GENE DISCOVERY AND GLUTAMATE SIGNALING DEFECTS IN INTELLECTUAL DISABILITY AND AUTISM

by

Tejasvi S Niranjan

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

March, 2015

Abstract

Neurodevelopmental disorders are a common class of brain disorders that affect up to 1 in 6 children in the industrialized world. They include a range of diseases, including Attention Deficit Hyperactivity Disorder, Autism Spectrum Disorders, and various forms Intellectual Disability. Many of these disorders display overlapping clinical phenotypes, including reductions in intellectual quotient, learning delays, difficulties in social behavior, stereotypical behaviors, and deficits in verbalization and communication. Frequently, other comorbidities may be present, such as epilepsy or craniofacial defects.

Many of these disorders are strictly genetic in their etiology, as determined by a consistent pattern of Mendelian inheritance in affected families. Other, more complex neurodevelopmental disorders, such as schizophrenia, major depression, bipolar disorder, and autism, show strong evidence that genetics is a substantial cause, along with a role for environmental factors.

The focus of this dissertation will be on elucidating genetic and molecular mechanisms involved in the etiology of two neurodevelopmental disorders, X-Linked Intellectual Disability and Autism Spectrum Disorders. Throughout this dissertation, a series of important observations and concepts will be discussed regarding the challenges faced when studying neurodevelopmental disorders of genetic etiology. These challenges are based in the intersecting complexities of how genetic variation influences neural mechanisms and how neural mechanisms influence intellectual function and behavior.

In order to address these challenges, I have developed computational tools to improve our ability to identify potential disease-causing variants and genes. One of these

tools is an effective method to identify causal genes in XLID, through improvements in the quality of sequenced variant calls, and through effective methods of variant filtering using a combination of datasets.

Lastly, I have employed a series of targeted genomics and functional studies to determine how genetic variation can modify neural function and behavior. This series of studies will discuss the role of glutamate signaling defects in autism etiology, with a focus on Glutamate Receptor Interacting Proteins (*GRIP1/2*) as autism susceptibility genes.

As a whole, these studies should provide a framework demonstrating how old and new genomics techniques can be used effectively to find disease-causing variants and genes in neurodevelopmental disorders of increasing complexity. Importantly, this work should reinforce our appreciation of the complexity of neural and genetic systems, and that any computational inference should be diligently investigated by functional work to identify a molecular mechanism for disease.

Advisor: Tao Wang

Reader: Sarah Wheelan

Preface

Isaac Newton once wrote in a letter, “If I have seen further it is by standing on the shoulders of giants.” Not just this dissertation, but all the wonderful experiences of my life are due entirely to the great people who have come before me, and the greater people who have carried me here.

Friends. Family. Mentors. Eventually they all are one and the same.

First and foremost, I must thank the members of my lab. To Abby Adamczyk, I appreciate all the time you took to diligently teach me how to maintain a lab and keep myself organized. To Becca Rose, I will never take for granted the absolute necessity of a lab technician, as a source of daily stability, research success, and above all, friendship. To Mei Han, I could ramble endlessly about how you have become an anchor for our lab, but what has inspired me most has been your ability to do everything: run experiments, write grants and papers, and start a family. You remind me to enjoy work and enjoy life. To Rebeca Mejias, I simply could not have done this without you. You are a priceless mentor, and I envy all the students you will have to come.

No amount of space is sufficient to express my gratitude to my advisor, Tao Wang. As the saying goes, “Give a man a fish, he will eat for a day. Teach a man to fish, he will eat for life.” Tao took this otherwise unassuming and unfocused aspiring scientist, and gave him the tools to become perpetually better: as a learner, as a teacher, and as a discoverer. Such a task seems impossible to me in retrospect, but I am blessed that I found Tao to lead me these last six years, as it seems only he could have done the

impossible. I don't know what the future has in store for me, Tao, but I hope I do you proud.

Sarah Wheelan has been a guiding star for me. She first taught me how to code almost a decade ago. I have come this far because she was always there to help me along, from my scientific infancy as an undergraduate, all the way through my doctoral research as a member of my thesis committee. Thank you so much, Sarah! In this same line, I would not have come to love genetics and the process of discovery without the right mentors to set me on this path long ago. Thank you dearly to Jef Boeke, Kyle Cunningham, and Victor Corces for instilling an early passion. You have always been in the back of my mind.

A special thanks to Rick Huganir, also a member of my thesis committee, and the members of his lab, especially Rich Johnson and Kamal Sharma. Your guidance on experimental protocols, generosity in resources, and insights on neurobiology have been invaluable in the work of this dissertation.

It has been the privilege of a lifetime to be a student of the Johns Hopkins Institute of Genetic Medicine and its graduate program in Molecular Biology and Human Genetics. This program would not function without the amazing people who run it. Thank you Sandy Muscelli, for keeping me straight and always being a source of humor and joy. Thank you Kirby Smith, on behalf of all my classmates and myself, for guiding us through the intricacies of comprehensive exams, rotations, graduation requirements, and ethanol. And I would also like to thank Hans Bjornsson, Jill Fahrner, Ada Hamosh, and Hal Dietz for allowing me to shadow them in the Harriet Lane Clinic, and to learn the

intricacies of clinical care. But of course, special thanks to the head of my thesis committee, David Valle. Your guidance and leadership has been fundamental to the success of this program, as well as my academic career. It has been a privilege to be one of your students.

There are so many more amazing scientists that I must thank. Margaret Taub and Hector Bravo, thank you for taking this computationally illiterate young man, and teaching him that big data is not to be feared, but enjoyed. Thank you to the members of the Valle lab, especially Cassandra Obie and Gary Steel; to the member of the Kim lab in South Korea; and to the members of the Greenwood Genetics Centers, especially Charles Schwartz, for resources and guidance.

Not enough words can express my gratitude to the friends and classmates I've made along the way. To my year, you have shown me endless friendship, compassion, and patience. To Samantha Maragh and Eric Stevens, I would call you best friends, but really we have adopted each other as family. We will go a long way together.

The following friends have made life energetic, colorful, and adventurous: Julie, Max R., Max C., Jon, Eric R., Amanda, Greg, Genevieve, Shaughn, Allen, Jerry, Kevin, Nina, Xuan, Anna, and Brian. And of course, en masse, all the hum-gen classmates who have made this program so vibrant.

I am blessed with a large, close family, but my parents take the cake. Mom and Dad, thank you for all your support throughout the years, the ceaseless energy you put in trying to raise me, and of course (I am a Human Geneticist after all), thank you for the genes. To my sister, Sumedha, and my cousin, Akhila: we have gone through so many

experiences together, learned so much from one another, and will continue to grow with each other. No matter what I do or where I go next, I am so happy to know that we are on this ridiculous journey together.

Above all, thank you to every child born a little more different, a little less perfect, a little bit uncertain of their own potential. We are each greater than our variances.

"Of science and the human heart, there is no limit…" - Miracle Drug, *U2*

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

The study of neurodevelopmental disorders is a uniquely challenging and engaging field. Neurodevelopmental disorders sit at the crossroads of two of the most complex biological disciplines, genetics and neuroscience; we invariably require an understanding and appreciation of both fields to make progress. In 1866, Gregor Mendel published his findings on plant hybridization, setting the foundation for modern genetics and the role of genetic variation in living organisms [1]. In 1888, Santiago Ramón y Cajal published his seminal work providing the first decisive evidence for the Neuron Doctrine of brain anatomy and function [2]. Though we have come far in the last 150 years, the most easily appreciated observation is how much farther we have to go.

1.1 Mendelian Inheritance and X-Linked Intellectual Disability

The first observation of a heritable disease is attributed to Archibald Garrod in 1902, when he identified Alkaptonuria as an inborn error of metabolism following a pattern of inheritance consistent with the Laws of Mendel [3]. Since then, thousands of cases of genetic disorders following Mendelian inheritance patterns have been identified, including over 6,000 unique Mendelian phenotypes recorded in the database Online Mendelian Inheritance in Man (OMIM) [4]. Within these records is a subclass of Mendelian traits and diseases that follow an X-linked pattern of inheritance. This pattern of inheritance, stemming from a causal variant located on the X chromosome, is strictly characterized by the mating of an unaffected father and a carrier mother producing sons, half of whom are affected, and daughters, half of whom are carriers. It is classically

characterized by the observation of a trait that "skips a generation," whereby an affected father produces a carrier daughter, who in turn produces an affected son. However, this pedigree is rarely seen for serious X-linked disorders, due to the negative selection placed on the grandfather. For X-linked disorders, the causative mutation generally can only persist in the family by propagation from carrier mother to carrier daughter. In these instances, carrier females are either unaffected or too minimally affected to place a negative pressure on reproduction and propagation of the underlying mutation. Examples of pedigrees for an X-linked recessive trait are described in Figure 1-1.

