variant prioritisation results. Therefore, by using predictive metrics to choose informative variants for functional characterization, we increase our yield of informative results. This method is likely more effective in the case of genes with numerous ASD-associated variants reported in the literature – in the case of a gene with only very few such variants, adding more and more computationally-predicted variants will do little to yield valuable information regarding ASD in particular – although some information regarding the validity of the computational predictions may be gained.

One further limitation in this study, and many other functionalization studies, is in the extent to which we can extrapolate disease causation from the functional results. By functionalizing ASD missense variants of unknown significance found in ASD candidate genes, we can infer (or rule out) the *probable* ASD causation of the variants tested. To conclusively determine human ASD causality of any particular variant would require testing the variant's function in a human system – obviously a difficult criteria to hold. Human induced pluripotent stem cells (iPSCs) gene edited with the CRISPR/CAS9 system have previously been used to functionally characterize human NDD variants (Duan et al., 2017). An iPSC assay integrated into a pipeline such as ours could provide important context to functional characterization efforts. At present, our multiplatform pipeline integrates evidence from bioinformatics methods and multiple biological assays, and therefore functional characterization results are supported by multiple lines of evidence.

3.4.3 SYNGAP1 variants

Functional characterization work is still underway for the prioritized SYNGAP1 variants. Given the lack of agreement between CADD and SNAP2 predictions in SYNGAP1, it remains to be seen which metric will better predict the functional effects of the tested variants. Functionalization results could inform subsequent variant selection in SYNGAP1, and possibly in other genes for which computational metrics display low correlation.

3.4.4 Computational metrics: agreement across genes

Comparisons of damage prediction scores have, to date, typically been on a genome-wide basis. Such comparisons typically report reasonably high rates of agreement between scores – for a genome-wide example, SIFT ranks scores are correlated with both CADD and Polyphen2 rank scores ($r = 0.63$ for both comparisons) (Liu et al., 2011). Our findings indicate that a more fine-

grained, gene-level perspective on the matter of agreement reveals important discrepancies between scores that should be considered when using multiple methods to score variants as deleterious. Indeed, in a large number of genes, computational metrics of damage prediction demonstrate very little agreement with each other (e.g. 24% of genes have CADD vs SNAP2 correlations 0.15 or below). All pairs of tools showed lower median per-gene correlations than genome-wide correlation, except for two pairs: SNAP2 and SIFT, and SNAP2 and PolyPhen2 . This might be considered an example of Simpson's paradox – where a trend exists in a group of data but then disappears when considering subsets of the data (Blyth, 1972; Simpson, 1951). Another factor is likely helping to create this phenomenon – because my analysis did not include intronic or intergenic variants, I omitted a large area of high agreement between scores (methods that score intronic and intergenic variants generally rank them as tolerated). These areas of the genome undoubtedly drive a large component of the agreement between scores. This omission notwithstanding, because the majority of variant discovery and variant prioritisation efforts focus on exonic sequences, this observation is important to consider for both end-users of the variant scoring tools and developers of variant prediction tools themselves.

The reason for this disagreement remains an open question to study. As variant prediction methods differ in terms of both training data and methodology (features and algorithms used), it is likely that the disagreement stems from these method-specific differences. Neither gene length, protein length, gene expression, nor number of isoforms correlated strongly with the six pairwise comparisons. It remains to be seen whether some other such gene-level characteristic can explain a gene's performance in this respect. However, analysis with ErmineJ demonstrated a non-random ordering (in terms of GO enrichment and gene multifunctionality) of genes and their mean rank correlations. This suggests that some effect (biological or an algorithmic artefact) may be driving

the gene-by-gene differences between variant scoring methods. While more research is necessary to pinpoint the root of the discrepancies, the fact that a signal exists suggests a common root may exist, and can therefore be corrected for.

As far as what this disagreement means for researchers investigating individual variants or genes, it appears to depend on the gene under study. In PTEN, a gene for which CADD and SNAP2 demonstrate reasonable agreement, we showed that functional characterization results corroborate the computational predictions of effect remarkably well. Whether this is true for other such genes will require further study, but this seems a safe assumption. In genes where computational predictors have low correlation, the disagreement can be interpreted (and managed) in a variety of ways. Firstly, while one may be tempted to add more computational predictors to the task at hand, with the aim of creating a "most votes" method to establish likely pathogenicity, it is difficult to concretely determine which method or methods, out of the multitude, are the correct ones to choose or drop to resolve ambiguity. Furthermore, many variant impact prediction methods share commonalities in their approach, so may not provide independent information. Because in many tasks for which variant predictors are used, the ground truth of pathogenicity is not known prior to prediction, functional characterization of variants is therefore the most straightforward way to establish pathogenicity when computational methods fail.

As is evidenced by our investigation into the functional effects of PTEN missense variants, no single score or combination of scores is sufficiently accurate to encapsulate the range of functional variation of missense variants. Thus, our results support the combined consideration of multiple sources of evidence (computational methods of prediction, population frequency, prior disease association of similar variants, functional characterization assays) for exploring potentially pathogenic variation. While it is likely that many genes will lend themselves well to this sort of

variant prioritization procedure, it is worth noting that our observations regarding variant damage metric correlations implies that not every gene will be as amenable to this style of variant prioritization.

Chapter 4: Conclusion

In Chapter 2 I report the discovery of candidate ASD variants in a cohort of ASD individuals. Several candidate variants were found in genes previously well-characterized as ASD genes by way of genomic evidence (SCN2A, NIPBL, CUL3, ADNP, and ASXL3). Other candidate variants were found in genes supported for ASD association by functional information, or supported for association in other NDDs (WDR45, ARID2, CACNA1B, SPTBN2, DGKD, RAPGEF4, and PLXNA2). These variants were supported as likely pathogenic by several threads of evidence - inheritance, rarity in the human population, computational predictors of damage, and gene-level depletion of damaging variation, in addition to the ASD-specific factors previously mentioned. Our experience, and the broader ASD variant discovery literature, strongly encourages the use of trio WGS in variant discovery efforts. Ascertaining the variants in NIPBL, WDR45, ARID2 and the two SCN2A variants as *de novo* concretely established their role as likely pathogenic. ASD has been previously linked to disruptions in synaptic function and chromatin remodeling, among others. Our results recapitulate these findings, as these genes have been previously implicated in these functions.

For the variants for which inheritance remains unknown, all of the variants we establish as high-priority candidates for ASD are found in genes with prior ASD or NDD association. Many ASD candidate variants have been reported in the past which have unknown inheritance status. While the genes containing these variants may not be considered as conclusively implicated in ASD as the ones for which we found *de novo* variants, variant discovery efforts such as this adds weight, variant by variant, to a gene's involvement in ASD, and to our understanding of the genetic etiology of ASD as a whole.

In Chapter 3 I report the prioritization of ASD variants of unknown significance for functional characterization in model organism assays, and the functionalization of a set of PTEN variants in yeast. Our preliminary results indicate that computational prioritization can provide a valid source of evidence to use to identify genes, and subsequently variants of interest for functional characterization. The results of the SGA experiment in yeast generally agree with our hypotheses for how each category would perform, and furthermore the specific computational prediction for the effect of each variant.

The functional characterization of variants of unknown significance demonstrates a robust method to explore the range of functional variation in ASD and other disease-associated variants. Using bioinformatics-based methods of variant prioritization, we can focus our efforts to functionally characterize variants, and can enhance the interpretability of our results. One important consideration which became evident during this study is the lack of consistent gene-wise agreement of popular variant scoring methods. At this point nothing conclusive can be said regarding the provenance of the disagreement, although preliminary analysis with ErmineJ suggests several interesting patterns. Further study must be done to determine why this disagreement occurs, and how we might best mitigate its effects in variant prioritization procedures.

