An investigation into the utility of guilt by association machine learning algorithms for the prioritization of autism spectrum disorder candidate risk genes

by

Margot Patricia Rainbow Gunning
B.Sc.H, Queen's University, 2017

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

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies
(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

December 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis/dissertation entitled:

An investigation into the utility of guilt by association machine learning algorithms for the prioritization of autism spectrum disorder risk genes

submitted by Margot Patricia Rainbow Gunning in partial fulfillment of the requirements for

the degree of Master of Science

in Bioinformatics

Examining Committee:

Paul Pavlidis, Michael Smith Laboratories/Psychiatry
Supervisor

Kurt Haas, Cellular and Physiological Sciences
Supervisory Committee Member

Elodie Portales-Casamar, Clinical Research Informatics Lead, BCCHR
Supervisory Committee Member

Joerg Gsponer, Biochemistry/Molecular Biology
Additional Examiner

Abstract

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impairments in social interaction and communication, and restrictive repetitive behaviours or interests, with extreme phenotypic and genetic heterogeneity. Currently, genetic association studies have identified 90 risk genes with high confidence out of an estimated 1000. Researchers have begun to use machine learning methods leveraging heterogeneous biological network data in attempts to aid in discovery of ASD risk genes. However, the real-world utility of these studies is questionable: network-based machine learners are often biased towards well studied genes because they operate on a principle called "guilty by association." In this thesis, I evaluate and compare genetic and computation approaches to ASD risk gene prioritization. I demonstrate that network-based computational approaches are adding little additional useful information compared to genetic approaches for prioritization. Furthermore, I demonstrate that gene expression profiles, and generic measures of disease gene likelihood may provide less biased contextual information that can be used to supplement genetic association data to prioritize ASD risk genes. Lastly, I discuss how data quality and data dependence impacts evaluation of machine learning algorithms and genetic association studies.

Lay Summary

Autism spectrum disorder (ASD) is an extremely heterogeneous neurodevelopmental disorder associated with social and communication deficits. Over the past decades, inroads have been made into delineating the complex genetic nature of ASD. However, we still have much to learn as only a fraction of the estimated genetic risk factors have been identified with high confidence. In this thesis, I evaluate different types of ASD gene discovery and prioritization methods, including genetic and computation approaches, to answer the question: can use of other non-genetic types of biological data aid in ASD gene discovery?

Preface

I was responsible for all word described, under the direction of my thesis supervisor, Dr. Paul Pavlidis, with the following exceptions: Manuel Belmadani was responsible for setting up the Ensembl Variant Effect Predictor used in Chapter 2. Shams Bhuiyan and Nathaniel Lim provided me with gene level annotation files for publication numbers and multifunctionality rankings used in Chapters 2 and 3.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ASD – autism spectrum disorder

AGRE – Autism Genetic Resource Exchange

ASC – Autism Sequencing Consortium

CADD – Combined Annotation Dependent Depletion

CNV – copy number variant

GBA – guilt by association

HGVS – Human Genome Structural Variation

iHart – Hartwell Autism Research and Technology Initiative

LGD – likely gene disrupting

LOF – loss of function

ML – machine learning

PCA –principal component analysis

pLI – probability loss-of-function intolerance

PTV – protein truncating variant

SFARI – Simons Foundation Autism Research Initiative

SPARK – Simons Powering Autism Research for Knowledge

SV – structural variant

SSC – Simons Simplex Collection

SNV – single nucleotide variant

TADA – Transmission and *de novo* association

WES – whole exome sequencing

WGS – whole genome sequencing

Acknowledgements

I thank Dr. Paul Pavlidis for his support and guidance throughout my Master's degree. I thank him for his continual lessons in all aspects of scientific research, and analysis, and for his patience, particularly during the writing process.

I thank my committee members: Dr. Kurt Hass, and Dr. Elodie Portales-Casamar for providing valuable feedback and support.

I thank Dr. Sanja Rogic, Shams Bhuiyan, Alex Morin, Eric Chu and the Pavlidis lab for providing me with daily counsel about the scientific process and advice about research questions. I thank Manuel Belmadani for his frequently sought help and technical expertise.

This work was supported by grants from the SFARI Foundation (Kurt Haas, PI), an NSERC Discovery Grant (Paul Pavlidis, PI), a scholarship from the UBC Bioinformatics Graduate Program via the NSERC CREATE program in High-Dimensional Biology and a CGS-M scholarship.

I thank the SFARI Foundation for providing access to additional variant data in SFARIGene.

To my friends, Bella, Caitlin, Emily, Gabby, Libbie, Lindsay, Matt, Sami: Thank you for constant support, and giving me opportunities to laugh and grow. And my friends from the Hawks organization: thank you providing me with an escape when needed.

To my family: Thank you for always picking up the phone, and being there for me. To my parents, and brother: There are not enough words to express how much I love you. Thank you for your unwavering support and enduring devotion.

Dedication

To my family.

Chapter 1: Introduction

1.1: Motivation

A central problem in autism spectrum disorder (ASD) research is identifying genes harbouring variants affecting risk. Advances in genetic and genomics technology have allowed for tremendous progress in identifying genetic variants associated with ASD. Over the past decade, a key finding from sequencing studies has been that rare variation, particularly rare *de novo* variation, plays a major role in contributing to ASD risk, especially in families with one child with an ASD diagnosis (simplex families) (De Rubeis et al., 2014; He et al., 2013; Iossifov et al., 2012; Levy et al., 2011; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2011, 2015, 2012). Based primarily on *de novo* variation identified in large-scale sequencing studies of simplex families, the number of ASD risk genes has been estimated to be approximately 1000 (De Rubeis et al., 2014; He et al., 2013; Sanders et al., 2015). To date, 90 ASD genes have been identified with high confidence based on replicated and statistically significant recurrence of damaging variation in ASD probands compared to controls. Modern genetics studies require large sample sizes, so progress in identifying more ASD risk genes can be slow and expensive. This challenge has led to increased effort in leveraging high-throughput heterogeneous biological data beyond genetics to identify risk genes not only for ASD, but also other complex genetic diseases.

In this thesis, I describe the work I have done to help address challenges in ASD gene discovery. Chapter 1 describes the key concepts guiding this thesis, including current knowledge about the phenotypic and genetic heterogenic nature of ASD, and methods used to identify and characterize ASD variants and genes. Chapter 2 describes my collection and annotation of

current ASD genetic risk factors. Chapter 3 delves into my work evaluating the performance and utility of genetics and non-genetics based ASD gene discovery and prioritization methods. Chapter 4 presents my conclusions and implications for future research.

1.2: What is ASD?

1.2.1: A brief history lesson and today's diagnosis

Autism spectrum disorder (ASD) was first formally described in the 1940s by Leo Kanner and Hans Asperger. Today, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) defines ASD as a neurodevelopmental disorder, characterized by deficits in social communication and social interaction, and restricted, repetitive patterns of behaviour, interests or activities which begin in childhood (Hyman, 2013). Sub-diagnoses, such as childhood disintegrative disorder and Asperger's disorder, have been removed. This removal reflects the theory of a broad spectrum of autistic behaviour first put forth by Wing and Gould (1979) based on their observations of extreme variability in symptom presentation and severity (Gillberg, 1992; Hyman, 2013; Lobar, 2016, p. 5; Volkmar & Reichow, 2013; Wing & Gould, 1979). Clinical diagnosis, treatment and management of ASD is further complicated by co-occurring medical conditions such as intellectual disability (approximately 30%), epilepsy and other seizure disorders (approximately 30%), and attention deficits and attention-deficit/hyperactivity disorder (approximately 30-40%) (Bauman, 2010; Croen et al., 2015; Devlin & Scherer, 2012; Jeste & Geschwind, 2014; Lyall et al., 2017; Matson & Cervantes, 2014). Currently, ASD is reported in approximately 1/59 children, with about a 4:1 ratio of boys: girls (Baio, 2018).

1.2.2: The etiology of ASD

The etiology of ASD has been much debated, and, at different times, suggested to be psychological, environmental and biological. Today, the only confirmed risk factor for ASD is genetics (S. E. Folstein & Piven, 1991; Geschwind, 2011; Trottier, Srivastava, & Walker, 1999). Multiple lines of evidence support a genetic etiology of ASD. Firstly, twin studies have shown high concordance rates in monozygotic twins (approximately 80-90%) compared to that of

dizygotic twins (approximately 10-30%) (Bailey et al., 1995; S. Folstein & Rutter, 1977; Rosenberg et al., 2009). Next, risk of ASD occurring in siblings and other first-degree relatives of probands is higher compared to the general population (approximately 3-20%), and first degree relatives often show subclinical symptomology at higher rates compared to the general population (Betancur, 2011; Caglayan, 2010; Geschwind, 2011; Losh et al., 2009; Ozonoff et al., 2011; Trottier et al., 1999). Furthermore, many genetic conditions, such as Fragile X Syndrome, and Tuberous Sclerosis, are often accompanied by ASD, or autistic features (Betancur, 2011; Caglayan, 2010; Geschwind, 2011; Lyall et al., 2017). Lastly, common, rare inherited and *de novo* structural variants (SVs) and single nucleotide variants (SNVs) have been associated with syndromic and non-syndromic (sporadic) ASD (Geschwind, 2011; Grove et al., 2019; Jeste & Geschwind, 2014; Neale et al., 2012; Sebat et al., 2007).

Currently, it is thought that common variants with low to modest effect contribute additively to the etiology of ASD, but, few common variants have been robustly associated from GWAS studies (Grove et al., 2019; Vorstman et al., 2017). In contrast, large-scale sequencing studies have demonstrated that rare *de novo* and inherited, highly penetrant variants across multiple genes confer high risk to ASD (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012; Vorstman et al., 2017). Because no single gene harbouring rare variants has been shown to account for even 1% of ASD cases, large-scale sequencing studies are needed to identify candidates. Based on computational estimates, and the observation that doubling of sample sizes is still approximately doubling the number of candidate ASD genes discovered, the number of additional ASD risk genes yet uncovered may by in the hundreds (He et al., 2013; Neale et al., 2012; Sanders et al., 2011, 2012; Satterstrom et al., 2019). The desire of the field of computational biology to help fill in the gap between gene discovery via

association and the expected number of genes contributing to ASD risk provides the main context of my thesis work.

1.3: Characterizing and predicting the impact of human genetic variation

Establishing genotype-phenotype relationships requires a deep understanding of the human genome and the variation therein. The diploid human genome is made up of some six billion base pairs, and the average human genome contains about four to five million differences from the reference human genome (1000 Genomes Project Consortium et al., 2015). Variation comes in various forms, from small changes, such as single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels), to larger structural variants (SVs), such as copy number variants (CNVs) and inversions. Single nucleotide polymorphisms are the most abundant form of human genetic variation, and are present at high frequency (more than 1-5%) in the human population, the majority of which have little to no impact on gene function (1000 Genomes Project Consortium et al., 2010). While most variation is neutral, variation with functional impacts can increase phenotypic diversity, including differences in disease risk (1000 Genomes Project Consortium et al., 2010).

1.3.1: Establishing genotype-phenotype associations is complex and multifactorial

There are many types of sequence variants that can be investigated with respect to effects on phenotypes. Within each class of variant, the potential for affecting gene function can vary widely. The first class of variants are structural variants (SVs), defined as those which affect chromosome structure at a relatively large scale (i.e. generally more than 1000 bases). Due to their larger size, SVs tend to have a high potential for affecting gene function. SVs can include copy number variants, duplications, deletions, insertions, inversions and translocations that involve at least 1000 bases of DNA. Large-scale SVs can be detected by karyotyping and/or chromosomal microarray technology. The other class of variants are "localized." Genomic variants involving fewer than roughly1000 base pairs can be classified in a variety of different

ways depending on specific changes in DNA and the location in the genome. These types of alterations can be detected using SNP array technologies, and next-generation sequencing technology. Variants in protein-coding regions can have a range of effects depending on how the function of the protein is impacted. Loss of function variants (also known as protein-truncating variants [PTVs] or likely gene-disrupting variants [LGDs]) include frameshift, stop gain and splice site variants. Smaller scale indels (i.e. less than roughly 1000 base pairs) can fall into this category of variation. These variants have a greater likelihood of affecting gene function because they change the sequence reading frame by shifting the codon grouping, introducing premature stop codons, or causing the loss of exons or inclusion of introns. Missense variants cause changes in one amino acid, and can have a range of effects on gene function, depending on how the new amino acid affects the overall functioning of the protein. Synonymous mutations, on the other hand, do not change the coded amino acid, and thus, are less likely to affect the function of the protein. Variants falling in non-coding genomic regions are less likely to cause functional changes compared to variants in protein-coding regions and are less often found to be associated with disease partly because they are harder to study, and their effects harder to predict due to lack of a clear-cut code. However, changes to regulatory regions, such as promoters and enhancers, can cause functional deficits. The typical genome has roughly 149-182 protein truncating variants, about 2100-2500 SVs, approximately 10-12000 peptide-sequence variants and around 459 000-565 000 variants in regulatory regions, all in comparison to the reference genome (1000 Genomes Project Consortium et al., 2015). When assessing variants found in patients diagnosed with complex genetic diseases, variant class can help discern the likelihood of functional impact and variant significance.

Inheritance patterns and allele frequency are important factors to consider when interrogating a variant's potential functional impact. It is estimated that there are roughly 100 *de novo* SNVs per genome per generation, and that, on average, only one is expected to be exonic (i.e. higher likelihood to impact gene function) (Besenbacher et al., 2015; Iossifov et al., 2012; Kong et al., 2012; Neale et al., 2012; O'Roak et al., 2012). Deleterious variants are expected to have lower allele frequencies than neutral, or weakly deleterious variants due to natural selection (1000 Genomes Project Consortium et al., 2010; Karczewski et al., 2019; Lek et al., 2016). Within a proband, rare *de novo* variants are more likely to be deleterious compared to rare inherited and common variation because of less stringent evolutionary selection (Veltman & Brunner, 2012). Effect size, and penetrance can also be used to assess variant impact. Effect size is the magnitude of the effect a genotype has on a phenotype; penetrance refers to the fraction of people with a given genotype who have the associated phenotype. High-impact variants often have high effect size and penetrance, meaning that the variant has a large impact on observed phenotype, and most people with the variant have the observed phenotype. By considering a variant's population allele frequency, and effect on observed phenotype, different categories of variants emerge, some of which are easier to study and detect than others (Figure 1.1).

Rare Mendelian or monogenic disorders are largely caused by deleterious genetic variants characterized by large effect size, high penetrance, and rare allele frequencies (Manolio et al., 2009) (Figure 1.1). Family-based linkage studies have historically been used to identify Mendelian disease genes. Linkage studies localize genomic/disease risk loci from co-inheritance of genetic markers, such as SNPs, and phenotypes in families across several generations (Bush & Moore, 2012; Laird & Lange, 2006) (Figure 1.1). Family-based linkage studies were used to

identify variants contributing to Cystic fibrosis, Huntington disease, and syndromic forms of ASD (discussed later in this chapter) (Bush & Moore, 2012).

The "common disease – common variant" theory postulates that a common variant in the population can contribute a small increase in disease risk, and explain a small proportion of heritability in common diseases (Bush & Moore, 2012; Manolio et al., 2009; Tam et al., 2019) (Figure 1.1). Genome-wide association studies (GWAS) assay millions of SNPs in large case/control populations and compare allele frequencies to assess how specific alleles at specific loci contribute to the phenotype of interest (Laird & Lange, 2006) (Figure 1.1). While these types of studies have identified many common variants associated with many complex diseases, it soon became clear that there was "missing heritability:" common variation identified by GWAS was not able to explain the estimated heritability of complex diseases (Manolio et al., 2009; McClellan & King, 2010).

Complex genetic diseases cannot be explained entirely by a small number of rare variants with large effect sizes, or by a limited number of common variants with moderate effect. Rare variants present in less than roughly 1% of the population with low to modest effect sizes cannot be detected by family-based linkage studies or GWAS because they do not have a large enough impact on phenotype nor do they occur at high enough frequencies to be tagged by genotyping arrays (Bush & Moore, 2012; Laird & Lange, 2006; Manolio et al., 2009) (Figure 1.1). This category of rare, moderate effect size variation is thought to play a large role in the genetic architecture of many complex diseases, including ASD (discussed later in this chapter).

Development of whole genome sequencing (WGS) and whole exome sequencing (WES) technology allowed for identification and greater characterization of common and rare variation across human populations as a whole, as well as of rare variation in human disease (1000

Genomes Project Consortium et al., 2015). While recurrence of rare variation within the same gene across different samples (i.e. independent events) provides increased support for association with a disease of interest, testing for association with disease is not as firmly established compared to family-based linkage analysis and GWAS. As previously mentioned, sequencing studies have shown that rare variation plays a large role in non-syndromic forms of ASD, so a lack of a "gold standard" genetic association test for rare variants with modest effect is problematic (De Rubeis et al., 2014; He et al., 2013; Iossifov et al., 2012; Levy et al., 2011; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2011, 2015, 2012).

An increasingly common test for association between rare genetic variation and disease, particularly in the field of autism genetics research, is the transmission and *de novo* association (TADA) test (De Rubeis et al., 2014; Feliciano et al., 2019; He et al., 2013; Ruzzo et al., 2019; Sanders et al., 2015; Satterstrom et al., 2019). The main advantage of this test is that it diverts from an allele-level approach, employed in family-based linkage analyses and GWAS, to a gene-level approach by allowing for recurrence of multiple types of variants to be collapsed, which maximizes power to find risk genes (He et al., 2013; Manolio et al., 2009). TADA analyses require data from *de novo* variants and/or inherited variants identified by large scale sequencing studies of simplex and multiplex families and case-control cohorts. Using this data, TADA builds a likelihood model based on allele frequency, relative risks of different classes of variation, and mutation rates to estimate a gene's likelihood of being involved in the phenotype.

TADA refers to a family of methods (fundamentally, statistical models), and differences in how it is used and parameterized are relevant to the comparisons across studies I present in this thesis. TADA has been applied using only *de novo* variants (TADA-Denovo), or the full model (TADA), which incorporates *de novo* variation, as well as inherited and case-control

variation (He et al., 2013). Additionally, TADA models require the user to provide parameter estimates for relative risk, fraction of risk genes, allele frequencies and mutation rates. Estimating these parameters can be problematic, yet important for determining the outcome of the analysis. The mutation rates are known for loss of function (LoF), missense and synonymous variants, and the fraction of ASD risk genes is usually estimated at 1000 based on simulations involving *de novo* LoF count and recurrence and accounting for allele frequency, mutation rate, relative risk and sample size. However, relative risk and allele frequency priors need to be estimated, and can vary between TADA analyses (De Rubeis et al., 2014; He et al., 2013; Ruzzo et al., 2019). Relative risk is often calculated based on burden of *de novo* and inherited mutations of each class observed within the data. However, there are notable exceptions to how relative risk is computed (as discussed in Chapter 3) (He et al., 2013; Satterstrom et al., 2019). The weighting of mutational classes generally follows the schema *de novo* LoF > *de novo* likely damaging missense > inherited LoF > inherited likely damaging missense (De Rubeis et al., 2014; He et al., 2013). One implication of the lack of well-established and validated methods for gene-level association studies of rare variants it that two studies of the same cohort can get different results, even if they both use a method labeled "TADA."

Most recent sequencing studies employing a TADA test for association with ASD provide an association score for each gene in the genome, and identify a subset a genes significantly associated with ASD under their model, at some expected false discovery rate (De Rubeis et al., 2014; Ruzzo et al., 2019; Sanders et al., 2015; Satterstrom et al., 2019). The genome-wide association scores allow us to compare prioritization of ASD risk gene candidates based on genetic association to other genome-wide prioritization scores based on other types of non-genetics data.

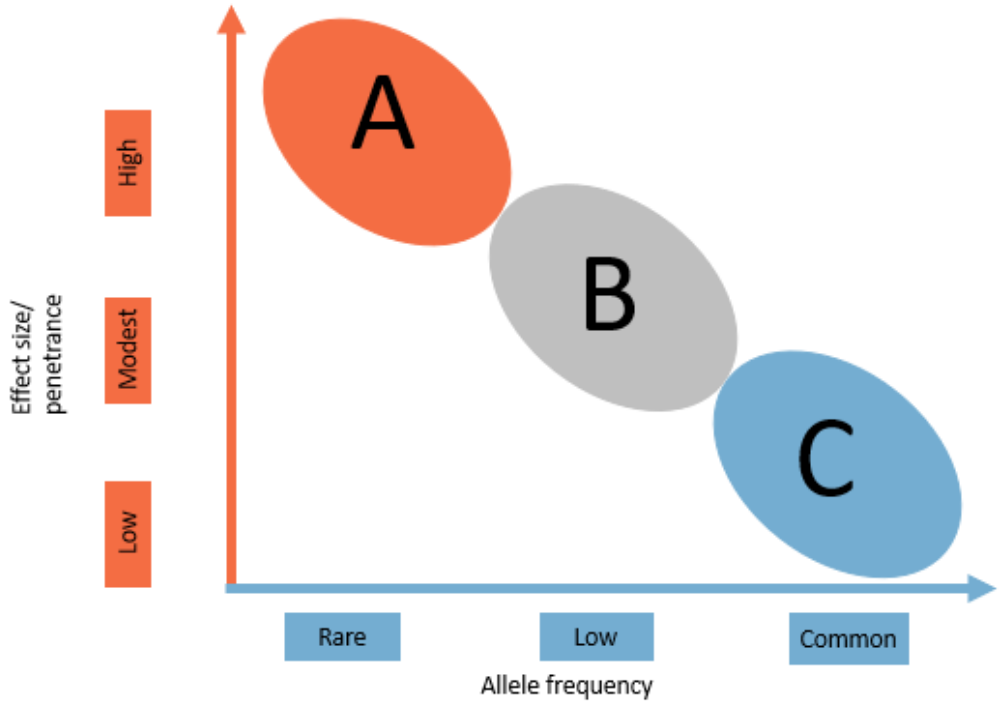

Figure 1.1: Detectability of different variant classes involved in human disease(McCarthy et al., 2008). A) Rare variants with high penetrance/effect size causing monogenic/Mendelian disease; usually found with family-based linkage and association studies. B) Low-frequency variants with moderate effects on disease, often found by whole genome/exome sequencing studies and associated with disease via association models such as TADA. C) Common variant with low impact implicated in common disease are often found by genome-wide association studies. Rare variants with small effect size (below the diagonal line) are hard to detect. High impact, common variation (above the diagonal line) are usually not found to impact common disease, but do influence non-disease phenotypes, such as eye colour and height.

1.3.2: Assessing functional impact of variation through use of computational tools

Within some classes of variation, there can be a range of effects depending on how and where the DNA sequence was altered. Many computational tools have been developed to predict variant deleteriousness. One popular approach is Combined Annotation Dependent Depletion (CADD), which is used for predicting the effects of SNVs and indels. CADD uses local sequence and evolutionary information to predict variant pathogenicity and incorporates the function prediction tools Sorting Intolerant from Tolerant (SIFT) and PolyPhen-2 (Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019). SIFT and PolyPhen-2 use sequence information and physical amino acid properties to predict the effects of amino acid substitutions in proteins as deleterious or tolerated, and probably damaging or benign, respectively (Adzhubei et al., 2010; Ng & Henikoff, 2003). CADD outputs a raw score, and a normalized score. Generally, the normalized CADD score is used for variant annotation because it takes into consideration all potential SNVs in the human genome. Higher normalized CADD scores correspond to a higher likelihood that the variant is deleterious. I utilized these scores when annotating variants reported in ASD probands for use in gene and variant prioritization.

As discussed, the impact of a variant is related to its allele frequency within the human population. Different catalogs of variant allele frequencies across populations have been created to aid in investigating how deleterious genetic variation relates to human disease. The 1000 Genomes Project was set up to investigate a broad range of human genetic variation using a combination of WGS, WES and microarray genotyping, including 99% of SNP variants with frequency above 1% in different populations (1000 Genomes Project Consortium et al., 2015). More recently, the Genome Aggregation Database (gnomAD; previously known as ExAC) used

roughly 15 000 whole genomes and approximately 120 000 exome sequences to report allele frequencies from unrelated individuals. I utilized these databases for allele frequency annotation.

I also used four gene-wise measures of constraint against missense and LoF variation calculated by ExAC and gnomAD in my analysis (Karczewski et al., 2019; Lek et al., 2016). The four gene-wise scores were created based on comparing the observed number of rare variants (frequency less than < 0.1%) per gene against the expected number of variants. Missense z-scores greater than three indicate that a gene has fewer than expected missense variants and, therefore, is considered to be under higher constraint. The observed/expected LoF (o/e LoF) measures depletion of LoF variation within a gene and scores less than 0.35 indicate a stronger intolerance to LoF variation. An extension of o/e LoF score is the probability loss of function intolerance score (pLI) which is corrected for gene length, and ranges from 0-1. Higher scores (pLI>0.9) reflect a greater likelihood that a gene is intolerant to LoF variation (Karczewski et al., 2019; Lek et al., 2016). The o/e LoF score is a more continuous measure of LoF constraint because it measures depletion of LoF variation across a spectrum of selection compared to the more bimodal pLI (Karczewski et al., 2019; Lek et al., 2016). The pLI score is more useful for classifying genes as likely intolerant to, or likely tolerant of LoF variation whereas the o/e LoF score is more interpretable across the spectrum i.e. a score of 0.35 means that 35% of expected LoF variants have been observed within a gene (Karczewski et al., 2019; Lek et al., 2016). I used these metrics in my project because genes which meet the given thresholds are more likely to have detrimental effects when mutated. Therefore, the constraint metrics can be thought of as generic proxies for how likely it is for a gene to be involved in *any* disease.

Gene expression data may also be used to interpret human genetic variation in disease. It is well known that genes vary in their expression across tissues and time, and that gene expression is under genetic control and thus influenced by genetic variation (Cummings et al., 2019; GTEx Consortium, 2017; Karczewski et al., 2019; X. Li et al., 2017; Stranger et al., 2017; Yang et al., 2018). Therefore, for diseases known to affect specific tissues or cells during specific developmental time periods, the use of spatiotemporal and/or tissue-specific gene expression may provide a useful biological context for interpretation of candidate variants and genes (Barrett et al., 2013; Cummings et al., 2019; GTEx Consortium, 2017; Karczewski et al., 2019; X. Li et al., 2017; L. Liu et al., 2014; Miller et al., 2014; Stranger et al., 2017; Yang et al., 2018). However, in theory, disease genes could be acting in any tissue, at any time, therefore, it is important to consider how narrowing the scope of gene expression analysis may impact disease gene and variant interpretation.

In summary, establishing genotype-phenotype relationships is a complex problem. There are specific methods best suited for identifying and associating specific variant classes with disease. To help in variant interpretation and prioritization, we can annotate variants identified in ASD probands with population allele frequencies, predictions of damaging consequences and other disease-relevant information.

1.4: The heterogeneous genetic etiology of ASD

1.4.1: Increasing insight into genetic etiology of ASD through technology and analysis advancements

As more sophisticated genomic technologies and analysis methods were developed, the field gained greater insight into the complex, heterogenous genetic architecture of ASD. Early stages of ASD genetics research were dominated by family-based linkage studies of rare syndromic forms of ASD, that is ASD in the presence of a known genetic syndrome with defined somatic and neurobehavioral abnormalities, such as Fragile X Syndrome (Fernandez & Scherer, 2017; Vorstman et al., 2017). However, syndromes account for a small number of ASD cases, and identifying other variant classes associated with ASD did not occur until large-scale GWAS and sequencing studies were conducted. As previously discussed, it is thought that common variants with low to modest effect contribute additively to the etiology of ASD, possibly accounting for a large proportion of ASD risk, but, few common loci have been robustly and significantly associated with ASD (Vorstman et al., 2017). Large-scale sequencing studies have demonstrated that rare *de novo* and inherited, highly penetrant variants across multiple different genes and samples confer high risk to ASD (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012; Vorstman et al., 2017). While there has been substantial progress in delineating the genetic etiologies of ASD, it is estimated that only roughly 20% of cases have a genetic diagnosis (Jeste & Geschwind, 2014; Neale et al., 2012). I will now discuss in greater detail the contribution of common, rare inherited and *de novo* variation to ASD etiology.

It has been estimated that 10-20% of individuals with ASD have some identifiable genetic syndrome or disorder or cytogenetically visible chromosomal abnormalities, such as 15q11-13 duplications (Betancur, 2011; Devlin & Scherer, 2012; Pinto et al., 2010). Fragile X Syndrome and Tuberous Sclerosis are among the most common ASD associated syndromes.

Family-based and case-control karyotyping and chromosomal microarray (CMA) studies were used early on to investigate the role of CNVs in ASD. Early studies established that CNVs play a role in ASD etiology through multiple lines of evidence: 1) A higher rate of *de novo* CNVs was found in ASD simplex vs. multiplex families, and ASD families and/or probands vs. controls; 2) *De novo* and inherited CNVs were found in sporadic and syndromic forms of ASD; 3) Many CNVs associated were found to have variable penetrance depending on proband sex and parental transmission of variation; and 4) Recurrence of rare *de novo* multigenic CNVs were found in unrelated ASD probands (Betancur, 2011; Cook Jr & Scherer, 2008; Devlin & Scherer, 2012; Pinto et al., 2010; Sanders et al., 2011; Sebat et al., 2007; Shen et al., 2010).

The low-cost of sequencing has shifted the focus of ASD genetics research almost entirely towards identifying low/rare frequency inherited and *de novo* variants with moderate to high effect size. Initial WES studies utilized small ASD cohorts (approximately 200-300 families) made up of mostly trios (one affected child, two parents) and reported similar findings to the CNV analyses above: 1) Recurrence of rare *de novo* variants were found in the same gene across different samples (i.e. independent events); 2) *De novo* variants, particularly LoF variants, were found at a significantly higher rate in cases vs. controls (approximately 2: 1), and sometimes at a trending higher rate in female vs. male probands; and 3) Rare *de novo* variants were found in both syndromic and sporadic ASD with a range of effect size and penetrance (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). These studies further established that ASD is a genetically heterogeneous disorder.

Across the early WES ASD studies, CHD8, KATNAL2, SCN2A, NTNG1, NRXN1 were implicated as ASD risk genes based on recurrence of independent variant events (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Ronemus, Iossifov, Levy, & Wigler, 2014;

Sanders et al., 2012). At the time of these findings, multiple classes of variants in SCN2A had already been associated with ASD in the presence and absence of seizures, and multiple epilepsy phenotypes; since then, SCN2A has also been associated with intellectual disability, schizophrenia and other disorders (Sanders et al., 2012; Vorstman et al., 2017). It is important to note that this is not an isolated occurrence: many other ASD risk genes with rare *de novo* CNVs and SNVs have been implicated in other neuropsychiatric and neurodevelopmental conditions, such as schizophrenia and intellectual disability (Sanders et al., 2012; Vorstman et al., 2017).

Early ASD WES studies were limited in scope because they focused almost exclusively on *de novo* variants and ignored other classes of rare variation. As previously mentioned, the TADA method was developed in order to incorporate *de novo* and rare variation information from family and case-control studies to increase power to identify disease risk genes. Similar to other association studies, those which employ TADA models also need large sample sizes in order to achieve good power for discovery of risk genes (He et al., 2013). Currently, there are five published genetic association studies using TADA models and WES and WGS sequencing data from the Simons Simplex Collection (SSC), the Autism Sequencing Consortium (ASC), the Autism Genetic Resource Exchange (AGRE), the Hartwell Autism Research and Technology Initiative (iHart) cohort, the Simons Powering Autism Research for Knowledge (SPARK) cohort, and multiple other ASD cohorts from across the globe (De Rubeis et al., 2014; Feliciano et al., 2019; Ruzzo et al., 2019; Sanders et al., 2015; Satterstrom et al., 2019). In aggregate, these studies have identified upwards of 100 ASD candidate risk genes.

The five published TADA studies have complex relationships with one another. The power to detect ASD risk gene candidates has increased in the more recent TADA studies because the sample size has greatly increased. However, the studies cannot be considered as

independent association studies because there is substantial overlap in their cohorts. Furthermore, the studies differ slightly in their implementation of the TADA model. For example, Sanders et al., 2015 incorporates small *de novo* deletions in addition to SNVs, and Satterstrom et al., 2019 modified the relative risk estimates of their TADA model to incorporate measures of constraint against different classes of variation. In addition to confirming previously known ASD risk genes, such as CHD8 and SCN2A, TADA studies have identified a plethora of risk genes based on recurrence of rare *de novo* and inherited LoF and probably damaging missense variants. For example, the analysis from the iHart consortium identified 14 novel ASD risk genes, including PCM1 due to four inherited LoF mutations in different samples (Ruzzo et al., 2019); Satterstrom et al. (2019) identified 31 novel ASD risk genes, of which, DPYSL2 had three likely damaging missense mutations in ASD probands; and the SPARK pilot analysis identified 13 novel ASD risk genes, with the inclusion of BRSK2 due to two additional LoF mutations being identified (Feliciano et al., 2019). These studies, among many others, support a more substantive role of rare *de novo* and inherited SNVs in the etiology of ASD (C Yuen et al., 2017; Iossifov et al., 2014).