Given that males possess only one copy of the X chromosome, X-linked conditions that would normally only be possible through homozygous recessive inheritance must be expressed in this hemizygous state. As such, males are far more frequently affected by X-linked disorders than females. More importantly, this hemizygous state makes it easier for X-linked recessive conditions to present at a rate far higher than for autosomal recessive conditions. In fact, X-linkage accounts for 16% of male Intellectual Disability cases, even though the X chromosome accounts for only 2.5% of the male genome [5]. This increase is likely due both to the ease by which X-linked recessive conditions can present in a hemizygous state and to the fact the X-linked conditions are more easily recognizable and investigated than autosomal recessive conditions.

Intellectual Disability is diagnosed in an individual meeting at least three criteria: an Intellectual Quotient (IQ) below 70, a substantial deficit in two or more adaptive behaviors, such as self-care, communication, or interpersonal skills, and an onset before

age 18 [6]. The first observation for Intellectual Disability with X-linkage (XLID) was made in the late 1960s [7]. Since then, thousands of cases have been described, but the molecular basis remains largely unknown. Broad genomic approaches have identified some 100 X-linked genes that may account for half of XLID cases, but the total number of responsible genes may be hundreds more [5]. In spite of the fact that half of XLID cases do not have a definitive molecular diagnosis, X-linked Intellectual Disability remains one of the better studied and characterized neurodevelopmental disorders.

1.2 A Spectrum of Complexity

At its most basic, a Mendelian disorder can trace its genetic source to a single mutation in a single gene, hence referred to as a single-gene disorder. In 1989, researchers discovered the first gene, *CFTR*, to be associated with a Mendelian disease, Cystic Fibrosis, including pinpointing the most frequently occurring defect [8]. Since then, an estimated 70% of Mendelian traits recorded in OMIM now have a known molecular basis [4]. This journey has been a challenging one, given the size of the diploid human genome (~6E9 bp) and the number of genes (~22,000). Many mapping strategies have been developed and employed to find the underlying genetic defects for single-gene traits, but none has been more successful than next-generation sequencing of human exomes. The first successful use of exome sequencing was demonstrated as proof-of-concept in 2009 using four patients diagnosed with Freeman-Sheldon Syndrome [9]. Sequencing generated hundreds of thousands of variants for analysis; through a combination of computational filters and by looking for genes with a burden of predicted

damaging mutations, the researchers identified *MYH3* as the likely causal gene for these four patients. Since then, some 2,000 additional exome sequencing studies have been performed, providing a molecular basis for countless Mendelian disorders. With these advances, a broad new set of tools and frameworks have been introduced to handle the large amounts of data produced by next-generation sequencing [10].

In spite of these successes, many challenges remain in finding the underlying genetic causes for all Mendelian disorders. Foremost among these challenges are Mendelian disorders with high locus heterogeneity. Whereas some Mendelian diseases can be strictly associated to a single gene (e.g.: Freeman-Sheldon Syndrome and *MYH3*, Cystic Fibrosis and *CFTR*), other diseases, such as Intellectual Disability, may result from mutation of one of many different genes, referred to as locus heterogeneity. However, only one gene will ever be responsible in a single individual. Though the observed pattern of inheritance for the disease in the family is still Mendelian, we can no longer view all affected families in aggregate as being affected by a single-gene disorder. Rather, we must now view these individuals as suffering from a disorder that can be caused by one of many different genes, and is clinically indistinguishable among the different genetic defects. X-linked Intellectual Disability, which may be caused by up to 200 X-linked genes, is an excellent example of such a Mendelian disorder with high locus heterogeneity [5].

However, the complexity of neurodevelopmental disorders is not restricted to the individual effect of a single gene or mutation. As we tackle common, complex diseases, we must then contend with challenges presented by variable expressivity, mutations of

variable effect size, including modifier mutations, and environmental factors. Autism Spectrum Disorders epitomize such complexity.

1.3 Autism Spectrum Disorder: A Multifactorial Disorder

Autism is a common neurodevelopmental disorder with a prevalence of at least one in 100 children. It typically presents by age three. Required clinical presentation includes deficits in reciprocal social interactions, repetitive or stereotypical behaviors, and restrictions in verbal communication. Many comorbidities may also present, including microcephaly and macrocephaly, epilepsy, and intellectual delay/disability. Males have a five-fold increased risk of developing disease over females, indicative of a possible, but poorly understood protective genetic background present in females [11].

There is a great deal of evidence indicating that genetics is important in autism risk. Analyses of various twin concordance studies estimate the heritability of autism risk between 36% (broad autism diagnosis) to 95% (strict autism diagnosis) [12,13]. Another study, based on a large Swedish epidemiological cohort, estimated heritability at 50%, with the majority of risk contributed by common variants [14]. For comparison, heritability for other psychiatric or neurodevelopmental disorders are also relatively high. Heritability has been estimated between 70-85% for schizophrenia [15,16], 40% for major depression [17], and as high as 70% for bipolar disorder [18,19]. These results likely reflect a complex and overlapping contribution of many different genetic pathways to individual behavioral phenotypes. Additional support for autism as a genetic disease comes from the observation that there are a number of hereditary syndromes, such as

Fragile X and Rett Syndrome, including multiple microdeletion and microduplication syndromes, which display autism traits [20,21].

That numerous genetic diseases display autistic phenotypes also adds support for the spectrum of phenotypic and genotypic complexity in autism. Take for example Fragile X Syndrome, caused by mutation of the *FMR1* gene on the X chromosome. Fragile X is the most frequent cause of intellectual disability, resulting from a trinucleotide repeat expansion near an *FMR1*-controlling CpG island that becomes hypermethylated in Fragile X patients [22]. While it is X-linked, Fragile X displays a dominant pattern of inheritance with reduced penetrance and affects both sexes. Fragile X patients also have a characteristic facies [23]. As such, Fragile X's presentation and Mendelian pattern of inheritance differs slightly from that of other XLID cases, where males are affected more frequently due to their hemizygous state and the causative mutations are recessively inherited. A separate inheritance pattern, more specific presentation, and specific molecular testing allows physicians to clinically separate Fragile X cases from remaining XLID cases. However, *FMR1* can be mutated without involving the trinucleotide repeat and produce a Fragile X-like XLID, making *FMR1* mutation a cause for both Fragile X intellectual disability and clinically inseparable XLID [23]. *FMR1* testing is therefore a necessary inclusion for genetic screens for any occurrence of XLID. At the same time, *FMR1* is an important regulator of neuronally-expressed genes, including *SHANK3*, *PTEN*, *TSC2*, and *mGluR5*, which have previously been implicated in autism [24]. *TSC2* is associated with tuberous sclerosis, for which a significant proportion of cases meet diagnostic criteria for autism [25]. As such, *FMR1*

mutations, depending on mutation effect and genetic background, can produce a pleiotropy of phenotype, from classic Fragile X Syndrome to XLID, and may even have a role in autism.

A similar scenario presents with Rett Syndrome, one of the most frequent intellectual disability syndromes in females. Caused by mutations in X-linked *MECP2*, a DNA methyl-binding protein, deleterious mutations are so serious that males are generally not viable and females display a severe intellectual disability as well as some autistic traits [26]. However, *MECP2* mutations cannot be excluded for male forms of XLID, as less deleterious mutations can cause XLID in males, and leave female carriers more or less asymptomatic. In fact, *MECP2* is frequently mutated in XLID cases, at a rate comparable to trinucleotide expansion of *FMR1* [27]. At the same time, *MECP2* mutations have been reported in autism cases and aberrant *MECP2* expression has been observed in the frontal cortex of autism brain samples [28]. These lines of evidence indicate that different mutations of *MECP2* can variably express autistic or intellectual disability traits.