In this thesis, I set out to find answers to two questions in ASD genetics. To the question "What genetic variation causes ASD?", I propose several novel candidate variants. While neither this effort, nor the cumulative results of ASD variant discovery proposes a comprehensive account of the genetic variation causing ASD, it is through incremental identification of candidate variation that we begin to more completely understand the genetic etiology underlying ASD. Further research efforts will hopefully build upon these results to ever increase our knowledge of ASD-

causal variation. To help in answering the question “What is the functional impact of reported ASD variants?”, I propose a computational method to more effectively explore the functional effects of ASD variants. Functionalization efforts such as the one reported are steadily filling in our knowledge of the biological mechanisms underlying not just ASD, but of all kinds of damaging genetic variation.

Bibliography

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

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

Anderson, S. R., & Romanczyk, R. G. (1999). Early intervention for young children with autism: Continuum-based behavioral models. *Journal of the Association for Persons with Severe Handicaps*, *24*(3), 162–173.

Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T. R., … Hallmayer, J. (2010). A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics*, *19*(20), 4072–4082. https://doi.org/10.1093/hmg/ddq307

Asperger, H. (1944). Die „Autistischen Psychopathen" im Kindesalter. *European Archives of Psychiatry and Clinical Neuroscience*, *117*(1), 76–136.

Association, A. P., & others. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Bacchelli, E., Blasi, F., Biondolillo, M., Lamb, J. A., Bonora, E., Barnby, G., … International Molecular Genetic Study of Autism Consortium (IMGSAC). (2003). Screening of nine candidate genes for autism on chromosome 2q reveals rare nonsynonymous variants in the cAMP-GEFII gene. *Molecular Psychiatry*, *8*(11), 916–924. https://doi.org/10.1038/sj.mp.4001340

Bailey, A., Couteur, A. L., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E., & Rutter, M. (1995). Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological Medicine*, *25*(1), 63–77. https://doi.org/10.1017/S0033291700028099

Baird, G., Simonoff, E., Pickles, A., Chandler, S., Loucas, T., Meldrum, D., & Charman, T. (2006). Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). *The Lancet*, *368*(9531), 210–215. https://doi.org/10.1016/S0140-6736(06)69041-7

Ballouz, S., Pavlidis, P., & Gillis, J. (2017). Using predictive specificity to determine when gene set analysis is biologically meaningful. *Nucleic Acids Research*, *45*(4), e20–e20. https://doi.org/10.1093/nar/gkw957

Barrows, C. M., McCabe, M. P., Chen, H., Swann, J. W., & Weston, M. C. (2017). PTEN loss increases the connectivity of fast synaptic motifs and functional connectivity in a developing hippocampal network. *The Journal of Neuroscience*, 0878-17. https://doi.org/10.1523/JNEUROSCI.0878-17.2017

Ben-Shalom, R., Keeshen, C. M., Berrios, K. N., An, J. Y., Sanders, S. J., & Bender, K. J. (2017). Opposing Effects on NaV1.2 Function Underlie Differences Between SCN2A Variants Observed in Individuals With Autism Spectrum Disorder or Infantile Seizures. *Biological Psychiatry*, *82*(3), 224–232. https://doi.org/10.1016/j.biopsych.2017.01.009

Besenbacher, S., Liu, S., Izarzugaza, J. M. G., Grove, J., Belling, K., Bork-Jensen, J., … Rasmussen, S. (2015). Novel variation and *de novo* mutation rates in population-wide *de novo* assembled Danish trios. *Nature Communications*, *6*, ncomms6969. https://doi.org/10.1038/ncomms6969

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

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364–366.

Boone, C., Bussey, H., & Andrews, B. J. (2007). Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, *8*(6), 437–449. https://doi.org/10.1038/nrg2085

Bourgeron, T. (2009). A synaptic trek to autism. *Current Opinion in Neurobiology*, *19*(2), 231–234. https://doi.org/10.1016/j.conb.2009.06.003

Bourgeron, T. (2015). From the genetic architecture to synaptic plasticity in autism spectrum disorder. *Nature Reviews Neuroscience*, *16*(9), 551–563. https://doi.org/10.1038/nrn3992

Bramswig, N. C., Caluseriu, O., Lüdecke, H.-J., Bolduc, F. V., Noel, N. C. L., Wieland, T., … Wieczorek, D. (2017). Heterozygosity for ARID2 loss-of-function mutations in individuals with a Coffin-Siris syndrome-like phenotype. *Human Genetics*, *136*(3), 297–305. https://doi.org/10.1007/s00439-017-1757-z

Butler, M. G., Dasouki, M. J., Zhou, X.-P., Talebizadeh, Z., Brown, M., Takahashi, T. N., … Eng, C. (2005). Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations. *J Med Genet*, *42*, 318–321. https://doi.org/10.1136/jmg.2004.024646

Buxbaum, J. D., Silverman, J. M., Smith, C. J., Kilifarski, M., Reichert, J., Hollander, E., … Davis, K. L. (2001). Evidence for a Susceptibility Gene for Autism on Chromosome 2 and for Genetic Heterogeneity. *The American Journal of Human Genetics*, *68*(6), 1514–1520. https://doi.org/10.1086/320588

Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *The American Journal of Human Genetics*, *86*(1), 6–22. https://doi.org/10.1016/j.ajhg.2009.11.017

Carvill, G. L., Heavin, S. B., Yendle, S. C., McMahon, J. M., O'Roak, B. J., Cook, J., … Mefford, H. C. (2013). Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nature Genetics*, *45*(7), 825–830. https://doi.org/10.1038/ng.2646

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

Chakrabarti, S., & Fombonne, E. (2005). Pervasive Developmental Disorders in Preschool Children: Confirmation of High Prevalence. *American Journal of Psychiatry*, *162*(6), 1133–1141. https://doi.org/10.1176/appi.ajp.162.6.1133

Clement, J. P., Aceti, M., Creson, T. K., Ozkan, E. D., Shi, Y., Reish, N. J., … Rumbaugh, G. (2012). Pathogenic SYNGAP1 mutations impair cognitive development by disrupting the maturation of dendritic spine synapses. *Cell*, *151*(4), 709–723. https://doi.org/10.1016/j.cell.2012.08.045

Codina-Solà, M., Rodríguez-Santiago, B., Homs, A., Santoyo, J., Rigau, M., Aznar-Laín, G., … Cuscó, I. (2012). Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders. https://doi.org/10.1186/s13229-015-0017-0

Consortium, 1000 Genomes Project, & others. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68.

Corsello, C. M. (2005). Early intervention in autism. *Infants & Young Children*, *18*(2), 74–85.

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Ercument Cicek, A., … Buxbaum, J. D. (2014a). Supple: Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, *515*, 1–21. https://doi.org/10.1038/nature13772

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

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, *24*(8), 2125–2137. https://doi.org/10.1093/hmg/ddu733

Duan, J., Zhang, H., Forrest, M., Moy, W., McGowan, H., Leites, C., … Gejman, P. V. (2017). Open Chromatin Dynamics In Ipsc-Derived Neurons And CRISPR/CAS9 Editing of open Chromatin Sequences At The MIR137 Schizophrenia Risk Locus Alters Synapto-Dendritic Architecture. *European Neuropsychopharmacology*, *27*, S429–S430. https://doi.org/10.1016/j.euroneuro.2016.09.484

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, *21*(16), 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

eGTEx, P. (2017). Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nature Genetics*.