Initially, as with other common diseases, GWAS provided support for the theory that common variation contributed substantially to the heritability of ASD, but few statistically significant, reproducible common variants have been found to be associated with ASD (Devlin & Scherer, 2012; Geschwind, 2011; Jeste & Geschwind, 2014). Common issues to all first-generation GWA studies were low statistical power due to low sample size, and inadequate control samples. However, this past year, a large ASD genome-wide association meta-analysis of roughly 18 000 ASD cases and approximately 28 000 controls robustly identified five genome-wide significant loci (Grove et al., 2019).

A wide spectrum of genetic risk factors have been associated with ASD, including rare *de novo* and inherited single nucleotide variants and copy number variants, common variants and known genetic syndromes. Substantial progress has been made in understanding the genetic etiology of ASD. But, more research needs to be done to further define the genetic causes of ASD on a population and individual level, and to identify impacted biological systems and cell types.

1.4.2: Databases collecting information on ASD genetic risk factors

In 2003, the Simon's Foundation launched the Simons Foundation Autism Research Initiative (SFARI) with the aim of increasing our biological understanding of ASD to improve diagnosis and treatment. A SFARI resource I used for my project was a database manually curated by MindSpec, Incorporated. called SFARIGene, which collects information on ASD genetic risk factors and genes (Abrahams et al., 2013; Banerjee-Basu & Packer, 2010). In SFARIGene, the Human Gene Module contains upwards of 1000 genes hypothesized to play a role in ASD, and contains information about reports, syndromes and variants associated with each gene. Each gene in the Human Gene Module is scored on a scale from 1-6 reflecting the amount of evidentiary support for a role in sporadic ASD based on human genetics studies and other functional studies, with 1 denoting genes with the highest confidence of association; additional scores are Syndromic (S) or No Score (not yet curated). Currently (August 2019), of 1089 genes listed, 90 are considered to be high confidence (category 1 and 2). In order to be considered a SFARI high confidence gene, there must be evidence of recurrent, "likely to be functional" variants in cases compared to controls with independent replication and/or be uniquely implicated with genome-wide significance in association studies. Many of the SFARI category 1 and 2 genes were identified by two of the first large-scale exome sequencing studies

using TADA from De Rubeis et al. (2014) and Sanders et al. (2015). It is probable that many genes identified in the most recent large-scale sequencing studies will be added to the SFARI database in the near future.

Other databases providing ASD genetic variation information I used in this project include VariCarta, MSSNG and ClinVar (Belmadani et al., 2019; C Yuen et al., 2017; Landrum et al., 2014). VariCarta is our in-house web-based database for the collection and cataloguing of genomic variants found in ASD probands. Currently (Fall 2019), 50 publications have been curated, and approximately 35 000 variant events have been identified from approximately 8 000 ASD probands (Belmadani et al., 2019). VariCarta harmonizes and standardizes variant reporting formats across the 50 publications, accounting for many variant reporting differences, but most importantly, for cohort overlap to avoid double counting the same variant event in the same subject (Belmadani et al., 2019). MSSNG is another ASD-specific resource with a goal of collecting genome sequences and extensive phenotype data of 10 000 ASD families to identify disease genes and pathways, and endophenotypes of ASD (C Yuen et al., 2017). ClinVar is a publicly accessible database collecting information about human genetic variation and health and disease (Landrum et al., 2014). While ClinVar is not specific to ASD, it does provide phenotypes and evidence associated with reported variation, as well as a clinical interpretation of variant pathogenicity (Landrum et al., 2014).

1.4.3: Can we go beyond genetics to find ASD genes

While genetic approaches have been successful in identifying ASD risk genes, many groups have proposed a different and potentially complementary tack, which is to use heterogeneous biological data in a machine learning framework. This class of approaches, generally termed guilt by association (GBA), has long been popular among computational

biologists for attempting to automatically annotate genes with functions or disease relevance (Pavlidis & Gillis, 2012, 2013). GBA has been a topic of close study in the Pavlidis lab for many years, culminating in a series of papers pointing out serious problems with the GBA paradigm, and highlighting its failure to actually influence gene function discovery. These doubts and the intersection of our research interests in ASD genetics and function prediction led me to use ASD gene prediction as a case study of the utility of GBA methods. In the next section, I describe the GBA paradigm and its application to ASD gene discovery.

1.5: Machine learning for prediction of autism spectrum disorder genes

1.5.1: Machine learning methods use guilt by association for gene function prediction

Predicting gene function generally operates on a principle called guilt by association (GBA). The GBA principle states genes with "associations" in the data are more likely to be "guilty" of sharing the same or similar functions. Therefore, using biological data, such as data from physical, genetic, biochemical, and/or evolutionary sources, and prior information about the genes involved in a function of interest, new genes can be ascribed a previously unannotated function. The GBA principle is not unique to gene function prediction and forms the basis of other computational analyses, such as testing for function enrichment in a gene set (Pavlidis & Gillis, 2012).

In this section, I first introduce a generalized machine learning setup before discussing how guilt by association is used to predict gene function. Supervised machine learning methods are most often used for gene function prediction. Supervised machine learning uses labelled data to fit a model relating the input (predictor or feature data) to the output (response variable), and then to predict responses for data not used in the fitting process. The requirements for supervised machine learning include: 1) Labelled training data and unlabelled data for prediction; 2) Feature data; and 3) An algorithm to fit a model on the labelled training data and the feature data in order to make predictions about the unlabelled data (Lee, Ambaru, Thakkar, Marcotte, & Rhee, 2010; Pavlidis & Gillis, 2012). Cross validation is used to validate training performance before predictions about candidate genes are made. Now, I will describe how supervised machine learning is used for gene function prediction in more detail.

Labelled and unlabelled data. Three different gene sets are needed. The labelled training gene set is made up of a positive and negative gene set. The positive gene set consists of genes

known to be associated with the function of interest, for example, “chromatin remodelling” or “ASD.” The negative gene set is made up genes not associated with the function of interest. Lastly, the unlabelled candidate genes are the genes we want to make predictions about, and typically consist of the rest of the genes in the genome.

Feature data. The feature data is used as the basis of pattern recognition performed by the algorithm. In this context, an ideal hypothetical feature is one that genes associated with ASD have, and genes not associated with ASD lack. Often, and most relevant to my thesis, the associations among the genes are represented as a network. These networks can be built from gene expression, protein-protein interaction, genetic or physical perturbation data, and/or other disease and phenotype data. Networks are often represented as an adjacency matrix: rows and columns are nodes, and the entry in the matrix is the association between the two; the entry can be binary (0=no association, 1=association), commonly employed in protein-protein interaction networks, or a range between 0-1, commonly used for co-expression networks (Ballouz, S., Weber, M., Pavlidis, P., & Gillis, J., 2017). In the hypothetical ideal case, in such a network, ASD-associated genes would be associated with each other, and lack associations with non-ASD genes. Obviously in that case, prediction would be trivial, and the reality is very far from this ideal. This is why more complex algorithms are needed.

Algorithm. The algorithm uses the feature data to differentiate between the positive and negative gene sets so it can make predictions about the unlabelled candidate gene set. The relationship between the feature data and the response variable is unknown. The algorithm uses the training data to recognize patterns within the network to estimate this unknown relationship so it can make predictions about which of the candidate genes are the most similar to the positive training set. Algorithms utilized by studies I investigated in my thesis range from more simple

logistic regression classifiers, to more complex models such as support vector machines, and random forests.

As previously discussed, the GBA principle postulates that genes with similar functions are more likely to be associated within the data which allows for new or previously unknown gene functional relationships to be inferred. In gene function prediction, the chosen algorithm can employ GBA on the feature data in a direct, or indirect manner. For example, direct GBA on a network can take the form of neighbour voting whereby the likelihood of a candidate gene having the function of interest is the fraction of its neighbours which have the function of interest (Gillis & Pavlidis, 2011b). In a network, this fraction can be calculated as the number of positive training genes a candidate gene is connected to out of its total number of connections (node degree). In other words, candidate genes with a high number of ASD positive neighbours are more likely to be "guilty" of being ASD genes themselves because they have "associations" with the ASD positive genes.

Many different techniques can be used to extend GBA to incorporate indirect or broad patterns in the feature data for function prediction (Gillis & Pavlidis, 2011b). Generally speaking, many of these techniques search the feature data in multidimensional space to find relations or discriminant boundaries within the data that can best separate the positive and negative training sets (Lanckriet, Deng, Cristianini, Jordan, & Noble, 2004). For example, support vector machines (SVMs) view the feature data as a multidimensional vector. If the feature data is a network, the network adjacency matrix is filtered so that the rows are the training genes, and the columns are the rest of the genes in the network, with the cells of the matrix representing the gene-pair association or interaction scores. The resulting matrix is viewed in the same number of dimensions as columns in the filtered network adjacency matrix

(i.e. features). In the multidimensional feature space, the SVM attempts to fit a hyperplane(s) which provides the optimal separation of positive and negative training examples. Different transformations called kernels i.e. linear or polynomial kernels, and/or tuning parameters are applied to the feature space to find the hyperplane which gives the maximum distance between the positive and negative training sets. Given an unlabelled candidate data point, the SVM can classify the item into a binary (or multi-class) output category by determining on which side of the hyperplane they are on. Roughly speaking, candidate genes on the side of the discriminant boundary with the majority of the ASD positive training genes are more likely to be positive ASD genes themselves essentially because they are "nearer" (associated with) ASD genes compared to non-ASD genes in the feature space.

Training evaluation. When evaluating the performance of a supervised machine learning algorithm, it is of primary interest how well the algorithm is able to generalize to new data. In other words, we want to know how accurate the predictions are when applied to previously unseen data not used to train the algorithm. The algorithm can be evaluated on training data as one set, however, good performance on training data does not ensure good performance on testing data. Therefore, cross validation is often used to assess performance of the algorithm by iteratively partitioning the training datasets into a training subset and a testing subset, also known as a hold-out or validation subset. The algorithm is trained on the training subset, following which, it can be applied to the testing subset and prediction performance can be assessed. There are different partitioning processes including iterative random partitioning of the data into equal groups, having only a single gene in the validation set, and/or splitting the data into a set number of folds where a subset of each fold is used as the validation set.

Ideally, the chosen algorithm will balance model bias with variance (Bias-Variance Trade-off). Model bias originates from trying to estimate complex real-life problems from simple models. Models with high bias can miss patterns due to a failure to accurately recapitulate the real-life data because of erroneous assumptions made by the model whereas models with low bias fit real-life data more closely. Variance refers to how much the estimate of the relationship between the input feature data and the response variable would change if a different training data set was used. Generally speaking, as algorithms become more flexible, the bias decreases because the model is following the data points more closely and there is an increase in performance on the training subset. However, the variance will also start to increase, meaning that changing the training data would have a larger impact, and that the performance on the testing subset will suffer. This is described as overfitting because the algorithm is trying too hard to find patterns in the training subset which do not generalize to testing subset, resulting in good training performance, but poor testing performance.

Receiver operating characteristic (ROC) curves and precision-recall (PR) curves are two very common metrics for assessing performance of the algorithm on training and testing subsets. In classification problems, there are four possible outcomes for each instance (gene). If the gene belongs to the class of interest (ASD), and it is classified as positive, it is a true positive; if it is classified as negative, it is counted as a false negative. If the gene is negative (not an ASD gene), and is classified as negative, it is a true negative; if it is classified as a positive, it is counted as a false positive.

Receiver operator curves plot the true positive rate vs. the false positive rate at various score thresholds. The true positive rate (recall, sensitivity, power of detection) is the proportion of true positives that are correctly classified. The false positive rate (fall-out, probability of false

alarm, 1-specificity) is the proportion of true negatives that are incorrectly classified. The overall performance of a classifier can be measured as the area under the receiver operator curve (AUROC). Ideally, the ROC curve will be in the top left corner of the plot, meaning the classifier is sensitive and specific, with a larger area under the curve indicating a better classifier. Typically, "good classifiers" will have an AUROC of over 0.8, whereas classifiers which perform no better than random chance have an AUROC of 0.5. During training, most supervised machine learning studies will employ cross-validation, reporting the average AUROC performance across all folds as a measurement of performance.

AUROC give a measure of performance that can be misleading for real-life applications where only the top predictions are likely to be inspected. For this reason, precision recall curves are often preferred as they are more sensitive to classification errors in the top ranks. Precision (positive predictive value) measures the proportion of classified positives which are true positives. As previously stated, recall measures the proportion of true positives that are correctly classified. A "highly relevant classifier" will make no or few false positive classifications when recovering all of the true positives.

Output and evaluation. The algorithm yields a ranking of the candidate genes, with the genes at the top of the ranking being those predicted to be most likely involved in the function of interest. Often the output rankings are subject to independent validation tests, in addition to cross validation during training. Ideally, one has a "gold standard" gene set independent of the data used during algorithm training which can be used for evaluation with ROC and PR curves. However, as most conditions have a small number of known positives (associated genes), many published evaluations look at how well their predictions are able to recover either genes with a

lower level of association with the function of interest, or genes found to be associated with the function of interest using previously unseen data.

1.5.2: Problems with guilt by association and gene networks for gene function prediction

As previously mentioned, prior work done in the Pavlidis lab has questioned the utility of GBA-based machine learning for gene function prediction (Gillis & Pavlidis, 2011a, 2012; Pavlidis & Gillis, 2012, 2013). A major problem with this type of GBA is the underlying quality of the heterogenous biological data used as input features. Biological networks themselves are often biased toward well studied genes. Well studied genes often have high numbers of associated functional annotations (multifunctional), and these genes are often highly connected (hubs) within the network (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). The issue with "multifunctionality bias" is that it drives GBA performance: GBA tends to ascribe new functions to genes which are highly connected within the network rather than learning additional, novel information from the connection patterns (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). Given that the same biological network can be used in many prediction situations with different training gene sets, one would expect some level of specificity for the training task. However, due to the aforementioned biases, GBA tends to prioritize highly connected genes within the network for any/all tasks for which the algorithm is trained on (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012; Qiao et al., 2010). Furthermore, genes which are not multifunctional or hubs within the network may only be so because they are understudied and lack any documented true associations with other genes (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). In summation, GBA ML studies tend to yield gene rankings correlated with multifunctionality and hubness, resulting in high prioritization of generically "disease-related" genes because they are well

studied. Due to a lack of disease-specificity, and bias towards well studied genes, GBA has difficulty identifying novel and disease-specific relationships between genes.

Protein-protein interaction and/or co-expression networks alone, or in combination with each other or with other types of interaction or functional data, are often used to create networks. Protein-protein interaction networks can be constructed from multiple different databases, such as STRING or BioGrid, which collect information about known and predicted protein-protein interactions by aggregating across many different types of experiments and experimental conditions (Oughtred et al., 2019; Szklarczyk et al., 2019). PPI networks are often strongly biased towards well studied genes, and the effects of multifunctionality and hubness can explain a large amount of the predictive performance of PPI networks (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). In comparison, co-expression networks can offer a more context-specific datatype which demonstrate less literature bias (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). In theory, gene expression networks could be better for GBA because there would be less literature bias to drive performance (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). However, without real-world evaluation, the utility of gene expression networks for gene function prediction is unknown. Correcting for multifunctionality and hubness within PPI and gene expression networks is difficult: filtering to alter node degree can result in loss of information, and using a designated threshold based on strength or reliability of association may also be misleading as the underlying data is biased (Gillis & Pavlidis, 2011a).

The implication of this work is that GBA can seem to work in cross validation while providing predictions of little actual value. To show true value, GBA methods must be validated in real-life situations. Overall, we are aware of few instances whereby GBA prediction has produced a *bona fide* new disease association for a gene. Multiple ASD gene prediction methods

have been developed, including genetic association tests, GBA machine learning methods using heterogeneous biological data, and generic measures of disease gene likelihood. By comparing GBA ML methods to accepted approaches for genetic association and other disease gene prediction methods, we were able to investigate the validity of GBA ML methods for ASD risk gene prediction.

1.6: Thesis outline

In this thesis, I will explore two main questions prompted by the two main research areas highlighted above. Mainly, what are the characteristics of current ASD risk genes, and how useful are guilt by association methods for prioritization of ASD risk genes?

Firstly, I describe the collection and curation of information about current ASD risk genes and their variants. As delineated above, the genetic architecture of ASD is complex and heterogeneous. Many factors have to be assessed in order to discover ASD risk genes and to implicate the causal variants within a gene. In Chapter 2, I present my efforts to collapse ASD risk gene information from multiple sources, including NDD-specific resources such as SFARI, and our in-house ASD-specific database VariCarta.

In the Chapter 3, I report my investigations into how GBA-based machine learning studies use heterogenous biological network data in attempts to prioritize ASD risk genes. I aimed to answer the following questions: What are the differences among the published and currently in-production learners? Can these learners accurately predict known ASD risk genes, and candidate ASD risk genes? Are these learners biased towards generic measure of gene function, for example multifunctionality? Are these learners adding any additional information we cannot get from genetic association studies, or generic measures of a gene's likelihood of being involved in any disease, such as probability loss-of-function intolerance score (pLI)?

In the conclusion (Chapter 4), I summarize insights gained, and propose future avenues for exploration.

Chapter 2: Collecting information on current ASD candidate genes

2.1: Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder with a genetically heterogenous etiology. Many different classes of variation have been implicated in ASD. Currently, research is aimed at identifying rare, highly penetrant *de novo* variants in ASD probands. Recently, many WGS and WES studies have identified many ASD risk genes based on rare *de novo* and inherited LoF (frameshift, nonsense, splice site alterations) or likely damaging missense variants. However, relatively few ASD probands have a genetic diagnosis, and much work needs to be done to identify more genetic causes of ASD.

We are involved in a collaboration project focused on studying *de novo* missense variants in ASD risk gene candidates from SFARI categories 3 and 4. The aim of this collaboration is to identify more high-confidence ASD risk genes. In order to help, we collected current information on ASD risk genes across multiple data sources to identify characteristics that could be useful for prioritizing risk genes candidates for further experimental study.

2.2: Materials and Methods

2.2.1:ASD gene set

ASD genes were collected from two sources: SFARI and recent genetic association studies. SFARIGene collects information curated by MindSpec, Incorporated about genes and genetic risk factors of ASD. Genes are categorized by amount and quality of evidence for association with ASD from1-6, with 1 being highest confidence; included also is a syndromic category. Currently, there over 1000 genes in SFARI, and 90 are considered high confidence with gene scores of 1 and 2. Three recent genetic association studies employing TADA analyses have identified 58 novel genes with statistically significant associations with ASD (Feliciano et al., 2019; Ruzzo et al., 2019; Satterstrom et al., 2019). As many genes in SFARI categories 1 and 2 were discovered using TADA analyses, it is likely that most of these genes will be categorized there in the near future (De Rubeis et al., 2014; Sanders et al., 2015) (Table 2.1).

Table 2.1: ASD risk gene sets

ASD risk gene set	Description
SFARI high confidence (SFARI-HC)	**90** genes in SFARI categories 1, 1S, 2 and 2S are considered to be high confidence ASD risk genes.
Novel high confidence (Novel-HC)	**58** novel ASD risk genes identified from three recent TADA genetic association studies. We consider these novel findings to be high confidence ASD risk gene candidates.
Rest of SFARI database (restSFARI)	**938** genes in SFARI with score 3, 3S, 4, 4S, 5, 6, or S (syndromic). These genes have substantially less evidence of a role in ASD. We consider these to be ASD risk gene candidates for prioritization for further study

2.2.2: Gene mapping

Gene symbols and Entrez gene identification numbers from the ASD risk gene sets were mapped to official gene symbols and Entrez gene identification numbers provided by the NCBI (Sayers et al., 2019). All protein-coding genes from NCBI RefSeq were used as a background comparison (Sayers et al., 2019).

2.2.3: ASD variant sources

VariCarta(Belmadani et al., 2019)*:* VariCarta is a publicly accessible database maintained by the Pavlidis lab which collects information of human genetic variants found in ASD probands and reported in peer-reviews publications. I downloaded the most up-to-date variant file staged for public release on 2019-08-29. Coordinates are reported in GRCh37. It is unknown if all sources in VariCarta utilized the same diagnostic criteria for ASD, or the same procedure for variant calling. However, VariCarta attempted to filter for subjects with a clear diagnosis of ASD where available, and harmonized variant reporting into a standardized format.

MSSNG (C Yuen et al., 2017): The Pavlidis Lab has access to the MSSNG database, and I accessed the 2017-02 variant calls from our servers on 2019-08-29. Coordinates are in GRCh37. MSSNG called variants using GATK (McKenna et al., 2010), and annotated with Annovar (K. Wang, Li, & Hakonarson, 2010).

Satterstrom et al. (2019)*:* A recent TADA analysis of upwards of 20 000 WES samples currently published on BioRxiv. Clinical evaluation and widely accepted ASD interviews and diagnostic tools were used to diagnose patients. I downloaded supplemental tables pertinent to this project, including sample information (Table S1), variant calls (Table S2) and TADA results (Table S4) on 2019-09-03. Coordinates are in GRCh37. They called variants using GATK (McKenna et al., 2010), and annotated with VEP (McLaren et al., 2016).

Ruzzo et al. (2019)*:* A recent TADA analysis incorporating new samples from roughly 2000 multiplex families from the iHart consortium. Clinical evaluation and widely accepted ASD interviews and diagnostic tools were used to diagnose patients. I downloaded supplemental tables for sample information (Table S1) and TADA results (Table S3) on 2019-09-03. Coordinates are in GRCh37. They called variants using GATK (McKenna et al., 2010), and annotated with Annovar (K. Wang et al., 2010).

Feliciano et al. (2019)*:* A recent TADA analysis incorporating new samples from roughly 400 trios from the SPARK consortium. Families self-reported ASD diagnoses, and the SPARK consortium did essential phenotype curation and/or clinical evaluation and interviews using accepted diagnostic tools. I downloaded supplemental tables for variant calls (Tables S1, S3, and S6) and TADA results (Table S7) on 2019-09-03. Coordinates are in GRCh37. They called variants using GATK (McKenna et al., 2010) and FreeBayes (Garrison & Marth, 2012), and annotated with Annovar (K. Wang et al., 2010).

SFARI: I downloaded an updated version of the SFARIGene Module from the web tool on 2019-08-08 (Abrahams et al., 2013). This file lists genes within the Human Gene Module, their gene score and number of associated reports. Not included in this file is information about the reported variants within each gene. Variant level data was provided courtesy of the SFARI foundation on 2019-08-06. It is unknown if all sources in SFARI are using the same diagnostic criteria for ASD, or the same procedure was used for variant calling. The variants in SFARIGene were manually curated by MindSpec Incorporated, and the ASD-status of the variant publication is not provided i.e. if the report was from an ASD-specific study or not.

I faced many problems with the SFARI data mainly due to inconsistent report or incomplete information (Table 2.2). Human Genome Structural Variation (HGVS) nomenclature

inconsistencies I came across in SFARI included allele or residue changes being reported without their positions or as reference/alternate alleles, such as A>G or delTGG, or L/S or p.Met1?. Further, variants reported as a protein HGVS change without their corresponding cDNA change could not be annotated using VEP due to failure to perform unambiguous reverse annotation. There were an estimated 1181 variants in SFARI that could not by annotated due to various inconsistencies in reporting.

Table 2.2: Examples of variants reported in SFARI with non or inconsistent HGVS format

Symbol (SFARI score)	Allele Change	Residue Change	Examples of proper HGVS notation from publication or VEP	Variant type	Inheritance	PMID
SCN2A (1)	N/A	p.Thr1711Leufs Ter8	SCN2A:c.5130_5131insCTCCCCCCCCCCCCCCCCCTA	Frameshift	De Novo	28379373
BCAS1 (4)	A>G	L/S	BCAS1:c.458T>C BCAS1:p.L153S	Missense	De Novo	25533962
HOXA1 (S)	delTGG	N/A	chr7:g.27101863delTGG	Inframe_del	Familial (6); de novo (2); unknown (27)	21624971
TSC2 (3S)	G>A	N/A	chr16:g.2138546G>A TSC2 :c.5359G>A TSC2 :p.G1787S	Missense	Unknown	29271092

ClinVar (Landrum et al., 2014): ClinVar is a publicly accessible database aggregating information about human genetic variation and health. I downloaded updated GRCh37 variant calling files, a per-variant summary tab delimited file and a variant citation tab delimited file from the ClinVar FTP on 2019-08-21. ClinVar is not specific to any one phenotype, so while some variants are reported to be associated with ASD, they are likely associated with other phenotypes as well. As such, it is possible that ClinVar could have incomplete information for some of their variant reports. For example, ClinVar reports a DYNC1H1 variant (14:102452354:C>T) with Yuen et al., 2017 (PMID:28263302) as an associated citation, which is an ASD WGS study, but ClinVar does not report ASD as an associated phenotype (Table 2.3). While this report is included in VariCarta, there is no record of the variant event occurring in VariCarta. Upon further inspection, the variant does not appear to exist in the Yuen publication.

Table 2.3: Example of incomplete information in ClinVar.

Variant Reported in ClinVar	Variant origin in ClinVar	PubMedIDs in ClinVar	Phenotypes in ClinVar	PubMedID in VariCarta	Variant in VariCarta or PubMedID
DYNC1H1: 14:102452354:C>T	Germline; Unknown; *De novo*	28263302 21820100 21076407 …	Hereditary disease Charcot-Marie-Tooth disease Spinal muscular atrophy…	28263302	No

2.2.4: Variant level annotations

Inheritance and Validation

Each of the above variant sources report inheritance of variant events. I simplified inheritance categories into *de novo*, inherited (familial, maternal, paternal etc.), and other (mosaic, unknown, etc.). VariCarta, MSSNG and the recent genetic association also report variant validation status, and validated inheritance. In cases where the validated inheritance differs from initial inheritance patterns identified from sequencing, I kept the validated inheritance category. MSSNG reported inheritance status for each variant in a string format, for example, 0,0:0,1:0,1. The 0,0 genotype represents a mother with a homozygous reference genotype, the first 0,1 genotype represents a father with a heterozygous genotype (1 being an alternate allele), and the second 0,1 represents a child with a heterozygous genotype. I considered proband genotypes not inherited from parents to be *de novo*; variants on the X chromosome of male probands were considered to be hemizygous, and fall in the other category of inheritance (C Yuen et al., 2017). ClinVar does not report specific inheritance modes for associated phenotype(s), as such, I did not simplify the inheritance of the variant event. Inheritance categories in ClinVar include germline, somatic, *de novo*, biparental, inherited, maternal, paternal, unknown, uniparental and/or a combination of these depending on reports and phenotypes associated with variant event. When aggregating and harmonizing across multiple variant sources, some variant events were found to be reported with unknown or mixed inheritance types, sometimes within the same sample across different publications (Table 2.3, 2.4). This presented us with challenges for counting variants, and likely resulted in miscounting of the number of *de novo* and inherited variants.

Table 2.4: Examples of unresolved inheritance pattern reported in the same sample across multiple studies

Variant	Sample	Source	Inheritance pattern
21:38877659:G>A	M12327	28831199 27824329	*De novo*
21:38877659:G>A	M12327	28191889	Unknown

VEP

I used a local installation of the Ensembl Variant Effect Predictor v97.3 (VEP) for variant annotation (McLaren et al., 2016). The installation differs from the web-based tool because it is set up to use a merged cache of human GRCh37 files from both Ensembl and RefSeq rather than the current default GRCh38 with Ensembl files only. In the local installation, multiple plugins had been previously setup for annotations not housed in the GRCh37 cache. I customized the local installation to ensure that protein identifiers and Human Genome Structural Variation (HGVS) cDNA, protein and genomic nomenclature were added as annotations, and to ensure that variants in the HGVS format could be used as inputs by adding local installations of GRCh37 fasta files and necessary flags. See Tables 2.6, 2.7 and Appendix A.1 for details on the GRCh37 cache, installed plugins, and flags used for VEP annotation.

Variants from each source were annotated with VEP to ensure consistency across the sources. Variant reporting format varied among the sources. SFARI reports variants in cDNA and protein changes in HGVS format, for example, CHD8:c.1096C>T, or CHD8:pGln366Ter. For variants reported in non HGVS or inconsistent HGVS format, I made attempts to rectify the nomenclature. I sorted cDNA and protein changes by chromosome and positional change and annotated with VEP independently. One issue we ran into with SFARI was that VEP can return different annotations for variants reported in cDNA or protein HGVS formats compared to

variants reported in genomic coordinates (sourced directly from original publications) (Table 2.5). This is important to keep in mind when selecting SFARI genes and variants for further study because it is possible that the annotated protein change is not the same change reported in an ASD proband, and it is possible that the variant was not found specifically in a proband with ASD.

Table 2.5: Example of the different VEP annotations resulting from different variant input formats.

Input	Source	HGVSg	HGVSc	HGVSp	Transcript
ADAMTS18: c.3300_3303dupGAAA	SFARI	16:g.77325264_77325265insTTTC	c.3300_3301insGAAA	p.Lys1101GlufsTer74	NM_199355 NP_955387
16:77325261:G>GTTTC	Original Publication 29346770	16:g.77325261_77325262insTTTC	c.3303_3304insGAAA	p.Pro1102GlufsTer73	NM_199355 NP_955387

The rest of the variant sources, VariCarta, MSSNG, the recent genetic association studies, and ClinVar, report genomic coordinates of variant events. I formatted the input as chromosome _ position _ referenceAllele / alternateAllele, and sorted by chromosomal location. VEP was run in four batches to account first for number of variants being annotated, and second to separate input format: VariCarta, MSSNG and recent genetic association studies; ClinVar; and two for each HGVS input type for SFARI (cDNA and protein HGVS, respectively).

After VEP was run, I performed additional manual filtering and annotation. For each input variant, VEP produces multiple variant outputs annotated with predicted transcript, gene and protein effects. In most cases, variant inputs have predicted effects in multiple transcripts and, in the case of regulatory variants, multiple transcripts and genes, necessitating further

filtering and annotation. First, I filtered the output variant calls so that only variants annotated to be affecting genes from the ASD gene set were kept. Second, we want to know, of the predicted effects for each input variant, which is likely to be the most damaging. I selected the most damaging variant based on two criteria: 1) predicted consequence (Table 2.8), and 2) transcript biotype. Transcript biotypes can be grouped into four main categories: protein coding, pseudogene, long noncoding and short noncoding. Variants in protein coding transcripts, including nonsense-mediated decay and non-stop decay, were considered likely to be more damaging than those found in pseudogene and noncoding transcripts. I simplified the most damaging variant class into three categories: 1) The loss-of-function category included splice site, frameshift and stop gain or stop loss variants; 2) The missense category only included missense variants; and 3) The other included a wide range of different variant types including in-frame indels, synonymous, intronic and intergenic variations. Lastly, for each variant, I selected the maximum reported allele frequency as the maximum value reported by either the 1000s Genomes Project or gnomAD.