Individuals with *MECP2* mutations also share phenotypic overlap with patients diagnosed with Angelman Syndrome, an imprinting disorder associated with deletion of the 15q11-13 region, resulting in loss of function of the *UBE3A* gene [29]. Such an observation is plausible given the discovery of a convergence in biological pathways between *MECP2* and *UBE3A* [30]. Angelman Syndrome itself also shares extensive phenotypic similarities with autism. Microdeletion and microduplication of the 15q11-13 region are statistically significantly enriched in patients diagnosed with autism without

specific diagnosis of Angelman. Additional microdeletion and microduplication syndromes, including the 16p11, 7q11, 17q12, and 22q11 regions, are also associated with autism. However, these syndromes lend more credence to the extensive pleiotropy of genetic defects expected in autism. 22q11 is best known as the region deleted in DiGeorge Syndrome, Shprintzen Syndrome, and Velocardiofacial Syndrome. 7q11 is associated with Williams Syndrome [31]. And while deletion of 16p11 is strongly associated with autism or developmental delay, duplication of the region is also strongly associated with autism, developmental delay, schizophrenia, and bipolar disorder [32].

Autism represents the extreme spectrum of complexity, both in variable expression of phenotype resulting from individual defects, and from a broad locus heterogeneity. Given that so many individual disorders present with autistic phenotypes, it is an acceptable assumption and generally agreed conclusion, that true autism is not a single disease, but a spectrum of individual disorders resulting from the complex interaction of many mutations of variable effect in each individual person. Hence, autistic behavior in any given individual is best diagnosed as an Autism Spectrum Disorder.

1.4 Common vs. Rare Variant Hypotheses

Though autism has a high prevalence, the frequency and effect size of the underlying mutations remains unclear. During the early stages of the genomics era, it was widely hypothesized that common diseases would be the result of common mutations that exist at an appreciable frequency in the general population [33]. This has certainly been true for some diseases, such as Crohn's Disease and early-onset Alzheimer's Disease

[34,35]. The underlying hypothesis for these studies assumes that a common disease is the result of many common variants, each of small effect, acting in concert to produce disease in a given individual. A number of genomic approaches, including Genome-Wide Association Studies (GWAS) and Linkage Studies, have been implemented to identify common causative variants for autism. Common variation has been estimated to account for up to half of autism risk [36,14].

An alternative hypothesis, the common disease-rare variant hypothesis, states that common disease is the result of many rare variants, each of large effect, acting in concert to produce disease in a given individual [37]. These rare variants may even be private to individual families. Several studies have successfully associated *de novo* Copy Number Variations (CNVs) and *de novo* Single Nucleotide Polymorphisms (SNPs) to autism risk [38-40]. However, these rare events account for risk in a very small percentage of cases. It has yet to be determined if rare, transmitting variants play a more substantial role.

Identifying common or rare variants is an important challenge that will likely require employing progressively larger sample sizes to achieve sufficient statistical power. However, due to the higher costs of using larger sample sizes, it is equally important to devise novel methods that may identify some important etiologic variants with smaller cohorts. Such “low-hanging fruit” may guide more targeted approaches for finding the remaining variation responsible for all heritable autism risk.

1.5 High-throughput Genomic Approaches to Identify Mutations

Earliest forms of genome-scale analysis of variation involved mapping of easily genotyped marker variants that were determined to be in linkage with a phenotype of interest in a pedigree. Presumably, these linkage markers would be proximal to the pathologic mutation and gene of interest [41]. Increased densities of markers would permit higher resolution mapping of genetic variation to phenotype. As increased marker densities became available through SNP arrays, it became possible to perform large-scale Genome Wide Association Studies (GWAS) for common, complex disorders caused by common variation. Through GWAS, one could test the Common Variant-Common Disease hypothesis using the principle of indirect association of marker genotypes in linkage with causal mutations to a phenotype of interest in a test group versus a control group [42]. These efforts have been highly successful for certain diseases, where the causative common variants have fairly large effect sizes that increase disease risk relative to the general population by at least 1.2-fold [43]. However, GWAS have had limited success in identifying major causes of autism [36]. In some instances, certain genes and gene networks have been implicated as increasing risk in a minority of cases, but these have not yet been fully replicated [44-46]. Presuming that a sufficient fraction of risk is due to common variation of very small effect, these limited results could be explained by insufficient power within each study to detect association [14].

Where linkage and GWAS approaches have produced limited results, advances in next-generation high-throughput sequencing may make the difference. Indeed, recent large-scale efforts in exome sequencing of autism cohorts has produced a large list of

potential autism susceptibility genes, by statistical analysis of gene networks or by identifying a significant burden of *de novo* mutations in genes [47,48,40].

Advances in high-throughput sequencing have produced exponential increases in sequencing capacity with a concomitant decrease in cost per base pair sequenced [49]. During the course of the sequencing projects that have been conducted in the context of this dissertation, sequencing capacity has increased 100-fold, but maintained roughly the same cost per experiment (Figure 1-2). These advances have allowed researchers to expand the number of samples in a given study, thereby improving statistical power.

Additional advances have been made in sequencing chemistries, substantially reducing the error rate generated from raw sequencing data, which can result in false positive results. Examples of how some of these errors present in the Illumina sequencing platform, how they are corrected, and how they have changed over chemistries, are discussed in a later chapter. Methods to identify and control sequencing errors are important, not only because they enable us to reduce the false positive rate, thereby minimizing the number of spurious variants that must be followed, but because they enable us to reduce the number of aligned sequenced short reads required to call a variant. Where before it was necessary to sequence a single genomic position at least 50 times to generate a confident variant call, if that site can be sequenced with only 25x coverage, the remaining 25 reads can be devoted to sequencing more regions or more samples.

The ability to sequence specific genomic regions is also important, as, in spite of the exponential advances in sequencing capacity, it is still expensive to sequence an

entire human genome. With the best of current technology, it still costs several thousand dollars to sequence a single human genome, takes up an eighth of the machine's capacity, and requires over a week-and-a-half to generate the raw data [50]. Performing this process for a thousand samples will be very expensive and time consuming. Additionally, it will generate a tremendous amount of data, only a portion of which is useful for studying disease etiology.

To overcome this challenge, many studies, including the ones discussed in this dissertation, make use of targeted sequencing, the most common form of which is exome sequencing. Initially, these targeting protocols involved using PCR amplification of target regions defined by specific, custom oligonucleotide pairs. Each PCR product for each sample would need to be amplified in a separate reaction, limiting scalability for large cohorts or genomic regions [51]. To improve scalability and cost, PCR pooling strategies and hybridization procedures were developed. Pooling strategies will be discussed further in a later chapter.

For hybridization protocols, a sample's DNA is hybridized to custom oligonucleotides fixed on a microarray slide. The sample DNA is then be washed off, leaving the oligonucleotide templates on the glass slide. The DNA in the wash is then used in sequencing library preparation [52].

Another targeting protocol, liquid hybridization, uses free-floating RNA templates in solution (generated on an RNA microarray) to select sample regions. The RNA templates are themselves bound to a capture-epitope, such as biotin. In this manner, a genomic region of interest will bind strongly to its complementary biotin-conjugated

RNA oligonucleotide template, producing an RNA-DNA duplex with stronger binding between strands than observed between RNA-RNA and DNA-DNA duplexes. The RNA-DNA duplex can then bind a streptavidin bead via the biotin-conjugate. The streptavidin bead, which is magnetic, can be extracted using a strong Neodymium magnet. This procedure, used in the Agilent SureSelect and Illumina TruSeq target enrichment platforms, is easily scalable to hundreds of samples in parallel [53-55].

Using targeted sequencing, studies have been able to focus on smaller, more functional (or more easily understood) portions of the genome, such as coding regions. Coding regions account for less than 2% of the human genome. By restricting studies to such small regions, it possible then to expand the number of samples analyzed 50-fold, at roughly the same cost of sequencing. The disadvantage to such an approach of course is that this restricted focus risks missing truly etiologic variants in the regions that are not analyzed. This can only be overcome when the speed and cost of whole-genome sequencing becomes comparable to that of targeted sequencing.