Elsayed, S. M., Heller, R., Thoenes, M., Zaki, M. S., Swan, D., Elsobky, E., … Bolz, H. J. (2014). Autosomal dominant SCA5 and autosomal recessive infantile SCA are allelic conditions resulting from SPTBN2 mutations. *European Journal of Human Genetics*, *22*(2), 286–288. https://doi.org/10.1038/ejhg.2013.150

Farzin, F., Perry, H., Hessl, D., Loesch, D., Cohen, J., Bacalman, S., … Hagerman, R. (2006). Autism spectrum disorders and attention-deficit/hyperactivity disorder in boys with the fragile X premutation. *Journal of Developmental & Behavioral Pediatrics*, *27*(2), S137–S144.

Fenske, E. C., Zalenski, S., Krantz, P. J., & McClannahan, L. E. (1985). Age at intervention and treatment outcome for autistic children in a comprehensive intervention program. *Analysis and Intervention in Developmental Disabilities*, *5*(1–2), 49–58.

Fischbach, G. D., & Lord, C. (2010). The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. *Neuron*, *68*(2), 192–195. https://doi.org/10.1016/j.neuron.2010.10.006

Fombonne, E. (2003a). Epidemiological Surveys of Autism and Other Pervasive Developmental Disorders: An Update. *Journal of Autism and Developmental Disorders*, *33*(4), 365–382. https://doi.org/10.1023/A:1025054610557

Fombonne, E. (2003b). The Prevalence of Autism. *JAMA*, *289*(1), 87–89. https://doi.org/10.1001/jama.289.1.87

Geschwind, D. H. (2009). Advances in autism. *Annual Review of Medicine*, *60*, 367–380. https://doi.org/10.1146/annurev.med.60.053107.121225

Geschwind, D. H. (2011). Genetics of autism spectrum disorders. *Trends in Cognitive Sciences*, *15*(9), 409–416. https://doi.org/10.1016/j.tics.2011.07.003

Gillis, J., Mistry, M., & Pavlidis, P. (2010). Gene function analysis in complex data sets using ErmineJ. *Nature Protocols*, *5*(6), 1148–1159. https://doi.org/10.1038/nprot.2010.78

Gillis, J., & Pavlidis, P. (2011). The Impact of Multifunctional Genes on "Guilt by Association" Analysis. *PLOS ONE*, *6*(2), e17258. https://doi.org/10.1371/journal.pone.0017258

Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., … Hakonarson, H. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, *459*(7246), 569–573. https://doi.org/10.1038/nature07953

Godlee, F., Smith, J., & Marcovitch, H. (2011). *Wakefield's article linking MMR vaccine and autism was fraudulent*. British Medical Journal Publishing Group.

Haack, T. B., Hogarth, P., Kruer, M. C., Gregory, A., Wieland, T., Schwarzmayr, T., … Hayflick, S. J. (2012). Exome sequencing reveals de novo WDR45 mutations causing a phenotypically distinct, X-linked dominant form of NBIA. *American Journal of Human Genetics*, *91*(6), 1144–1149. https://doi.org/10.1016/j.ajhg.2012.10.019

Hamdan, F. F., Daoud, H., Piton, A., Gauthier, J., Dobrzeniecka, S., Krebs, M.-O., … Michaud, J. L. (2011). De Novo SYNGAP1 Mutations in Nonsyndromic Intellectual Disability and Autism. *Biological Psychiatry*, *69*(9), 898–901. https://doi.org/10.1016/j.biopsych.2010.11.015

Hamdan, F. F., Daoud, H., Rochefort, D., Piton, A., Gauthier, J., Langlois, M., … Michaud, J. L. (2010). De Novo Mutations in FOXP1 in Cases with Intellectual Disability, Autism, and Language Impairment. *The American Journal of Human Genetics*, *87*(5), 671–678. https://doi.org/10.1016/j.ajhg.2010.09.017

Hayflick, S. J., Kruer, M. C., Gregory, A., Haack, T. B., Kurian, M. A., Houlden, H. H., … Hogarth, P. (2013). β-Propeller protein-associated neurodegeneration: a new X-linked

dominant disorder with brain iron accumulation. *Brain: A Journal of Neurology*, *136*(Pt 6), 1708–1717. https://doi.org/10.1093/brain/awt095

Hecht, M., Bromberg, Y., & Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC Genomics*, *16*(Suppl 8), S1. https://doi.org/10.1186/1471-2164-16-S8-S1

Hoffjan, S., Ibisler, A., Tschentscher, A., Dekomien, G., Bidinost, C., & Rosa, A. L. (2016). WDR45 mutations in Rett (-like) syndrome and developmental delay: Case report and an appraisal of the literature. *Molecular and Cellular Probes*, *30*(1), 44–49. https://doi.org/10.1016/j.mcp.2016.01.003

Howlin, P., & Moore, A. (1997). Diagnosis in Autism: A Survey of Over 1200 Patients in the UK. *Autism*, *1*(2), 135–162. https://doi.org/10.1177/1362361397012003

Iossifov, I., Levy, D., Allen, J., Ye, K., Ronemus, M., Lee, Y., … Wigler, M. (2015). Low load for disruptive mutations in autism genes and their biased transmission. *Proceedings of the National Academy of Sciences*, *112*(41), E5600–E5607. https://doi.org/10.1073/pnas.1516376112

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

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., … Wigler, M. (2012). De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron*, *74*(2), 285–299. https://doi.org/10.1016/j.neuron.2012.04.009

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

Kanner, L. (1949). Problems of Nosology and Psychodynamics of Early Infantile Autism*. *American Journal of Orthopsychiatry*, *19*(3), 416–426. https://doi.org/10.1111/j.1939-0025.1949.tb05441.x

Kanner, L., & others. (1943). Autistic disturbances of affective contact. *Nervous Child*, *2*(3), 217–250.

Kircher, M., Witten, D. M., Jain, P., O, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. https://doi.org/10.1038/ng.2892

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kogan, M. D., Blumberg, S. J., Schieve, L. A., Boyle, C. A., Perrin, J. M., Ghandour, R. M., … Dyck, P. C. van. (2009). Prevalence of Parent-Reported Diagnosis of Autism Spectrum Disorder Among Children in the US, 2007. *Pediatrics*, *124*(5), 1395–1403. https://doi.org/10.1542/peds.2009-1522

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., … Stefansson, K. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, *488*(7412), 471–475. https://doi.org/10.1038/nature11396

Kosmicki, J. A., Samocha, K. E., Howrigan, D. P., Sanders, S. J., Slowikowski, K., Lek, M., … Daly, M. J. (2017). Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nature Genetics*. https://doi.org/10.1038/ng.3789

Krumm, N., O'Roak, B. J., Shendure, J., & Eichler, E. E. (2014). A de novo convergence of autism genetics and molecular neuroscience. *Trends in Neurosciences*, *37*(2), 95–105. https://doi.org/10.1016/j.tins.2013.11.005

Krumm, N., Turner, T. N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., … Eichler, E. E. (2015). Excess of rare, inherited truncating mutations in autism. *Nature Genetics*, *47*(6). https://doi.org/10.1038/ng.3303

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

La Malfa, G., Lassi, S., Bertelli, M., Salvini, R., & Placidi, G. F. (2004). Autism and intellectual disability: a study of prevalence on a sample of the Italian population. *Journal of Intellectual Disability Research*, *48*(3), 262–267. https://doi.org/10.1111/j.1365-2788.2003.00567.x

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., … Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