Table 2.6: Data included in the GRCh37 VEP cache

Source	Description	Version
1000 Genomes	Identify genetic variants present at 1% of the populations studied. Provides global allele frequencies.	Phase 3 (1000 Genomes Project Consortium et al., 2015)
COSMIC	Resource for somatic mutations in human cancers. Allows for identification of existing or collocated variants with the input.	86 (Tate et al., 2019)
ClinVar	Gives clinical significance of variant if the alternate allele is reported to be associated with a specific phenotype in ClinVar	10-2018 (Landrum et al., 2014)
ESP	Provides allele frequencies from the NHLBI GO Exome Sequencing Project aimed at discovering genes associated with heart, lung and blood disorders.	03-11-2014 (https://esp.gs.washington.edu/drupal/)
HGMD-PUBLIC	Public version of the Human Gene Mutation Database collects information about gene lesions involved in inherited disease. Allows for identification of existing or collocated variants with the input.	2017.4 (Stenson et al., 2017)
Assembly	Human genome assembly used.	GRCh37.13
dbSNP	SNPs and indels imported from dbSNP.	151 (Sherry et al., 2001)
GENCODE	Provides annotations from human and mouse with a high level of biological evidence.	19
gnomAD	Provides allele frequency from 125 748 whole exomes.	r.2.1, exomes (Karczewski et al., 2019)
PolyPhen2	Provides scores predicting impact of amino acid substitutions on protein structure and function	2.2.2 (Adzhubei et al., 2010)
RefSeq	Predicted RefSeq transcripts, along with associated gene symbols and variant effect prediction scores.	01-2015 (O'Leary et al., 2016)
Regulatory build	Annotates with overlapping regulatory regions	1.0
SIFT	Provides scores predicting effects of amino acid substitution on protein structure and function.	5.2.2 (Ng & Henikoff, 2003)

Table 2.7: Plugins installed for VEP and utilized for annotation

Plugin	Description	Version (associated assembly)
CADD	Retrieves CDD raw and phred scores for SNVs and indels	1.4 (GRCh37) (Rentzsch et al., 2019)
Conservation	Variant position GREP conservation score from Ensembl Compara database	EPO 35-way mammalian alignment
dbscSNV	Retrieves splicing variant information	1.1 (GRCh37) (X. Liu, Jian, & Boerwinkle, 2011)
Downstream	Predicts downstream effects of a frameshift variant on protein sequence; splicing changes not predicted	Unknown
ExAC	Retrieves ExAC allele frequencies	ExAC.r0.3.1.vep.vcf (Lek et al., 2016)
Exac_pLI	Retrieves ExAC probability loss-of-function intolerance score	ExAC.r0.3 (Lek et al., 2016)
LoFTool	Ranked genic intolerance and disease susceptibility based on ratio of LoF mutations to synonymous mutations while adjusting for mutation rate and conservation	From Fadista J et al. (2017) (GRCh37)

Table 2.8: VEP hierarchy of variant consequences. Modified from Ensembl Variation – Calculated variant consequences (https://uswest.ensembl.org/info/genome/variation/prediction/predicted_data.html)

Consequence	Description
transcript ablation	Deleted transcript feature
splice acceptor variant	Splice variant in 2-base region of 3' end of intron
splice donor variant	Splice variant in 2-base region of 5' end of intron
stop gained	Premature stop codon and shortened transcript
frameshift variant	Disrupt translation reading frame by changing triplet code
stop lost	Disrupt terminator codon for elongated protein
start lost	Canonical start site disrupted
transcript amplification	Amplification of a region with a transcript
inframe insertion	Non-synonymous insertion not disrupting triplet code
inframe deletion	Non-synonymous deletion not disrupting triplet code
missense variant	Alter 1 or more base resulting in a different amino acid
protein altering variant	Predicted change in protein coding sequence
splice region variant	Variant in slicing region but not within the splice sites
incomplete terminal codon variant	Change of one base of the termination codon of an incompletely annotated transcript
start retained variant	Change the base of a start codon without losing start site
stop retained variant	Change the base of a termination codon without losing terminator
synonymous variant	Variant causing no change to the encoded amino acid
coding sequence variant	Sequence variant changing coding sequence
mature miRNA variant	Transcript variant located in mature miRNA
5 prime UTR variant	Variant in 5' UTR
3 prime UTR variant	Variant in the 3' UTR
noncoding transcript exon variant	Variant changing non-coding exon sequence in a non-coding transcript
intron variant	Transcript variant in an intron
NMD transcript variant	Variant in a non-sense mediated decay transcript
noncoding transcript variant	Variant in a non-coding RNA gene
upstream gene variant	Variant in 5' upstream sequence
downstream gene variant	Variant in 3' upstream sequence
TFBS ablation	Deleted of region including transcription factor binding site
TFBS amplification	Amplification of region including a transcription factor binding site
TF binding site variant	Variant located within a transcription factor binding site
regulatory region ablation	Deleted region including a regulatory region
regulatory region amplification	Amplification of a region with a regulatory region
feature elongation	Variant that causes the extension of a genomic feature
regulatory region variant	Variant in a regulatory region
feature truncation	Variant reducing a genomic feature
intergenic variant	Variant in the intergenic region, between genes

Ensembl and RefSeq FTP

I downloaded the fasta files for coding and peptide sequences of Ensembl and RefSeq transcripts from the Ensembl GRCh37 FTP and the RefSeq FTP on 2019-09-29 (Flicek et al., 2014; O'Leary et al., 2016; Sayers et al., 2019; Zerbino et al., 2018). I used BiomaRt with Ensembl 37 to find sequences for transcripts not documented in the FTPs (Durinck et al., 2005; Durinck, Spellman, Birney, & Huber, 2009). For transcripts with coding and protein sequences, I calculated coding and protein length, the GC content of coding transcripts, and molecular weight of peptide transcripts using the "mw" function from the R Peptides package (Osorio, D., Rondon-Villarreal, P., & Torres, R., 2014, 2015).

Joining variant annotations

Information from the variant sources, VEP and Ensembl/RefSeq annotation steps were joined by input variant so all information pertinent to a variant was collected (reported inheritance, publication identification numbers, variant level annotations etc.). For SFARI, variants were first joined by cDNA identifiers, followed by protein changes if cDNA changes were unavailable. VEP cannot always unambiguously annotate cDNA and protein changes, thus, some SFARI variants were unable to be mapped and annotated (Table 2.2).

2.2.5: Gene level annotations

Variant recurrence and counting

We want to know how many variants each gene in the ASD gene set has, and what categories these variants fall into. A problem faced when pooling variant information from multiple sources is double counting of the same variant from the same samples, and of the same variant across multiple samples. Double counting can result in inaccurate assessment of genic association with ASD. To help combat this, variant event identifiers were used. VariCarta has

variant event identification numbers based on sample identification number, genomic coordinates and the reference and alternative alleles. MSSNG and the recent genetic association studies do not report variant event IDs, but they do report sample identification numbers for variant events. As such, if the sample identification number and variant event were already reported in VariCarta, I assigned the variant event the identification number provided in VariCarta. Otherwise, I built an identifier from the sample identification number and variant event. I considered a variant event to be recurrent if it was reported in different samples. I counted the number of non-recurrent and recurrent missense and LoF variant events present in each gene in the gene set.

SFARI does not report sample identification numbers, but it does report the publication associated with the variant. Therefore, to ensure variants are not double counted, if the publication was included in VariCarta or was one of the recent genetic association studies, I eliminated the variant event from the SFARI count. I considered the same variant event reported in multiple publications by SFARI to be recurrent because I could not ascertain if the variant events were reported in the same or different samples without manually checking each associated publication. I counted the number of non-recurrent and recurrent missense and LoF variant events present in each gene in the gene set.

ClinVar does not report sample identification numbers nor is the specific mode of inheritance for associated phenotype reported. Therefore, to avoid double counting, if the variant event was previously identified, I filtered the variant event out prior to counting. I counted additional missense and LoF variants associated with ASD in ClinVar.

LoF and missense variation constraint measures

I downloaded measures of per-gene constraint against missense and LoF variation from the gnomADv2.1.1 database on 2019-07-18 (Karczewski et al., 2019). The missense-z score measures the deviation of a genes' number of observed missense variants from the expected number (Karczewski et al., 2019; Lek et al., 2016). The observed/expected LoF score measures the deviation of a genes number of observed LoF variants from the expected number (Karczewski et al., 2019; Lek et al., 2016). The probability loss-of-function intolerance (pLI) score is an extension of the o/e LoF score. pLI measures a gene's likelihood to be intolerant of LoF variants, and, unlike the observed/expected LoF score, is corrected for gene length and ranges between 0-1 (Karczewski et al., 2019; Lek et al., 2016). Measures of constraint against LoF and missense variation act as a proxy for likelihood of involvement in disease because genes found to be under high constraint (missense $z>3$, observed/expected LoF < 0.35, pLI>0.9) are more likely to cause detrimental effects when mutated.

Other gene level annotations

I built a binary protein-protein interaction network using BioGridv3.5.169 human physical interaction data and counted the node degree or the number of interactors a gene has using the R package EGAD (Ballouz, S. et al., 2017; Oughtred et al., 2019). The per-gene multifunctionality rank was previously computed using Gene Ontology annotations by Nathaniel Lim of the Pavlidis lab, and roughly reflects the number of functions a gene has been annotated with. The per-gene number of publications was calculated previously by Shams Bhuiyan of the Pavlidis lab.

See Figure 2.1 for schematic overview of variant and gene level annotation pipeline.

Code availability

Code and publicly available raw data used in this analysis are currently available on request.

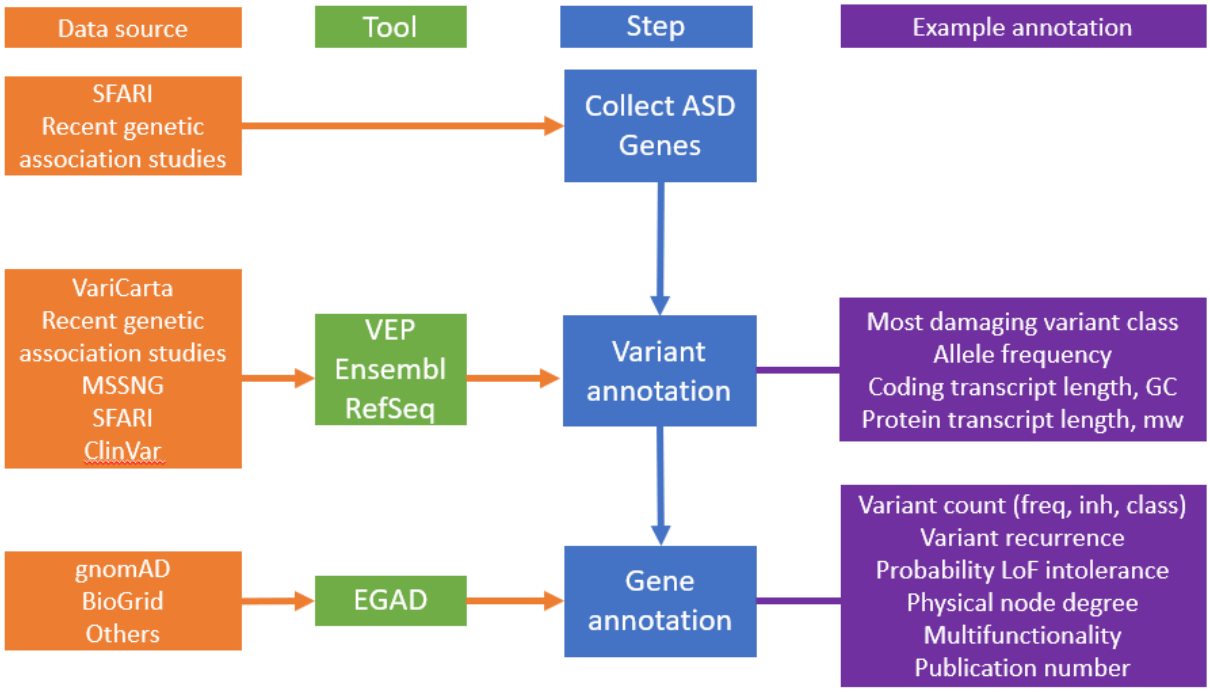

Figure 2.1: Overview schematic of gene and variant annotation for ASD genes. The ASD gene set was created from SFARI and significant findings from recent genetic association studies. Variant sources include VariCarta, the recent genetic association studies, MSSNG, SFARI and ClinVar. VEP was used to ascribe different annotations to variants, including predicted variant consequence, allele frequencies, damaging predictions. RefSeq and Ensembl FTPs were used to calculate coding transcript length and GC content, and protein transcript length and molecular weight (mw). Gene level annotations included variant count based on variant class (LoF and missense), inheritance, and recurrence, as well as measures of constraint against LoF and missense variation, such as probability of loss-of-function intolerance, and others, such as the number of physical interactors (physical node degree), number of functions (multifunctionality) and the number of publications.

2.3: Results

2.3.1: Variant annotation collection statistics

The goal of collecting variant and gene level annotations for SFARI genes and other genes found to be associated with ASD through genetic association studies was to aid our collaborators in their prioritization of SFARI category 3 and 4 genes for further experimental study.

By aggregating and harmonizing data across five variant sources, 2868 rare (<1%) *de novo* missense and 2128 rare *de novo* LoF variant events were found in 690 and 399 genes, respectively (Table 2.9). We enumerated fewer rare inherited missense (1460 in 354 genes) and LoF (842 in 387 genes) variant events (Table 2.9). Another large category of variant reports was rare missense events with "other" inheritance categories, including unknown, mosaic and multiple inheritance patterns reported for the same variant event (1797 events in 433 genes) (Table 2.9). ClinVar had the fewest number of variant events associated with ASD, with only an additional 213 missense events across 29 genes and 88 LoF events across 30 genes (Table 2.9). However, these numbers of additional ASD missense and LoF variants serve more as an estimation because it appears that some variants reported in ClinVar have incomplete information (Table 2.3). Lastly, SFARI had an estimated 1181 variant entries in 367 genes reported as missense or LoF that could not be mapped with VEP due to various variant reporting inconsistencies (Table 2.9).

Establishing accurate variant profiles and counts within a gene is essential for identifying ASD risk genes. Choosing ASD risk genes and variants for further study is a complex task, and requires the consideration of many factors, including variant class, allele frequency and

inheritance patterns. Comparing variant profiles of high and low-confidence ASD risk genes may be able to aid in identifying risk genes and variants for further study.

Table 2.9: Gene and Variant collection statistics

Category	Number of variant events	Number of genes
Number rare *de novo* missense	2868	690
Number of rare inherited missense	1460	354
Number of rare other missense	1797	433
Number rare *de novo* LoF	2128	399
Number of rare inherited LoF	842	387
Number of rare other LoF	487	223
Number additional ASD missense ClinVar	213	29
Number additional ASD LoF ClinVar	88	30
Number of unmapped missense/LoF SFARI vars	1181	367

2.3.2: Gene annotation collection statistics

I annotated genes with multiple generic gene annotations, including measures of constraint because measuring the deviation of observed variant counts from the expected number provides insight into how detrimental it would be if a gene was mutated. These scores can act as proxies for how likely it is for a gene to be involved in any disease. Missense z-scores over 3, observed/expected LoF (o/e LoF) scores below 0.35 and probability loss of function intolerance (pLI) scores above 0.9 indicate that a gene is more likely to be depleted of missense or LoF variation, and can be considered to be under high constraint. We found that the SFARI and novel high-confidence genes were enriched for genes with high pLI and missense z-scores, and low o/e LoF scores compared to the rest of the protein coding genes in the genome (Figure 2.2, Table 2.10). These results suggest that generic measures of disease gene likelihood have some predictive power for identifying ASD genes and they may be useful for prioritizing the set of lower confidence ASD risk gene candidates. This is by no means a novel finding, and these constraint scores have already been used in ASD gene prioritization studies, as discussed later on and in Chapter 3.

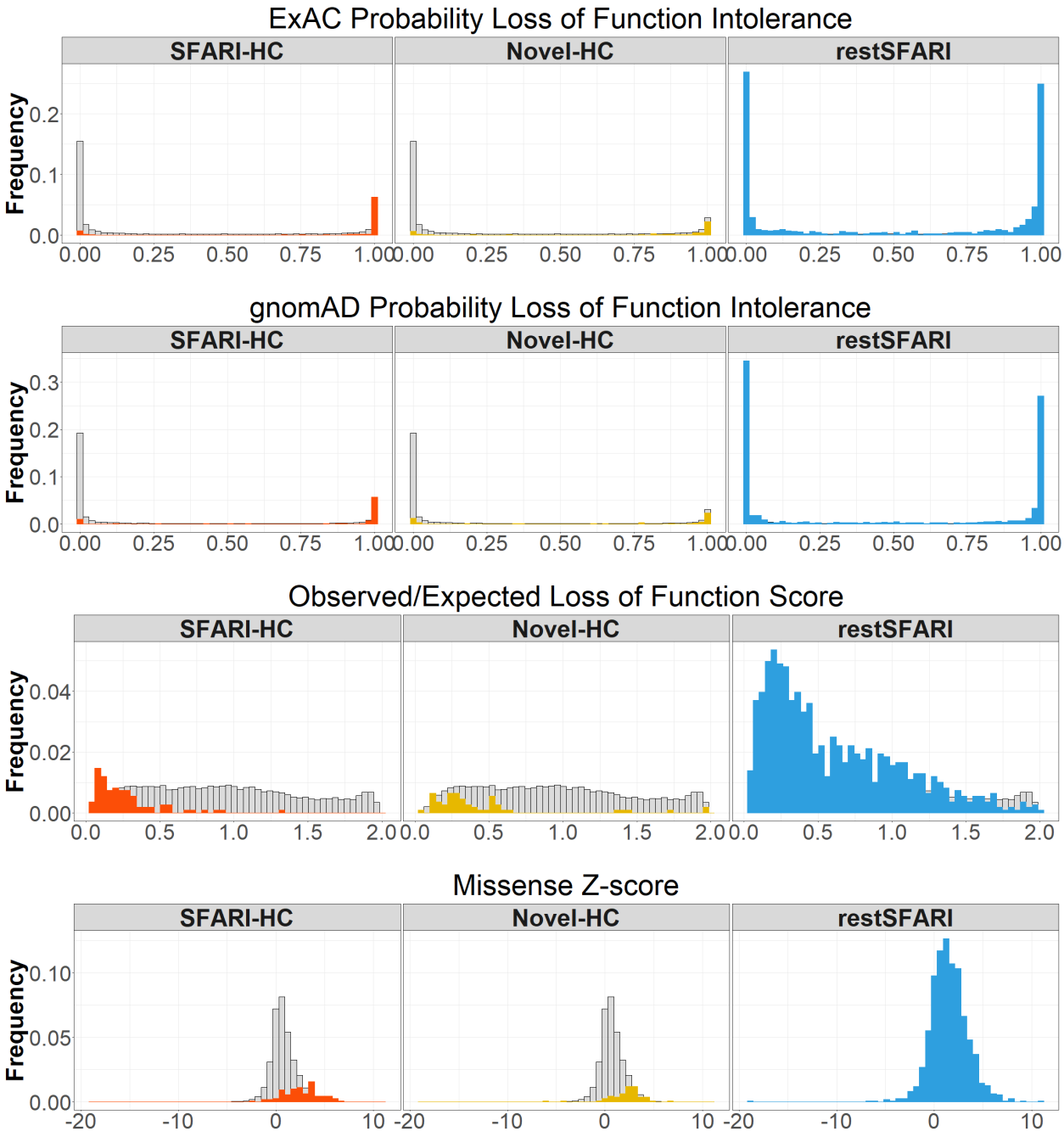

Figure 2.2: LoF and missense variant constraint score distributions for ASD gene sets. ASD gene sets are plotted against the background of roughly 19 000 protein coding genes in the genome (grey). Constraint scores measure deviation of observed variant counts from expected variant counts. High pLI (>0.9) or low observed/expected LoF (<0.35) score and high missense z-score (>3), imply a gene is depleted of LoF or missense variation, and is under increased constraint.

Table 2.10: Median values of constraint scores for ASD gene sets

	Rest of protein coding genes in genome	SFARI high confidence	Novel high confidence	Rest of SFARI genes
gnomAD pLI	0.0013	0.99*	0.97*	0.32*
ExAC pLI	0.021	0.99*	0.97*	0.58*
Observed/expected LoF score	0.91	0.20*	0.30*	0.45*
Missense z score	0.61	2.6*	2.4*	1.4*

* denotes a significant difference (Bonferroni adjusted p-value < 2e-11; Wilcox test) between the median of the ASD gene set and the background of the rest of the protein coding genes in the genome.

I also annotated genes with other generic annotations including the number of physical interaction partners (physical node degree), the number of functions (multifunctionality) and the number of associated publications. We found that the SFARI and novel high-confidence genes were enriched for well studied genes with a higher number of physical interaction partners, functions and publications compared to the background of the rest of the protein coding genes in the genome (Figure 2.3, Table 2.11). These results highlight a potential bias in the ASD gene sets toward well studied genes. The utility of constraint and other gene annotation scores for prioritizing ASD risk gene candidates will be investigated further in the following chapter.

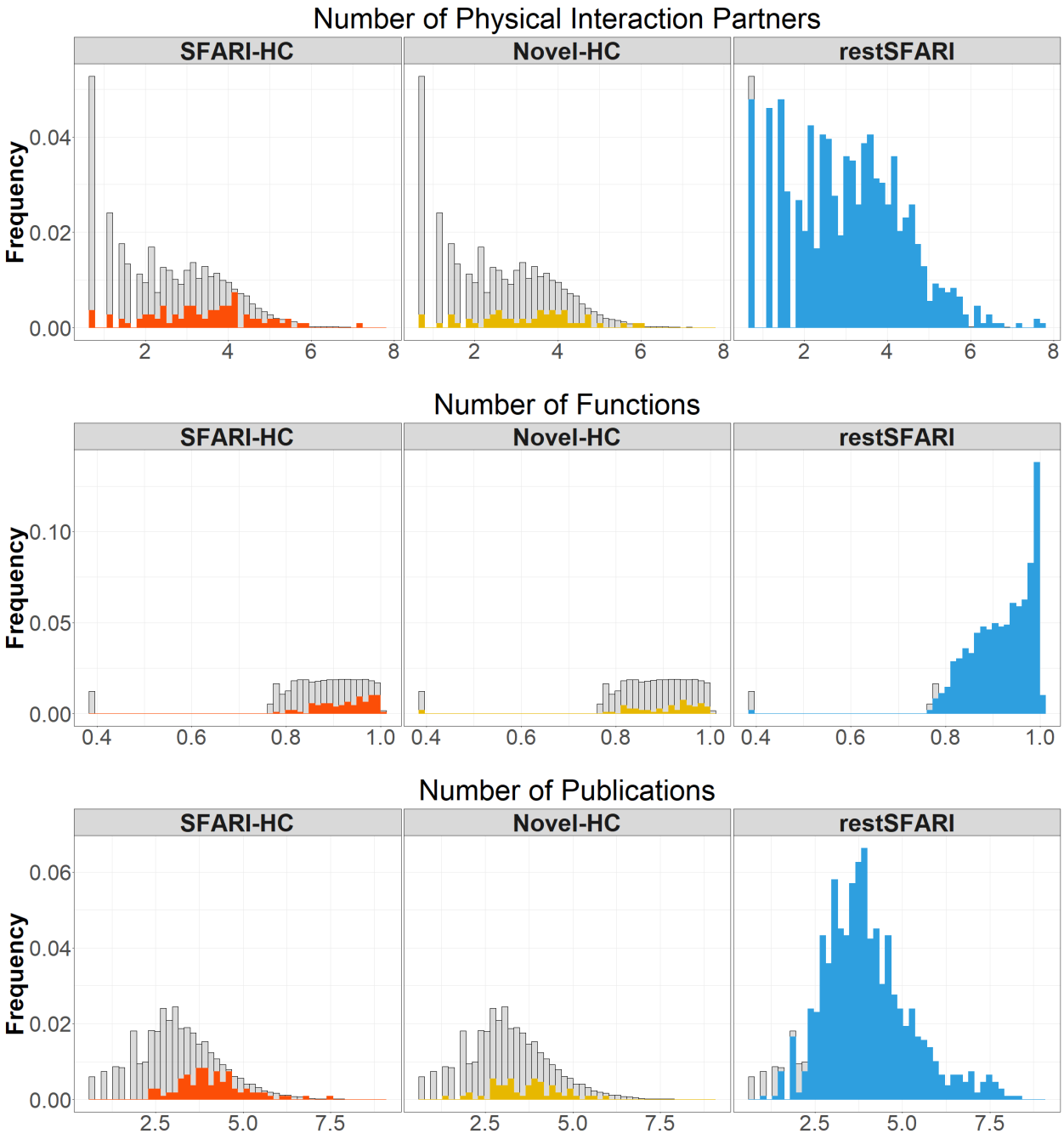

Figure 2.3: Distribution of number of physical interaction partners, number of functions and number of publications for ASD gene sets. ASD gene sets against the background of roughly 19 000 protein coding genes in the genome (grey). Number of physical interaction partners (physical node degree) and number of publications are plotted in a log+1 transformation.

Table 2.11: Median values of number of publications, multifunctionality and physical node degree for gene sets.

	Rest of genes in genome	SFARI high confidence	Novel high confidence	Rest of SFARI genes
Number of physical interaction partners (physical node degree)	12	31*	30*	20*
Number of functions (multifunctionality)	0.83	0.94*	0.92	0.93*
Number of publications	21	52*	40*	45*

* denotes a significant difference (Bonferroni adjusted p value < 4e-3; Wilcox test) between the median of the ASD gene set and the background of the rest of the protein coding genes in the genome.

2.4: Discussion

Currently, 90 genes have been associated with ASD with high confidence, but there are hundreds more which have been associated with ASD with lower degrees of certainty. We are part of a collaboration project aimed at identifying more high-confidence ASD genes. Specifically, the collaboration is focused on studying rare *de novo* missense variants found in SFARI category 3 and 4 genes. To aid in our collaborators lab's prioritization of SFARI genes, we enumerated missense and LoF variants found in ASD gene candidates, and collected a multitude of variant and gene level annotations, such as allele frequencies, *in silico* damaging predictions, measures of constraint against different classes of variation and gene multifunctionality.

An important factor in establishing a high confidence genic association with ASD is recurrence of variant events within the same gene across multiple samples. Given that a large proportion of ASD risk has been estimated to be from rare *de novo* and inherited variation, multiple sequencing studies are often needed to established variant recurrence (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). An issue that comes from using multiple sequencing studies is variant event count inflation caused by reporting the same variant in the same individual multiple times. As demonstrated by VariCarta , some instances of sample overlap are easy to deal with, for example, changing – (dash) to _ (underscore), or removal of .p1 suffix identifiers (Belmadani et al., 2019). However, there are instances where how to correct for sample overlap is unclear, or where it is not possible due to lack of information. For example, the recent large-scale genetic association study from the iHart consortium re-sequenced 119 ASD samples and their biological parents (trios) which were originally classified as cases (without parents) in the De Rubeis et al. (2014) genetic association

study (Ruzzo et al., 2019). The iHart sample identification numbers are prefixed with iHart (eg.iHart1002) whereas the De Rubeis sample identification numbers are unique based on source of variant from multiple different contributors, making direct comparison impossible (De Rubeis et al., 2014; Ruzzo et al., 2019). The iHart study and De Rubeis et al. (2014) use TADA models, which rely on integrating different information to estimate a genes likelihood of being involved in disease, including per gene counts of variation across multiple samples (He et al., 2013). Therefore, proper control for cohort overlap and variant reporting format is important to avoid double counting. The iHart consortium was able to control for double counting the same variant event in the same sample by obtaining sample cohort overlap from the De Rubeis et al. (2014) publication (Ruzzo et al., 2019). However, there are other genetic association studies using TADA models which have instances of double counted variants due to sample misidentification and lack of variant format harmonization (Du et al., 2019). Additionally, SFARI and ClinVar do not report sample identification numbers for variants, making it impossible for me to ascertain if a variant event is reported in the same or different subjects across multiple studies.

Other possible sources of errors in variant event counting are inconsistent variant reporting and incomplete or incorrect annotation. Variants with unresolved or unknown inheritance patterns (Table 2.2, 2.4), reported in formats that cannot be annotated with VEP (Table 2.2), reported with incomplete phenotype information (Table 2.3) or reported in NDD-related, but non ASD, phenotypes (SFARI) could all cause miscounting of variant events found in ASD probands. SFARI is not fully ASD specific, meaning that many variants are reported in other disorders or syndromes which are associated with ASD, or in studies where ASD is not the primary focus. Having accurate variant counts is essential establishing genic association with ASD, and for prioritizing ASD candidate genes for further study.

The utility of constraint scores and other gene level annotations for prioritization will be investigated in more detail in Chapter 3; however, I will discuss implications for prioritization briefly here. Due to the current focus in identifying highly penetrant, rare *de novo* variation associated with ASD, there has been increasing amount of evidence to suggest that ASD genes are more likely to be under high evolutionary constraint against LoF variation, and thus, more likely to be "disease genes" (De Rubeis et al., 2014; Kosmicki et al., 2017; Lek et al., 2016; Ruzzo et al., 2019; Satterstrom et al., 2019). Our results confirm this because we found that high-confidence ASD genes are more intolerant to LoF variation compared to the rest of the protein coding genes in the genome (Figure 2.2, Table 2.10). However, not all genes associated with ASD have evidence of constraint against LoF variation. From this, we can conclude that while constraint measures may be useful for prioritizing ASD genes, evidence of constraint against variation does not imply a particular disease status for a gene (Karczewski et al., 2019; Ruzzo et al., 2019; Satterstrom et al., 2019). Furthermore, our results demonstrate that high-confidence ASD genes have a higher number of associated publications, functions and interaction partners compared to the rest of the protein coding genes in the genome (Figure 2.3, Table 2.11). Superficially, these results do not appear to be worrisome: ASD high-confidence genes are likely well studied because they are associated with ASD. However, being "well studied" does not mean that all the functional annotations collected are reliable, and/or specific to ASD. The implications of bias in gene sets for gene function prediction using guilt by association will be explored in detail in Chapter 3.

Despite the discussed limitations, particularly of SFARI and ClinVar, the variants we have collected and annotated here represent the cleanest set of variants found in ASD probands that we are aware of. While counting variants is an important step, recurrence is not all that is

needed to establish a risk gene. Currently, researchers are investigating how other genic features, like constraint against LoF variation, may be used to help characterise and prioritize ASD risk genes. Comparing ASD risk gene candidates, like lower-confidence ASD risk genes in SFARI category 3 and 4, to current high-confidence ASD risk genes in terms of their variant profiles, and other defining characteristics, like high LoF constraint, may help prioritize ASD risk genes and variants for further study.

Chapter 3: Evaluation of ASD gene prioritization studies

3.1: Introduction

Genetic association studies identify ASD risk genes by collecting case individuals affected with ASD and control individuals not affected with ASD, and looking for correlations between disease status and underlying genetic variation. There are multiple statistical models used to evaluate the strength of the genetic evidence supporting a gene's contribution to ASD. Currently, statistical association approaches for rare variants are still being developed, but some form of genetic association and/or linkage analysis is the only way that is widely accepted and proven for identifying disease risk genes (Botstein & Risch, 2003; Dean, 2003; Gilissen, Hoischen, Brunner, & Veltman, 2012). To date, genetic association studies have identified only a subset of the estimated 1000 ASD risk genes.

Researchers have suggested that machine learning algorithms which leverage heterogeneous biological network data can be used to aid in the search for more ASD risk genes. However, as discussed previously, work in the Pavlidis lab has shown that using gene networks for gene function prediction based on guilt by association (GBA) is complicated by biasing factors from the underlying networks, resulting in a lack of real-world utility (Gillis & Pavlidis, 2011a, 2012; Pavlidis & Gillis, 2012). In the following investigation, we aim to assess the reliability and usability of guilt by association machine learning studies to answer the question: do they provide additional useful information for ASD gene prioritization? To answer this question, I collected twelve published ASD gene prioritization studies. Six studies are considered "gold standard" methods because they are using tests for genetic association, whereas the other six utilize GBA with heterogeneous network data (Table 3.2). I used a set of current SFARI high-confidence ASD risk genes and a set of newly discovered (novel) ASD risk genes

for performance evaluation (Table 3.1). Additionally, I looked at how the prioritization studies agreed with generic network features, and other generic gene annotations, such as measures of constraint, because to be considered successful, ASD-specific GBA ML prioritization studies should out-perform generic methods and be competitive with genetics-based methods.

3.2: Materials and Methods

3.2.1: Gene sets

I assembled two sets of ASD candidate risk genes for use in algorithm evaluation. I assembled the SFARI high-confidence (SFARI-HC) ASD risk gene set from current (August 2019) high-confidence ASD genes from SFARI categories 1 and 2 (n=90). Many of these genes were initially identified by the genetic association studies from De Rubies et al. (2014) and Sanders et al. (2015). Different subsets of these high-confidence genes were used for training of the GBA machine learning algorithms. We hypothesized that both the genetic association studies and machine learning studies would perform well when evaluated using genes they had identified, or had been trained on. Therefore, we considered this evaluation a "sanity check". The novel high-confidence (novel-HC) ASD risk gene set I assembled was made up of recently identified ASD risk gene candidates from three large-scale TADA studies: Ruzzo et al. (2019), Feliciano et al. (2019) and Satterstrom et al. (2019). These three studies were built based on the background of the De Rubeis et al. (2014) and Sanders et al. (2015) studies, and given that many of the SFARI high-confidence genes were identified by the earlier TADA studies, we expect many of these novel findings will be considered high-confidence SFARI genes in the near future. Most of the novel significant findings were not used in the training of the GBA machine learning algorithms. We considered this evaluation a "testing scenario" because the ultimate use case of the machine learning algorithms is to highly prioritize and predict novel ASD genes.