1.6 Computational Approaches to Identify Mutations

Variant detection from high-throughput sequencing is built on algorithms that align short reads to a reference genome and identify areas of genetic variation in relation to that reference. Numerous software platforms for alignment and variant calling have rapidly evolved [56]. This includes faster, more accurate alignment algorithms, such as Bowtie2, BWA and SOAP, and improvements in quality control functions, such as base recalibration, indel realignment, and duplicate removal, through software such as the

GATK and Picard [57]. Variant callers have improved the reliability and quality of variant calls with better algorithms and by training themselves across multiple samples [58].

These technical improvements, however, do not solve the problem of identifying just the handful of variants relevant to disease in an individual from the thousands of variants produced by sequencing. Where possible, it is useful to filter out variants that are not relevant to the disease being studied. Such filtering methods include using publicly available databases and family-based data to remove variants that are likely non-pathological. Other methods, particularly useful in prioritizing variants, include predicting variant function, through analyses of conservation of amino acid changes, identifying splicing changes, and predicting non-genic effects of variants on regulatory elements, miRNA binding sites, splicing enhancer sites, etc. [59-61]. These methods will be discussed in greater depth in the next chapter.

For much larger studies of complex disorders, a number of genome-wide statistical approaches can be implemented to identify mutation burdens in specific genes, gene sets, or interaction pathways [62,63]. For family-based studies, additional statistical approaches are available including genotype-phenotype correlations and linkage analysis. In all sequencing studies, it is now possible to conduct genome-wide rare variant association tests, with consideration of pedigrees if family-based data is included in the study design [64].

1.7 Functional Studies of Mutations

Once potential pathologic variants have been enriched and prioritized, they must be studied functionally to determine a molecular mechanism of disease and to identify potential therapeutic targets. The type of functional study performed is contingent on the nature of the mutation identified. Variants in non-coding regions will generally require initial expression profile assays to determine changes in gene transcript levels in specific cellular and tissue types. Coding changes will generally require initial assays identifying changes in protein structure or function. The options for experimentation are vast, modifiable, and as a general subject, too broad for discussion in this dissertation. Relevant functional studies for variants identified in this dissertation will be discussed with specifics in forthcoming chapters. These studies include protein-level *in vitro* assays, such as Yeast-Two-Hybrid, cellular-level assays, such as neuronal morphology and synaptic receptor recycling assays, and organismal-level assays, such as histological and behavioral tests of model organisms.

Two model organisms are of particular note for this dissertation. Zebrafish (*Danio rerio)* has emerged as an excellent model organism for studying developmental biology. Zebrafish possess a number of innate characteristics that make them a highly versatile tool, including embryonic transparency, allowing for *in vivo* imaging of living tissue during early development, and rapid generation scales with up to 100 embryos in three months from a single mating. Additionally, microinjection techniques have made it possible to quickly generate transgenic, knockdown, and exogenous plasmid expression

lines to study the function of individual genes, and more importantly, individual mutations [65].

The mouse (*Mus musculus*) has been one of the most robust model organisms for studying human disease, due to its strong similarity to humans in genomics (90% similarity between mouse and human genomes), development, and anatomical features [66]. Though the ease of use and cost-effectiveness of using mice is not on par with zebrafish, their closer biological similarity to humans allows for a more accurate translation of observation made between our species and theirs. To this end, an expansive array of experimental protocols are available for studying neurodevelopmental disease in mice, from *in vivo* electrophysiology, optogenetics, constitutive and conditional transgenic lines, and a wide range of behavioral protocols to test changes in memory, motor function, and psychiatric endophenotypes [67]. These organisms will be discussed in greater detail in the forthcoming chapters in the context of studying the function of specific genes in X-Linked Intellectual Disability and autism.

1.8 Establishing a Molecular Mechanism of Disease

In order to establish a gene as relevant for a particular disease, it is necessary to follow a strictly logical path. The gene or gene pathway must be determined to be mutated in disease cases. The nature of these mutations, though not immediately apparent upon their discovery, must act differently than the nature of mutations identified in the same gene or gene pathway for unaffected controls. The disease-associated mutations must perturb a particular biological pathway in the appropriate tissues and organs. This

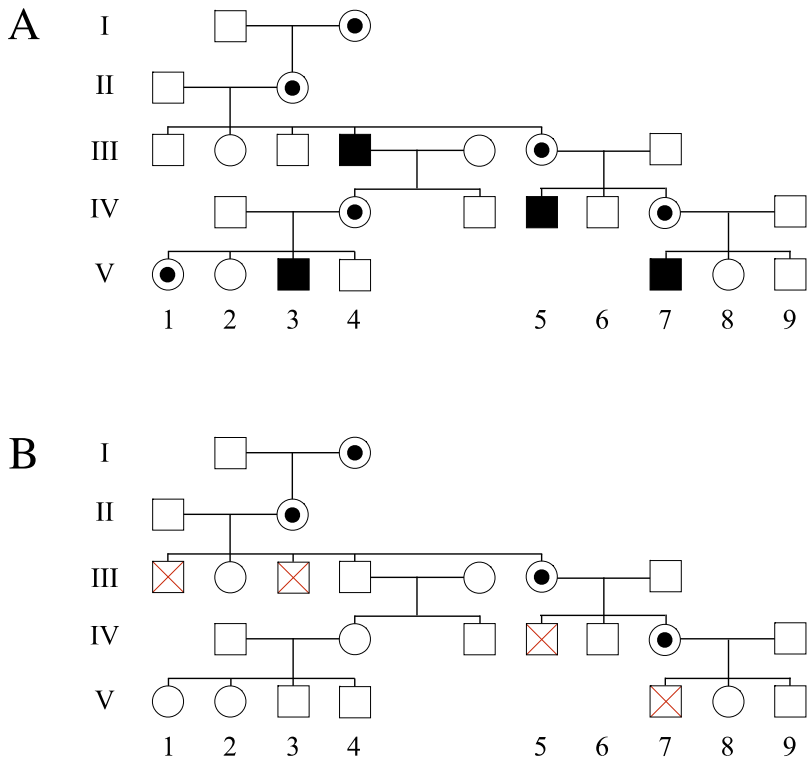
altered pathway then influences physiology and behavior in a pattern consistent with disease phenotype. Genetic evidence leads to functional evidence, which leads to clinical evidence, thereby establishing a molecular mechanism of disease.

In the following chapters, using this model, I will detail the technical and functional work in discovering variants and determining their role in X-Linked Intellectual Disability and in autism. Chapter Two, adapted from the publication "Affected Kindred Analysis of Human X Chromosome Exomes to Identify Novel X-Linked Intellectual Disability Genes," by Niranjan et al., describes an effective computational method for analyzing next-generation sequencing data to identify potential pathologic variants in a Mendelian disorder with high locus heterogeneity [68]. Additional data will provide functional support for some of the genes implicated.

Chapter Three, adapted from the publication "Effective Detection of Rare Variants in Pooled DNA Samples Using Cross-pool Tailcurve Analysis," by Niranjan et al., will delve deeper into the computational techniques for analyzing raw next-generation sequencing data, and how such techniques can be adapted to screen for mutations in larger sample cohorts [51]. Chapter Four discusses functional work performed in analyzing mutations identified in Chapter Three, and the role that these mutations may play in modulating glutamate signaling in autism through the synaptic scaffolding genes, Glutamate Receptors Interacting Proteins (*GRIP1/2)*.