Leach, N. T., Sun, Y., Michaud, S., Zheng, Y., Ligon, K. L., Ligon, A. H., … Morton, C. C. (2007). Disruption of Diacylglycerol Kinase Delta (DGKD) Associated with Seizures in Humans and Mice. *American Journal of Human Genetics*, *80*(4), 792–799.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … Macarthur, D. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature Publishing Group*, *536*. https://doi.org/10.1038/nature19057

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Cai, T., Jiang, Y., Chen, H., He, X., Chen, C., … Wu, J. (2015). Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Molecular Psychiatry*, *21*, 290–297. https://doi.org/10.1038/mp.2015.40

Liu, X., Jian, X., & Boerwinkle, E. (2011). dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation*, *32*(8), 894–899. https://doi.org/10.1002/humu.21517

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

Long, M., Abdeen, N., Geraghty, M. T., Hogarth, P., Hayflick, S., & Venkateswaran, S. (2015). Novel WDR45 mutation and pathognomonic BPAN imaging in a young female with mild cognitive delay. *Pediatrics*, *136*(3), e714–e717. https://doi.org/10.1542/peds.2015-0750

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (2000). The autism diagnostic observation schedule—generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223. https://doi.org/10.1023/A:1005592401947

Lord, C., Rutter, M., & Couteur, A. L. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *24*(5), 659–685. https://doi.org/10.1007/BF02172145

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., Bishop, S., & others. (2012). *Autism diagnostic observation schedule: ADOS-2*. Western Psychological Services Los Angeles, CA.

Mah, S., Nelson, M. R., DeLisi, L. E., Reneland, R. H., Markward, N., James, M. R., … Braun, A. (2006). Identification of the semaphorin receptor PLXNA2 as a candidate for susceptibility to schizophrenia. *Molecular Psychiatry*, *11*(5), 471–478. https://doi.org/10.1038/sj.mp.4001785

Majithia, A. R., Tsuda, B., Agostini, M., Gnanapradeepan, K., Rice, R., Peloso, G., … Altshuler, D. (2016). Prospective functional classification of all possible missense variants in PPARG. *Nature Genetics*, *48*(12), 1570–1575. https://doi.org/10.1038/ng.3700

Mandell, D. S., Novak, M. M., & Zubritsky, C. D. (2005). Factors Associated With Age of Diagnosis Among Children With Autism Spectrum Disorders. *Pediatrics*, *116*(6), 1480–1486. https://doi.org/10.1542/peds.2005-0185

Martínez-Cerdeño, V. (2017). Dendrite and spine modifications in autism and related neurodevelopmental disorders in patients and animal models. *Developmental Neurobiology*, *77*(4), 393–404. https://doi.org/10.1002/dneu.22417

Matson, J. L., & Kozlowski, A. M. (2011). The increasing prevalence of autism spectrum disorders. *Research in Autism Spectrum Disorders*, *5*(1), 418–425. https://doi.org/10.1016/j.rasd.2010.06.004

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), nrg2344. https://doi.org/10.1038/nrg2344

McCarthy, S. E., Gillis, J., Kramer, M., Lihm, J., Yoon, S., Berstein, Y., … Corvin, A. (2014). De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular Psychiatry*, *19*(6), 652–658. https://doi.org/10.1038/mp.2014.29

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Miles, J. H., Takahashi, T. N., Hong, J., Munden, N., Flournoy, N., Braddock, S. R., … Farmer, J. E. (2008). Development and validation of a measure of dysmorphology: Useful for autism subgroup classification. *American Journal of Medical Genetics, 146A*(9), 1101–1116. https://doi.org/10.1002/ajmg.a.32244

Nakashima, M., Takano, K., Tsuyusaki, Y., Yoshitomi, S., Shimono, M., Aoki, Y., … Matsumoto, N. (2016). WDR45 mutations in three male patients with West syndrome. *Journal of Human Genetics*, *61*(7), 653–661. https://doi.org/10.1038/jhg.2016.27

Neale, B. M., Kou, Y., Liu, L., Ma ’ayan, A., Samocha, K. E., Sabo, A., … Daly, M. J. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, *485*. https://doi.org/10.1038/nature11011

Ng, P. C., & Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. *Genome Research*, *11*(5), 863–874. https://doi.org/10.1101/gr.176601

Nijman, S. M. B. (2011). Synthetic lethality: General principles, utility and detection using genetic screens in human cells. *Febs Letters*, *585*(1), 1–6. https://doi.org/10.1016/j.febslet.2010.11.024

Ohba, C., Nabatame, S., Iijima, Y., Nishiyama, K., Tsurusaki, Y., Nakashima, M., … Matsumoto, N. (2014). De novo WDR45 mutation in a patient showing clinically Rett syndrome with childhood iron deposition in brain. *Journal of Human Genetics*, *59*(5), 292–295. https://doi.org/10.1038/jhg.2014.18

Olsson, I., Steffenburg, S., & Gillberg, C. (1988). Epilepsy in Autism and Autisticlike Conditions: A Population-Based Study. *Archives of Neurology*, *45*(6), 666–668. https://doi.org/10.1001/archneur.1988.00520300086024

O'Roak, B. J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J. J., Girirajan, S., … Eichler, E. E. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, *43*(6), 585–589. https://doi.org/10.1038/ng.835

O'Roak, B. J., Stessman, H. A., Boyle, E. A., Witherspoon, K. T., Martin, B., Lee, C., … Eichler, E. E. (2014). Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nature Communications*, *5*, 5595. https://doi.org/10.1038/ncomms6595

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

Pardo, C. A., & Eberhart, C. G. (2007). The Neurobiology of Autism. *Brain Pathology*, *17*(4), 434–447. https://doi.org/10.1111/j.1750-3639.2007.00102.x

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., … Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, *466*(7304), 368–372. https://doi.org/10.1038/nature09146

Rands, C. M., Meader, S., Ponting, C. P., & Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the

Human Lineage. *PLOS Genetics*, *10*(7), e1004525. https://doi.org/10.1371/journal.pgen.1004525

Redfern, R. E., Redfern, D., Furgason, M. L. M., Munson, M., Ross, A. H., & Gericke, A. (2008). PTEN Phosphatase Selectively Binds Phosphoinositides and Undergoes Structural Changes. *Biochemistry*, *47*(7), 2162–2171. https://doi.org/10.1021/bi702114w

Redon, S., Benech, C., Schutz, S., Despres, A., Gueguen, P., Le Berre, P., … Ferec, C. (2017). Intragenic deletion of the WDR45 gene in a male with encephalopathy, severe psychomotor disability, and epilepsy. *American Journal of Medical Genetics Part A*, *173*(5), 1444–1446. https://doi.org/10.1002/ajmg.a.38180

Robinson, E. B., Samocha, K. E., Kosmicki, J. A., McGrath, L., Neale, B. M., Perlis, R. H., & Daly, M. J. (2014). Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(42), 15161–15165. https://doi.org/10.1073/pnas.1409204111

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

Rodríguez-Escudero, I., Oliver, M. D., Andrés-Pons, A., Molina, M., Cid, V. J., & Pulido, R. (2011). A comprehensive functional analysis of PTEN mutations: implications in tumor- and autism-related syndromes. *Human Molecular Genetics*, *20*(21), 4132–4142. https://doi.org/10.1093/hmg/ddr337

Rogic, S., Holmes, N., Jacobson, M., Belmadani, M., Tan, P. P. C., & Pavlidis, P. (Manuscript in preparation). MARVdb: Meta-Analysis of Rare Variants.