3.2.2: ASD gene prioritization studies and generic measures of disease gene likelihood

In this section, I describe the twelve studies that I evaluated in this work. Each study scored genes based on the author's assessment of their probability of contributing to ASD risk. All studies also provided lists of genes they considered to be high-confidence ASD risk gene candidates based on a thresholding of their rankings. I obtained these scores from the supplemental tables of the publications. I also evaluated three measures of constraint against LoF variation because they can be thought of as generic measures of disease gene likelihood.

I mapped gene symbols and Entrez gene identification numbers provided by each study to NCBI official gene symbols and Entrez gene identification numbers, and kept only protein-coding genes (Sayers et al., 2019). I used the average score when a gene was listed more than once in a study. I ranked the scores from each study so that 1 was the highest possible score, indicating higher assessed likelihood of being involved in ASD, and 0 was the lowest possible score. The probability loss of function scores from ExAC and gnomAD were already in the proper scale, and I ranked the scores from highest, indicating genes likely to have high constraint against LoF variation, to lowest. The scale of the observed/expected LoF score is opposite to the pLI scale and does not range from 0-1. I ranked the o/e LoF score from lowest to highest. Lastly, for protein-coding genes not assessed in each study, I set the score to be 0, or in the case of the o/e LoF score, the highest observed value (2).

Studies are organized into four categories based on the approach they used. I gave each study a short name that I use to refer to them throughout the remainder of the thesis.

<u>Genetic association studies:</u>

De Rubeis (De Rubeis et al., 2014) used WES from approximately 13 000 samples from trios and case-controls to identify *de novo* and inherited LoF variants, and *de novo* likely damaging

missense variants (Mis3 by PolyPhen2). They used a TADA analysis to identify 33 ASD risk genes at FDR < 0.1. Samples from the Autism Sequencing Consortium (ASC), from Simons Simplex Consortium (SSC) (O'Roak et al. (2012), Sanders et al. (2012), and Iossifov et al. (2012)), and other cohorts were used. Association scores were provided for 18 735 genes.

Sanders (Sanders et al., 2015) used WES from approximately 17 000 samples from trios and case-controls to identify *de novo* and inherited LoF variants, *de novo* likely damaging missense variants (Mis3 by PolyPhen2), and small *de novo* deletions. They employed a TADA analysis to identify 65 ASD risk genes at FDR < 0.1. They sequenced roughly 2500 SSC families in addition to using SSC samples from Levy et al. (2011), Iossifov et al. (2014) and Dong et al. (2014), and ASC samples from De Rubeis et al. (2014), and samples from Pinto et al. (2014), among others. Association scores were provided for 18 665 genes.

iHart (Ruzzo et al., 2019) used WGS from 2308 individuals from 493 multiplex Autism Genetic Resource Exchange (AGRE) families to identify *de novo* and inherited LoF variants and *de novo* likely damaging missense variants (Mis3 by PolyPhen2). They used their data and the Sanders data, and the Sanders TADA model to identify 69 ASD risk genes with FDR < 0.1, including 14 novel findings. Association scores were provided for 18 472 genes. 424 samples included in their analysis were listed as cases in the original ASC TADA analysis from De Rubeis et al. (2014); 119 samples and their parents were re-sequenced for determination of inheritance patterns, and subsequent correction for variant counts prior to their TADA analysis (Ruzzo et al., 2019).

Spark (Feliciano et al., 2019) was the pilot study for the Simons Powering Autism Research for Knowledge (SPARK) project. They identified inherited and *de novo* likely damaging missense mutations (CADD >= 25) in 465 SPARK trios. They combined their *de novo* variants with *de*

novo variants from 4773 other simplex ASD trios from the ASC (De Rubeis et al. (2014)) and SSC (Iossifov et al. (2104); Krumm et al. (2015)), among other sources, for a TADA analysis. They identified 67 genes with FDR < 0.1, with 13 novel findings. They provided scores for the 2249 genes found to have additional variation in SPARK families (Feliciano et al., 2019).

Satterstrom (Satterstrom et al., 2019) is the most recent and largest-scale genetic association study, with well upwards of 20 000 samples. They used samples from the SSC (Iossifov et al. (2012); Iossifov et al. (2014); O'Roak et al. (2012); Sanders et al. (2012)), the ASC (De Rubeis et al. (2014) and others), others from the AGRE and many other cohorts around the world. They used WES to identify *de novo* and case-control LoF, and *de novo* missense mutations (predicted by MPC, the "missense, PolyPhen-2, constraint score"), and employed TADA analysis to identify 102 ASD risk genes at FDR < 0.1. They considered 31 significant genes to be novel findings. Association scores were provided for 17 484 genes. Importantly, they changed the TADA method from the studies above by using the pLI score from ExAC and the MPC score to estimate the priors for the relative risk of LoF and missense variant classes. At time of this writing, (November 2019) Satterstrom et al. (2019) is only available as a preprint and has not been peer-reviewed.

Iossifov (Iossifov et al., 2015) computed a "Likely Gene-Disruptive" (LGD) score based on recurrence of LGD variants, the difference in frequency of LGD variants between ASD probands and unaffected siblings (ascertainment differential), and the load of LGD variation in ASD probands. They used data from WES of 2471 families from the SSC (Iossifov et al. (2014)), and exome variants from approximately 6000 neurotypicals from the Exome Variant Server (Iossifov et al., 2015). The theory behind the LGD score is similar to the TADA test and to generic measures of constraint against LoF and missense variation because they use recurrence of

variants across multiple samples and models of expected LGD variation in a typical gene to increase power to find disease genes (He et al., 2013; Iossifov et al., 2015; Lek et al., 2016). First, they calculated a gene's "vulnerability score" based on: 1) A likelihood model of expected LGD variants in a typical gene built from synonymous variation data in the parents of the SSC and the control neurotypicals; and 2) The proportion of causal ASD genes estimated from the ascertainment differential for LGD variation between affected probands and unaffected sibling controls (Iossifov et al., 2015). By combining the "vulnerability score" with the observed load of LGD variants in proband WES data, they created a heuristic prioritization score for ASD genes (Iossifov et al., 2015). They provided scores for 23 953 genes (Iossifov et al., 2015). They compared their gene rankings to a score called Residual Variation Intolerance Score (RVIS), which is a measure of constraint derived from missense mutations.

I stress that the studies I list above, particularly the studies using TADA analyses, are among the most important in terms of identifying what are generally considered high-confidence ASD genes, using methods that human geneticists are generally comfortable with. I include them in my study primarily to help establish a baseline to which the approaches listed below can be compared.

GBA machine learning studies

Studies in this class do not use information from ASD genetic association studies, but they use machine learning algorithms to distinguish ASD from non-ASD risk genes using other types of non-genetics data.

Princeton (Krishnan et al., 2016) used a support vector machine (SVM) trained on a human brain-specific functional interaction network built from multiple protein-protein interaction databases, gene expression datasets, and other regulatory and genetic and chemical perturbation

data (Greene et al., 2015). Their training labels included 549 positive genes weighted by the strength of evidence of association with ASD (E1,2,3,4), and a set of 1189 manually curated non-mental health disease genes. Their feature space was a gene-gene matrix where the cells represented the probability of a gene-gene interaction calculated from their brain-specific functional interaction network. Using their gene-gene matrix, and their ASD-positive and ASD-negative training gene sets, they fit a linear SVM with a penalty parameter to control misclassification of their evidence-weighted labels (i.e. lower misclassifications of E1, high-confidence labels). They ran 5-fold cross-validation 50 times on different subsets of their evidence-weighted training labels and found that the model with all evidence-weighted labels had the best performance for separating positive and negative genes. In theory, the SVM fit a linear plane in the high-dimensional feature space which was able to maximize the separation between positive and negative training genes. For each candidate gene, the distance between the candidate gene and discriminant hyperplane (i.e. prediction from computing the linear kernel function) was converted to a probability using regression; an average probability was taken across each of the 5 cross-validation folds. Prediction scores were provided for 25 825 genes, and they identified their top decile of genes as likely ASD risk gene candidates. Their published evaluation and validation of their ranking system included: 1) Calculating enrichment of genes with *de novo* mutations in independent ASD sequencing studies in their top decile (Sanders et al. (2012), O'Roak et al. (2012), Iossifov et al. (2012), Neale et al. (2012), Iossifov et al. (2014), and De Rubeis et al. (2014)); 2) Calculating enrichment of experimentally determined targets of ASD-related proteins and pathways, such as FMRP and MAPK signalling, in their top decile; and 3) Calculating enrichment of genes found to be associated with intellectual disability, schizophrenia, and other developmental disorders in their top decile. From their main

evaluation, they found that their evidence-weighted labels had significantly better performance during cross validation than other combinations of training labels, and that there was significant enrichment of genes found to have *de novo* likely damaging variation in independent ASD sequencing studies in their top decile of genes. Overall, they concluded that their method was able to prioritize many new ASD candidate risk genes, and claimed that their top ranked genes had the potential to speed up ASD gene discovery.

ASD_frn (Duda et al., 2018) used a random forest classifier with a brain-specific functional interaction network. They built their network from human, rat and mouse gene expression datasets from non-cancer related brain experiments, multiple protein-protein interaction databases, and protein docking and phenotype annotations. They utilized 143 ASD genes from SFARI 1, 1S, 2, 2S and Sanders as positive training labels, and 1176/1189 of the negative non-mental health genes from Princeton. Their feature space was a gene-gene matrix where the cells represented the probability of a gene-gene interaction based on their brain-specific functional interaction network. Using their gene-gene matrix, and their ASD-positive and ASD-negative training gene sets, they trained 5 different machine learning models with 5-fold cross-validation, and found that their random forest model had the best performance based on the average AUROC from the 5-folds. Random forests are built from multiple decision trees which segment the feature matrix into a number of simple regions by recursive binary splitting. In each decision tree of a random forest, each split in each tree uses a random sample of features, and at each successive split, the best splitting rule is chosen so that the two new regions are as pure as possible. In other words, the feature and its threshold which give the best separation between the positive and negative training data is chosen at each split point in each tree. The leaves at the bottom of a decision tree are called terminal nodes. Predictions are made for candidate items

(genes) based on which decision path it follows, and the proportion of positive and negative training observations in the terminal node. In other words, after allowing a candidate gene to follow a decision path and entre a terminal node, if the majority of the genes in the terminal node are positive training genes, the candidate gene will be predicated as a positive. Prediction scores were provided for 21 114 genes, and they identified their top 2111 genes as likely ASD risk gene candidates. Their published evaluation and validation of their ranking system included: 1) Calculating the enrichment of genes with recurrent and non-recurrent *de novo* LoF mutations in ASD probands and unaffected siblings from the SSC (Iossifov et al. (2014)) and MSSNG (Yuen et al. (2017)) in their top decile; and 2) Calculating the enrichment of genes found to be involved in Alzheimer's disease, Parkinson's disease and ataxia. From their evaluation, they found significant enrichment of genes with *de novo* LoF mutations in SSC and MSSNG probands in their top decile, and an absence of significant enrichment for genes involved in other brain-related disorders. Overall, they concluded that their method predicted genes with evidence of ASD association and was able to propose numerous novel genes they claimed had a high likelihood of contributing to ASD.

DAMAGES (Zhang & Shen, 2017) used cell-type specific expression data from 24 mouse central nervous system cell types from 6 regions, and measures of constraint against LoF and missense variation from ExAC to try to identify ASD risk genes. First, they created a DAMAGES (D) score built from gene expression profiles of 145 genes found to have *de novo* LGD variants in probands and unaffected siblings from Iossifov et al. (2012), Neale et al. (2012), O'Roak et al., (2012) and Sanders et al. (2012) using Principal Component Analysis (PCA). Regression analysis was used to evaluate how each principal component from the PCA analysis was able to predict a gene's variation source as proband or sibling control. Next, they used

logistic regression to combine the D score with measures of constraint against LoF and missense variation to create an ensemble score (E). Their training labels for their logistic regression classifier were 36 genes found to have 2 or more *de novo* LGD mutations in ASD probands, and 156 genes with only 1 or more *de novo* LGD mutations in sibling controls. Their feature space consisted of the D score (PCA-based gene expression profiles) and ExAC constraint scores. They used logistic regression to estimate the effect size of each feature, and then calculated an ensemble (E) score for each candidate gene predicting its likelihood of being a haploinsufficient ASD gene. The mouse genes they used to create the D score were mapped to human orthologs so the constraint scores could be added to create the E score. They identified the top 117 genes by E score as likely ASD risk gene candidates. I kept E scores for 15 881 genes with single, unambiguous mappings to human genes. Their published evaluation and validation of their ranking system included: 1) Calculating enrichment of genes with LGD mutations from sequencing studies published after 2012 (De Rubeis et al. (2014), Iossifov et al. (2014)), and 438 SFARI genes by category (S,2,3,4,5,6) in the top ranking of the D score; 2) Comparing the D score and Ensemble score to constraint measures alone, and ranks provided by Princeton by calculating a modified precision recall statistic. From their evaluations, they found enrichment of genes found to have *de novo* likely damaging variation in independent ASD sequencing studies, and that their method have favourable performance compared to other studies. Overall, they concluded that their gene expression signatures reflected haploinsufficiency in ASD, and claimed that it was able to predict whether or not likely damaging variants confer increased risk to ASD.

RF_Lin (Lin, Rajadhyaksha, Potash, & Han, 2018) employed a random forest classifier using gene-level constraint measures from ExAC and a weighted network built from BrainSpan and

InWeb protein-protein interaction data as features (T. Li et al., 2017; Miller et al., 2014). Their training labels are the same employed in the ASD_frn method above. See ASD_frn for description of random forests. In theory, their random forest was able to split the feature space (constraint, weighted network information) into regions which could separate their positive training genes from the negative training genes, and thereby predict which candidate genes were the most similar to positive training genes. Prediction scores were provided for 17 099 genes, and they identified their top 2111 genes as likely ASD risk gene candidates. They did not provide scores for their training labels. Their published evaluation and validation of their ranking system included: 1) Calculating enrichment of genes found to have *de novo* LoF or missense mutations from 2517 SSC families (Iossifov et al. (2014)) and MSSNG (Yuen et al. (2017)) in their top decile; 2) Comparing their ranking system to ExAC pLI, Iossifov, Sanders, Princeton and DAMAGES by calculating decile enrichment of 130 SFARI category 3 genes, and 43 genes with recurrent *de novo* LoF mutations identified in Stessman et al. (2017), Wang et al., (2016), Yuen et al., (2017), and Li et al., (2017); and 3) Comparing their ranking system to ExAC pLI, Iossifov, Sanders, Princeton and DAMAGES by calculating the AUROC with their labelled and unlabelled data, with the 130 SFARI category 3 genes, and with the 43 genes with recurrent *de novo* LoF mutations. From their evaluations, they found significant enrichment of genes with *de novo* LoFs in SSC and MSSNG probands, and that their method showed higher enrichment of 173 candidate ASD genes in their top decile compared to other methods. Overall, they concluded that their method demonstrated that spatiotemporal gene expression and constraint metrics predicted ASD risk genes, and claimed that their method provided many new candidate genes with strong evidence of contributing to ASD.

Genetics-GBA machine learning studies

The studies in this section used a combination of ASD-specific features and other features to build their models. The ASD-specific features come from genetic association data in the studies described above. Information from the two classes of features are integrated prior to training a machine learning algorithm to distinguish ASD from non-ASD risk genes, using high-confidence ASD genes from genetic association studies as their positive training set.

DAWN (L. Liu et al., 2014) targeted their search to genes found to be expressed in the prefrontal and motor-somatosensory (PFC-MSC) neocortex during the 10-24 weeks post-conception phase based on previous findings from Willsey et al.(2013). Willsey et al., (2013) found that the PFC-MSC from 10-24 weeks post-conception was a potential nexus for risk based on coalescence of gene expression during that time. They built a co-expression network from BrainSpan data of the selected regions and time points using Weighted Gene Co-expression Network Analysis (WGCNA) and overlaid genetic association statistics from a TADA model to identify ASD risk gene candidates (Langfelder & Horvath, 2008; L. Liu et al., 2014). The TADA scores utilized were calculated from rare *de novo* likely damaging variants, rare transmitted likely damaging variants and case-control likely damaging variants from multiple sources, including Iossifov et al. (2012), Neale et al. (2012), O'Roak et al. (2012), Sanders et al. (2012), and the ARRA Autism Sequencing Consortium, among others. They used unsupervised model-based clustering and a hidden Markov random field to model the correlation of genetic association scores across the co-expression network to identify co-expressed nodes with high genetic evidence of association with ASD ("network ASD genes"). Next, they used a false discovery rate procedure to determine which of the "network ASD genes" were most likely to contribute to ASD ("risk ASD genes", rASD genes). They provided prediction scores for the 10

233 genes in the network which had exome data. They identified their top 127 genes (FDR < 0.05) to be likely ASD risk gene candidates. Their published evaluation and validation of their ranking system included: 1) Two permutation dilution experiments were conducted whereby the signal from genetic association or co-expression data was diluted to determine the sensitivity of the rASD gene list to either signal; and 2) Calculation of enrichment of *de novo* LoF mutations identified in a targeted sequencing study of 44 ASD candidate genes in 2448 ASD trios in their 127 rASD genes. From their evaluation experiments, they found that DAWN was sensitive to both the TADA signal and the co-expression signal, and that DAWN was able to identify genes found to have more *de novo* LoF variants in ASD probands. Overall, they concluded that there was a high likelihood that DAWN had identified true ASD risk genes.

forecASD (Brueggeman, Koomar, & Michaelson, 2018) is a stacked random forest ensemble classifier. They built the first layer from BrainSpan spatiotemporal gene expression data (Miller et al., 2014) and a protein-protein interaction matrix built from STRING (Szklarczyk et al., 2019). The second layer was built from the scores from layer 1, and scores from other studies, all of which are included in my study (Princeton, DAWN, DAMAGES, De Rubeis and Sanders). Their training data included 76 SFARI high-confidence ASD genes, and 1000 random non-SFARI genes (Brueggeman et al., 2018). See ASD_frn for description of random forests. In theory, each random forest layer they built was able to split the feature space (BrainSpan, STRING, genome-wide ASD prediction sores) into regions which could separate their positive from their negative training genes, allowing for candidate gene predictions to be made based on shared/similar associations to the positive training genes in the feature space. Prediction scores were provided for 17 957 genes, and they identified their top 1787 as likely ASD risk gene candidates. Their published evaluation and validation of their ranking system included: 1)

Calculating enrichment of genes with *de novo* likely disrupting mutations in MSSNG (Yuen et al., (2017)) and Spark (unpublished at time of forecASD development) probands, SFARI 3, 4, 5 and syndromic genes, and gene targets of CHD8 and FMRP in their top decile of ranks; 2) Calculating the AUROC for their score, and scores from Princeton, DAMAGES and Sanders SFARI category 1 and 2 genes, and SFARI category 3 genes; 3) Calculating enrichment of genes with *de novo* likely disrupting mutations in MSSNG (Yuen et al. (2017)) and Spark (unpublished at time of forecASD development) probands in the top decile of the scores from Princeton, DAMAGES and Sanders for comparison; and 4) Fitting logistic regression models to assess how much forecASD is adding to genetic TADA signals. From their evaluations, they found their method was better able to classify SFARI 1 and 2 genes and SFARI 3 genes, and they showed greater enrichment in their top decile for genes found to have recurrent *de novo* likely damaging variants in Spark and MSSNG probands compared to other studies. Further, they concluded that forecASD was able to provide biological context to TADA genetic signals important for prioritization. Overall, they concluded that their method was able to generalize to new data and claimed that they had created a valuable framework for combing both genetics and non-genetics data to prioritize ASD risk gene candidates which will be useful for when ASD gene discovery by genetic association slows.

See Table 3.2 for details.

Generic measures of disease gene likelihood

The scores in this section were developed without any disease specificity, and measure the depletion of LoF variation within a gene. Therefore, these scores act as generic proxies for the likelihood of a gene to be involved in *any* disease. I downloaded these scores from the gnomADv.2.1.1 database on 2019-07-18 (Karczewski et al., 2019).

exac_pLI measures the probability of a gene to be extremely intolerant of LoF variation. It's scale is ranges from 0-1, with genes over 0.9 representing those extremely intolerant to LoF variation and under higher constraint. The ExAC browser is no longer available, and has been updated with the gnomAD browser. The ExAC score incorporates roughly 60 000 exomes.

gnomAD_pLI also measures the probability of a gene to be extremely intolerant of LoF variation. It's scale is ranges from 0-1, with genes over 0.9 representing those extremely intolerant to LoF variation and under higher constraint. The gnomAD score incorporates roughly 120 000 exomes.

o/e_lof measures the deviation of the number of observed LoF variants within a gene to the expected number. This score differs from the above two because its scale is reversed, with scores below 0.35 indicating extreme depletion of LoF variation and higher constraint. This measure is recommended for identifying genes likely to be depleted of LoF variation because it is more interpretable than the pLI (i.e. a score of 0.4 indicates that 40% of the expected LoF variants within a gene have been observed), and gives a better representation across the spectrum of selection. Unlike the pLI, the o/e LoF score comes with a confidence interval because it does not consider uncertainty around variant counts due to sample size and/or too many observed LoF variants.

Table 3.1: Gene sets used for evaluation

ASD Gene set	Description
SFARI high-confidence ASD genes (SFARI-HC)	90 genes in SFARI are in category 1, 2, 1S or 2S. These genes are considered high-confidence ASD genes (August 2019). The S denotes the gene has been associated with a syndromic form of ASD, but the numerical designation indicates enough evidence to be highly associated with sporadic ASD.
Novel high-confidence ASD genes (novel-HC)	Three recent large-scale genetic association studies employing TADA analyses (Feliciano et al., 2019; Ruzzo et al., 2019; Satterstrom et al., 2019) identified 58 novel high-confidence ASD risk gene candidates which are currently in SFARI categories 1, 2, 1S or 2S..

Table 3.2: Genetic association and GBA-based ML ASD gene prioritization studies.

Paper	Genetics	GBA	Method	Description
De Rubeis (De Rubeis et al., 2014)	✓		TADA	WES from approximately 13 000 samples. Analysed *de novo* and inherited LoF variants and *de novo* likely damaging missense variants (Mis3 by PolyPhen2). Found 33 genes significantly associated with ASD (FDR < 0.1).
Sanders (Sanders et al., 2015)	✓		TADA	WES from approximately 17 000 samples. Analyzed *de novo* and inherited LoF variants , *de novo* likely damaging missense variants (Mis3 by PolyPhen2), and small *de novo* deletions. Found 65 genes significantly associated with ASD (FDR < 0.1).
iHart (Ruzzo et al., 2019)	✓		TADA	WGS from 2308 multiplex AGRE families. Analysed *de novo* and inherited LoF variants and *de novo* likely damaging missense variants ; combined with analysis from Sanders. Found 14 novel significant associations with ASD (FDR < 0.1).

Spark (Feliciano et al., 2019)	✓		TADA	WES from 436 ASD trios. Analysed *de novo* and inherited LoF and *de novo* likely damaging missense variants; combined with analysis with ~4000 other published ASD trios. Found 13 novel significant associations with ASD (FDR < 0.1).
Satterstrom (Satterstrom et al., 2019)	✓		TADA	WES from upwards of 20 000 samples. Analysed *de novo* and case-control LoF, and *de novo* missense (predicted by MPC, the "missense badness score") variants; cohort used includes samples from De Rubeis and Sanders. Found 31 novel significant associations with ASD (FDR < 0.1)
Iossifov (Iossifov et al., 2015)	✓		Likely gene disruptive score	Computed a likely gene-disruptive score based on the load of disruptive (LoF) mutations and gene vulnerability using WES from the approximately 2000 SSC families and approximately 6000 neurotypicals from Exome Variant Server database.
Princeton (Krishnan et al., 2016)		✓	Evidence-weighted SVM classifier	An evidence-weighted support vector machine built on a functional interaction network of human gene expression, protein-protein interaction data, regulatory and genetic and chemical perturbation data. Trained on 594 evidence weighted ASD genes, and 1189 non-mental health associated genes.
ASD_frn (Duda et al., 2018)		✓	Evidence-weighted random forest classifier	An evidence-weighted random forest built on a functional interaction network of human, mouse and rat brain gene expression, protein-protein interaction data, protein docking and phenotype annotations. Trained on 143 high-confidence ASD genes from SFARI and Sanders, and 1176/1189 of the Princeton negative genes.
DAMAGES (Zhang & Shen, 2017)		✓	Logistic regression classifier	Used a combination of regularized and logistic regression with gene-expression profiles built from 24

				mouse CNS cell types in 6 regions and constraint measures from ExAC. Training genes included approximately 200 genes found to have *de novo* LGDs in ASD probands and sibling controls across multiple sequencing studies.
RF_Lin(Lin et al., 2018)		✓	Random forest classifier	A random forest classifier using a weighted network of BrainSpan gene co-expression and PPI evidence from InWeb, as well as other network features, and measures of constraint against different classes of variation. Training labels used from ASD_frn.
DAWN(L. Liu et al., 2014)	✓	✓	Cluster analysis with co-expression and TADA	Used WGCNA to build a co-expression network from prefrontal and motor-somatosensory neocortex during the 10-24 weeks post-conception BrainSpan data, and overlaid association statistics from a TADA analysis from approximately 3000 ASD families.
forecASD(Brueggeman et al., 2018)	✓	✓	Ensemble stacked random forest classifier	A stacked random forest ensemble method utilizing BrainSpan data, STRING protein-protein interaction data, and genome-wide results from Princeton, DAWN, DAMAGES, Sanders and De Rubies. Trained on 76 SFARI high confidence genes and 1000 non-SFARI genes.

3.2.3: Evaluation

Recovery and Prioritization of ASD gene sets:

I evaluated how well the ranked scores were able to recover and prioritize the SFARI high confidence and novel high-confidence ASD gene sets described above using receiver operator (ROC) curves and precision-recall (PR) curves. When evaluating the ability of the scores to rank the novel high-confidence ASD gene set, I removed the SFARI high-confidence ASD genes and other ASD genes used in the training of the ML algorithms. This was done to ensure that the algorithms were not penalized for performance on ASD genes. I plotted ROC curves as true positive rate (TP/(TP+FN)) vs. false positive rate (FP/(FP+TN)), and PR curves as precision (TP/(TP+FP)) vs. recall (TP/(TP+TN)), respectively. The top ranks provided by the studies are their predictions as to the most likely ASD risk candidates. Therefore, the PR curves are the preferred evaluation metric because they are more sensitive to classification errors in the top ranks.

ROC and PR curves plot true positive rate vs. false positive rate, and precision vs. recall at every score threshold. For example, at a score threshold of 0.85, all genes at and above that score threshold are predicted as a positive, and every gene below is predicted as a negative. Next, the number of true positive predictions and false positive predictions are enumerated, and the TPR, FPR, and precision can be calculated.

Summary statistics and 95% confidence intervals:

I calculated area under the ROC (AUROC) using the "auc" function with the "trapezoid" method from the DescTools R package(al. ASem, 2019). I calculated precision at 20% recall (P20R) of total genes in the ASD gene sets, and precision at 43% recall (P43R) of total genes in the ASD gene sets. Precision at 20% recall was selected as a 'midrange' for display purposes,

and is a commonly reported point statistics among ML studies (Peña-Castillo et al., 2008). While we could have calculated the area under the precision recall curve, we are most interested in the rate of false positive predictions in top ranks provided by each study. Many genes have an ExAC and a gnomAD pLI score of 1.0. Over 20% of the high-confidence ASD genes have a pLI score of 1.0, meaning that we are, therefore, not measuring the precision at 20% recall. We measured the precision at 43% recall as well to have a consistent comparison for precision-recall across all studies. I used 2500 bootstrapped samples to calculate 95% confidence intervals for AUROC and P20R statistics. The sampling I did was stratified, and done with replacement. This means that I sampled from the ASD gene sets and the rest of the scored protein coding separately in each of the 2500 iterations to ensure balanced coverage, and that the same gene could be sampled more than once in each iteration. Therefore, in each bootstrapped sample, I kept only unique genes for evaluation. Studies whose performance measures confidence intervals did not overlap were considered significantly different from each other.

Correlation:

I measured the correlation between the ranked scores provided by each study, and other metrics using the Spearman Correlation Coefficient, and their pairwise complete observations. Other metrics measured include multifunctionality rank, node degree of the BioGrid (Oughtred et al., 2019) protein-protein interaction network, number of publications, SFARI numeric gene score, and measures of constraint against LoF and missense variation.

Overlap:

Each method provided a cut off for their set of likely ASD genes, and I calculated the overlap in their top gene sets as their shared number of genes.

3.2.4: forecASD performance broken down by feature sets

I obtained code for the forecASD classifier from https://github.com/LeoBman/forecASD, and re-ran it locally. A different version of randomForest was implemented in my analysis (version 4.6-14) compared to the version used in the original research (version 4.6-12) (Brueggeman et al., 2018; Liaw & Wiener, 2002). I refit the final ensemble model (03_ensemble_model.R) with different sets of the input features used in final ensemble model: the noClassifiers (noClass) model removed features from other classifiers; the noClassifiersPPI (noClassPPI) model eliminated the other classifiers, and the STRING score; noClassifiersPPIBS (noClassPPIBS) model eliminated the other classifiers, the STRING score, and the BrainSpan score; the PPIOnly model only used the STRING score; and the BrainSpanOnly model only used the BrainSpan score. Feature importance was measured by mean decrease in accuracy and mean decrease in Gini node impurity. Mean decrease in accuracy is measured by randomly permuting each feature, and measuring the out-of-bag (CV) accuracy of the resulting trees. Mean decrease in Gini measures how well the features can split the data from mixed labelled nodes into pure single class nodes. They did not provide code for their feature importance plots; I used "varImpPlot" from the randomForest package to plot feature importance (Brueggeman et al., 2018; Liaw & Wiener, 2002). When rerunning their provided code, I found that two columns in their meta data had been improperly labelled, D (DAMAGES) and D_ens (DAMAGES ensemble), necessitating re-labelling for plotting of feature importance. Lastly, I evaluated each adapted model using the two ASD gene sets and with the same metrics described above.

Code and data availability:

Code and raw data pulled from the supplemental tables of the studies used in this analysis are currently available on request.

3.3: Results

3.3.1: Systems–based GBA ML methods do not prioritize novel high-confidence ASD genes well compared to other disease gene prioritization methods

The most meaningful test we performed was investigating how well the GBA ML studies prioritized novel high-confidence ASD genes which were not used during method development, and comparing their performance to genetics-based approaches and generic approaches for disease gene prioritization. We began this investigation with some clear expectations for how certain methods would perform. In order to be considered a successful method, an ASD-specific GBA ML study should have comparable performance to the genetic association studies alone. Further, given that we confirmed that high-confidence ASD genes are enriched for genes under higher constraint against LoF variation in Chapter 2, we expected that ASD-specific GBA ML studies would outperform generic measures of disease gene likelihood i.e. the generic measures of disease gene likelihood would perform with lower precision compared to ASD-specific approaches. Lastly, the more recent genetic association studies (iHart and Spark, and Satterstrom) are built on the De Rubeis and Sanders studies in that they are using overlapping samples, and similar model parameters and variant classes in their TADA analyses. Therefore, we expected that the De Rubeis and Sanders studies would rank the novel-HC ASD genes at lower or borderline significant levels. We also hypothesized that the iHart and Satterstrom studies would show higher rankings of each other's hits.

The key question here is whether the machine learning algorithms highly prioritized the novel-HC genes. If they perform well, then we might conclude that they are competitive with genetics-based approaches. If they perform poorly, given that in the field of genetic diseases, the only accepted way to identify disease genes is through genetic association studies, then we might

conclude that these methods have little real-life use for their intended purpose (Botstein & Risch, 2003; Dean, 2003; Gilissen et al., 2012).