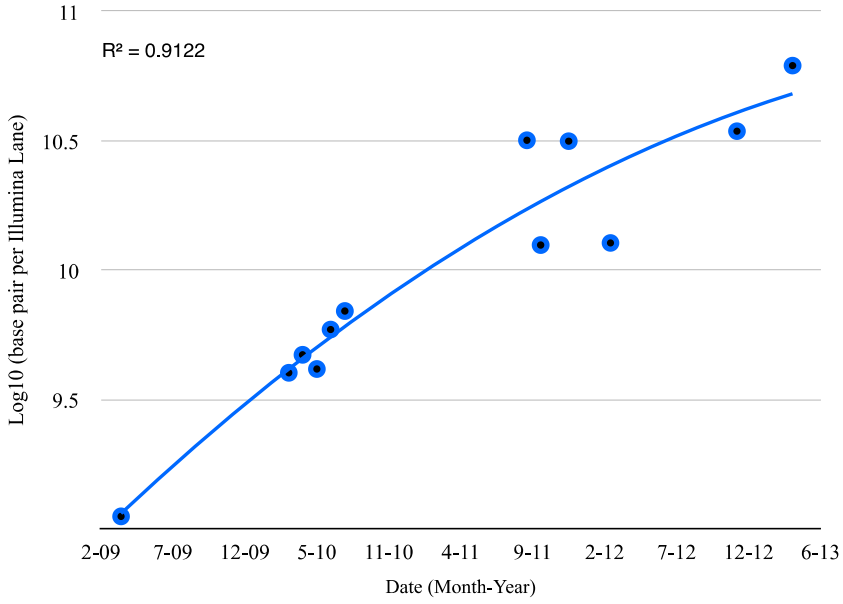
1.9 Figures: Chapter 1

Figure 1-1. Sample pedigrees of X-linked recessive traits



A five-generation pedigree is described. Females (circles) carrying the mutant allele are shown with a central dot. Affected males are shown with filled-in squares or red crosses. Panel A describes an X-linked recessive trait that is non-lethal. Hemizygous males inheriting the mutant allele from a carrier mother display the phenotype. Panel B describes an X-linked recessive trait that is lethal or reduces reproductive fitness. Affected males (red crossed squares) do not pass the trait on to later generations, but carrier females do.

Figure 1-2. Exponential increase in sequencing capacity over time



An exponential (log-scale base-10) increase in sequencing capacity using the Illumina platform is described. Each point represents the average number of base pairs (paired-end sequencing) obtained in a single lane of an Illumina sequencing machine (GA, GAII, HiSeq 2000) in individual sequencing experiments. From early 2009 to mid- and late 2013, a five year period, sequencing capacity increased 100-fold from ~1E9 bp per lane to almost 1E11 bp per lane. Though the rate of increase has tapered off more recently, this expansion in capacity has allowed for dynamic changes in methods of study design.

Chapter 2: Affected Kindred Analysis of Human X Chromosome Exomes to Identify Novel X-Linked Intellectual Disability Genes

X-linked Intellectual Disability (XLID) is a group of genetically heterogeneous disorders caused by mutations in genes on the X chromosome. Deleterious mutations in ~10% of X chromosome genes are implicated in causing XLID disorders in ~50% of known and suspected XLID families. The remaining XLID genes are expected to be rare and even private to individual families. To systematically identify these XLID genes, we sequenced the X chromosome exome (X-exome) in 56 well-established XLID families (a single affected male from 30 families and two affected males from 26 families) using an Agilent SureSelect X-exome kit and the Illumina HiSeq 2000 platform. To enrich for disease-causing mutations, we first utilized variant filters based on dbSNP, the male-restricted portions of the 1000 Genomes Project, or the Exome Variant Server datasets. However, these databases present limitations as automatic filters for enrichment of XLID genes. We therefore developed and optimized a strategy that uses a cohort of affected male kindred pairs and an additional small cohort of affected unrelated males to enrich for potentially pathological variants and to remove neutral variants. This strategy, which we refer to as Affected Kindred/Cross-Cohort Analysis, achieves a substantial enrichment for potentially pathological variants in known XLID genes compared to variant filters from public reference databases, and it has identified novel XLID candidate genes. We conclude that Affected Kindred/Cross-Cohort Analysis can effectively enrich

for disease-causing genes in rare, Mendelian disorders, and that public reference databases can be used effectively, but cautiously, as automatic filters for X-linked disorders [68].

2.1 Introduction

X-linked Intellectual Disability (XLID) is a group of genetically highly heterogeneous disorders with mutations in genes on the X chromosome [69-71]. With the characterization of relatively common XLID genes, it is expected that the majority of the remaining mutations in unknown XLID genes are very rare and even private to individual patients and families [72]. Identification of these XLID genes is essential to provide accurate molecular diagnosis for individual XLID families and to better understand the molecular basis of intellectual function and disability in humans [70,71]. The extreme rarity and vast genetic heterogeneity of the individual XLID disorders pose a significant challenge, because ~10% of more than 1,000 annotated X-linked genes have already been implicated to cause half of all XLID disorders [69]. If these 10% of genes reflect the "low-hanging fruit," then the remaining half of XLID cases without a known genetic cause will likely involve even more rare mutations affecting a broader range of genes, making them more difficult to isolate.

The rapid developments in high-throughput sequencing platforms that are coupled with effective targeted capture have made it possible to determine nearly all coding variants that present in an individual human genome [73,74]. Exome-based sequencing has become a powerful approach to elucidate the genetic basis of Mendelian disorders of

unknown etiology and provide gene diagnoses of specific disorders with high genetic and phenotypic heterogeneity [75-78]. This strategy has had some success in identifying a handful of additional causes for autosomal intellectual disability [79,80]. The most significant challenge in using this strategy is in differentiating disease-causing (pathological) mutations from the large quantity of non-causal (neutral) variants. One may expect in any large scale exome sequencing study for approximately 20,000-24,000 variants to be found in an individual exome, with ~10%, coming from the X chromosome [9,75].

Common strategies for identifying causal mutations in rare Mendelian disorders include sequencing proband patients with the same phenotype from multiple families [81,82], filtering out neutral variants using large databases such as dbSNP and the 1000 Genomes project [83-85], predicting functional relevance of variants using bioinformatics software such as SIFT [60] and PolyPhen-2 [61], conducting segregation analysis in proband families, and correlating known or predicted function of the candidate genes with the disease phenotype.

The success of these approaches relies on a number of factors: (1) the recruitment of multiple families with the same phenotype of interest, which can prove challenging for very rare Mendelian disorders with high locus heterogeneity and a wide spectrum of phenotypic expressivity; (2) the reliability of public variant databases, which are presumably generated from individuals lacking the disease under study, and would therefore only contain a pool of neutral variants for a given phenotype; (3) the reliability of bioinformatics tools in predicting variant significance; (4) a burden of pathological

mutations in the functional portions of genes that are targeted for sequencing, like coding exons and splice sites, with less dependence on mutations occurring in poorly covered regions like regulatory elements; and (5) the extent of knowledge available on a candidate gene in mechanistically linking its biological functions to the phenotype.

To identify rare and causal mutations in heterogeneous X-linked Mendelian disorders, it is essential to utilize filters to remove the majority of neutral variants and sequencing errors in order to focus on potential pathological variants. Public databases such as dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP), the 1000 Genomes dataset (http://www.1000genomes.org), and the Exome Variant Server (EVS, http://evs.gs.washington.edu/EVS) have been used extensively as discrete variant filters for many studies [83-86]. The expected rarity of individual causal mutations in novel XLID makes it reasonable to eliminate common polymorphisms above a given frequency using data from these databases. Furthermore, it is generally assumed that public databases such as 1000 Genomes consist of individuals devoid of the phenotype of interest, and thereby serve as a public "normal control" set. The free and easy assess to these large "control" datasets makes them the top choices for many small scale genetic studies [83-86].

Our study aims to identify causal mutations in novel XLID genes using X chromosome exome sequencing. We systematically evaluated variant data from dbSNP, the 1000 Genomes, and EVS as discrete filters to determine how effectively each could reduce the number of neutral variants from our sequenced cohort. In doing so, we recognized that these databases present with limitations in their current forms as

automatic filters to enrich for causal XLID genes. We therefore developed and optimized a strategy using affected male kindred pairs and affected unrelated males to enrich for potentially pathological variants and to remove neutral variants. Our study shows that this Affected Kindred/Cross-Cohort strategy achieves a substantial reduction in variants compared to the public database-dependent discrete filters alone. Importantly, our study shows that this strategy could achieve a significant enrichment for known and candidate XLID genes.

2.2 Results

2.2.1 Study Sample from X Chromosome Exome Sequencing

Genomic DNA samples from males with XLID (n = 82) were sequenced. Of the 82 samples, 30 are single male probands from unrelated XLID families; the remaining 52 samples constitute 26 affected pairs including full brothers, maternal male cousins, and maternal uncle and nephew pairs (Table 2-1). The relationships of the affected pairs were validated by a calculation of the relatedness between samples in the entire study cohort (Figure 2-1). Approximately 79.9% of the target regions are covered at ≥4x read depth across all samples with an average read depth of 49x. Variant calling on whole-genome aligned reads generated an average of 1,774 ± 239 variants per sample on or near target regions, and 14.7% of these variants are non-synonymous or are present at conserved splice junctions. Several discrete filters were then applied individually or cumulatively to enrich for potential causal variants. The total variant compositions prior to and after each

filtering step are shown in Table 2-2.