Rohatgi, S., Clark, D., Kline, A. D., Jackson, L. G., Pie, J., Siu, V., … Deardorff, M. A. (2010). Facial diagnosis of mild and variant CdLS: insights from a dysmorphologist survey. *American Journal of Medical Genetics. Part A, 0*(7), 1641–1653. https://doi.org/10.1002/ajmg.a.33441

Ronald, A., Happé, F., Bolton, P., Butcher, L. M., Price, T. S., Wheelwright, S., … Plomin, R. (2006). Genetic Heterogeneity Between the Three Components of the Autism Spectrum: A Twin Study. *Journal of the American Academy of Child & Adolescent Psychiatry, 45*(6), 691–699. https://doi.org/10.1097/01.chi.0000215325.13058.9d

Rosenberg, R. E., Law, J. K., Yenokyan, G., McGready, J., Kaufmann, W. E., & Law, P. A. (2009). Characteristics and Concordance of Autism Spectrum Disorders Among 277 Twin Pairs. *Archives of Pediatrics & Adolescent Medicine, 163*(10), 907–914. https://doi.org/10.1001/archpediatrics.2009.98

Sanders, S. J., Gulhan Ercan-Sencicek, A., Hus, V., Luo, R., Murtha, M. T., Moreno-De-Luca, D., … State, M. W. (2011). Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. *Neuron, 70*, 863–885. https://doi.org/10.1016/j.neuron.2011.05.002

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., … State, M. W. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*. https://doi.org/10.1016/j.neuron.2015.09.016

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., … State, M. W. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*. https://doi.org/10.1038/nature10945

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., & Reichenberg, A. (2017). The Heritability of Autism Spectrum Disorder. *Jama*, *318*(12), 1182–1184.

Schaaf, C. P., Sabo, A., Sakai, Y., Crosby, J., Muzny, D., Hawes, A., … Zoghbi, H. Y. (2011). Oligogenic heterozygosity in individuals with high-functioning autism spectrum disorders. *Human Molecular Genetics*, *20*(17), 3366–3375. https://doi.org/10.1093/hmg/ddr243

Schaaf, C. P., & Zoghbi, H. Y. (2011). Solving the Autism Puzzle a Few Pieces at a Time. *Neuron*, *70*(5), 806–808. https://doi.org/10.1016/j.neuron.2011.05.025

Schieve, L. A., Rice, C., Devine, O., Maenner, M. J., Lee, L.-C., Fitzgerald, R., … others. (2011). Have secular changes in perinatal risk factors contributed to the recent autism prevalence increase? Development and application of a mathematical assessment model. *Annals of Epidemiology*, *21*(12), 930–945.

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

SFARI. (2016, January 12). Making sense out of missense mutations. Retrieved November 7, 2017, from https://www.sfari.org/2016/01/12/making-sense-out-of-missense-mutations/

Shang, L., Cho, M. T., Retterer, K., Folk, L., Humberson, J., Rohena, L., … Chung, W. K. (2015). Mutations in ARID2 are associated with intellectual disabilities. *Neurogenetics*, *16*(4), 307–314. https://doi.org/10.1007/s10048-015-0454-0

Sherry, S. T., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1).

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 238–241.

Srivastava, D. P., Woolfrey, K. M., Jones, K. A., Anderson, C. T., Smith, K. R., Russell, T. A., … Penzes, P. (2012). An autism-associated variant of Epac2 reveals a role for Ras/Epac2 signaling in controlling basal dendrite maintenance in mice. *PLOS Biology, 10*(6), e1001350. https://doi.org/10.1371/journal.pbio.1001350

Tammimies, K., Marshall, C. R., Walker, S., Kaur, G., Thiruvahindrapuram, B., Lionel, A. C., … Fernandez, B. A. (2015). Molecular diagnostic yield of chromosomal microarray analysis and whole-exome sequencing in children with autism spectrum disorder. *JAMA, 314*(9), 895–903. https://doi.org/10.1001/jama.2015.10078

Tan, P. P. C., Rogic, S., Zoubarev, A., McDonald, C., Lui, F., Charathsandran, G., … Pavlidis, P. (2016). Interactive exploration, analysis, and visualization of complex phenome–genome datasets with ASPIREdb. *Human Mutation*, *37*(8), 719–726. https://doi.org/10.1002/humu.23011

Tavazoie, S. F., Alvarez, V. A., Ridenour, D. A., Kwiatkowski, D. J., & Sabatini, B. L. (2005). Regulation of neuronal morphology and function by the tumor suppressors Tsc1 and Tsc2. *Nature Neuroscience*, *8*(12), 1727–1734. https://doi.org/10.1038/nn1566

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

The Autism Spectrum Disorders Working Group of the Psychiatric Genomics Consortium. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Molecular Autism*, *8*(1), 21. https://doi.org/10.1186/s13229-017-0137-9

Tuchman, R., Cuccaro, M., & Alessandri, M. (2010). Autism and epilepsy: Historical perspective. *Brain and Development*, *32*(9), 709–718. https://doi.org/10.1016/j.braindev.2010.04.008

Turner, T. N., Coe, B. P., Dickel, D. E., Hoekzema, K., Nelson, B. J., Zody, M. C., … others. (2017). Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell*, *171*(3), 710–722.

Turner, T. N., Yi, Q., Krumm, N., Huddleston, J., Hoekzema, K., F. Stessman, H. A., … Eichler, E. E. (2017). denovo-db: a compendium of human de novo variants. *Nucleic Acids Research*, *45*(D1), D804–D811. https://doi.org/10.1093/nar/gkw865

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

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five Years of GWAS Discovery. *The American Journal of Human Genetics*, *90*(1), 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wakefield, A., Murch, S., Anthony, A., Linnell, J., Casson, D., Malik, M., … Walker-Smith, J. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, *351*(9103), 637–641. https://doi.org/10.1016/S0140-6736(97)11096-0

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

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J. T., Abrahams, B. S., … Hakonarson, H. (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*, *459*(7246), 528–533. https://doi.org/10.1038/nature07999

Wang, T., Guo, H., Xiong, B., Stessman, H. A. F., Wu, H., Coe, B. P., … Eichler, E. E. (2016). *De novo* genic mutations among a Chinese autism spectrum disorder cohort. *Nature Communications*, *7*, ncomms13316. https://doi.org/10.1038/ncomms13316

Wazana, A., Bresnahan, M., & Kline, J. (2007). The Autism Epidemic: Fact or Artifact? *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*(6), 721–730. https://doi.org/10.1097/chi.0b013e31804a7f3b

Wiggins, L. D., Robins, D. L., Adamson, L. B., Bakeman, R., & Henrich, C. C. (2012). Support for a Dimensional View of Autism Spectrum Disorders in Toddlers. *Journal of Autism and Developmental Disorders*, *42*(2), 191–200. https://doi.org/10.1007/s10803-011-1230-0

Williams, C. A., Dagli, A., & Battaglia, A. (2008). Genetic disorders associated with macrocephaly. *American Journal of Medical Genetics Part A*, *146*(15), 2023–2037.

Willsey, A. J., Sanders, S. J., Li, M., Dong, S., Tebbenkamp, A. T., Muhle, R. A., … State, M. W. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*, *155*(5), 997–1007. https://doi.org/10.1016/j.cell.2013.10.020

Yatsenko, S. A., Hixson, P., Roney, E. K., Scott, D. A., Schaaf, C. P., Ng, Y., … Lupski, J. R. (2012). Human subtelomeric copy number gains suggest a DNA replication mechanism for formation: beyond breakage – fusion - bridge for telomere stabilization. *Human Genetics*, *131*(12), 1895–1910. https://doi.org/10.1007/s00439-012-1216-9

Yıldız Bölükbaşı, E., Afzal, M., Mumtaz, S., Ahmad, N., Malik, S., & Tolun, A. (2017). Progressive SCAR14 with unclear speech, developmental delay, tremor, and behavioral problems caused by a homozygous deletion of the SPTBN2 pleckstrin homology domain.