Our main finding was that the systems-based GBA ML studies had comparable performance to the generic measures of disease gene likelihood, as is best depicted by their overlapping 95% confidence intervals for precision at 20% recall (i.e. $P20R_{ASD\_frn}$ = 0.93-4.92%; $P20R_{exac\_pLI}$ = 0.81-2.33%) (Figure 3.1D, Figure 3.2D, Table 3.3). While the studies had high AUROC statistics with overlapping 95% confidence intervals, these metrics are misleading because they are not sensitive to false positive predictions in top rankings, which are most relevant for prioritization studies (Figure 3.1B, Figure 3.2B, Table 3.3). This key finding suggests limited utility of GBA ML studies for ASD gene prioritization: use of a simple non-ASD specific measure of LoF constraint has comparable performance to complex ML approaches using complex networks with unknown reliability. Furthermore, we also found that the best performing GBA ML method was the genetics-GBA method forecASD ($P20R_{forecASD}$=5.04-11.11%), which had similar levels of performance to the genetic association studies developed before iHart, Spark and Satterstrom studies (i.e. $P20R_{Sanders}$=3.37-8.62%) (Figure 3.1D, Figure 3.3D, Table 3.3). The other genetics-GBA method, DAWN, has similar performance to systems-based GBA ML studies likely because they only provide predictions scores for roughly 10 000 genes in the genome (Figure 3.1, Table 3.3). Lastly, we found that the Satterstrom and iHart studies are not performing at 100% precision at 20% recall (i.e. $P20R_{iHart}$=4.06-60.98%) (Figure 3.3D, Table 3.3). This suggests that the two recent TADA studies do not agree on what genes are significantly associated with ASD. Additionally, the previous TADA studies have some performance, which would suggest that they were able to identify some of the novel genes at marginal levels of significance, and with the accumulation of

more data, these genes became significant in the newer studies (i.e. $P20R_{Sanders}$=3.37-8.62%) (Figure 3.3D, Table 3.3). Overall, these findings indicate limited utility of systems-based GBA ML studies for ASD gene prioritization, and that the TADA studies are not equivalent in their findings. Non- equivalence in TADA studies is further highlighted when investigating recovery of the novel high-confidence ASD genes as separate sets (See Appendix B.2,3,4,5 for details).

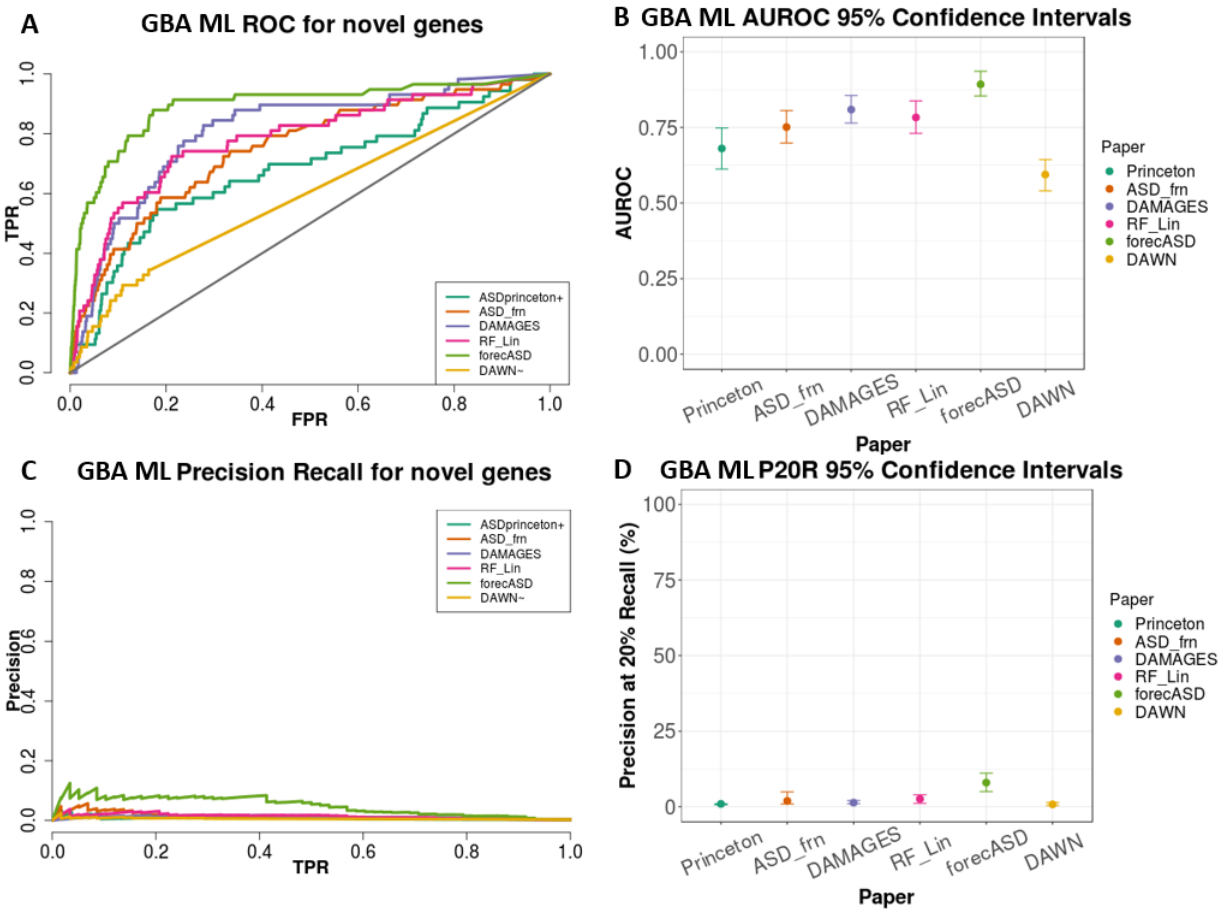

Figure 3.1: AUROC and PR statistics for GBA ML studies performance on novel-HC ASD genes. GBA ML studies have poor performance relative to genetics-GBA studies with lower AUROC (A) and low precision at 20% recall (C) in testing scenario with novel ASD genes (n=58). ML methods incorporating genetics information, particularly forecASD, have better relative performance. 95% confidence intervals were created from 2500 stratified bootstrap samples (B,D). TPR, True Positive Rate; FPR, False Positive Rate; AUROC, Area Under the Receiver Operator Curve; P20R, Precision at 20% Recall.

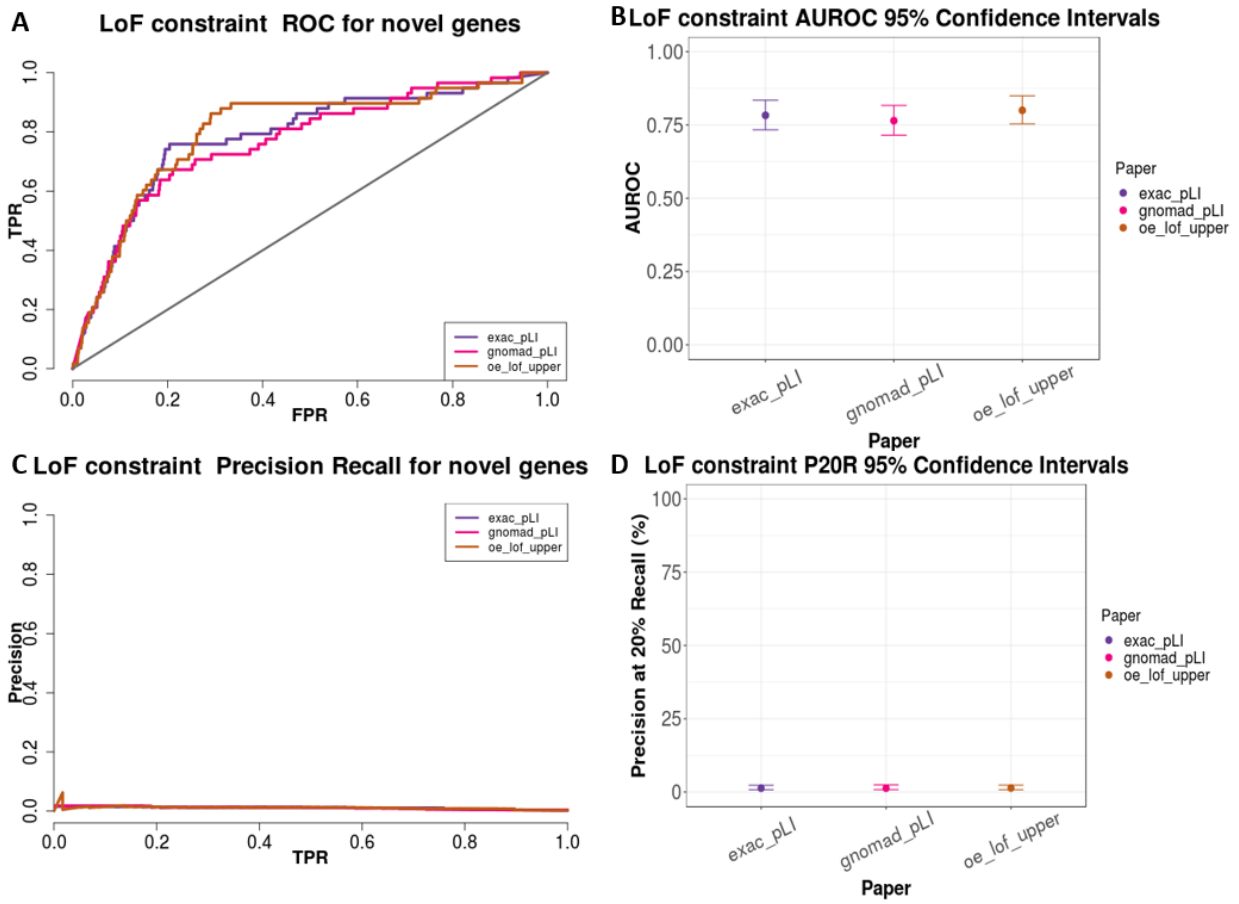

Figure 3.2: AUROC and PR statistics for generic LoF constraint measures performance on novel-HC ASD genes. LoF constraint measures have comparable performance to above systems methods with high AUROC (A) but low precision at 20% recall (C) in testing scenario with novel ASD genes (n=58). 95% confidence intervals were created from 2500 stratified bootstrap samples (B,D). TPR, True Positive Rate; FPR, False Positive Rate; AUC, Area Under the Receiver Operator Curve; P20R, Precision at 20% Recall.

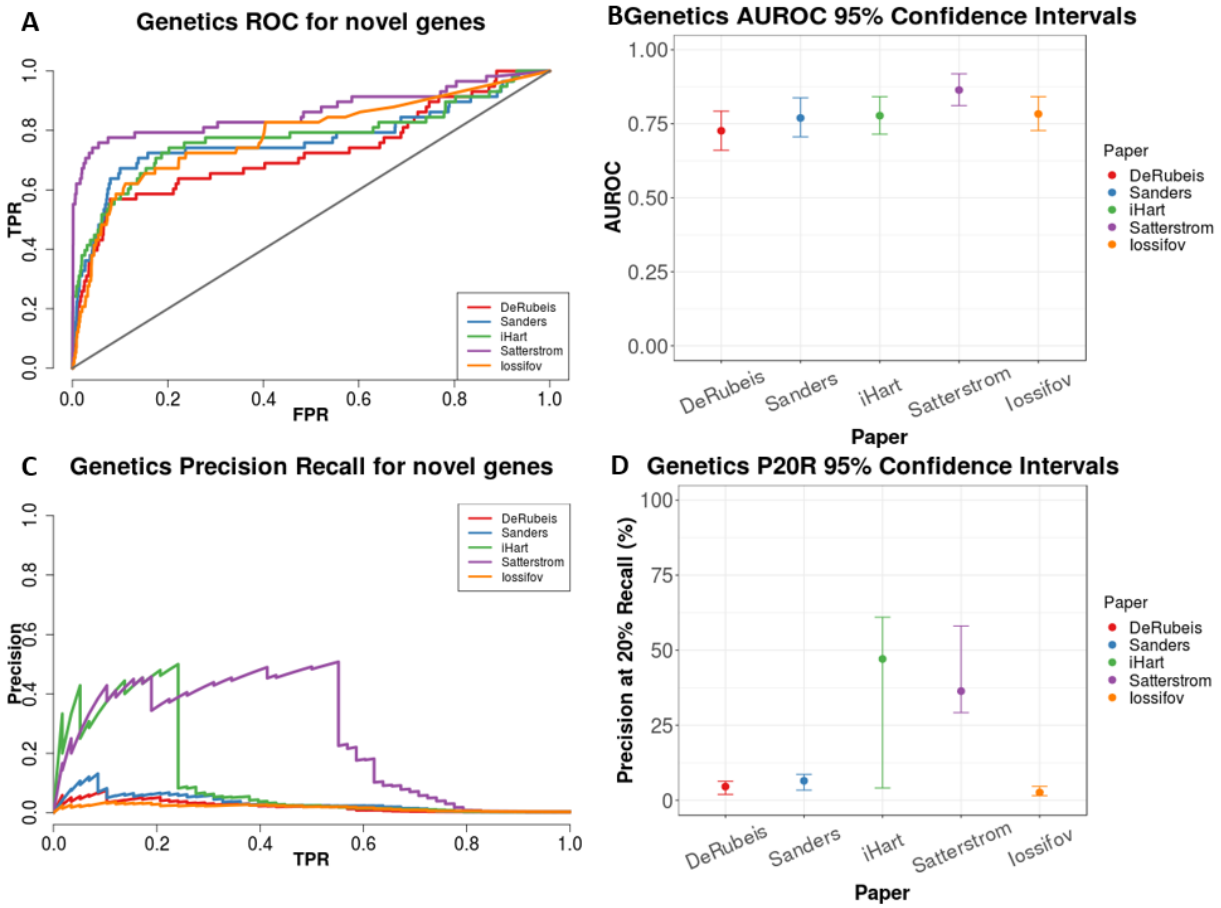

Figure 3.3: AUROC and PR statistics for genetics-based prioritization studies performance on novel-HC ASD genes. Genetic associations studies show a decline in performance with lower AUROC (A) and precision at 20% recall (C) in the testing scenario with novel ASD genes (n=58). Previous TADA studies have moderate performance on the novel gene set, suggesting that some genes which were found to have borderline association with ASD previously, are now ranked highly in newer TADA studies. 95% confidence intervals were created from 2500 stratified bootstrap samples (B,D). TPR, True Positive Rate; FPR, False Positive Rate; AUROC, Area Under the Receiver Operator Curve; P20R, Precision at 20% Recall.

Table 3.3: AUROC, Precision at 20% Recall and Precision at 43% recall performance statistics on novel ASD genes (* denote ties at recall of 20 or 43% of gene set)

Paper	AUROC	AUROC_CI	P20R	PR20_CI	P43R	PR43_CI
Princeton	0.68	0.61, 0.75	0.93	0.64, 1.02	0.92	0.59, 1.28
ASD_frn	0.75	0.7, 0.81	1.94	0.93, 4.92	1.03	0.73, 1.78
DAMAGES	0.81	0.76, 0.86	1.35	1.03, 2.07	1.58	0.93, 2.14
RF_Lin	0.78	0.73, 0.84	2.57	1.15, 3.93	1.76	1.28, 2.3
forecASD	0.89	0.85, 0.94	7.97	5.04, 11.11	6.14	3.82, 9.9
DAWN	0.59	0.54, 0.64	0.77	0.44, 1.4	0.62*	0.44, 0.82
exac_pLI	0.78	0.73, 0.83	1.32	0.81, 2.33	1.28	0.92, 1.67
gnomad_pLI	0.76	0.72, 0.82	1.33	0.85, 2.45	1.30	0.92, 1.7
oe_lof_upper	0.80	0.75, 0.85	1.38	0.8, 2.37	1.24	0.95, 1.62
DeRubeis	0.73	0.66, 0.79	4.55	1.95, 6.3	2.16	1.63, 3.69
Sanders	0.77	0.71, 0.84	6.47	3.37, 8.62	2.66	1.79, 5.1
iHart	0.78	0.71, 0.84	47.08	4.06, 60.98	3.02	1.64, 6.8
Satterstrom	0.86	0.81, 0.92	36.36	29.17, 58.08	45.88	37.42, 56.62
Iossifov	0.78	0.73, 0.84	2.59	1.53, 4.61	2.45	1.54, 3.24

3.3.2: Systems–based GBA ML studies do not perform well on SFARI high-confidence ASD genes

We also analyzed how well the GBA ML studies recovered SFARI high-confidence genes, many of which were used in the training of the ML algorithms, and comparted the results to other methods for disease gene prioritization The genes in the SFARI-HC set were discovered by different genetic association studies, many of which were first identified by the De Rubeis and Sanders studies (See Appendix B.1 for gene lists). Given the relationship between all the TADA studies, we expect the original and newer genetic association studies to highly prioritize SFARI high-confidence genes. The systems-based GBA ML studies used different subsets of

SFARI high-confidence genes, and other ASD associated genes, during training. Therefore, we would expect that these studies should highly prioritize SFARI-HC genes. Here, this is not a pure test of training performance because not all SFARI-HC genes were used during the training step. However, because the methods were developed at different times using different training gene sets, we opted for a consistent evaluation gene set. Lastly, based on findings from Chapter 2 and the previous section, we hypothesized that the generic measures of disease gene likelihood would have similar performance to the GBA ML studies on the SFARI-HC genes.

Our findings from this set of analyses parallel what we found for the novel-HC gene set. Mainly, we found that the GBA ML studies have comparable performance to the generic measures of disease gene likelihood with overlapping 95% confidence intervals for precision at 20% recall (i.e. $P20R_{ASD\_frn}$ = 3.87-9.38%; $P20R_{exac\_pLI}$ = 5.26-8.93%) (Figure 3.4D, Figure 3.5D, Table 3.4). RF_Lin did not provide predictions for their training genes, which partially explains its poorer performance relative to other studies ($AUROC_{RF\_Lin}$=0.32-0.47; $P20R_{RF\_Lin}$=0.9-3.41%) (Figure 3.5B,D, Table 3.4). Again we found that the genetics-GBA method forecASD had the best performance of the GBA ML studies with similar performance to genetic association studies (i.e. $P20R_{forecASD}$ = 75.95-100%; $P20R_{Sanders}$=91.67-100%; $P20R_{Satterstrom}$=68.75-100%) (Figure 3.4D, Figure 3.6D, Table 3.4). Again, these results show that systems-based GBA ML studies are providing little ASD-specific information above that provided by the generic measures of constraint against LoF variation. Overall, these findings highlight the limited utility of the systems-based GBA ML studies for prioritizing ASD risk genes.

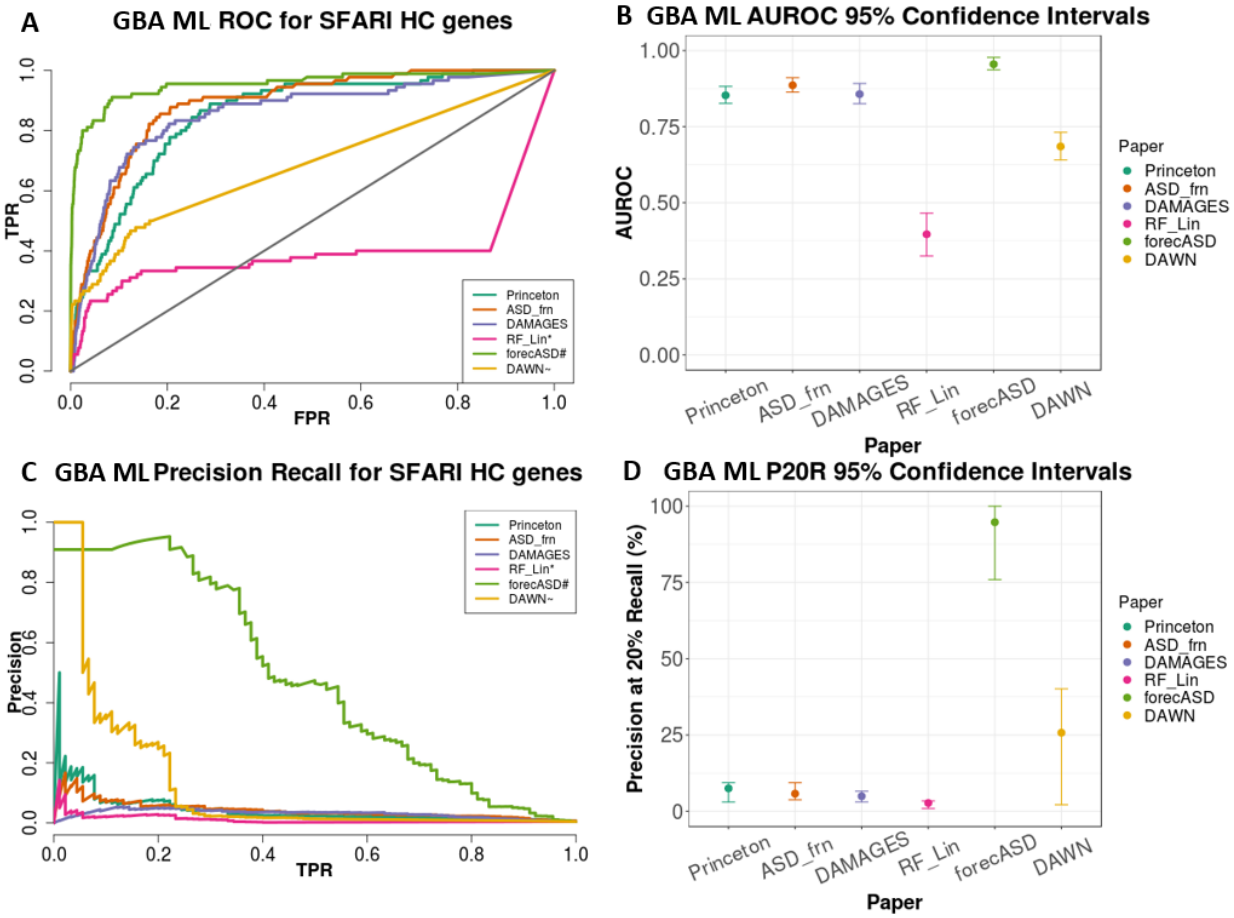

Figure 3.4: AUROC and PR statistics for GBA ML studies performance on SFARI-HC ASD genes. Systems-based GBA ML studies have high AUROC (A) and but low precision at 20% recall (C) in training scenario with SFARI high confidence genes (n=90). RF_Lin* does not provide estimates for training genes; DAWN~ does not provide estimates for all genes in genome; forecASD# has ties at the top of their rankings. 95% confidence intervals were created from 2500 stratified bootstrap samples (B,D). TPR, True Positive Rate; FPR, False Positive Rate; AUROC, Area Under the Receiver Operator Curve; P20R, Precision at 20% Recall.

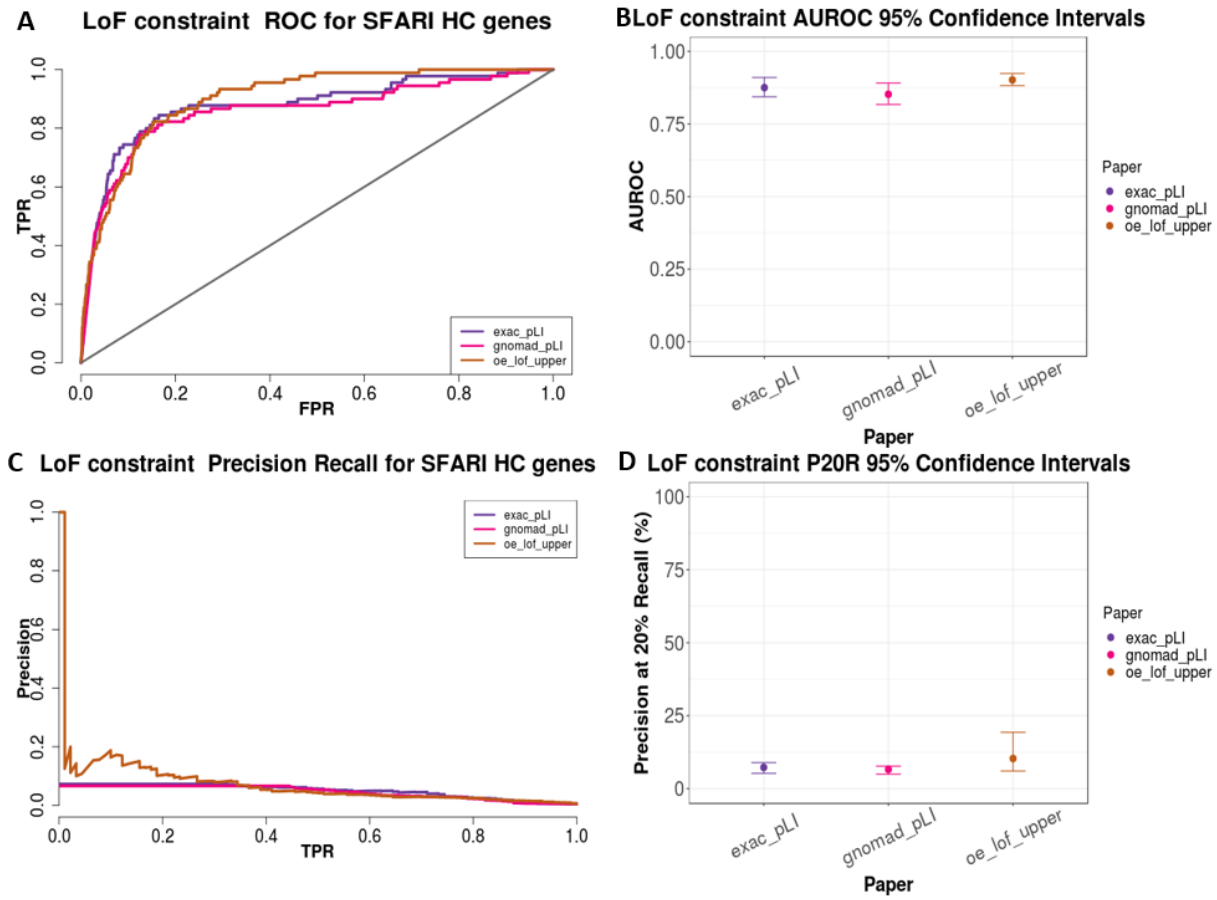

Figure 3.5: AUROC and PR statistics for LoF constraint scores performance on SFARI-HC genes. Generic LoF constraint measures have comparable performance to systems-based GBA methods with high AUROC (A) but low precision at 20% recall (C) in training scenario with SFARI high confidence genes (n=90). 95% confidence intervals were created from 2500 stratified bootstrap samples (B,D). TPR, True Positive Rate; FPR, False Positive Rate; AUC, Area Under the Receiver Operator Curve; P20R, Precision at 20% Recall.

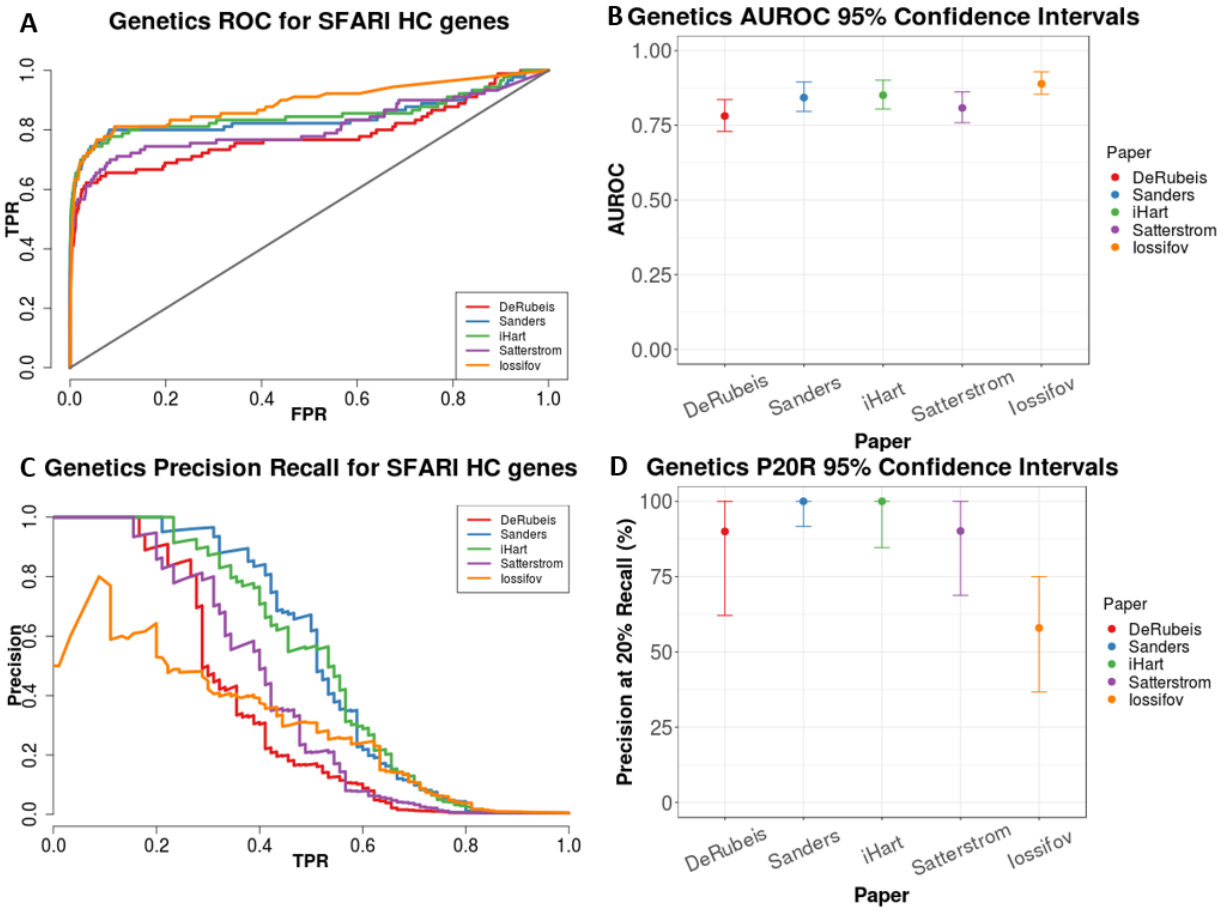

Figure 3.6: AUROC and PR statistics for genetics-based prioritization studies performance on SFARI-HC ASD genes. Genetics methods preform have high AUROC (A) and precision at 20% recall (C) in control experiment with SFARI high confidence genes (n=90). 95% confidence intervals were created from 2500 stratified bootstrap samples (B,D). TPR, True Positive Rate; FPR, False Positive Rate; AUC, Area Under the Receiver Operator Curve; P20R, Precision at 20% Recall.

Table 3.4: AUROC, Precision at 20% Recall and Precision at 43% recall performance statistics on SFARI-HC ASD genes (* denote ties at recall of 20 or 43% of gene set)

Paper	AUROC	AUROC_CI	P20R	PR20_CI	P43R	PR43_CI
Princeton	0.85	0.83, 0.88	7.52*	3.04, 9.38	2.44	1.87, 3.39
ASD_frn	0.89	0.86, 0.91	5.77	3.78, 9.38	3.78	2.64, 5.48
DAMAGES	0.86	0.83, 0.89	4.93	3.08, 6.58	3.61	2.86, 4.48
RF_Lin	0.40	0.32, 0.47	2.71	0.9, 3.41	0.26*	0.2, 0.55
forecASD	0.96	0.94, 0.98	94.74	75.95, 100	47.27	38.96, 77.99
DAWN	0.68	0.64, 0.73	25.74	2.15, 40.11	1.77	1.14, 2.47
exac_pLI	0.88	0.84, 0.91	7.29*	5.26, 8.93	6.11	4.49, 7.94
gnomad_pLI	0.85	0.82, 0.89	6.62*	4.97, 7.75	6.62*	4.31, 7.75
oe_lof_upper	0.90	0.88, 0.92	10.34	6.08, 19.3	4.85*	3.42, 8.23
DeRubeis	0.78	0.73, 0.84	90.00	62.06, 100	20.27	13.4, 43.11
Sanders	0.84	0.8, 0.9	100.00	91.67, 100	71.63	49.64, 92.59
iHart	0.85	0.8, 0.9	100.00	84.62, 100	62.91	47.92, 85.71
Satterstrom	0.81	0.76, 0.86	90.15	68.75, 100	35.29	17.98, 65.82
Iossifov	0.89	0.85, 0.93	57.97	36.66, 75	34.37	24.62, 46.71

3.3.3: Lack of agreement between systems and genetics-based methods

As previously discussed, GBA postulates that genes with shared associations are more likely to have shared functions or be involved in the same diseases. However, GBA often ascribes new functions to genes that are well characterized because they are highly studied, and have a high number of associated annotations, rather than learning additional information (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). In other words, based on previous work, we expect that the GBA methods would tend to rank generically "disease-related" genes highly because they are well-studied. And because this ranking is not ASD-specific and biased towards well-studied genes, they cannot readily identify novel and specific relationships. On the other hand, methods that do not recapitulate these generic rankings may perform badly because the main source of apparent performance of GBA methods is their ability to prioritize well-studied genes ("multifunctionality bias" as per Gillis and Pavlidis).