2.2.2 Assessment of Relatedness between Samples and Population Stratification

The relationships between affected pairs were validated by estimating relatedness across samples in the entire study cohort (Figure 2-1). Additionally, relatedness of samples was assessed by ascertainment of regions of IBD (shared segments of genotypes inherited Identical By Descent) between all samples, using a combination of an automated 5 MB sliding window detector across IBD genotypes and by manual curation (Figure 2-2). By this process, no IBD sharing was detectable between samples known to be unrelated. For a more refined analysis of genetic sharing, we plotted the correlation of known linked SNPs to genotype sharing (Figure 2-3). SNPs that are known to be in linkage were found to never cross into boundaries defined as regions of IBD from Figure 2-2. This suggests that recombination sites consistent with known linkage intervals define the boundaries for shared segments between related samples in our cohort. However, only higher resolution, full chromosome genotyping (not exome sequencing) can prove this conclusively. Based on these analyses, an estimation of relatedness across all sample genotypes, detection of regions of IBD across all samples, and definition of these IBD regions around known linked SNPs, the expected relatedness between samples is likely accurate.

Additionally, we looked for population stratification, in the event that large background population deviation may influence downstream analysis, particularly when using some public database-dependent filters. All samples in our cohort are expected to

be of European descent. We compared the degree of deviation of individual samples from the cohort genotype mean, as well as the load of SNPs at high frequency in the European American population, as obtained from EVS (Figure 2-4). As expected, the majority of samples clustered together with genotypes of primarily European ancestry. However, five samples showed slight deviations from the main cluster. When SNP loads were compared to EVS data from the African American population, the sample cluster deviation was reproduced, indicating that these five samples have a small, but detectable contribution of African ancestry. We do not believe this marginal stratification will affect downstream analysis, due in part to the limited degree of deviation, the large presence of both African and European ancestry populations in public databases, and our desire to find rare mutations, which are less likely to be affected by genetic background.

2.2.3 Variant Filtering Using Strand and Proximity Metrics

We have previously observed that many false positive variant calls can be efficiently and specifically removed by applying two pre-filters determined by the proportion of base call from opposing strands and by inter-variant proximity [51]. Firstly, during read alignment, sequenced reads may be aligned to either the positive (Crick) or negative (Watson) strand. Variant base calls are therefore made in relation to either the positive or negative strand. An excess of variant base calls from one strand over another often characterizes false positive variants. The strand-based pre-filter eliminates variant calls that are not represented by at least one variant base call on each strand. Secondly, false positive variant calls often aggregate in close proximity. Generally, we should not

observe more than two variants in a 1,000 bp stretch of DNA in a single individual [87]; even less should be observed for more evolutionarily conserved regions, like exons [88]. The proximity-based pre-filter very conservatively removes variant calls if they are present within 10 bp of each other. The consequence of these pre-filters is a substantial reduction in the frequency of erroneous variants that would otherwise confound downstream analysis (Figure 2-5). Both of these pre-filters are applied universally, prior to any other filter method, described below (Table 2-2, Row 1).

2.2.4 Variant Filtering Using dbSNP (Build 137)

To enrich potential disease-causing variants for further genetic and functional studies, we tested multiple filtering methods. Numerous published studies have made use of variants in dbSNP to filter out variants of non-clinical significance [81,82,89]. Build 137 of dbSNP includes variants of known pathological significance; we therefore selectively removed dbSNP variants that are present in the CLINVAR database (http://www.ncbi.nlm.nih.gov/clinvar; CLINSIG = 4 [probable-pathogenic] or 5 [pathogenic]). Prior removal of these variants of known pathological significance from dbSNP is important, as there may be overlap between these pathological variants and disease-causing variants in our study cohort. Retention of these pathological variants in dbSNP would result in their removal from our cohort, creating a false negative.

This abridged dbSNP dataset is still large and will likely successfully filter out many non-pathological variants. However, we expect using this method will be flawed to the extent that most pathological variants in dbSNP are not known nor annotated, and are

therefore retained in this abridged dbSNP. Such unannotated pathological variants may overlap with important variants of interest in our study; such variants would be unintentionally filtered out, resulting in a higher false negative rate. To assess the fraction of remaining unannotated variants that may be potentially pathological, we identified the predicted truncating mutations (nonsense and frame-shift) that occur transcriptionally upstream of known deleterious mutations. Of the 27,242 coding variants present in our abridged dbSNP, 100 (or 0.37%) fit the above description. This is a highly conservative estimation, as it does not account for additional downstream nonsense and frame-shift mutations. Importantly, it does not account for missense and splicing changes that likely constitute a much larger fraction of remaining deleterious variants in known disease genes.

Given the rarity of XLID, potential XLID variants that are present but not annotated as pathological in dbSNP should have a very low minor allele frequency (MAF). Variants below a specified MAF could be removed from our abridged dbSNP filter, thereby ensuring we are not filtering out potentially pathological variants in our cohort. A frequent cutoff for rare variants is an MAF of 1%. However, we determined a more accurate MAF cutoff could be derived for male-restricted X chromosome variants of interest, because the majority of MAFs for dbSNP are derived from the 1000 Genomes Project. There are 525 male X chromosomes and 1,134 (2 x 567) female X chromosomes in 1000 Genomes (1,659 X chromosomes total). An appropriate MAF cutoff would minimize the probability that an unannotated pathological variant exists only in unaffected female carriers and never exists in unaffected males in 1000 Genomes. We

calculated that if a variant is present in ≥12 of the 1,659 copies of the X chromosome (MAF ≤ 0.73%) in 1000 Genomes, the probability that all 12+ variant copies are present in only female carriers is ≤1% (Table 2-3). We therefore chose to remove from our abridged dbSNP dataset all variants with a 1000 Genomes MAF ≤ 0.73%, to mitigate the unintended loss of unannotated pathological variants from our study cohort during filtration.

This new, "Non-Clinical" dbSNP Filter, redacted of known pathological variants and variants of low MAF, when used on our XLID cohort, results in a substantial 93.6% reduction in the number of variants for further analysis (Table 2-2, Row 5).

2.2.5 Variant Filtering Using 1000 Genomes

Individuals in the 1000 Genomes came from ethnically diverse populations and are not expected to have severe intellectual disability. The master variant output for the 1000 Genomes project (Integrated Phase 1, version 3: 20101123) includes calls made from both males and females. Females could potentially possess an XLID mutation in the heterozygous state (carrier status) without clinical phenotype, while males are not expected to carry disease-causing mutations for XLID in the hemizygous state. We reason that variant data from males in the 1000 Genomes can be used as a filter to reduce the number of neutral variants in the X-exome sequenced from males in our XLID study. We thus generated a male-only variant dataset (n=525) by removing all X chromosome variant calls made only in the female portion of the 1000 Genomes.

Publicly accessible variant data output from the 1000 Genomes consist of low

coverage and exome variant calling including SNPs and indels at the current stage. This data was generated from multiple parallel pipelines at different sequencing and data analysis centers, and then merged into one VCF file [90]. For initial assessment of this dataset as a variant filter in our project, we analyzed the composition of variant genotypes for all the chromosomes from the 1000 Genomes project. Two unexpected discrepancies were noted in our analysis, which preclude the immediate use of the 1000 Genomes dataset as a filter.

Firstly, we observed the presence of ambiguous genotypes among male variant calls. An individual can have a genotype (sum of variant alleles at variant position) of 0 (no variant), 1 (heterozygous), or 2 (homozygous) as compared to the established reference. Because the X chromosome is in hemizygous state in males, we anticipate that genotypes for the male X chromosome should only be assigned as 0 or 1. However, we observed many ambiguous genotypes, including 2, 3, and 4, as well as non-integer genotypes ranging from 0.05 to 3.95 in a substantial fraction of variant calls (Table 2-4). Ambiguous genotypes account for ~9% of non-zero genotypes and have integer and non-integer values ranging from 0.05 to 4. A non-integer value for the number of alleles of a variant should not occur, with the exception of somatic mutations. The restricted presence of these ambiguous non-integer values to only male X chromosome variants, and the complete lack of such values among autosomal or female X chromosome variants, suggests that they are erroneous.