American Journal of Medical Genetics Part A, *173*(9), 2494–2499. https://doi.org/10.1002/ajmg.a.38332

Young, B. P., & Loewen, C. J. (2013). Balony: a software package for analysis of data generated by synthetic genetic array experiments. *BMC Bioinformatics*, *14*, 354. https://doi.org/10.1186/1471-2105-14-354

Yuen, R. K. C., Merico, D., Bookman, M., Howe, J. L., Thiruvahindrapuram, B., Patel, R. V., … Scherer, S. W. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience*, *20*(4), nn.4524. https://doi.org/10.1038/nn.4524

Yuen, R. K. C., Merico, D., Cao, H., Pellecchia, G., Alipanahi, B., Thiruvahindrapuram, B., … Scherer, S. W. (2016). Genome-wide characteristics of de novo mutations in autism. *Npj Genomic Medicine*, *1*, 16027-1-16027–10. https://doi.org/10.1038/npjgenmed.2016.27

Zhang, X. C., Piccini, A., Myers, M. P., Van Aelst, L., & Tonks, N. K. (2012). Functional analysis of the protein phosphatase activity of PTEN. *Biochemical Journal*, *444*(Pt 3), 457–464. https://doi.org/10.1042/BJ20120098

Appendices

Appendix A : Prioritized variants for functional characterization

A.1 PTEN prioritized variants

Table 4-1. Prioritized variants for PTEN.
Category column indicates which prioritization technique was used to prioritize the entry. Biochem = Biochemically validated variant. ClinVar = pathogenic variant as listed in the ClinVar database (link). ExAC = Population variants found in the ExAC database of 60,706 subjects. YA = Yeast Activity in SGA screen. PQ = Protein Quantification results. CADD score is Phred-scaled. Genomic change reported for hg19. Source column indicates either the publication in which the variant is found, or the database–specific ID, where applicable. A * in the Source column indicates the variant was reported as *de novo* in an ASD proband.

Variant	Category	Genomic Change (hg19)	Snap2	CADD	Yeast Activity	Protein Quant	Source
WT	NA	NA	NA	NA	0.91	1	NA
4A	Biochemical	NA	NA	NA	1.59	1.19	NA
M1I	ASD	10:89624229:G/T	-92	19.05	-0.01	0.03	Hobert, 2013
K6E	ASD	10:89624242:A/G	67	19.93	0.09	0.97	Vanderver, 2014
K6I	ASD	10:89624243:A/T	64	18.24	-0.09	0.76	Vanderver, 2014
N12T	ASD	10:89624261:A/C	-39	16.03	0.37	1.03	Frazier, 2014
R14G	ASD	10:89624266:A/G	52	16.13	-0.04	1.1	Frazier, 2014
R15S	ASD	10:89624271:A/T	65	20.9	0.1	0.52	Vanderver, 2014
D22E	ASD	10:89624292:C/G	18	16.96	0.02	0.91	Buxbaum, 2007
Y27C	ASD	10:89653782:A/G	17	16.92	0.06	0.67	Vanderver, 2014*
A34D	Predicted High Impact	10:89653803:C/A	86	32	-0.08	1.13	
A34P	Predicted High Impact	10:89653802:G/C	82	33	-0.03	1.09	
G36E	Predicted High Impact	NA	82	31	-0.17	0.02	
P38H	ASD	10:89653815:C/A	44	28.1	-0.21	0.84	Klen, 2013
G44D	ASD	10:89653833:G/A	65	31	-0.24	1.26	Frazier, 2014
R47K	Predicted High Impact	10:89653842:G/A	82	35	-0.09	1.1	
R47W	Predicted High Impact	10:89653841:A/T	90	31	-0.02	0.77	

Variant	Category	Genomic Change (hg19)	Snap2	CADD	Yeast Activity	Protein Quant	Source
F56C	Population	10:89685272:T/G	12	22	0.1	1.29	
Y65C	ASD	10:89685299:A/G	33	20.7	0.83	0.9	Vanderver, 2014
Y68H	ASD	10:89685307:T/C	86	23.3	0.03	0.29	Vanderver, 2014
Y68N	ASD	NA	87	23.7	-0.08	0.23	Vanderver, 2014
L70V	ASD	10:89685313:C/G	49	21.6	0.52	0.94	Hobert, 2013
E73K	Predicted High Impact	10:89690810:G/A	76	36	0.67	0.72	
T78A	ASD	10:89690825:A/G	-77	12.12	0.81	0.6	Schaaf, 2011*
A79T	Population	10:89690828:G/A	-95	13.23	0.96	1.34	
D92N	ASD	10:89692790:G/A	90	33		1.17	Iossifov, 2011*
H93Q	Cancer Variant	NA		20.03	0.81	1.28	
H93R	ASD	10:89692794:A/G	64	23.2	1.53	0.93	Butler, 2005*
H93Y	ASD	10:89692793:C/T	67	26.3	0.58	1.22	Butler, 2005
I101F	ClinVar	10:89692817:A/T	6	27.6	0.6	0.86	
I101T	ASD	10:89692818:T/C	44	23.4	-0.04	0.5	O'Roak, 2014*
D107G	ASD	10:89692836:A/G	39	26.1	0.34	0.57	O'Roak, 2012
D107V	ASD	10:89692836:A/T	25	25.2	-0.06	0.22	O'Roak, 2014*
N117S	Population	10:89692866:A/G	-92	10.49	0.91	0.95	
H118P	ASD	10:89692869:A/C	85	23.6	0.02	0.41	Orrico, 2009
A121E	Predicted High Impact	10:89692878:C/A	79	32	0.02	0.57	
A121P	Predicted High Impact	10:89692877:G/C	83	33	0.21	0.23	
H123Q	ASD	10:89692885:C/G	82	22	-0.08	0.93	McBride, 2010
H123Y	ClinVar	10:89692883:C/T	80	28.2	-0.4	0.9	
C124S	Biochemical	10:89692886:T/A	83	27.5	-0.22	1	
C124S-4A	Biochemical	NA			-0.48	0.42	
A126D	Predicted High Impact	10:89692893:C/A	76	32	-0.13	1.18	
A126P	Predicted High Impact	NA	79	32	-0.2	0.87	

Variant	Category	Genomic Change (hg19)	Snap2	CADD	Yeast Activity	Protein Quant	Source
G127R	Predicted High Impact	10:89692895:G/A	91	32	-0.28	0.74	
G129E	Biochemical	10:89692902:G/A	89	28.7	-0.23	1.55	Rodriguez-Escudero, 2011
G129R	ClinVar	10:89692901:G/A	92	31	-0.28	0.74	Rodriguez-Escudero, 2011
R130L	ASD	10:89692905:G/T	92	34	-0.15	1.11	Klen, 2013
R130Q	Predicted High Impact	10:89692905:G/A	81	36	-0.06	1.08	
R130X	ASD-LGD	NA		48	0.03	0.07	
T131I	ASD	10:89692908:C/T	80	27.8	-0.06	1.29	O'Roak, 2012*
G132D	ASD	10:89692911:G/A	92	31	-0.09	0.7	Frazier, 2014
M134I	ASD	10:89692918:G/A	-15	31	0.04	0.75	Hobert, 2013
M134T	ASD	10:89692917:T/C	-22	22	-0.03	0.06	Hobert, 2013
I135fs	ASD-LGD	NA			-0.4		O'Roak, 2012*
I135T	Population	10:89692920:T/C	33	24.2	0.57	0.92	
I135V	ClinVar	10:89692919:A/G	-72	20.7	0.49	0.91	
C136Mfs	ASD-LGD	NA			-0.11		O'Roak, 2012*
Y138L	Biochemical	NA	44		0.22	0.87	
L139X	ASD-LGD	10:89692932:T/A		48	0		Varga, 2009
A151P	Predicted High Impact	10:89692967:G/C	33	33	-0.04	0.8	
E157G	ASD	10:89692986:A/G	8	25.3	1.28	0.63	Iossifov, 2011
T167N	ASD	10:89711882:C/A	45	28.6	0.55	0.83	O'Roak, 2012*
Q171E	ASD	10:89711893:C/G	45	27.2	-0.08	0.75	Vanderver, 2014
R173H	ASD	10:89711900:G/A	54	35	0.29	0.42	Rodriguez-Escudero, 2011
R173P	Predicted High Impact	10:89711900:G/C	31	31	-0.04	0.92	
Y176C	ASD	10:89711909:A/G	-47	13.94	0.59	0.82	Orrico, 2009*
Y178X	ASD-LGD	10:89711916:T/G		50	0.04	0.11	
Y180H	Population	10:89711920:T/C	-78	11.78	-0.12	0.08	