To examine these questions, I compared the genetic association and GBA ML scores to generic network features and generic gene annotations (Figure 3.7). My results show that some of the GBA ML studies are indeed strongly biased. For example, the genetics-GBA study, forecASD, has moderate agreement with physical node degree (R_S =0.26) and number of publications (R_S=0.25), as do DAMAGES and RF_Lin. In the work of Gillis and Pavlidis, correlations of this magnitude were sufficient to explain much predictive performance. Yet, Princeton and ASD_frn did not appear to show bias (i.e. $R_{S:ASD\_frn,pnd}$=0.12, $R_{S:ASD\_frn,numPubs}$=-0.09). In comparison, the TADA analyses show no agreement with these generic features (i.e. $R_{S:iHart,pnd}$=-0.04, $R_{S:iHart,numPubs}$=-0.05). These findings offer an explanation for the poor performance of the systems-based GBA ML studies. Studies which are not biased towards well studied genes may be performing poorly because there is no bias to drive GBA performance

(Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). On the other hand, studies which are biased towards well studied genes may be performing poorly because GBA is assigning new functions to highly connected genes in the network, and not learning ASD-specific information (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012).

High agreement between generic measures of constraint and systems-based GBA ML studies further suggest that their predictions are generic, and not overly specific to ASD (Figure 3.7). Furthermore, the lack of agreement between Satterstrom and the other TADA analyses further highlights the non-equivalence between the genetic association studies and the need for TADA model validation (Figure 3.7). The genetic association studies showing agreement with constraint scores are Satterstrom and Iossifov. Satterstrom directly incorporates pLI into its TADA model, however, it displays a negative correlation (R_S=-0.32). Due to the Bayesian-nature of TADA collapsing multiple pieces of information to derive the per-gene association scores, the direct effects of pLI on the score are likely complex, and non-linear, meaning that a Spearman correlation may not adequately capture the relationship (Satterstrom et al., 2019). The Iossifov score is the most similar to pLI in its construction: both scores attempt to quantity the deviation of the observed number of LoF variants from an expectation of LoF variation derived from complex models incorporating rates synonymous variation, among many other factors (Iossifov et al., 2015; Karczewski et al., 2019; Lek et al., 2016). The Iossifov score is ASD-specific because they incorporate an estimate of the number of causal ASD genes, and measure the deviation of LoF variants observed in ASD probands from their calculated expectation whereas the LoF constraint scores were developed without any disease specificity (Iossifov et al., 2015). Lastly, moderate agreement between the SFARI gene score and generic measures of constraint and generic network features parallel findings from Chapter 2, further demonstrating

that high-confidence ASD genes have a relationship with constraint scores, and that they are likely well studied genes (Figure 2.2, 2.3, Table 2.9, 2.10; Figure 3.7).

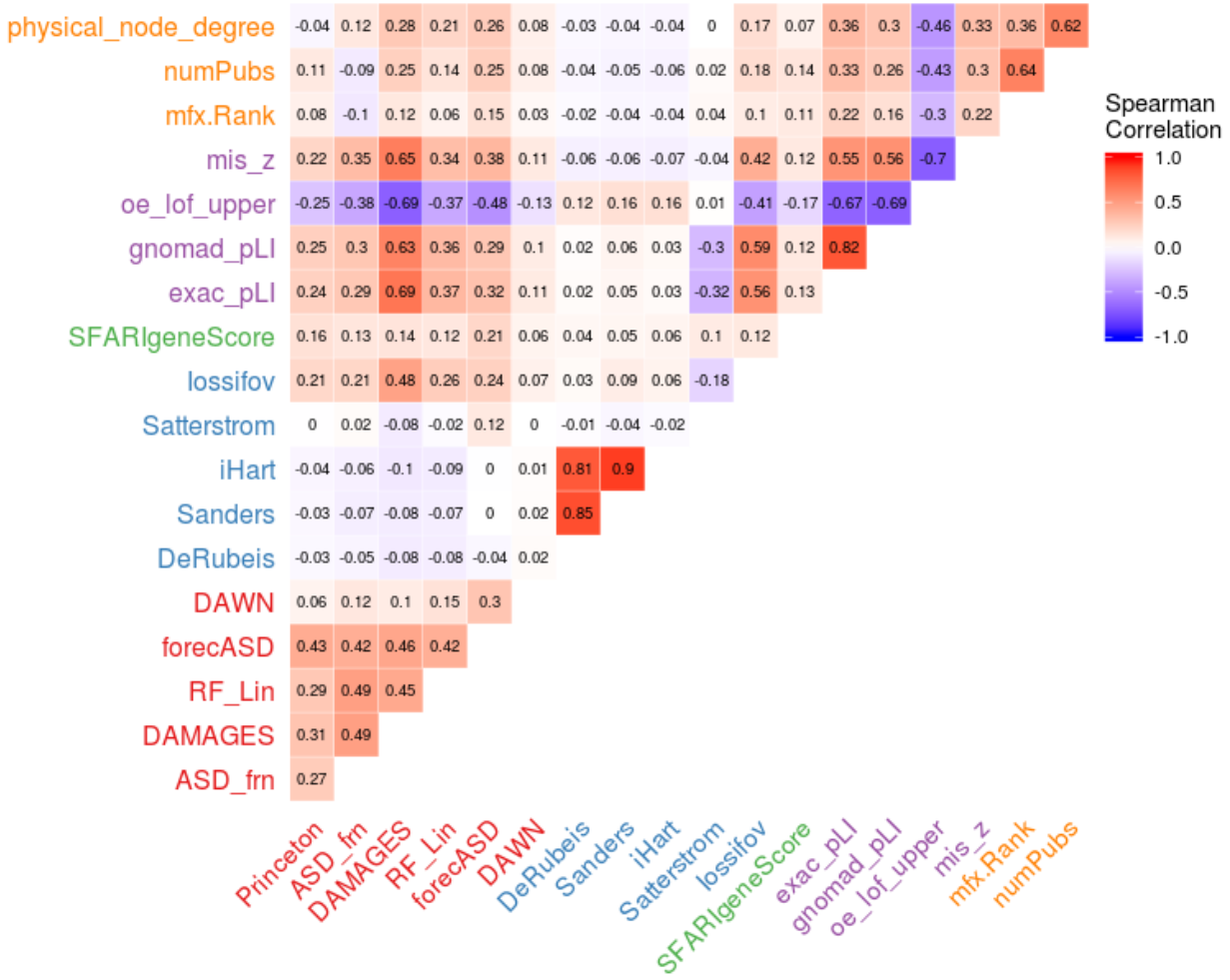

Figure 3.7: Spearman correlation heatmap for ASD prioritization and generic gene scores. Notable patterns include lack of correlation between genetic association methods, ML methods and other network features such as node degree and publication number; increased correlation between select ML methods and other network features; negative correlation between Satterstrom score and pLI despite its incorporation in the statistical framework.

3.3.4: Overlap in the subset of genes identified as likely ASD risk genes by each study

I next examined whether the low correlation among prioritization methods might still reflect commonality among the top-ranked genes (Table 3.5). For example, while forecASD and Princeton share 831 genes in their top rankings, forecASD is able to recover 82/90 SFARI high confidence genes from a potential 1803 compared to the 52/90 recovered from a potential 2467 by Princeton (Table 3.5, red highlights). Likewise, Princeton and ExAC pLI share 1045 genes in their top rankings, but ExAC pLI captures 75/90 from a potential 3220 (Table 3.5, red highlights). This again shows us that the systems-based ML studies are not performing as well as those with ASD-specific genetics information, and that they are providing little ASD-specificity above that provided by the generic measures of constraint. Lastly, we found that multiple genes identified in previous TADA analyses are no longer statistically significantly associated with ASD in Satterstrom (Table 3.6), and that the TADA analyses only share 18 genes in their top findings (Table 3.7). The differences in overlap of top findings between the TADA analyses further suggests that the differences between the TADA models need to be investigated more closely.

Table 3.5: Overlap of top ranked ASD genes. Counts highlighted in red are discussed in text.

	Princeton	ASD_frn	DAMAGES	RF_Lin	forecASD	DAWN	DeRubeis	Sanders	iHart	Satterstrom	Iossifov	exac_pLI	gnomad_pLI	oe_lof	SFARIHC
Princeton	**2467**	1014	70	842	831	38	16	28	34	55	108	1045	1026	997	52
ASD_frn		**2111**	74	985	842	42	20	35	36	62	121	1093	1023	1031	60
DAMAGES			**117**	89	90	9	4	7	7	12	25	116	115	116	12
RF_Lin				**2089**	854	30	6	5	8	43	129	1436	1335	1378	27
forecASD					**1803**	63	33	65	65	89	187	1109	1044	1052	82
DAWN						**127**	11	17	16	19	27	50	53	55	20
DeRubeis							**33**	26	23	21	24	26	24	25	25
Sanders								**65**	52	39	45	46	44	44	43
iHart									**69**	36	40	45	41	42	41
Satterstrom										**102**	53	89	81	83	38
Iossifov											**239**	204	198	203	56
exac_pLI												**3220**	2475	2477	75
gnomad_pLI													**3046**	2838	71
oe_lof														**2957**	72
SFARIHC															**90**

Table 3.6: SFARI-HC genes identified in previous TADA analyses no longer found to be significantly associated with ASD in Satterstrom. * denotes genes that are borderline significant in TADA-based studies (0.1 < FDR < 0.3).

Study	nGenes	SFARI-HC genes
DeRubeis	4	CUL3, KATNAL2, NAA15*, RELN*
Sanders	13	CUL3, ERBIN, ILF2, INTS6, KAT2B, KATNAL2, NAA15*, NCKAP1*, NLGN3, RANBP17, TNRC6B, TRIO*, WDFY3*
iHart	11	CUL3, DDX3X, ERBIN, GRIA1, ILF2, INTS6, KATNAL2, NCKAP1*, RANBP17, TNRC6B, WDFY3*

Table 3.7: Shared SFARI high confidence genes across the TADA analyses. SFARI gene score shown in brackets

Studies	nGenes	SFARI-HC genes
TADA	18	ADNP(1S), ANK2(1), ARID1B(1S), ASH1L(1), BCL11A(2S), CHD8(1S), DYRK1A(1S), GABRB3(2), GRIN2B(1), KMT2C(2S), KMT5B(1), MYT1L(1), POGZ(1S), PTEN(1S), SCN2A(1), SETD5(1S), SYNGAP1(1S), TBR1(1)

3.3.5: Feature importance in the forecASD algorithm

While forecASD had significantly better performance with SFARI and novel high-confidence ASD risk genes compared to other systems-based GBA ML studies, it is performing with low precision (P20R=5.01-11.11%). If a GBA ML method is to be considered successful, it must be able to generalize to new data, and highly true positives. However, our findings do suggest that forecASD may be able to extract additional predictive information from the non-genetic features it uses. To understand this in more detail, I examined the performance of forecASD using training feature sets made up of different combinations of the original features used in the model (Table 3.8).

Table 3.8: Features included in the feature-modified forecASD classifiers

Model \ Feature	BrainSpan Score	STRING Score	Other classifiers	De Rubies	Sanders
forecASD / redo	✓	✓	✓	✓	✓
noClassifiers	✓	✓		✓	✓
noClassifiersPPI	✓			✓	✓
noClassifiersPPIBS				✓	✓
PPIOnly		✓			
BrainSpanOnly	✓				

Other classifiers include DAWN(L. Liu et al., 2014), Princeton(Krishnan et al., 2016) and DAMAGES(Zhang & Shen, 2017). Features from the DAWN(L. Liu et al., 2014) classifier include the list of risk ASD genes (rASD), the network score produced by the screening stage, and the minimum FDR reported from the cleaning stage. Features from the DAMAGES(Zhang & Shen, 2017)classifier include both the D score and Ensemble score. There are 4 FDR values from the Sanders(Sanders et al., 2015) publication: tada_asc+ssc+del (most thorough with exome data and small *de novo* dels), tada_asc+ssc (both sources of exome data), tada_asc (ASC exome only), and tada_ssc (SSC exome only). The per-gene Bayes Factor from the De Rubeis(De Rubeis et al., 2014) study was also included (TADA_BF).

There are discrepancies when comparing the feature importance by decrease in Gini purity reported in the original forecASD preprint to what we found when rerunning the model. For example, the original forecASD paper reports the STRING score as the most informative feature, whereas we found that the most comprehensive score from Sanders was the most informative (Brueggeman et al., 2018) (Figure 3.8). When I reran the forecASD code, I found that scores outputted from the STRING, BrainSpan and ensemble random forests were the same to those reported by the original publication (i.e. Pearson $R_{redo,forecASD}=1$). Subsequently during my investigation of forecASD, I found that my rerun model and the original model had the same performance measures, and top gene predictions (Table 3.9, Table 3.10, Figure 3.11). The exact reason for the discrepancies are unknown because they did not provide code for their plots, however, we think that the discrepancies are from a mistake during plotting rather than an inability to recapitulate the forecASD model.

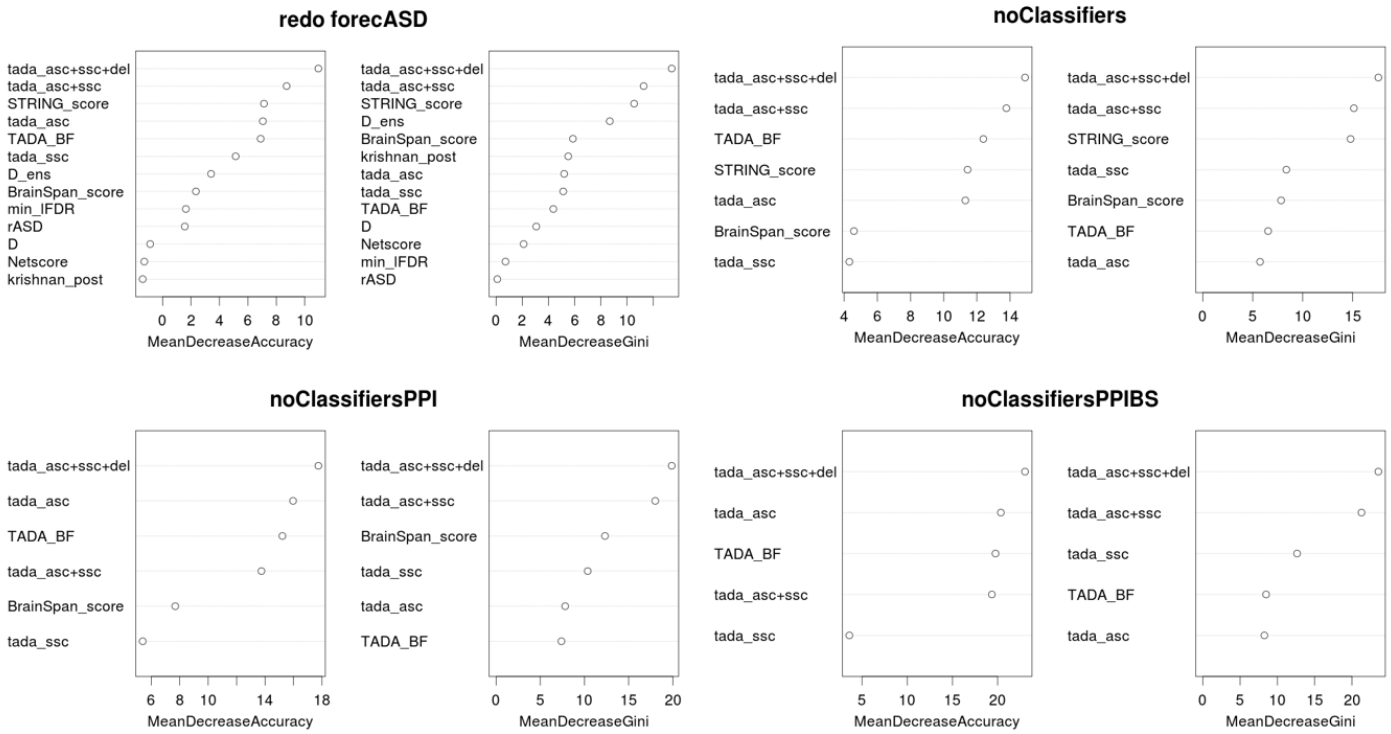

Figure 3.8: Feature importance of the adapted forecASD models. In each rerun model, the most comprehensive score from the Sanders, tada_asc+ssc+del, was ranked as the most important feature for discerning ASD from non ASD training genes, followed by other TADA-based statistics.

We evaluated the forecASD models on the novel-HC and SFARI-HC gene sets. We found that the forecASD models, except for the PPIonly model, did not have significantly different performance on novel-HC genes (Figure 3.9, Table 3.9). The overlapping 95% confidence intervals for precision at 20% recall for noClassPPIBS and BrainSpanOnly models would suggest that the gene expression data provides, at best, marginal additional information for predicating novel ASD genes (Figure 3.9D, Table 3.9). When investigating the SFARI-HC gene sets, we found that the models incorporating genetics information significantly outperformed the PPIOnly and BrainSpanOnly models (Figure 3.10, Table 3.10). Overall, these results show that forecASD is driven by ASD genetics data, which supports our previous finding that the STRING score was not the most informative feature in forecASD (Figure 3.8).

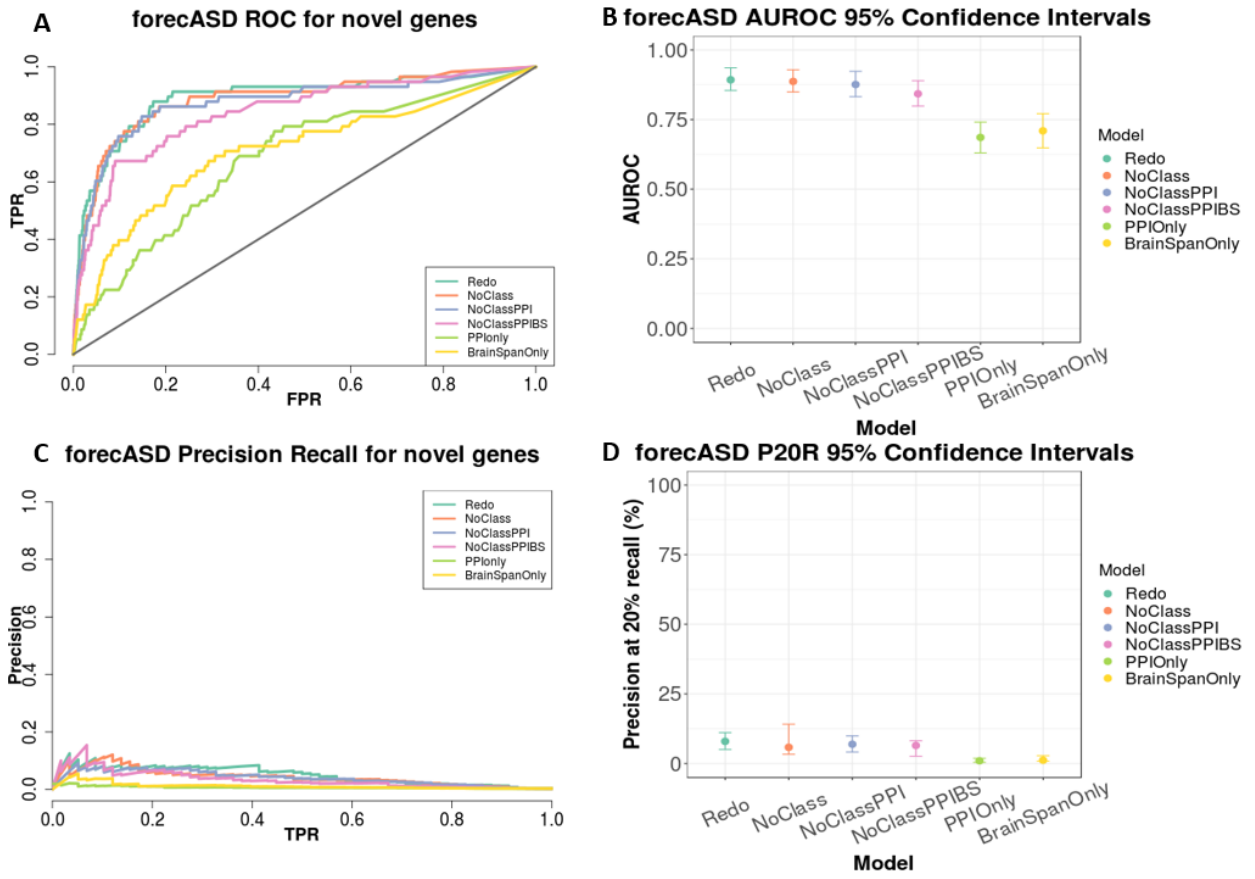

Figure 3.9: AUROC and PR statistics for adapted forecASD models performance on novel-HC ASD genes. Only the PPIonly model shows significantly different performance in testing scenario with novel ASD genes. 95% confidence intervals were created from 2500 stratified bootstrap samples. TPR, True Positive Rate; FPR, False Positive Rate; AUROC, Area Under the Receiver Operator Curve; P20R, Precision at 20% Recall.

Table 3.9: forecASD adaptions AUROC, Precision at 20% Recall and Precision at 43% recall performance statistics on novel-HC ASD genes (* denote ties at recall of 20 or 43% of gene set)

Method	AUROC	AUROC_CI	P20R	PR20_CI	P43R	PR43_CI
redo	0.89	0.85, 0.94	7.97	5.04, 11.11	6.14	3.82, 9.9
noClass	0.89	0.85, 0.93	5.83	3.33, 14.12	4.70	2.64, 6.42
noClassPPI	0.88	0.83, 0.92	6.96	4.16, 9.94	4.43	2.81, 6.66
noClassPPIBS	0.84	0.8, 0.89	6.49	2.62, 8.2	3.01	1.7, 4.88
PPIonly	0.69	0.63, 0.74	1.00	0.47, 1.86	0.59	0.44, 0.94
BrainSpanOnly	0.71	0.65, 0.77	1.21	0.89, 2.84	1.02	0.64, 1.74

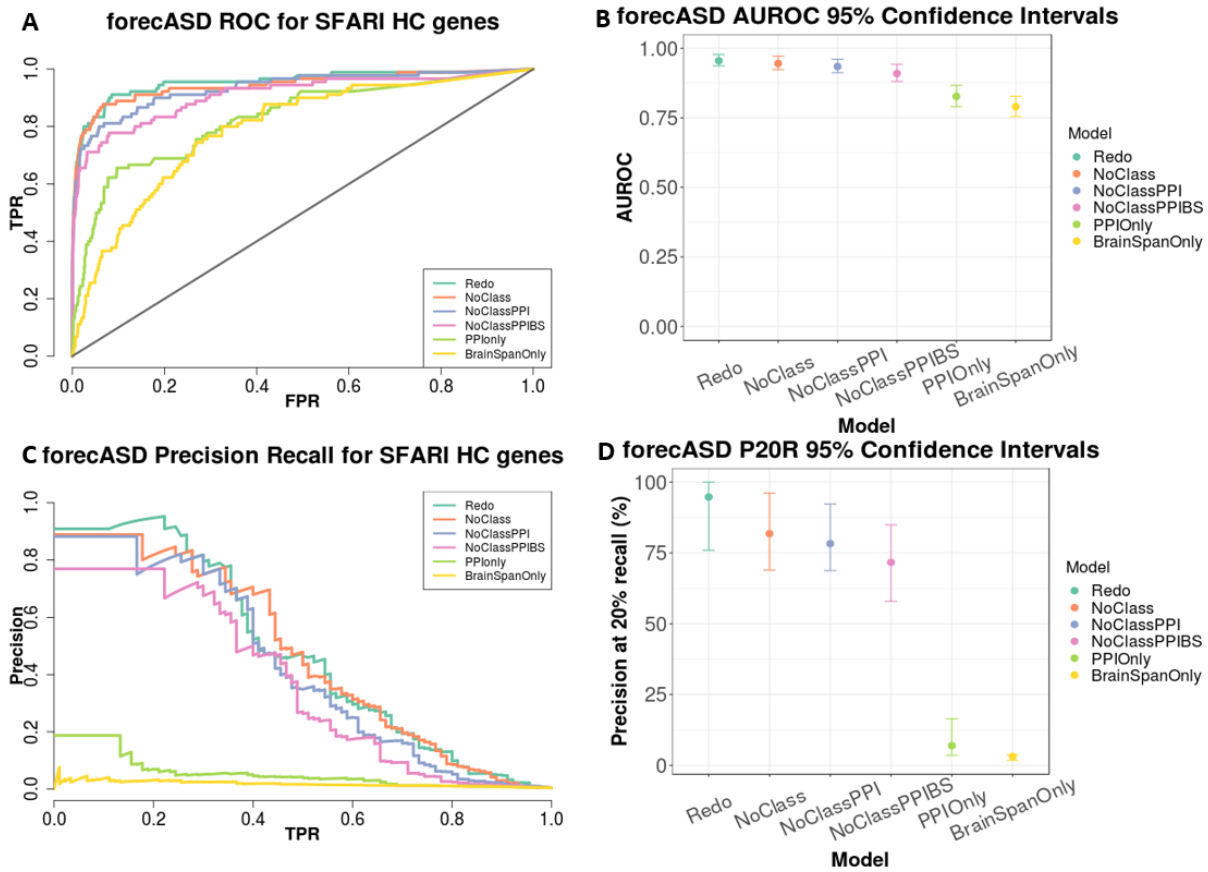

Figure 3.10AUROC and PR statistics for adapted forecASD models performance on SFARI-HC ASD genes. PPIonly is the only adapted model to show significantly different performance on with SFARI high confidence genes. 95% confidence intervals were created from 2500 stratified bootstrap samples. TPR, True Positive Rate; FPR, False Positive Rate; AUROC, Area Under the Receiver Operator Curve; P20R, Precision at 20% Recall.

Table 3.10: forecASD adaptations AUROC, Precision at 20% Recall and Precision at 43% recall performance statistics on SFARI-HC ASD genes (* denote ties at recall of 20 or 43% of gene set)

Method	AUROC	AUROC_CI	P20R	PR20_CI	P43R	PR43_CI
Redo	0.96	0.94, 0.98	94.74	75.95, 100	47.27	38.96, 77.99
noClass	0.95	0.92, 0.97	81.82	68.99, 96.15	64.99	37.39, 79.35
noClassPPI	0.93	0.91, 0.96	78.26	68.75, 92.31	47.27	29.8, 76.67
noClassPPIBS	0.91	0.88, 0.94	71.67*	57.89, 84.93	46.99*	22.37, 64.8
PPIonly	0.83	0.79, 0.87	6.95	3.56, 16.44	4.36	3.23, 6.61
BrainSpanOnly	0.79	0.76, 0.83	3.13	1.77, 3.71	1.94	1.34, 2.95

We found that the forecASD models show moderate to high agreement with one another (i.e. $R_{S:forecASD,noClass}$=0.77, $r^2_{forecASD,PPIonly}$=0.37) (Figure 3.11). Notably, we found that the model incorporating only genetics information (noClassPPIBS) showed lower agreement with measures of constraint against LoF variation (i.e. $R_{S:noClassPPIBS,exac\_pLI}$=-0.01, $R_{S:noClass,exac\_pLI}$=0.13), and appeared to have less bias towards well studied genes (i.e. $r^2_{noClassPPIBS,numPubs}$=0.06, $R_{S:noClass,numPubs}$=0.12) compared to other forecASD models (Figure 3.11). Furthermore, we found that while the BrainSpanOnly model and PPIOnly model had similar agreement with physical node degree, the BrainSpanOnly model showed lower agreement with number of publications and multifunctionality (i.e. $R_{S:BranSpanOnly,numPubs}$=0.08, $R_{S:PPIOnly,numPubs}$=0.21), which may suggest that gene expression data used had less literature bias compared to data from STRING and other ASD GBA ML classifiers (Figure 3.11).

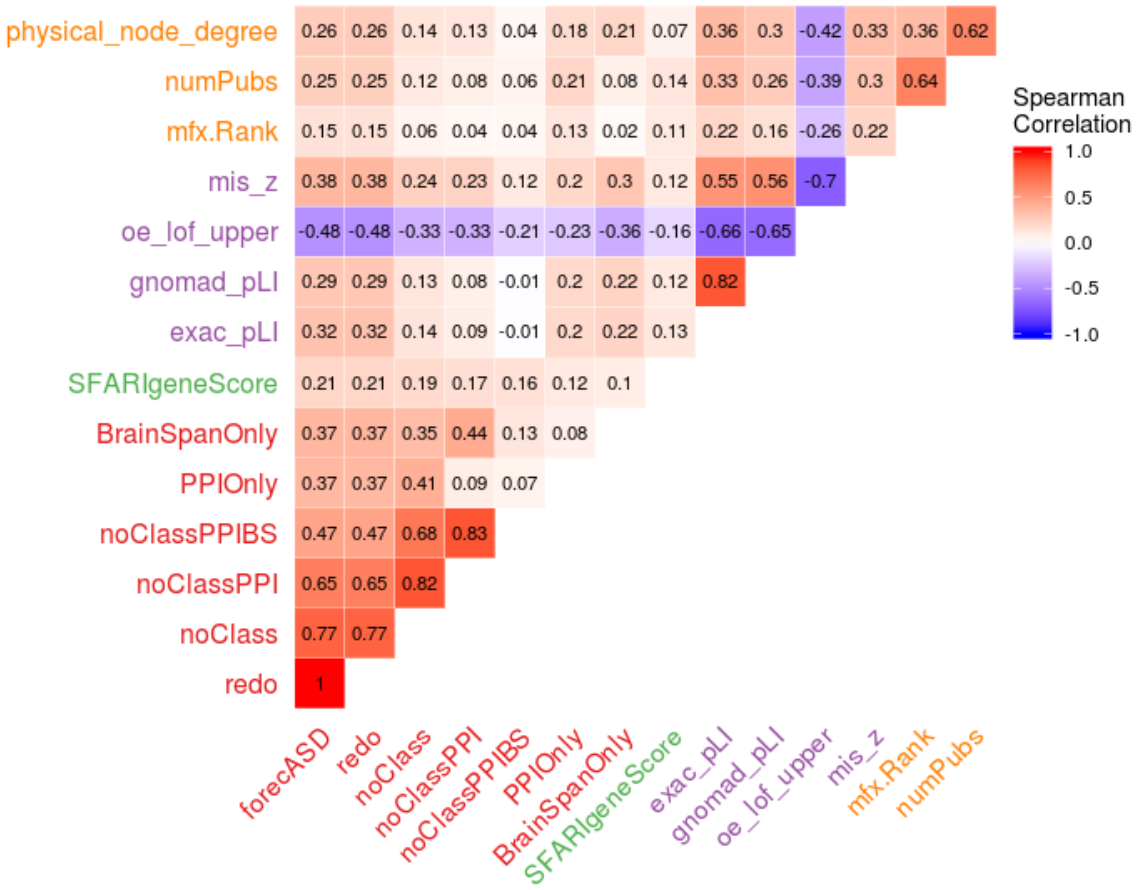

Figure 3.11: Spearman correlation heatmap for forecASD adaptations. The adapted forecASD models show decline in correlation with each other and with generic disease predictors and generic network features as more features are removed from the models, except for the model only using the STRING protein-protein interaction network.

Overall, we found that the forecASD models had similar performance on novel-HC and SFARI-HC gene sets, and that their performance is driven by their use of ASD-genetics data. Further, we found that the gene expression data utilized appeared to have less literature bias compared to the STRING data alone. Exploring the differences in the gene level predictions of the studies and forecASD models under investigation may provide additional insight into the utility of non-genetics features for gene prioritization.

3.3.6: Top overlap between adapted forecASD models and genetic association studies

Looking at the gene-level differences, we found that no study or model adaptation under investigation was able to highly prioritize every SFARI or novel high-confidence ASD risk gene (Table 3.11). Combining data from Sanders and DeRubeis (noClassPPIBS), and supplementing with gene expression data (noClassPPI) and other data (noClass) resulted in the loss of fewer novel-HC genes compared to genetic association studies alone (Table 3.11) Furthermore, we found that the PPIOnly forecASD model was able to capture more SFARI-HC genes compared to the BrainSpanOnly (BSOnly) model (56/90 vs. 34/90, respectively), however, the reverse was true for the novel-HC ASD genes (13/58 vs. 22/58, respectively) (Table 3.11, red highlights). This result could potentially reflect the effects of multifunctionality: GBA is often biased towards well studied genes, and as genes are associated with diseases, they become more studied, and accumulate more annotations. Therefore, the PPIOnly model could have recovered more SFARI-HC genes because they are well studied, and have more functional annotations compared to the novel-HC genes. These results demonstrate that while genetics data is the most informative feature for predicting ASD genes, genetic association analyses, like TADA, cannot capture every high-confidence ASD risk gene, and that gene expression data may be useful providing less biased contextual information potentially useful for prioritization.