Secondly, we assessed the concordance between 1000 Genomes variant calls and SNP genotyping (Omni Platform), to approximate the accuracy of the variant output.

Concordance for the X chromosome was calculated to be 68.6 ± 4.7%, which is much lower than expected [90]. To clarify this discrepancy, we generated variant calls directly from aligned sequence data of 162 random males in the 1000 Genomes using the Unified Genotyper and compared it to the Omni genotype data. Correlation between these 162 samples and their respective Omni genotypes was calculated at 98.8 ± 0.4%, which is comparable to what has been previously reported [90]. Given these apparent discrepancies, we have chosen not to use the current publicly accessible build of the 1000 Genomes as an automatic variant filter in our XLID project. Instead, we use the internally generated genotypes from the 162 male samples. This [1000G] Male-162 Internal Exome Filter is the least efficient discrete filter in our analysis, reducing the average variant count per sample by only 25.2% (Table 2-2, Row 3).

2.2.6 Variant Filtering using the Exome Variant Server Dataset

The Exome Variant Server (EVS) dataset was generated through the NHLBI Exome Sequencing Project, by sequencing more than 5,000 exomes for samples collected as "healthy" controls or samples diagnosed with a heart, lung, or blood disorder. Samples with an XLID should be rare, if present at all. Approximately half of EVS samples are male. For the X chromosome, male genotypes for the non-pseudoautosomal regions can be distinguished from female genotypes, because male genotypes are explicitly annotated as hemizygous or wildtype. This EVS (Male Only) Filter, much like the [1000G] Male-162 Internal Exome Filter, is based on the assumption that none of the variants in these "control" populations will contribute to XLID, and can therefore be subtracted from the

variant list in our XLID cohort.

This filter performs much better than using the [1000G] Male-162 Internal Exome Filter (40.3% reduction in variants), but does not outperform the dbSNP filter (Table 2-2, Row 4). For our analysis, this database is presumably the most reliable of the three tested, as it is consistent in its genotyping and the population in question should not have any overlapping disease traits with our analysis. However, we anticipate that the use of the EVS cohort, which is still a disease population (heart, lung, and blood disorders), will complicate other X-linked studies examining similar disease traits.

2.2.7 Variant Filtering by Relatedness

The rarity of XLID and the distinct inheritance (no male-to-male transmission) predict that that the remaining XLID mutations are likely very rare and even private to individual families and are therefore unlikely to be shared with unrelated families (with rare exceptions) [72]. Based on this prediction, we have developed and implemented a filter based on relatedness, which we call the Affected Kindred/Cross-Cohort Filter.

One component of this filter, referred to as the Shared Segment Filter, retains all variants present within regions observed as shared (Identical by Descent) between affected related samples in the cohort (Figure 2-5). This is a relatively simple step that results in a quick reduction in variants, while simultaneously providing confirmation of sample relatedness. When this step is applied alone, it results in a 29% reduction in potentially neutral variants (Table 2-2, Row 2).

The major component to the Affected Kindred/Cross-Cohort Filter involves a

more sensitive implementation based on relatedness than the Shared Segment Filter. In this step, variants shared between two affected individuals from the same family are retained, while variants shared between two unrelated individuals in the small study cohort are removed. We constructed this filter to accommodate for variants that are discordant between affected kindred pairs where one variant is absent due to insufficient coverage, by simply retaining the variant with sufficient coverage. A schematic describing all steps from alignment to filtering using the Affected Kindred/Cross-Cohort filter is provided (Figure 2-6).

Though the Affected Kindred/Cross-Cohort Filter is more sensitive than the Shared Segment Filter alone, it will not necessarily remove all the same passenger variants as the Shared Segment Filter. For example, shared common variants in the unshared chromosomal region of two affected kindred pairs will be retained by the Affected Kindred/Cross-Cohort Filter, because this filter is not aware if shared variants that are identical by state are also identical by descent. This step of the Affected Kindred/Cross-Cohort Filter results in the largest reduction (96.5%) in potentially neutral variants of all the individual filters (Table 2-2, Row 6).

Though the Affected Kindred/Cross-Cohort Filter relies solely on in-house data, it differs from other in-house filtering platforms that have been used previously. Importantly, this filter does not rely on using a separate negative control cohort of unaffected, unrelated individuals. Rather it relies on variant comparisons conducted within the same affected cohort, which is more akin to using an internal control. This has a number of unique advantages. Firstly, any reoccurring systematic errors that occur

during library preparation or sequencing will self-neutralize. Secondly, more sequencing capacity can be devoted away from unaffected samples to affected samples, improving the power to detect disease-causing mutations. If at least some of this additional capacity is devoted to sequencing affected related sample pairs, that should further improves detection power. The main advantage of such a filter without additional controls is having greater detection power, which is one of the primary challenges when studying a disease with substantial locus heterogeneity.

2.2.8 Comparison of Filters

Each filtering strategy above has its advantages and limitations. The advantage of dbSNP as a filter is that it is the largest annotated collection of variants available, and therefore removes the most variants from our cohort out of all the public database-derived filters (93.6% reduction; Table 2-2, Row 5). However, appropriate use of the database requires that it be modified to exclude as many known or possible pathological variants, or such variants risk being filtered out of the variant list of the study cohort. In this study, we removed known or probably pathological variants based on prior clinical annotation or MAF. Even so, this "Non-Clinical" dbSNP may still contain potentially functional variants relevant to our study cohort, and therefore this filter should be used cautiously.

Variants from males in the 1000 Genomes should serve as a valuable filter to remove neutral variants in our study. However, the publicly accessible dataset for males in the 1000 Genomes (n = 525) appears to contain ambiguous genotypes unique to

variants from X-linked genes, and is therefore not suitable as an automatic filter to enrich for causal variants for rare X-linked disorders. To better assess whether this dataset can be used as a filter in our study, we generated variant calls directly from aligned sequence data for 162 males in the 1000 Genomes using the Unified Genotyper. The resulting variant dataset shows an excellent correlation with SNP genotype data (Omni Platform) made publicly available through the 1000 Genomes Project. Filtering using this smaller variant dataset succeeded in removing 25.2% of potentially neutral variants (Table 2-2, Row 3). We expect that a similar dataset for all sequenced males in the 1000 Genomes project (n = 525) will likely achieve a greater reduction of neutral variants. Consistent with this expectation, filtering using the much larger EVS dataset (Table 2-2, Row 4) results in a 40.3% reduction in potentially neutral variants. We do not expect a full 1000 Genomes filter (males only) to outperform the "Non-Clinical" dbSNP filter, which consists of variants from both 1000 Genomes and alternate sources.

The Affected Kindred/Cross-Cohort Filter achieved the most substantial reduction in the average number of potentially neutral variants, demonstrating the robustness of this method (96.5% reduction; Table 2-2, Row 6). A large portion of the variants in dbSNP that have a MAF > 0.73% are re-discovered and removed given the sample size of our study, thus explaining why our filter outperforms the "Non-Clinical" dbSNP filter. Additionally, we have previously observed that false positive variant calls occur in parallel in multiple samples due to systematic errors in preparation or sequencing of the same library [51]. As an added advantage, our strategy removes these variants by comparing samples prepared in the same library, a process of self-neutralizing these

reoccurring systematic errors.

Compared to the discrete filters based on public variant databases, our Affected Kindred/Cross-Cohort Filter shows superior performance at the removal of non-pathological variants in this XLID study. Importantly, this performance is achieved using a cohort size ($n < 100$) much smaller than those used in the 1000 Genomes, EVS, or aggregated into dbSNP. Given this performance, we predict that the sequencing of a large cohort of unrelated "normal" controls is not necessary. An additional advantage is that the Affected Kindred/Cross-Cohort Filter works independently from public reference databases and will therefore not be affected by any of the limitations noted previously for the database-dependent filters.

However, there is the disadvantage that, because this filter relies on the rare and private nature of XLID mutations, a potential causal mutation may be lost in the event that two affected but unrelated individuals possess that same mutation. This may be a result of incorrect ascertainment of relatedness, or because a disease-causing mutation by chance arose independently in two unrelated families. These events can be minimized by the confirmation of relatedness between samples and by rescuing recurrent causal mutations using dbSNP database (see below).