Variant	Category	Genomic Change (hg19)	Snap2	CADD	Yeast Activity	Protein Quant	Source
M198I	Population	10:89711976:G/A	-85	8.539	0.98	1.2	Hobert, 2013
T202I	ASD	10:89711987:C/T	-14	28.5	0.64	0.74	Rodriguez-Escudero, 2011
I203V	Predicted Low Impact	10:89711989:A/G	-84	4.851	0.99	0.02	
C211W	ASD	10:89712015:C/G	73	25.5	0.71	0.85	O'Roak, 2012
N228S	Population	10:89717658:A/G	-78	7.946	0.99	1.07	
F241S	ASD	10:89717697:T/C	21	17.1	-0.02	0.18	Butler, 2005*
P246L	ASD	10:89717712:C/T	35	19.43	0.38	0.47	Vanderver, 2014
D252G	ASD	10:89717730:A/G	80	26.5	0.15	0.22	Butler, 2005*
V255A	ASD	10:89717739:T/C	36	22	0.21	0.67	Klein,2013
E256K	Predicted High Impact	10:89717741:G/A	35	35	0.83	1.52	Iossifov, 2011
N262S	Population	10:89717760:A/G	-63	11.48	1.23	1.18	
D268E	Population	10:89720653:C/G	-89	1.668	0.94	1	Butler, 2005
W274L	ASD	10:89720670:G/T	81	26.1	0.77	0.63	McBride, 2010
N276S	ASD	10:89720676:A/G	87	24.5	-0.26	0.06	Orrico, 2009*
F279I	Predicted High Impact	10:89720684:T/A	31	31	-0.05	1.19	Butler, 2005
F279L	Predicted High Impact	10:89720684:T/C	31	31	0.88	0.93	Butler, 2005
G285X	ASD-LGD	NA		48	1.09	1.17	Frazier, 2014
L295V	Predicted Low Impact	10:89720732:C/G	-67	4.363	1.15	0.4	Varga, 2009
Q298E	Population	10:89720741:C/G	-75	10.98	0.98	1.29	
E307Q	Population	10:89720768:G/C	-76	14.28	0.92	1	
A309S	Predicted Low Impact	10:89720774:G/T	-80	4.925	1.04	1.18	
L320X	ASD-LGD	10:89720808:T/G		48	0.03	0.01	Varga, 2009
K322E	Predicted High Impact	10:89720813:A/G	31	31	0.67	1.25	
D326N	ASD	10:89720825:G/A	74	35	0.25	0.93	Buxbaum, 2007*
K330E	Predicted High Impact	10:89720837:A/G	31	31	0.97	1.52	

Variant	Category	Genomic Change (hg19)	Snap2	CADD	Yeast Activity	Protein Quant	Source
R335X	ASD-LGD	10:89720852:C/T		48	-0.27	0.05	
Y346F	Predicted Low Impact	10:89725054:A/T	-55	0.328	0.98	0.94	Vanderver, 2014
T348S	Predicted Low Impact	10:89725059:A/T	-83	3.372	0.94	1.16	
P354Q	Population	10:89725078:C/A	-52	14.55	0.94	1.14	
N356D	Population	10:89725083:A/G	-59	13.42	1.05	0.81	
N356H	Population	10:89725083:A/C	-82	12.87	1	1.08	
P357S	Predicted Low Impact	NA	-81	3.123	1.03	0.9	
T363N	Population	10:89725105:C/A	-53	12.86	1.01	1.04	
Q396R	Population	10:89725204:A/G	-51	13.5	1	0.08	

A.2 SYNGAP1 prioritized variants

Table 4-2. Prioritized variants for SYNGAP1.
Category column indicates which prioritization technique was used to prioritize the entry. Biochem = Biochemically validated variant. ClinVar = pathogenic variant as listed in the ClinVar database (link). ExAC = Population variants found in the ExAC database of 60,706 subjects. CADD score is Phred-scaled. Genomic change reported for hg19. Source column indicates either the publication in which the variant is found, or the database–specific ID, where applicable. A * in the Source column indicates the variant was reported as *de novo* in an ASD proband.

Variant	*Category*	*Genomic Change*	*CADD*	*Snap2*	*ExAC MAF*	*Source*
R41H	ASD-associated missense	6:33391308 G/A	16.77	13	0.0000082	Codina-Sola, 2015*
R47Q	ASD-associated missense	6:33391326 G/A	15.24	13	0	DeRubeis, 2014
E75D	Population	6:33393610 G/C	7.074	-21	0.0000084	ExAC
L102V	Population	6:33399946 T/G	3.699	-60	0.0000084	ExAC
P111fs	ASD-associated LGD	6:33399974 CA/C	NA	NA	0	O'Roak, 2014
Y116X	ASD-associated LGD	6:33399990 C/A	22.5	NA	0	Mignot, 2016
E119D	Population	6:33399999 G/T	5.563	-31	0.000017	ExAC
R143X	ASD-associated LGD	6:33400501 C/T	26.4	NA	0	Mignot, 2016*
T144fs	ASD-associated LGD	6:33400504 ACGAA/A	NA	NA	0	Parker, 2015*
R152fs	ASD-associated LGD	6:33400528 CGGACC/C	NA	NA	0	Mignot, 2016*
R164X	ASD-associated LGD	6:33400564 C/T	34	NA	0	Mignot, 2016*
R170Q	ASD-associated missense	6:33400583 G/A	21	4	0	McRae, 2016*
R170Q	ASD-associated missense	6:33400583 G/A	21	4	0	DDD
D201E	Population	6:33403022 T/G	6.31	-55	0.0000082	ExAC
C233Y	ASD-associated missense	6:33403326 G/A	23.4	57	0	Kosmicki, 2017*
W267X	ASD-associated LGD	6:33405482 G/A	37	NA	0	Carvill, 2013*
P276A	Population	6:33405508 C/G	3.262	-59	0.0000082	ExAC
S296P	ASD-associated missense	6:33405568 T/C	18.13	-31	0	MSSNG*
L327P	ASD-associated missense	6:33405662 T/C	20.4	25	0	Parker, 2015*
D332N	ASD-associated missense	6:33405676 G/A	26.1	-35	0	MSSNG*
T347N	Population	6:33405722 C/A	0.472	-59	0.0000084	ExAC
Q361X	ASD-associated LGD	6:33405763 C/T	37	NA	0	DeRubeis, 2014*