Table 3.11: Top overlap of forecASD adaptations and genetic association studies. Counts highlighted in red discussed in text. BSOnly, BrainSpanOnly model.

	forecASD	noClass	noClassPPI	noClassPPIBS	PPIOnly	BSOnly	DeRubeis	Sanders	iHart	Satterstrom	Iossifov	exac_pLI	gnomad_pLI	oe_lof	SFARI-HC	Novel-HC
forecASD	**1803**	1287	1008	858	689	510	33	65	65	89	187	1109	1044	1052	82	41
noClass		**1787**	1197	1075	731	441	33	65	67	85	179	783	758	757	79	42
noClassPPI			**1785**	1300	311	521	33	64	66	84	174	708	670	682	73	42
noClassPPIBS				**1785**	280	276	32	63	65	77	162	582	531	545	70	38
PPIOnly					**1784**	307	20	38	38	53	99	683	683	660	56	13
BSOnly						**1799**	12	21	25	43	100	725	716	713	34	22
DeRubeis							**33**	26	23	21	24	26	24	25	25	0
Sanders								**65**	52	39	45	46	44	44	43	0
iHart									**69**	36	40	45	41	42	41	14
Satterstrom										**102**	53	89	81	83	38	32
Iossifov											**239**	204	198	203	56	6
exac_pLI												**3220**	2475	2477	75	35
gnomad_pLI													**3046**	2838	71	33
oe_lof														**2957**	72	35
SFARI-HC															**90**	0
Novel-HC																**58**

3.3.5: Potential ASD risk genes for further study

Throughout this chapter, we have shown that GBA ML studies without ASD-genetics data have limited utility for ASD gene prioritization because they provide little advantage compared to generic measures of constraint against LoF variation. We have further found that the driving force behind the best performing GBA ML study, forecASD, was in fact ASD-genetics data. As such, when looking for ASD genes to study further, an obvious place to start looking would be within the subset of genes found to be significantly associated with ASD by genetic association which are not yet considered to be SFARI-HC genes: there are 107 genes in DeRubeis, Sanders, iHart, Spark and Satterstrom studies which are not yet considered to be SFARI high-confidence ASD genes. Given that we confirmed LoF and missense constraint scores are useful for identifying ASD risk genes, we filtered the set of 107 genes by the o/e LoF metric, and by the missense z-score. We chose o/e LoF because it is now recommended over the pLI because it is more continuous across the spectrum of selection, and the missense z-score because we are part of a collaboration looking at the role of *de novo* missense variants in ASD. There are 32 genes with o/e LoF scores less than 0.35 and missense z-scores over 3. Of the 32 genes, 8 had over 4 *de novo* missense variants according to enumeration conducted in Chapter 2. All of the 8 genes are currently in SFARI as category 3, 4, 3S, 4S, or 5 genes, and could be considered higher-value targets for further study (Table 3.12).

Table 3.12: Possible ASD genes for further study. GA rare dn mis: rare *de novo* missense variants reported in ASD probands from genetic association studies from VariCarta or the 4 recent TADA analyses; S rare dn mis: additional rare *de novo* missense variants reported in SFARI. Note that the TADA scores are in their original format with FDR < 0.1 being significant (genes in bold).

Gene symbol	SFARI score	De Rubeis	Sanders	iHart	Satterstrom	ExAC pLI	gnomAD pLI	o/e LoF	mis_z	GA rare dn ms	S rare dn mis
CACNA1E	**3**	**0.87**	**0.94**	**0.95**	**0.03**	**1.00**	**1.00**	**0.12**	**5.81**	**5**	**13**
CREBBP	**5**	**0.94**	**0.72**	**0.81**	**0.02**	**1.00**	**1.00**	**0.07**	**3.90**	**7**	**23**
DYNC1H1	**3**	**0.51**	**0.57**	**0.75**	**0.003**	**1.00**	**1.00**	**0.08**	**10.97**	**7**	**6**
MYO5A	**3**	**0.84**	**0.65**	**0.07**	**0.27**	**0.99**	**0.94**	**0.30**	**3.10**	**8**	**0**
PPP2R5D	4S	0.92	0.46	0.46	0.01	1.00	1.00	0.18	3.65	3	6
SCN1A	3S	0.33	0.13	0.17	0.05	1.00	1.00	0.07	5.22	6	39
STXBP1	3S	0.77	0.75	0.82	0.01	1.00	1.00	0.09	4.26	5	30
TLK2	4S	0.77	0.15	0.22	0.00	1.00	1.00	0.11	4.49	3	9

3.4: Discussion

3.4.1: Limited utility of systems-based GBA ML studies for ASD risk gene prioritization

Evaluating and comparing the performance and biases of GBA ML studies, genetic association studies and generic measures of constraint against LoF variation provided us with insight into the reliability and utility of GBA ML studies for ASD gene prioritization. Our investigation has shown that GBA ML methods which do not incorporate genetic information have limited utility because they provide no useful information above that provided by genetic association data and generic measures of disease gene likelihood. Further, investigation of forecASD, the genetics-GBA ML method with the best relative performance, demonstrated that genetics data drives its performance, and that gene expression data may provide less-biased contextual information for prioritization of ASD risk genes for further study.

3.4.2: Non-equivalence of genetic association studies

Genetic association studies identify *bona fide* ASD genes, however, not all studies agree on which genes are significantly associated with ASD. We found that the different TADA analyses share only 18 SFARI-HC genes in their significant findings (Table 3.7), and that many of the SFARI-HC genes identified by previously by TADA analyses as significantly associated with ASD are no longer significant in Satterstrom, including three SFARI category 1 non-syndromic genes CUL3, KATNAL2, and RELN (Table 3.6). While this is slightly disconcerting, it is important to remember that the TADA analyses find genes statistically associated with ASD under their specific model of *de novo* and inherited variation used on their specific sample cohort. Therefore, analyses only show that "missing genes" are missing because they not statistically significantly associated with ASD under their statistical model of variation.

Of the TADA analyses, the Satterstrom analysis differs the most from the others. The Satterstrom study changes the TADA method in a substantial way by adding the ExAC pLI score as a continuous metric for assessing relative risk of LoFs, and a tiered "missense badness, PolyPhen-2, constraint" (MPC) score for assessing relative risk of missense variants (Satterstrom et al., 2019). We are anticipating an update from Satterstrom et al. which further changes the TADA model by incorporating the observed/expected LoF metric rather than the ExAC pLI to provide a more suitable metric to delineate ASD risk genes across a range of selective pressures (Karczewski et al., 2019; Lek et al., 2016; Satterstrom et al., 2019). This change, ideally, will reflect more closely the complex genetic architecture of ASD (Karczewski et al., 2019; Lek et al., 2016; Satterstrom et al., 2019). Other expected changes include incorporating CNV data and adding more samples from the Spark consortium (Satterstrom, F.K. et al., 2019).

Without direct comparison of the different TADA models on the same samples, I can only speculate as to the direct cause of the differences between Satterstrom and the other TADA analyses. It is possible that their lack of agreement could be explained by the increased number of samples in Satterstrom alone. However, the iHart study uses the same TADA model as the Sanders study, but with an increased number of samples from multiplex families, and these two studies show the highest agreement (r^2=0.90) and overlap of significant genes (52/80) (Ruzzo et al., 2019). As such, it is more likely that the incorporation of pLI and MPC in the Satterstrom TADA method have a larger impact on the observed differences.

While the TADA-based genetic association studies can identify *bona fide* ASD risk genes, the lack of agreement between scores with different underlying TADA models and their specific gene-level predictions suggests that gene discovery is sensitive to how assumptions made about the genetic architecture of ASD shape the calculation of model parameters. While

updates in the TADA model may indeed reflect the genetic architecture of ASD, validation of model adaptations needs to occur in order to establish reliability. Establishing a "gold standard" genetic association model, and a "gold standard" ASD gene set, will require more comprehensive evaluation of current TADA statistical frameworks.

3.4.3: Systems-based GBA ML studies are comparable to generic measures of LoF constraint

Claimed use cases of the GBA ML studies include prediction and/or prioritization of ASD risk genes, framing WES/WGS results for further exploration in resequencing or mechanistic studies, and/or uncovering new and delineating possible pathways implicated in ASD etiology (Brueggeman et al., 2018; Duda et al., 2018; Krishnan et al., 2016; Lin et al., 2018; L. Liu et al., 2014; Zhang & Shen, 2017). While we assessed the first two example use cases, the third is more difficult to evaluate. The ML studies provide a subset of genes which are the most likely ASD gene candidates based on their classifiers. The utility of performing any sort of pathway enrichment analysis on these subsets of genes to delineate ASD associated pathways is limited because the true association status of the gene to ASD is unknown, meaning that the pathways identified may not be specific to ASD. Overall, for GBA ML study to be considered successful in identifying novel ASD genes, it should highly prioritize known ASD genes and provide additional, specific and unbiased predictions above that which could be obtained from generic measures of constraint. We have shown that the systems-based ML studies do not do this.

As discussed previously, GBA works on the principle that genes with "shared associations" are more likely to be involved in the same functions or diseases. However, previous work has found that GBA is often biased towards well studied genes (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). Well studied genes are often highly connected in networks (i.e.

"hubs") because they have a high number of associated interaction partners, functions or other annotations. This "multifunctionality bias" drives GBA performance because GBA often assigns new functions to genes that are highly connected within the network rather than learning additional new information from the connection patterns (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). Therefore, we expect that methods employing GBA would tend to rank generically "disease-related" genes highly because they are well-studied. And because this ranking is not specific to any disease/function at hand, and biased towards well-studied genes, the methods cannot readily identify novel and specific relationships. Conversely, methods which are not biased towards generic rankings may perform badly because the main source of apparent performance of GBA methods is their ability to prioritize well-studied genes ("multifunctionality bias" as per Gillis and Pavlidis).

Overall, we found that the GBA ML studies without genetics data have low precision when ranking high-confidence ASD risk genes, and are comparable to generic measures of constraint against LoF variation (Table 3.3, 3.4). However, we found that two studies, Princeton and ASD_frn, are not biased towards generic rankings of the number of physical interaction partners and functions (Figure 3.7). These two studies both built complex functional interaction networks from multiple types of data, including protein-protein interaction and gene expression data. They used Gene Ontology annotations to define "gold standards" of functional relationships and Bayesian frameworks to weight and integrate their biological data (Duda et al., 2018; Greene et al., 2015; Krishnan et al., 2016; The Gene Ontology Consortium, 2019). Their poor performance could be due to their GO functional categorization not aligning well with the multiple biological data types and/or not providing any useful ASD-specific information. However, it is much more likely that these studies do not perform well because there is no/little

multifunctionality bias in their networks to drive GBA performance (Gillis & Pavlidis, 2011; Pavlidis & Gillis, 2012). The other GBA ML studies showcase the effects of GBA more clearly. RF_Lin and DAMAGES were found to be more biased towards well studied genes with high numbers of interactors and functions (Figure 3.7). This would suggest that their underlying biological networks have multifunctionality bias. Therefore, their poor performance is likely due to GBA assigning new functions to highly connected genes instead of finding ASD-specific information within the network. However, further investigation into each study is required to delineate how multifunctionality bias is affecting their performance.

With regard to measures of constraint, we have confirmed that they are able to identify ASD genes, albeit with low precision, and that they agree with generic network features and annotations. Many previous studies have found ASD genes, particularly those found with high numbers of recurrent *de novo* variants, to be enriched for genes under high evolutionary constraint, and LoF constraint has previously been reported to be positively correlated with the number of physical interaction partners (De Rubeis et al., 2014; Karczewski et al., 2019; Lek et al., 2016; Ruzzo et al., 2019; Satterstrom et al., 2019). From this, we can confirm that measures of constraint against loss of function variation measure generic susceptibility to disease, and that high constraint does not automatically guarantee a particular disease status, necessitating incorporation with data specific to the disease at hand to increase precision (Cummings et al., 2019; Karczewski et al., 2019; Lek et al., 2016). Furthermore, while these measures are also correlated with numbers of interaction partners, functions and publications, they may point towards more biologically relevant information, such as the ability of a gene to influence different phenotypic traits, rather than number of connection partners based on network structure ("hubness") (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012).

The implication of this analysis is that supplementing ASD-specific information with measures of constraint may provide a more fruitful avenue forward compared to creating GBA ML methods using biased biological networks. We can see this already being done by the Satterstrom TADA analysis by their incorporation of the pLI and MPC into the method in attempts to provide more detailed information about variant classes with higher burden in ASD probands (Satterstrom et al., 2019).

3.4.4: Gene expression data may provide less biased contextual information for prioritization

We found that one genetics-GBA study, forecASD, had comparable, and sometimes superior, performance to some TADA analyses alone (Table 3.3, 3.4). We investigated forecASD more closely by removing different features from the classifier and exploring the overall performance and gene-level differences between the adapted models to determine the utility of their non-genetics features. By fitting different models made up of different feature sets, we found that the genetics data was driving the performance of forecASD, and that gene expression data provides marginal, if any, additional predictive performance (Table 3.9, 3.10). While expression data may not provide any additional predictive performance, it may provide less biased contextual information for prioritization of ASD genes because it appears to have less literature and multifunctional bias (Figure 3.11). In other words, genes found to be significantly associated with ASD by genetic association studies could be prioritized for further study by their gene expression levels within the brain. ASD, after all, is a neurodevelopmental disorder, and studies have shown that some ASD genes have preferential expression in brain tissue (De Rubeis et al., 2014; L. Liu et al., 2014; Satterstrom et al., 2019; Velmeshev et al., 2019). However, in theory, ASD genes could be acting anywhere at and time, and therefore, we could miss ASD genes by limiting the scope to specific spatiotemporal regions of the brain, as seen in the DAWN

analysis (Table 3.3, 3.4). Nonetheless, an important point of future research is to investigate the relationship between gene expression data, and other useful genic features, like constraint against LoF variation, and how they can be used in conjunction with genetic association data to prioritize genes for further study.

3.4.4: Potential ASD risk genes for further study

To be useful, the results of the prioritization studies should be actionable in that they give clinically relevant information or are able to help elucidate the genetic architecture of ASD. The genetic association studies identify genetic variants in ASD probands and have the ability to provide families with genetic diagnoses. We have shown that the systems-based GBA ML studies have little, if any, relevance for ASD gene prioritization. Given that constraint against LoF variation has been shown to be useful for identify ASD risk genes, we used constraint scores and variant profiles to identify possible genes for further study from a subset of likely candidates from genetic association studies.

We identified eight genes as potential candidates for further study based on genetic association, constraint and variant data (Table 3.12). Four of these genes, MYO5A, CACNA1E, CREBBP, and DYNC1H1 were identified in the top findings of the Satterstrom and iHart studies, and thus could be higher-value targets. The three genes identified by Satterstrom, CACNA1E, CREBBP, DYNC1H1, were not considered novel findings because they have been associated with other neurodevelopmental disorders (Satterstrom et al., 2019). However, all four have not been studied extensively in relation to ASD. MYO5A is a motor protein involved in transportation of different cargo vesicles in brain cells and melanosomes and has been found to be involved in two diseases affecting pigmentation, Griscelli disease and Elejalde syndrome, which differ mainly by their levels of immunological and neurological impairment (Anikster et

al., 2002; Masters, Kendrick-Jones, & Buss, 2017; Pastural et al., 1997; Sanal et al., 2002). CACNA1E encodes a subunit of R-type voltage-gated calcium channels thought to be involved in modulation of neuron firing patterns (Helbig et al., 2018; Heyne et al., 2018). *De novo* missense mutations in CACNA1E have been associated with developmental epileptic encephalopathies characterized by severe impairment, macrocephaly and movement impairments (Helbig et al., 2018). CREBBP is a ubiquitously expressed transcriptional coactivator thought to be involved in many different processes via interaction with different transcription factors and chromatin remodeling activity (Negri et al., 2016; Stevens, 2019). Variation in CREBBP has been associated with Rubinstein-Taybi Syndrome, characterized by mental retardation, growth deficits, microcephaly, and other physical abnormalities, and Menke-Hennekam syndrome, characterized by deficits in cognition and facial dysmorphisms (Menke et al., 2016, 2018; Negri et al., 2016; Petrif et al., 1995; Rubinstein & Taybi, 1963; Stevens, 2019). DYNC1H1 is part of the cytoplasmic dynein molecular motor complex and is involved in processes such as organelle and protein sorting and movement. DYNC1H1 variation has been associated with malformations of brain cortical development and neuropathy, including Charcot-Marie-Tooth disease and spinal muscular atrophy (Harms et al., 2012; Poirier et al., 2013; Weedon et al., 2011). The other four genes identified for possible further study are also in SFARI categories 3 and 4, but also have been identified as potential syndromic ASD genes. Given that these eight genes have been associated with other neurodevelopmental disorders and syndromes, identifying candidate for genes involved solely in sporadic ASD will likely be very challenging.

3.4.5: Establishing fair evaluation sets is difficult due to lack of data independence

Aside from the more common problems faced in bioinformatics projects, such as mapping gene symbols and identification numbers across multiple sources, a larger issue faced in

my project is lack of independence between gene sets and genetic association studies. Lack of independence means that creating validation gene sets for consistent and fair evaluation across multiple studies is difficult.

ASD genes are discovered through genetic association studies and used to build machine learning tools which aim to identify more ASD genes. TADA analyses are built up from one another by utilizing overlapping sample cohorts and similar model parameters. This means that genes which are found to be significantly associated with ASD in newer genetic association studies may have in fact been implicated at a lower level of significance in previous studies. Therefore, their gene discovery should not be thought of as independent events.

Genes found to be significantly associated with ASD in newer genetic association studies, while they may not have been "borderline significant" in previous studies, may have been previously linked to ASD. For example, the iHart study considers the significant findings MOY5A and RAPGEF4 (FDR < 0.1) to be novel high confidence genes despite being in SFARI as category 3 and 4 genes prior to significant association. This is important for the Princeton study because they included five genes considered to be novel significant findings (FDR < 0.1) in the iHart, Spark or Satterstrom studies in their training labels as genes with low evidence of association with ASD. However, even with the inclusion of these genes as training labels, Princeton does not highly prioritize other novel ASD genes well, highlighting the inability of GBA ML studies to provide ASD-specific predictions.

It is unlikely that we will be able to construct a gene set independent of bias of previous genetic association studies, and multifunctionality, that can be used as a fair validation step for the new GBA machine learning studies being produced for prioritization purposes. The implication of this is that we have likely overestimated the performance of the GBA ML studies.

Given that GBA performance is often driven by multifunctionality bias, if we are assessing GBA-based studies using a gene set that is itself biased towards these features, then it is likely that we will see higher values in performance metrics. Genes found to be associated with disease often become more and more studied, and, therefore, accumulate more and more annotations. While these annotations may be biologically relevant, GBA-based studies using heterogenous biological data will continue to struggle in providing disease-specific information, and we will continue to overestimate performance.

GBA ML studies try to avoid the data dependence problem when evaluating their methods by looking at enrichment or recovery different gene sets. For example, the studies under investigation looked at gene sets including: 1) genes found to have *de novo* likely damaging variation in independent sequencing studies, 2) genes found to have lower evidence of an association with ASD, 3) genes found to be involved in ASD-related pathways , and/or 4) genes found to be involved in other brain-related disorders (Brueggeman et al., 2018; Duda et al., 2018; Krishnan et al., 2016; Lin et al., 2018; L. Liu et al., 2014; Zhang & Shen, 2017). There are multiple issues with these types of evaluation. Firstly, calculating enrichment faces the same problems as AUROC: while it can show good recovery of genes, it does not provide information as to where in the top decile of ranks those genes lie. Furthermore, calculating modified statistics (Princeton, DAMAGES) is not helpful because they often cannot be compared directly across studies. Secondly, evaluating genes found to have variation in ASD probands but not controls in independent sequencing studies is problematic because recurrence and burden are not the only requirements for finding genic association with ASD. Further, many of the independent sequencing studies used by GBA ML studies do not perform rigorous test for statistically significant association, mainly because a “gold standard” test for rare variant disease association

has yet to be established. It is likely that generic measures of constraint against LoF variation would likely produce similar enrichment. Lastly, looking at genes with lesser evidence of association with ASD, involved in ASD associated pathways or involved in related disorders provide, at best, anecdotal evidence of ASD-specificity.

Chapter 4: Conclusions

In this thesis, I collected gene and variant level annotations of ASD risk genes in an attempt to provide useful information for our collaborators to prioritize candidate ASD risk genes for further experimental study. I also looked at how ASD genes are currently identified by genetic association and prioritized by ASD-specific and generic computational methods.

In Chapter 2, I reported variant and gene level annotations of genes associated with ASD. The genes we annotated were from SFARI and significant findings from recent TADA analyses. Variant annotations collected include transcript, gene and protein effects and identifiers, variant consequences, allele frequencies and damaging predictions. Gene level annotations collected included measures of constraint against different classes of variation, and the per-gene number of physical interaction partners, molecular functions and publications.

Enumerating variant events across heterogeneous studies is a complex task because sequencing studies, whether targeted or whole exome or genome, employ different sequencing technologies, variant calling, annotation software, variant nomenclature and tests for association. Establishing genic association with ASD is based partly on recurrence of variant events within the same gene across multiple samples. Therefore, ensuring count inflation does not occur is an important analytical step (Belmadani et al., 2019). Standardization of not only sample cohort, and variant reporting formats, but also variant calling pipelines, annotation software and association tests, will help with ease of harmonization across studies.

In Chapter 3, I detailed investigations into the utility of GBA machine learning studies for the prioritization of ASD risk genes. I collected twelve ASD gene prioritization studies, four systems-based GBA ML studies, two genetics-GBA ML studies and five genetic association studies. Using both known and novel ASD gene sets, we looked at the ability of the different

studies to recover these genes. We further compared these studies to generic measures of constraint against LoF variation from gnomAD, and other measures of generic gene annotations to establish the specificity of the studies.

Genetic association studies identify *bona fide* ASD risk genes. Tests employing sophisticated statistical frameworks to establish association, such as TADA, while popular, are still only models of the underling inherited and *de novo* variation. Assumptions made about the genetic architecture of ASD need to be validated to ensure that the underlying architecture in the population being investigated is accurate.

Systems-based machine learning studies using guilt by association for ASD gene prioritization do not perform well, and demonstrate similar performance levels to generic measures of constraint against loss of function variation. Guilt by association is often driven by multifunctionality bias in the underlying biological network, resulting in generic predictions (Gillis & Pavlidis, 2011a; Pavlidis & Gillis, 2012). As more heterogeneous biological data is collected, GBA-based methods will likely struggle even more to find disease-specific connection patterns with the networks. While measures of constraint are based on models of expected variation, their interpretation is much more clear: genes with fewer than expected variants are under higher constraint, and the estimations will only become more accurate as the number of available high-quality sequences increases.

Some of the GBA ML studies under investigation claim to provide frameworks for studying other complex brain disorders (Duda et al., 2018; Krishnan et al., 2016; L. Liu et al., 2014). However, we have demonstrated that these types of GBA ML methods have limited utility for predicting genes associated with ASD. Therefore, it is likely that they would not perform any better if their method was re-trained on other disease gene sets. Guilt by association

is a general principle applied in many different biological experiments and medical contexts. We are not suggesting that guilt by association methods should be altogether forgotten. Rather, we are suggesting that for gene function prediction, guilt by association is not the way forward to discover novel disease genes.

In my view, an important avenue for research is to further develop and validate the statistical models used to assess genetic association for rare variants. In particular, the recent changes to TADA in the Satterstrom study have a fairly drastic effect on which genes are associated with ASD (compared to Saunders et al., for example), but the method has apparently not been validated. When appropriately validated, we would gain a deeper understanding of how the spectrum of constraint and selection against certain mutation classes in genes can be used in increase precision in gene discovery.

My results also suggest that while gene expression data does not provide any additional predictive performance compared to genetic association studies, it may be useful for providing less biased contextual information for prioritization of ASD risk genes for further study. Investigating more closely at the relationship between tissue or cell-type specific expression, and constraint data may help in creating a more comprehensive ranking for ASD risk genes, including SFARI category 3 and 4 genes for collaborators.

Bibliography

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., … McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

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*(1), 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

al. ASem. (2019). DescTools: Tools for Descriptive Statistics. (Version R package version 0.99.30) [R]. Retrieved from https://cran.r-project.org/package=DescTools.

Anikster, Y., Huizing, M., Anderson, P. D., Fitzpatrick, D. L., Klar, A., Gross-Kieselstein, E., … Hurvitz, H. (2002). Evidence that Griscelli Syndrome with Neurological Involvement Is Caused by Mutations in RAB27A, Not MYO5A. *The American Journal of Human Genetics*, *71*(2), 407–414. https://doi.org/10.1086/341606

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

Baio, J. (2018). Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *MMWR. Surveillance Summaries*, *67*. https://doi.org/10.15585/mmwr.ss6706a1

Ballouz, S., Weber, M., Pavlidis, P., & Gillis, J. (2017). EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics*, *33*(4), 612–614.

Banerjee-Basu, S., & Packer, A. (2010). SFARI Gene: An evolving database for the autism research community. *Disease Models & Mechanisms*, *3*(3–4), 133–135. https://doi.org/10.1242/dmm.005439

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(Database issue), D991–D995. https://doi.org/10.1093/nar/gks1193

Bauman, M. L. (2010). Medical comorbidities in autism: Challenges to diagnosis and treatment. *Neurotherapeutics*, *7*(3), 320–327. https://doi.org/10.1016/j.nurt.2010.06.001

Belmadani, M., Jacobson, M., Holmes, N., Phan, M., Pavlidis, P., & Rogic, S. (2019). VariCarta: A comprehensive database of harmonized genomic variants found in ASD sequencing studies. *BioRxiv*, 608356. https://doi.org/10.1101/608356

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*(1), 1–9. 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

Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, *33*(3), 228–237. https://doi.org/10.1038/ng1090

Brueggeman, L., Koomar, T., & Michaelson, J. J. (2018). Forecasting autism gene discovery with machine learning and genome-scale data. *BioRxiv*, 370601. https://doi.org/10.1101/370601

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, *8*(12). https://doi.org/10.1371/journal.pcbi.1002822

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

Caglayan, A. O. (2010). Genetic causes of syndromic and non-syndromic autism. *Developmental Medicine & Child Neurology*, *52*(2), 130–138. https://doi.org/10.1111/j.1469-8749.2009.03523.x

Cook Jr, E. H., & Scherer, S. W. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature*, *455*, 919–923. https://doi.org/10.1038/nature07458

Croen, L. A., Zerbo, O., Qian, Y., Massolo, M. L., Rich, S., Sidney, S., & Kripke, C. (2015). The health status of adults on the autism spectrum. *Autism: The International Journal of Research and Practice*, *19*(7), 814–823. https://doi.org/10.1177/1362361315577517

Cummings, B. B., Karczewski, K. J., Kosmicki, J. A., Seaby, E. G., Watts, N. A., Singer-Berk, M., … MacArthur, D. G. (2019). Transcript expression-aware annotation improves rare variant discovery and interpretation. *BioRxiv*, 554444. https://doi.org/10.1101/554444

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

Dean, M. (2003). Approaches to identify genes for complex human diseases: Lessons from Mendelian disorders. *Human Mutation*, *22*(4), 261–274. https://doi.org/10.1002/humu.10259

Devlin, B., & Scherer, S. W. (2012). Genetic architecture in autism spectrum disorder. *Current Opinion in Genetics & Development*, *22*(3), 229–237. https://doi.org/10.1016/j.gde.2012.03.002

Du, Y., Li, Z., Liu, Z., Zhang, N., Wang, R., Li, F., … Wu, J. (2019). Nonrandom occurrence of multiple de novo coding variants in a proband indicates the existence of an oligogenic model in autism. *Genetics in Medicine*, 1–11. https://doi.org/10.1038/s41436-019-0610-2

Duda, M., Zhang, H., Li, H.-D., Wall, D. P., Burmeister, M., & Guan, Y. (2018). Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. *Translational Psychiatry*, *8*(1), 1–9. https://doi.org/10.1038/s41398-018-0098-6

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 (Oxford, England)*, *21*(16), 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Feliciano, P., Zhou, X., Astrovskaya, I., Turner, T., Wang, T., Brueggeman, L., … Chung, W. (2019). *Exome sequencing of 457 autism families recruited online provides evidence for novel ASD genes* [Preprint]. https://doi.org/10.1101/516625

Fernandez, B. A., & Scherer, S. W. (2017). Syndromic autism spectrum disorders: Moving from a clinically defined to a molecularly defined approach. *Dialogues in Clinical Neuroscience*, *19*(4), 353–371.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., … Searle, S. M. J. (2014). Ensembl 2014. *Nucleic Acids Research*, *42*(D1), D749–D755. https://doi.org/10.1093/nar/gkt1196

Folstein, S. E., & Piven, J. (1991). Etiology of Autism: Genetic Influences. *Pediatrics*, *87*(5), 767–773.

Folstein, S., & Rutter, M. (1977). Infantile autism: A genetic study of 21 twin pairs. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *18*(4), 297–321.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv:1207.3907 [q-Bio]*. Retrieved from http://arxiv.org/abs/1207.3907

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

Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. A. (2012). Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*, *20*(5), 490–497. https://doi.org/10.1038/ejhg.2011.258

Gillberg, C. L. (1992). The Emanuel Miller Memorial Lecture 1991. Autism and autistic-like conditions: Subclasses among disorders of empathy. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *33*(5), 813–842.

Gillis, J., & Pavlidis, P. (2011a). The impact of multifunctional genes on "guilt by association" analysis. *PloS One*, *6*(2), e17258. https://doi.org/10.1371/journal.pone.0017258

Gillis, J., & Pavlidis, P. (2011b). The role of indirect connections in gene networks in predicting function. *Bioinformatics*, *27*(13), 1860–1866. https://doi.org/10.1093/bioinformatics/btr288

Gillis, J., & Pavlidis, P. (2012). "Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks. *PLOS Computational Biology*, *8*(3), e1002444. https://doi.org/10.1371/journal.pcbi.1002444

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., … Troyanskaya, O. G. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, *47*(6), 569–576. https://doi.org/10.1038/ng.3259

Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., … Børglum, A. D. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics*, *51*(3), 431–444. https://doi.org/10.1038/s41588-019-0344-8

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. https://doi.org/10.1038/nature24277

Harms, M. B., Ori-McKenney, K. M., Scoto, M., Tuck, E. P., Bell, S., Ma, D., … Baloh, R. H. (2012). Mutations in the tail domain of DYNC1H1 cause dominant spinal muscular atrophy. *Neurology*, *78*(22), 1714–1720. https://doi.org/10.1212/WNL.0b013e3182556c05

He, X., Sanders, S. J., Liu, L., Rubeis, S. D., Lim, E. T., Sutcliffe, J. S., … Roeder, K. (2013). Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. *PLOS Genetics*, *9*(8), e1003671. https://doi.org/10.1371/journal.pgen.1003671

Helbig, K. L., Lauerer, R. J., Bahr, J. C., Souza, I. A., Myers, C. T., Uysal, B., … Mefford, H. C. (2018). De Novo Pathogenic Variants in CACNA1E Cause Developmental and Epileptic Encephalopathy with Contractures, Macrocephaly, and Dyskinesias. *The American Journal of Human Genetics*, *103*(5), 666–678. https://doi.org/10.1016/j.ajhg.2018.09.006

Heyne, H. O., Singh, T., Stamberger, H., Jamra, R. A., Caglayan, H., Craiu, D., … Lemke, J. R. (2018). De novo variants in neurodevelopmental disorders with epilepsy. *Nature Genetics*, *50*(7), 1048–1053. https://doi.org/10.1038/s41588-018-0143-7

Hyman, S. L. (2013). New DSM-5 includes changes to autism criteria. *AAP News*, E130604-1. https://doi.org/10.1542/aapnews.20130604-1

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

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., … MacArthur, D. G. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, 531210. https://doi.org/10.1101/531210

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 using population reference samples. *Nature Genetics*, *49*(4), 504–510. https://doi.org/10.1038/ng.3789

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C. L., Wong, A. K., Tadych, A., … Troyanskaya, O. G. (2016). Genome-wide prediction and functional characterization of the genetic

basis of autism spectrum disorder. *Nature Neuroscience*, *19*(11), 1454–1462. https://doi.org/10.1038/nn.4353

Laird, N. M., & Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, *7*(5), 385–394. https://doi.org/10.1038/nrg1839

Lanckriet, G. R. G., Deng, M., Cristianini, N., Jordan, M. I., & Noble, W. S. (2004). Kernel-based data fusion and its application to protein function prediction in yeast. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 300–311. https://doi.org/10.1142/9789812704856_0029

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(Database issue), D980–D985. https://doi.org/10.1093/nar/gkt1113

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., & Rhee, S. Y. (2010). Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nature Biotechnology*, *28*(2), 149–156. https://doi.org/10.1038/nbt.1603

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

Levy, D., Ronemus, M., Yamrom, B., Lee, Y., Leotta, A., Kendall, J., … Wigler, M. (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, *70*(5), 886–897. https://doi.org/10.1016/j.neuron.2011.05.015

Li, J., Wang, L., Guo, H., Shi, L., Zhang, K., Tang, M., … Xia, K. (2017). Targeted sequencing and functional analysis reveal brain-size-related genes and their networks in autism spectrum disorders. *Molecular Psychiatry*, *22*(9), 1282–1290. https://doi.org/10.1038/mp.2017.140

Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., … Lage, K. (2017). A scored human protein–protein interaction network to catalyze genomic interpretation. *Nature Methods*, *14*(1), 61–64. https://doi.org/10.1038/nmeth.4083

Li, X., Kim, Y., Tsang, E. K., Davis, J. R., Damani, F. N., Chiang, C., … Montgomery, S. B. (2017). The impact of rare variation on gene expression across tissues. *Nature*, *550*(7675), 239–243. https://doi.org/10.1038/nature24267

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Lin, Y., Rajadhyaksha, A. M., Potash, J. B., & Han, S. (2018). A machine learning approach to predicting autism risk genes: Validation of known genes and discovery of new candidates. *BioRxiv*, 463547. https://doi.org/10.1101/463547

Liu, L., Lei, J., Sanders, S. J., Willsey, A. J., Kou, Y., Cicek, A. E., … Roeder, K. (2014). DAWN: A framework to identify autism genes and subnetworks using gene expression and genetics. *Molecular Autism*, *5*(1), 22. https://doi.org/10.1186/2040-2392-5-22

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

Lobar, S. L. (2016). DSM-V Changes for Autism Spectrum Disorder (ASD): Implications for Diagnosis, Management, and Care Coordination for Children With ASDs. *Journal of Pediatric Health Care*, *30*(4), 359–365. https://doi.org/10.1016/j.pedhc.2015.09.005

Losh, M., Adolphs, R., Poe, M. D., Couture, S., Penn, D., Baranek, G. T., & Piven, J. (2009). Neuropsychological profile of autism and the broad autism phenotype. *Archives of General Psychiatry*, *66*(5), 518–526. https://doi.org/10.1001/archgenpsychiatry.2009.34

Lyall, K., Croen, L., Daniels, J., Fallin, M. D., Ladd-Acosta, C., Lee, B. K., … Newschaffer, C. (2017). The Changing Epidemiology of Autism Spectrum Disorders. *Annual Review of Public Health*, *38*, 81–102. https://doi.org/10.1146/annurev-publhealth-031816-044318

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Masters, T. A., Kendrick-Jones, J., & Buss, F. (2017). Myosins: Domain Organisation, Motor Properties, Physiological Roles and Cellular Functions. In B. M. Jockusch (Ed.), *The Actin Cytoskeleton* (pp. 77–122). https://doi.org/10.1007/164_2016_29

Matson, J. L., & Cervantes, P. E. (2014). Commonly studied comorbid psychopathologies among persons with autism spectrum disorder. *Research in Developmental Disabilities*, *35*(5), 952–962. https://doi.org/10.1016/j.ridd.2014.02.012

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), 356–369. https://doi.org/10.1038/nrg2344

McClellan, J., & King, M.-C. (2010). Genetic Heterogeneity in Human Disease. *Cell*, *141*(2), 210–217. https://doi.org/10.1016/j.cell.2010.03.032

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

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Menke, L. A., Belzen, M. J. van, Alders, M., Cristofoli, F., Ehmke, N., Fergelot, P., … Hennekam, R. C. M. (2016). CREBBP mutations in individuals without Rubinstein–Taybi syndrome phenotype. *American Journal of Medical Genetics Part A*, *170*(10), 2681–2693. https://doi.org/10.1002/ajmg.a.37800

Menke, L. A., Gardeitchik, T., Hammond, P., Heimdal, K. R., Houge, G., Hufnagel, S. B., … Hennekam, R. C. (2018). Further delineation of an entity caused by CREBBP and EP300 mutations but not resembling Rubinstein–Taybi syndrome. *American Journal of Medical Genetics Part A*, *176*(4), 862–876. https://doi.org/10.1002/ajmg.a.38626

Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., … Lein, E. S. (2014). Transcriptional landscape of the prenatal human brain. *Nature*, *508*(7495), 199–206. https://doi.org/10.1038/nature13185

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*. https://doi.org/10.1038/nature11011

Negri, G., Magini, P., Milani, D., Colapietro, P., Rusconi, D., Scarano, E., … Gervasini, C. (2016). From Whole Gene Deletion to Point Mutations of EP300-Positive Rubinstein-Taybi Patients: New Insights into the Mutational Spectrum and Peculiar Clinical Hallmarks. *Human Mutation*, *37*(2), 175–183. https://doi.org/10.1002/humu.22922

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733-745. https://doi.org/10.1093/nar/gkv1189

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

Osorio, D., Rondon-Villarreal, P., & Torres, R. (2014). Peptides: Calculate indices and theoretical physicochemical properties of peptides and protein sequences (Version R Package Version 2.2). Retrieved from http: //CRAN.R-project.org/package=Peptides

Osorio, D., Rondon-Villarreal, P., & Torres, R. (2015). Peptides: A package for data mining of antimicrobial peptides. *The R Journal*, *7*(1), 4–14.

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., … Tyers, M. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, *47*(D1), D529–D541. https://doi.org/10.1093/nar/gky1079

Ozonoff, S., Young, G. S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., … Stone, W. L. (2011). Recurrence risk for autism spectrum disorders: A Baby Siblings Research Consortium study. *Pediatrics*, *128*(3), e488-495. https://doi.org/10.1542/peds.2010-2825

Pastural, E., Barrat, F. J., Dufourcq-Lagelouse, R., Certain, S., Sanal, O., Jabado, N., … Basile, G. de S. (1997). Griscelli disease maps to chromosome 15q21 and is associated with mutations in the Myosin-Va gene. *Nature Genetics*, *16*(3), 289–292. https://doi.org/10.1038/ng0797-289

Pavlidis, P., & Gillis, J. (2012). Progress and challenges in the computational prediction of gene function using networks. *F1000Research*, *1*, 14. https://doi.org/10.12688/f1000research.1-14.v1

Pavlidis, P., & Gillis, J. (2013). Progress and challenges in the computational prediction of gene function using networks: 2012-2013 update. *F1000Research*, *2*, 230. https://doi.org/10.12688/f1000research.2-230.v1

Peña-Castillo, L., Tasan, M., Myers, C. L., Lee, H., Joshi, T., Zhang, C., … Roth, F. P. (2008). A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biology*, *9*(Suppl 1), S2. https://doi.org/10.1186/gb-2008-9-s1-s2

Petrif, F., Giles, R. H., Dauwerse, H. G., Saris, J. J., Hennekam, R. C. M., Masuno, M., … Breuning, M. H. (1995). Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP. *Nature*, *376*(6538), 348–351. https://doi.org/10.1038/376348a0

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

Poirier, K., Lebrun, N., Broix, L., Tian, G., Saillour, Y., Boscheron, C., … Chelly, J. (2013). Mutations in *TUBG1*, *DYNC1H1*, *KIF5C* and *KIF2A* cause malformations of cortical development and microcephaly. *Nature Genetics*, *45*(6), 639–647. https://doi.org/10.1038/ng.2613

Qiao, Y., Harvard, C., Tyson, C., Liu, X., Fawcett, C., Pavlidis, P., … Rajcan-Separovic, E. (2010). Outcome of array CGH analysis for 255 subjects with intellectual disability and search for candidate genes using bioinformatics. *Human Genetics*, *128*(2), 179–194. https://doi.org/10.1007/s00439-010-0837-0

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

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

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

Rubinstein, J. H., & Taybi, H. (1963). Broad Thumbs and Toes and Facial Abnormalities: A Possible Mental Retardation Syndrome. *American Journal of Diseases of Children*, *105*(6), 588–608. https://doi.org/10.1001/archpedi.1963.02080040590010

Ruzzo, E. K., Pérez-Cano, L., Jung, J.-Y., Wang, L., Kashef-Haghighi, D., Hartl, C., … Wall, D. P. (2019). Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. *Cell*, *178*(4), 850-866.e26. https://doi.org/10.1016/j.cell.2019.07.015

Sanal, O., Ersoy, F., Tezcan, I., Metin, A., Yel, L., Ménasché, G., … de Saint Basile, G. (2002). Griscelli Disease: Genotype–Phenotype Correlation in an Array of Clinical Heterogeneity. *Journal of Clinical Immunology*, *22*(4), 237–243. https://doi.org/10.1023/A:1016045026204

Sanders, S. J., Ercan-Sencicek, A. G., 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*(5), 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*, *87*(6), 1215–1233. 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*, *485*(7397), 237–241. https://doi.org/10.1038/nature10945

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., Rubeis, S. D., An, J.-Y., … Buxbaum, J. D. (2019). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *BioRxiv*, 484113. https://doi.org/10.1101/484113

Satterstrom, F.K., Fu, J., Brand, H., Kosmicki, J. A., Wang, H., Zhao, X., … Autism Sequencing Consortium. (2019, October). *Insights into the genetic architecture of autism from exome and genome sequencing of over 60,000 individuals*. Presented at the American Society of Human Genetics, Houtson, Texas. Retrieved from https://eventpilotadmin.com/web/page.php?page=Session&project=ASHG19&id=147005

Sayers, E. W., Beck, J., Brister, J. R., Bolton, E. E., Canese, K., Comeau, D. C., … Ostell, J. (2019). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz899

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 (New York, N.Y.)*, *316*(5823), 445–449. https://doi.org/10.1126/science.1138659

Shen, Y., Dies, K. A., Holm, I. A., Bridgemohan, C., Sobeih, M. M., Caronna, E. B., … Miller, D. T. (2010). Clinical Genetic Testing for Patients With Autism Spectrum Disorders. *Pediatrics*, *125*(4), e727–e735. https://doi.org/10.1542/peds.2009-1684

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., … Cooper, D. N. (2017). The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, *136*(6), 665–677. https://doi.org/10.1007/s00439-017-1779-6

Stessman, H. A. F., Xiong, B., Coe, B. P., Wang, T., Hoekzema, K., Fenckova, M., … Eichler, E. E. (2017). Targeted sequencing identifies 91 neurodevelopmental disorder risk genes with autism and developmental disability biases. *Nature Genetics*, *49*(4), 515. https://doi.org/10.1038/ng.3792

Stevens, C. A. (2019). *Rubinstein-Taybi Syndrome*. Retrieved from https://www.ncbi.nlm.nih.gov/sites/books/NBK1526/

Stranger, B. E., Brigham, L. E., Hasz, R., Hunter, M., Johns, C., Johnson, M., … Montgomery, S. B. (2017). Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease The eGTEx Project. *Nature Genetics*, *49*(12), 1664–1670. https://doi.org/10.1038/ng.3969

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., … Mering, C. von. (2019). STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(Database issue), D607–D613. https://doi.org/10.1093/nar/gky1131

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., … Forbes, S. A. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

Trottier, G., Srivastava, L., & Walker, C. D. (1999). Etiology of infantile autism: A review of recent advances in genetic and neurobiological research. *Journal of Psychiatry and Neuroscience*, *24*(2), 103–115.

Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., … Kriegstein, A. R. (2019). Single-cell genomics identifies cell type–specific molecular changes in autism. *Science*, *364*(6441), 685–689. https://doi.org/10.1126/science.aav8130

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

Volkmar, F. R., & Reichow, B. (2013). Autism in DSM-5: Progress and challenges. *Molecular Autism*, *4*, 13. https://doi.org/10.1186/2040-2392-4-13

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

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

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*(1), 1–10. https://doi.org/10.1038/ncomms13316

Weedon, M. N., Hastings, R., Caswell, R., Xie, W., Paszkiewicz, K., Antoniadi, T., … Ellard, S. (2011). Exome Sequencing Identifies a DYNC1H1 Mutation in a Large Pedigree with Dominant Axonal Charcot-Marie-Tooth Disease. *The American Journal of Human Genetics*, *89*(2), 308–312. https://doi.org/10.1016/j.ajhg.2011.07.002

Wing, L., & Gould, J. (1979). Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. *Journal of Autism and Developmental Disorders*, *9*(1), 11–29. https://doi.org/10.1007/BF01531288

Yang, R. Y., Quan, J., Sodaei, R., Aguet, F., Segrè, A. V., Allen, J. A., … Xi, H. S. (2018). A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. *BioRxiv*, 311563. https://doi.org/10.1101/311563

Yuen, R. K., 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), 602. https://doi.org/10.1038/nn.4524

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098

Zhang, C., & Shen, Y. (2017). A Cell Type-Specific Expression Signature Predicts Haploinsufficient Autism-Susceptibility Genes. *Human Mutation*, *38*(2), 204–215. https://doi.org/10.1002/humu.23147

Appendix A: Additional information for Chapter 2

A1: Gene level annotations

Column	Source	Description
Physical node degree	BioGrid	The number of physical interaction partners a gene has from BioGrid
Multifunctionality	In-lab	Roughly the number of molecular functions a gene has enumerated from GO
Number of publications	In-lab	The number of publications associated with a gene
exac_pLI	gnomAD	The probability of gene for being intolerant to heterozygous loss-of-function variation. Corrected for gene length. Proxy for likelihood of being involved in disease. Computed from ExAC exomes
gnomad_pLI	gnomAD	The probability of gene for being intolerant to heterozygous loss-of-function variation. Corrected for gene length. Proxy for likelihood of being involved in disease. Computed from gnomAD exomes
observed/expected loss-of-function score	gnomAD	Measure of deviation of observed loss-of-function variation from expected. Not corrected for gene length. Proxy for likelihood of being involved in disease across a greater spectrum of selection against LoF variation. Computed from gnomAD.
missense_z score	gnomAD	Measure of deviation of observed missense variation from expected. Not corrected for gene length. Proxy for likelihood of being involved in disease across a greater spectrum of selection against LoF variation. Computed from gnomAD.

A2: Flags used for running VEP

Flag	Flag type	Description
--verbose	Basic	Prints out warning and error flags for variants as they run
--config	Basic	Load configuration options from common and specific config file. Common config file specifies data directories, virtual environments, resource and runtime parameters. The specific config file made for this set of analyses specified plugin directories and VEP annotation flags
--species	Basic	Specific species as homo sapiens
--assembly	Basic	GRCh37
--input_file	Basic	Input variants to be annotated. In HGVS format for SFARI variants, or in chromosome_position_ref/alt format
--output_file	Basic	Output file name
--force_overwrite	Basic	Fail with an error if the output file already exists to avoid overwriting files
--fork	Basic	Enables forking, specified at 10 for improving runtime
--cache	Cache	Enables use of cache (see table 2.1)
--merged	Cache	Enables use of Ensemble and RefSeq transcription annotation. Source column reports source of each transcript
--fasta	Cache	Specify FASTA file directory to use for lookup of reference sequences. Needed specifically for retrieval of HGVS annotation (--hgvs). Fasta file is the GRCh37.75 primary assembly from Ensembl
--buffer_size [number]	Cache	Sets internal buffer size for number of variants read into memory at the same time. Low numbers for less memory and longer runtime. Set as 500.
--port [number]	Database	Specifies port to Ensembl database (3337) to be used when the import format is HGVS
--plugin	Other annotation sources	Plugins specified include CADD, dbscSNV, exac_pLI, G2P, LOFTOOL (see Table 2.2)

--use_given_ref	Other annotation sources	Avoids use of transcript reference, and uses the provide reference allele from the input
--vs.	Output format options	Output file format as VCF. Data fields separated by |, and order is written in VCF header.
--sift [p|s|b]	Output options	Gives SIFT prediction term (deleterious, tolerated) and score of effect of amino acid substitution on protein. Specify b for both score and prediction.
--polyphenol [p|s|b]	Output options	Gives Polypheme prediction term (probably damaging, benign) and score of effect of amino acid substitution on protein. Specify b for both score and prediction.
--distance [bp_distance]	Output options	Modifies distance up/downstream between a variant and a transcript with annotation of upstream_gene_variant or downstream_gene_variant. Specified as 5000 base pairs.
--overlaps	Output options	Report proportion and length of transcript overlapped by SV
--regulatory	Output options	Reports overlaps with regulatory regions, and if a variant overlaps with a transcription factor binding site. Specified as Regulatory Feature or MotifFeature
--hgvs	Identifiers	Add HGVS nomenclature based on Ensembl stable identifiers with version numbers. If –cache, have to add --fasta. Coding and protein sequence names both given, and reported offset of the HGVS annotation and the variant
--hgvsg	Identifiers	Add HGVS genomic nomenclature based on input chromosome. If –cache, have to add --fasta.
--transcript_version	Identifiers	Add version numbers to transcript identifiers
--protein	Identifiers	Add protein identifier
--symbol	Identifiers	Add gene symbol, symbol source, and HGNC ID if applicable
--uniprot	Identifiers	Adds best match accession for translated protein products from SWISSPROT, TREMBL and UNIPARC
--tsl	Identifiers	Adds transcript support level. Note usually only available for GRCh38

--appris	Identifiers	Adds APPRIS isoform annotation for transcript. Note usually only available for GRCh38
--canonical	Identifiers	Adds flag indicating canonical transcript for gene
--mane	Identifiers	Adds flag indicating if the transcript is the MANE Select transcript (match with RefSeq). Note usually only available for GRCh38
--biotype	Identifiers	Adds biotype of the transcript or regulatory feature
--check_existing	Co-located variants	Checks for known variants overlapping with input. Compared on allele-specific basis. Outputs Existing_variation, CLIN_SIG, SOMATIC, PHENO
--clin_sig_alleles	Co-located variants	Returns clinical significance for alternate allele when associated with a phenotype as CLIN_SIG
--af	Co-located variants	Returns global allele frequency (AF; population_AF) from 1000 Genomes Phase 3 data
--af_gnomad	Co-located variants	Returns allele frequencies from gnomAD exome populations (gnomad_AF; gnomad_population_AF)
--pubmed	Co-located variants	Adds PubMed identification numbers for publications citing existing variant
--failed [0|1]	Co-located variants	Set to 0 to exclude variants that have failed.

A3: Variant level annotations

Column	Source	Description
Chrom	VEP	Chromosome of input variant
Position	VEP	Position of reference allele of input variant
ID	VEP	Input variant format.; chom_pos_ref/alt or HGVS
Ref	VEP	Reference allele of input variant
Alt	VEP	Alternate allele of input variant
Allele	VEP (default)	Alternate (consequence) allele as determined by VEP
Consequence	VEP (default)	VEP predicted variant consequences on identified transcripts. Based on combination of allele and transcript type. See table 2.4 for details.
Impact	VEP (default)	Impact rating of high (disruptive to protein),moderate (may change effectiveness), modifier (hard to predict, no evidence; often for noncoding vars), or low (harmless, benign).
Symbol	VEP (default)	Gene symbol
Gene	VEP (default)	Identification number associated with gene and Source i.e. Ensembl gene ID, RefSeq (Entrez) gene ID
Feature_type	VEP (default)	Type of transcript i.e. Transcript, MotifFeature, RegulatoryFeature
Feature	VEP (default)	Ensembl or RefSeq transcript versioned identification number
Biotype	VEP (default)	Biotype of transcript or regulatory feature i.e. protein coding
Exon	VEP (default)	Affected exon number
Intron	VEP (default)	Affected intro number
HGVSc	VEP (default)	HGVS nomenclature based on Ensembl/RefSeq identifiers for cDNA
HGVSp	VEP (default)	HGVS nomenclature based on Ensembl/RefSeq identifiers for protein
cDNA_position	VEP (default)	Position of variant in cDNA of identified transcript
CDS_position	VEP (default)	Position of variant in coding sequence of identified transcript
Protein_position	VEP (default)	Position of variant in protein of identified transcript
Amino_acids	VEP (default)	Affected amino acids in coding transcripts
Codons	VEP (default)	Affected codons in coding transcripts
Existing_variation	VEP (default)	Existence of known co-located variants; compared on coordinates and alleles

Symbol_source	VEP (default)	Source of gene symbol i.e. HGNC, Entrez, Ensembl
HGNC_ID	VEP	Associated HGNC gene identification number if source is HGNC
Canonical	VEP	Yes or blank, indicating if Ensembl considers the reported Feature (transcript) as canonical
ENSP	VEP	Ensembl or RefSeq protein identification number
SWISSPROT	VEP	SwissProt protein identification number
TREMBL	VEP	TrEMBL protein identification number
UNIPARC	VEP	UniParc stable and unique protein identification number
RefSeq_match	VEP	Indicates if and how RefSeq transcript differs from underlying genome
Source	VEP	String indicating source of report, Ensembl/RefSeq
SIFT	VEP	Prediction from SIFT as tolerated/deleterious_confidence(score)
PolyPhen	VEP	Prediction from PolyPhen2 as benign, possibly_damaging, probably_damaging (score)
HGVS_offset	VEP	Number of bases the variant has shifted relative to input genomic coordinates
HGVSg	VEP	Genomic HGVS nomenclature based on Ensembl/RefSeq identifiers based on input chromosome
AF	VEP	1000s Genomes allele frequency
AFR_AF	VEP	1000s Genomes allele frequency; African
AMR_AF	VEP	1000s Genomes allele frequency; African American
EAS_AF	VEP	1000s Genomes allele frequency; East Asian
EUR_AF	VEP	1000s Genomes allele frequency; European
SAS_AF	VEP	1000s Genomes allele frequency; South Asian
AA_AF	VEP	1000s Genomes allele frequency; American
EA_AF	VEP	1000s Genomes allele frequency; European American
gnomAD_AF	VEP	gnomAD allele frequency
gnomAD_AFR_AF	VEP	gnomAD allele frequency; African
gnomAD_AMR_AF	VEP	gnomAD allele frequency; American
gnomAD_ASJ_AF	VEP	gnomAD allele frequency; Ashkenazi Jewish
gnomAD_EAS_AF	VEP	gnomAD allele frequency; East Asian
gnomAD_FIN_AF	VEP	gnomAD allele frequency; Finish

gnomAD_NFE_AF	VEP	gnomAD allele frequency; Non-Finish European
gnomAD_OTH_AF	VEP	gnomAD allele frequency; Other
gnomAD_SAS_AF	VEP	gnomAD allele frequency; South Asian
CLIN_SIG	VEP	Clinical significance as reported by ClinVar; compared on alleles and coordinates
SOMATIC	VEP	Origin as reported by ClinVar; compared on alleles and coordinates. Reported as a string of 0 and 1s with 1=somatic origin
PHENO	VEP	Phenotype indicator as reported by ClinVar; compared on alleles and coordinates. Reported as a string of 0 and 1s associated with above, with 1=Phenotype with origin
PUBMED	VEP	PubMed identification number
OverlapBP	VEP	Base pair overlap
OverlapPC	VEP	Percentage overlap
MOTIF_NAME	VEP	Name of motif if variant found to overlap with regulatory region
MOTIF_POS	VEP	Position of motif if variant found to overlap with regulatory region
HIGH_INF_POS	VEP	Binary (N,Y) if variant falls in high information position within a transcription factor finding site
MOTIF_SCORE_CHANGE	VEP	Change in score of motif if variant overlaps regulatory region; motif scores range from 0-1 with 1 being a strong binding site for transcript factor
CADD_PHRED	VEP	CADD raw score that has been scaled based on rank of each variant relative to all possible substitutions in reference genome
CADD_RAW	VEP	CADD raw score output from the SVM with negative values meaning the variant is likely observed, and positive values meaning the variant is likely simulated
Ada_score	VEP	dbscSNV prediction from ADA-boost ensemble method for variant occurring in splice consensus region; score ranges from 0-1, with 0.6 as cut-off for likely damaging/likely not damaging effects
Rf_score	VEP	dbscSNV prediction from random forest ensemble method for variant occurring in splice consensus region; score ranges from 0-1, with 0.6 as cut-off for likely damaging/likely not damaging effects

exac_pLI	VEP	Probability loss-of-function constraint score from ExAC measuring deviation of observed variant counts from expected variant counts
LoFtool	VEP	Gene intolerance score and susceptibility to disease based on loss-of-function variants from ExAC
Coding_seq_length	Ensembl/RefSeq	Length of the coding sequence of the transcript
Coding_seq_GC	Ensembl/RefSeq	GC content of the coding sequence of the transcript
Protein_length	Ensembl/RefSeq	Length of the peptide sequence of the transcript
Protein_mw	Ensembl/RefSeq	Molecular weight of the peptide sequence of the transcript

Appendix B: Additional information for Chapter 3

B.1: SFARI high confidence ASD genes

SFARI category	Genes
1, 1S	ADNP, ANK2, ARID1B, ASH1L, ASXL3, CHD2, CHD8, CUL3, DSCAM, DYRK1A, GRIN2B, KATNAL2, KMT2A, KMT5B, MYT1L, NAA15, POGZ, PTEN, RELN, SCN2A, SETD5, SHANK3, SYNGAP1, TBR1, TRIP12
2, 2S	ANKRD11, BAZ2B, BCKDK, BCL11A, CACNA1D, CACNA1H, CACNA2D3, CEP41, CIC, CNOT3, CNTN4, CNTNAP2, CTNND2, CUX1, DDX3X, DEAF1, DIP2C, ERBIN, FOXP1, GABRB3, GIGYF2, GRIA1, GRIP1, ILF2, INTS6, IRF2BPL, KAT2B, KDM5B, KDM6A, KMT2C, LEO1, MACROD2, MBOAT7, MECP2, MED13, MED13L, MET, NCKAP1, NCOR1, NLGN3, NRXN1, PHF3, PTCHD1, RANBP17, RIMS1, SCN9A, SET, SHANK2, SLC6A1, SMARCC2, SPAST, SRCAP, SRSF11, TAOK2, TBL1XR1, TCF20, TNRC6B, TRIO, UBN2, UPF3B, USP15, USP7, WAC, WDFY3, MAGEL2

B.2: Novel ASD gene sets from iHart, Spark and Satterstrom

Paper	nGenes	Gene Set (SFARI Score)
iHart	14	BTRC, CCSER1, CMPK2, FAM98C, METTL26, MLANA, MYO5A(3), PCM1, PRKAR1B, RAPGEF4(4), SMURF1, TMEM39B, TSPAN4, UIMC1
Spark	13	BRSK2, CPZ, DMWD, DPP6(4), EGR3, FEZF2(4), ITSN1, KDM1B, NR4A2(4), PAX5(3), RALGAPB(3), RNF25, SH3RF3
Satterstrom	31	AP2S1, CELF4(3), CORO1A, DPYSL2(3), EIF3G(4), ELAVL3(3), GRIA2, HDLBP(4), HECTD4(3), KIAA0232, LDB1, LRRC4C, MAP1A, MKX, NCOA1, NUP155, PHF12, PPP5C, PRR12(S), RFX3(4), RORB, SATB1, SRPRA, TAOK1, TEK, TM9SF4, TRIM23, UBR1, VEZF1, ZMYND8, PPP1R9B

B.3: AUROC, Precision at 20% Recall and Precision at 43% recall performance statistics on iHart novel ASD genes (* denote ties at recall of 20% of gene set)

Paper	AUC	AUC_CI	P20R	PR20_CI	P43R	PR43_CI
Princeton	0.57	0.4, 0.74	0.18	0.04, 0.27	0.12	0.04, 0.29
ASD_frn	0.69	0.56, 0.82	0.16	0.06, 0.53	0.16	0.08, 0.26
DAMAGES	0.65	0.52, 0.77	0.24	0.05, 0.32	0.12	0.06, 0.38
RF_Lin	0.58	0.43, 0.73	0.14	0.04, 0.29	0.12	0.04, 0.22
forecASD	0.87	0.77, 0.97	4.81	0.86, 13.59	2.23	0.4, 6.92
DAWN	0.53	0.42, 0.65	0.15	0, 0.34	0.1461*	0, 0.24
exac_pLI	0.52	0.38, 0.66	0.10	0.03, 0.16	0.07	0.04, 0.16
gnomad_pLI	0.49	0.37, 0.61	0.06	0.02, 0.1	0.07	0.04, 0.1
oe_lof_upper	0.56	0.42, 0.7	0.08	0.03, 0.12	0.11	0.05, 0.14
DeRubeis	0.90	0.85, 0.96	3.95	0.5, 6.67	1.10	0.33, 5.72
Sanders	0.98	0.96, 0.99	10.02	2.55, 13.81	6.35	2.29, 16.71
iHart	1.00	1, 1	32.66	14.29, 66.67	37.50	25, 53.57
Satterstrom	0.69	0.54, 0.84	1.12	0.15, 1.69	0.77	0.05, 1.71
Iossifov	0.69	0.53, 0.85	0.65	0.16, 1.62	0.62	0.06, 1.03

B.4: AUROC, Precision at 20% Recall and Precision at 43% recall performance statistics on Spark novel ASD genes (* denote ties at recall of 20% of gene set)

Paper	AUROC	AUROC_CI	P20R	PR20_CI	P43R	PR43_CI
Princeton	0.61	0.43, 0.79	0.15	0.03, 0.14	0.13	0.03, 0.23
ASD_frn	0.75	0.62, 0.9	0.35	0.15, 0.61	0.36	0.1, 0.6
DAMAGES	0.84	0.76, 0.92	0.33	0.12, 0.43	0.32	0.15, 0.53
RF_Lin	0.81	0.7, 0.93	0.50	0.12, 1.04	0.34	0.15, 0.78
forecASD	0.89	0.8, 0.98	1.13	0.45, 1.77	1.40	0.72, 1.84
DAWN	0.61	0.48, 0.74	0.18	0.06, 0.45	0.16*	0.06, 0.29
exac_pLI	0.79	0.69, 0.9	0.17	0.07, 0.29	0.23	0.13, 0.31
gnomad_pLI	0.78	0.68, 0.89	0.17	0.07, 0.28	0.22	0.12, 0.3
oe_lof_upper	0.75	0.64, 0.87	0.23	0.08, 0.35	0.23	0.08, 0.37
DeRubeis	0.80	0.7, 0.9	0.32	0.09, 0.63	0.24	0.12, 0.52
Sanders	0.86	0.75, 0.96	1.30	0.29, 2.11	0.67	0.31, 2.09
iHart	0.86	0.77, 0.94	0.97	0.26, 1.51	0.62	0.16, 1.57
Satterstrom	0.73	0.57, 0.89	0.72	0.2, 3.18	0.47	0.08, 1.18
Iossifov	0.80	0.66, 0.94	0.52	0.22, 1.48	0.71	0.23, 0.92

B.5: AUROC, Precision at 20% Recall and Precision at 43% recall performance statistics for Satterstrom novel ASD genes (* denote ties at recall of 20% of gene set)

Paper	AUROC	AUROC_CI	P20R	PR20_CI	P43R	PR43_CI
Princeton	0.75	0.68, 0.83	0.56	0.35, 3.15	0.67	0.41, 0.92
ASD_frn	0.78	0.71, 0.85	2.77	0.5, 6.92	0.81	0.37, 1.42
DAMAGES	0.87	0.83, 0.92	1.06	0.57, 1.39	1.07	0.66, 1.41
RF_Lin	0.86	0.82, 0.92	1.77	0.76, 2.45	1.38	0.86, 1.86
forecASD	0.90	0.86, 0.95	3.56	2.37, 5.38	3.69	0.95, 6.16
DAWN	0.61	0.54, 0.69	0.39	0.23, 0.83	0.39*	0.25, 0.6
exac_pLI	0.90	0.88, 0.93	1.15*	0.61, 1.73	1.09	0.72, 1.57
gnomad_pLI	0.89	0.86, 0.93	1.42*	0.61, 1.96	1.08	0.7, 1.57
oe_lof_upper	0.91	0.89, 0.93	1.24	0.57, 1.75	0.90	0.63, 1.44
DeRubeis	0.60	0.5, 0.71	0.89	0.49, 1.96	0.50	0.12, 1.26
Sanders	0.64	0.53, 0.75	0.77	0.45, 1.38	0.95	0.21, 1.19
iHart	0.64	0.54, 0.75	0.61	0.39, 1.02	0.61	0.29, 1
Satterstrom	1.00	1, 1	34.31	22.54, 50.53	36.11	28, 48.68
Iossifov	0.82	0.75, 0.89	1.65	0.7, 2.33	1.01	0.65, 1.77