Variant	*Category*	*Genomic Change*	*CADD*	*Snap2*	*ExAC MAF*	*Source*
W362R	ASD-associated missense	6:33405766 T/C	18.56	58	0	Berryer, 2013*
W362R	ASD-associated missense	6:33405766 T/A	19.12	58	0	Berryer, 2013*
M377I	Population	6:33405813 G/A	0.953	-18	0.0000097	ExAC
S385W	Clinvar Benign	6:33405836 C/G	6.758	79	0	
V400L	Population	6:33405880 G/C	5.558	-83	0.0000084	ExAC
Y417fs	ASD-associated LGD	6:33405933 TAA/T	NA	NA	0	O'Roak, 2014*
K418fs	ASD-associated LGD	6:33405934 AAA/A	NA	NA	0	Mignot, 2016*
V434I	Clinvar Uncertain	6:33405982 G/A	18.64	-84	0	
A438V	Population	6:33405995 C/T	4.565	-63	0.0000165	ExAC
M477V	Population	6:33406238 A/G	6.628	-71	0.0000082	ExAC
R485P	Biochemical	6:33406263 G/C	21.7	25	0	Pena, 2008
R485A	Biochemical	NA	NA	NA	0	Ahmadian, 2003
R485K	Biochemical	NA	NA	NA	0	Ahmadian, 2003
N487T	Biochemical	6:33406269 A/C	19.56	-55	0	Pena, 2008
G511R	ASD-associated missense	6:33406340 G/C	19.05	6	0	McRae, 2016*
S535T	Population	6:33406624 G/C	8.321	-60	0.0000411	ExAC
S557fs	ASD-associated LGD	6:33406690 C/CA	NA	NA	0	Yuen, 2016
H558del	ASD-associated LGD	6:33406693 ACTGGT/A	NA	NA	0	Yuen, 2016
P562L	ASD-associated missense	6:33408514 C/T	33	17	0	Berryer, 2013*
P562L	MRAD5	6:33408514 C/T	33	17	0	Berryer, 2013*
R573L	Pred High Impact	6:33408547 G/T	31	53	0	NA
F594fs	ASD-associated LGD	6:33408610 TC/C	NA	NA	0	DeRubeis, 2014*
L595G	Biochemical	NA	NA	NA	0	Ahmadian, 2003
L595A	Biochemical	NA	NA	NA	0	Ahmadian, 2003
R596M	Biochemical	NA	NA	NA	0	Ahmadian, 2003
R596A	Biochemical	NA	NA	NA	0	Ahmadian, 2003
R596K	Biochemical – Negative	NA	NA	NA	0	Ahmadian, 2003
R596L	Pred High Impact	6:33408616 G/T	34	64	0	NA
L607fs	ASD-associated LGD	6:33408650 CTT/C	NA	NA	0	Iossifov, 2012

Variant	*Category*	*Genomic Change*	*CADD*	*Snap2*	*ExAC MAF*	*Source*
K628E	Biochemical	6:33408711 A/G	25.8	2	0	Ahmadian, 2003
N635D	Biochemical	6:33408732 A/G	20.6	-29	0	Ahmadian, 2003
P701S	Population	6:33409137 C/T	3.25	-45	0.00000825	ExAC
P728fs	ASD-associated LGD	6:33409425 CC/C	NA	NA	0	Berryer, 2013*
P734S	Population	6:33409442 C/T	8.419	-35	0.00000833	ExAC
S738fs	ASD-associated LGD	6:33409455 GTGAG/G	NA	NA	0	Mignot, 2016*
P741S	Population	6:33409463 C/T	5.766	-58	0.0000083	ExAC
P743S	Population	6:33409469 C/T	5.436	-54	0.0000084	ExAC
M759V	Population	6:33409517 A/G	9.559	-24	0.0000169	ExAC
R775Q	Clinvar Uncertain	6:33410259 G/A	10.96	16	0.0000247	
S788A	Biochemical	6:33410691 T/G	18.68	-54	0	Walkup, 2014
S788D	Biochemical	NA	NA	NA	0	Walkup, 2014
T790A	Biochemical	6:33410697 A/G	18.48	-61	0	Walkup, 2014
P794fs	ASD-associated LGD	6:33410711 A/AC	NA	NA	0	MSSNG*
P794L	Population	6:33410710 C/T	6.553	-29	0.0000166	ExAC
G800A	Population	6:33410728 G/C	0.261	-16	0.0000165	ExAC
S817A	Biochemical	6:33410778 T/G	17.27	-46	0	Walkup, 2014
L877fs	ASD-associated LGD	6:33410958 C/CT	NA	NA	0	O'Roak, 2014*
A888G	Population	6:33410992 C/G	3.782	-64	0.00000834	ExAC
I899V	Population	6:33411024 A/G	4.637	-67	0.00000840	ExAC
M904V	Population	6:33411039 A/G	4.513	-62	0.00001677	ExAC
R922X	ASD-associated LGD	6:33411093 C/T	39	NA	0	Parker, 2015*
L925fs	ASD-associated LGD	6:33411102 CT/C	NA	NA	0	Parker, 2015*
P941S	Population	6:33411150 C/T	2.315	-3	0.00000828	ExAC
G949S	Clinvar Uncertain	6:33411174 G/A	1.426	21	0.0002	
R967X	ASD-associated LGD	6:33411228 C/T	40	NA	0	DeRubeis, 2014*
E970Q	Population	6:33411237 G/C	7.879	-5	0.00000826	ExAC
D987E	Population	6:33411290 C/A	6.318	-38	0.00000825	ExAC
V992I	Population	6:33411303 G/A	5.574	-84	0.00000825	ExAC
R1019C	ASD-associated missense	6:33411384 C/T	16.89	40	0.00003303	Kosmicki, 2017*
I1039T	Population	6:33411445 T/C	6.322	-17	0.00000834	ExAC
G1063S	Population	6:33411516 G/A	2.428	-20	0.00000915	ExAC
Q1093X	ASD-associated LGD	6:33411606 C/T	40	NA	0	Parker, 2015*

Variant	*Category*	*Genomic Change*	*CADD*	*Snap2*	*ExAC MAF*	*Source*
L1109V	Population	6:33411654 C/G	7.861	-55	0.00000977	ExAC
I1115T	Clinvar Benign	6:33411673 T/C	4.569	-14	0.0101	
I1115T	Population	6:33411673 T/C	4.569	-14	0.0101	ExAC
S1121G	Population	6:33411690 A/G	9.526	-3	0.0000138	ExAC
I1133V	Population	6:33411726 A/G	7.645	-81	0.0000610	ExAC
S1165L	MRAD5	6:33412306 C/T	21.6	3	0	Fieremans, 2016
N1213S	Population	6:33414407 A/G	7.788	-65	0.0000330	ExAC
E1228fs	ASD-associated LGD	6:33414451 GAACA/G	NA	NA	0	DeRubeis, 2014*
E1271G	ASD-associated missense	6:33415637 A/G	19.92	-28	0	MSSNG*
E1286D	Population	6:33415683 A/T	2.636	-76	0.0001	ExAC
P1301H	Clinvar Benign	6:33419553 C/A	12.91	-29	0	
P1301T	Population	6:33419552 C/A	9.154	-50	0.0000087	ExAC
T1305A	Population	6:33419564 A/G	0.371	-39	0.0000446	ExAC
P1321fs	ASD-associated LGD	6:33419612 CAG/C	NA	NA	0	MSSNG*
P1321S	Population	6:33419612 C/T	9.036	-54	0.0000271	ExAC
P1324fs	ASD-associated LGD	6:33419622 CA/CCC	NA	NA	0	MSSNG*