2.2.9 Combination of Filters

Given these noted advantages and limitations, we tested one more filter that combines features from all of them. Firstly, the Shared Segment Filter and the Affected Kindred/Cross-Cohort Filter are applied sequentially. Secondly, we forcibly retain/re-

introduce any variants lost in the Affected Kindred/Cross-Cohort Filter that are annotated as pathological in dbSNP and have a MAF < 1%. This retention would prevent the loss of known, rare disease-causing mutations that are shared between unrelated samples in the study cohort. Lastly, the three discrete public database-dependent filters ("Non-Clinical" dbSNP, [1000G] Male-162 Internal Exome, and Exome Variant Server) are used. Application of the database-dependent filters does not remove a significant number of neutral variants (already achieved in the first step), but it is nonetheless a simple series of filters to apply. A schematic of this combined filter is provided (Figure 2-7).

The combination of these filters together results in 98.5% reduction in potentially neutral variants (Table 2-2, Row 7). Importantly, the number of variants for many samples is reduced to single-digit levels, thereby making downstream predictive and functional analyses much easier. Additionally, the re-introduction of pathological variants annotated in dbSNP succeeds in retaining a nonsense mutation, R37X, in *ATRX*. This mutation is a known cause for XLID, has been annotated as such in dbSNP, and occurs independently in our cohort in two unrelated individuals and one sibling pair. Because R37X in *ATRX* is shared between unrelated samples in our cohort, it was initially lost after application of the Affected Kindred/Cross-Cohort Filter. While such an event is rare due to broad allelic heterogeneity, retention of such a variant is preferable; these samples would otherwise be unnecessarily subjected to additional mutation discovery screens.

2.2.10 Variant Validation by Sanger Sequencing

From the final variant list of the combined filter, we selected 19 potentially

deleterious (17 coding and 2 splicing) variants and performed Sanger sequencing in 28 samples from the proband families. All 19 variants were positively confirmed in the respective 28 samples.

2.2.11 Enrichment of Potential Causal Mutations

Our current enrichment system targets 975 genes on the X chromosome, of which 103 (10.6%) are known to cause XLID when mutated [69,91]. In order to determine if our combined filtering system could enrich for XLID genes, we prioritized genes by mutation burden. The presence of at least one splicing or non-synonymous coding change in a sample would elevate the priority of a gene. Additional weight was given to nonsense and frame-shift mutations. No weight was given to intronic or synonymous mutations.

The resulting list consists of 89 X-linked genes (Table 2-5), of which 24 (27%) are previously associated with XLID. Though kindred pairs are responsible for only 33.6% of variants that were used to derive this list, they provided for a much greater enrichment (3.35-fold) of XLID genes compared to sporadic cases (1.39-fold). The expected fraction of XLID genes in the list should be 10.6%, not 27%. This 2.55-fold enrichment (kindred pairs and sporadic cases combined) for XLID genes over expected is statistically significant (p < 1E-5; hypergeometric test), demonstrating that our filtering system can be used quickly and effectively to re-identify known XLID genes. Given this enrichment, we hypothesize that the remaining portion of the list is also enriched for novel XLID genes. This approach has proven instrumental in identification of novel XLID genes, including one gene that was prioritized at the top of our list, *ZC4H2*; this

gene was recently implicated in Arthrogryposis Multiplex Congenita and Intellectual Disability [92]. A sample list of enriched genes and filtered variants is provided (Table 2-6).

2.3 Discussion

We developed a strategy for rapid filtering of high-throughput sequencing data to identify disease-causing mutations in a rare X-linked Mendelian disorder with extensive locus and allelic heterogeneity. In sequencing affected samples with known relatedness in a small cohort of fewer than 100 samples, our Affected Kindred/Cross-Cohort Filter removes likely non-pathological variants to a level greater than that achieved using the largest publicly available variant dataset (dbSNP) as a filter. Variant sets from public databases or large control cohorts, though easily applied, are not required for effective filtration. This feature is important due to the intrinsic limitations of some of the public datasets. However, these public databases can be used if modified appropriately to compensate for the intrinsic limitations, as described above.

Using a combination of filters, we found a statistically significant enrichment for known XLID genes, strongly indicating that our method can be used to enrich for known disease-causing genes. Multiple novel candidate genes were also identified in this study, many of which are likely etiologic based on known biological function (Table 2-5 and Table 2-6).

PLXNA3, or Plexin A3, with Semaphorin, is involved in chemotactic signaling, a pathway involved in normal targeting of axonal projections in the central nervous system

[93]. The Plexin-Semaphorin pathway has been previously implicated in an intellectual disability syndrome [94].

GRIPAP1, or *GRIP*-associated protein 1 (also known as *GRASP1*), is a neuronally expressed gene, with a role in glutamatergic AMPA receptor signaling, by regulating receptor trafficking and distribution through the Ras pathway [95]. *De novo* deleterious mutations in genes for glutamatergic signaling have been previously implicated in Intellectual Disability [96]. Additionally, glutamatergic receptors are dysregulated with loss of *FMR1* function, which is also a known cause of XLID [97,98]. *GRIPAP1* is located at Xp11, a region subject to duplication that has been previously associated with autism with severe intellectual disability [99,100].

EphrinB1 is ligand for Eph-related receptor tyrosine kinases and is involved in regulation of neuronal axon guidance [101]. Mutation of *EphrinB1* is a primary cause of Craniofrontonasal Syndrome, which can include symptoms of learning disability [102]. An *EphrinB1* deficient mouse model of Craniofrontonasal Syndrome shows cortical abnormalities and learning deficits [103].

OGT, or O-linked N-acetylglucosamine Transferase, is essential for post-translation modification of serine and threonine residues. Additionally, *OGT* forms a complex with TET proteins in the nucleus to regulate chromatin [104]. This complex includes *HCFC1*, a target of *OGT*, which is another gene implicated in XLID [105].

Our results demonstrate a robust reduction in variants and a significant enrichment for known and putative disease causing genes. In complementation with other mutation prioritization options like functional variant prediction analysis, the Affected

Kindred/Cross-Cohort Filter can identify causal mutations in rare, heterogeneous X-linked disorders such as XLID.

2.4 Materials and Methods

2.4.1 Study Sample

Genomic DNA samples from males with XLID (n = 82) were sequenced. Of the 82 samples, 30 are sporadic cases and are unrelated to other individuals in the study cohort. The remaining 52 samples constitute 26 kindred pairs, with relationships described in Table 2-1. X-linkage was determined by at least one of four criteria: 1) the responsible locus was mapped by linkage analysis or a similar method to the X chromosome; 2) affected males in two or more generations in their pedigree show a pattern of inheritance consistent with X-linkage; 3) two or more affected males in the same generation in their pedigree are consistent with X-linked inheritance, with evidence for a skewed (90:10) pattern of X-inactivation in females; 4) the presentation of disease is consistent with the clinical diagnosis of a known, well-defined XLID syndrome for which the causative gene is unknown. All the samples were previously tested as negative for Fragile X Syndrome, cytogenetic abnormalities, and known inborn errors of metabolism. An informed consent was obtained from all families enrolled in this study at the Greenwood Genetic Center, SC, and/or the Johns Hopkins University. The Institutional Review Board from the respective institutions approved this study.

2.4.2 Library Preparation and Sequencing

Sequencing libraries were prepared using the TruSeq™ DNA Sample Preparation kit following a standard protocol from the manufacturer (Illumina). Twelve or 24 individually indexed libraries were pooled at equal molar ratio and enriched for the X-exome using a SureSelect Human X Chromosome Exome Kit (Agilent). Each pooled library was quantified by qPCR using a KAPA library quantification kit (KAPA Biosystems) and sequenced in one lane of HiSeq2000 using 75bp pair-end sequence module.

2.4.3 Sequence Data Analysis

Bowtie2 was used to align fastq reads using the [--very-sensitive-local] parameter and other Bowtie2 parameters were kept at default [106]. PCR duplicates were removed using Picard (http://picard.sourceforge.net). Indel realignment and base recalibration was conducted using GATK [107]. Unified Genotyper (GATK) was used for variant calling with the ploidy parameter setting at "1" (haploid) due to the hemizygous nature of the male X chromosome [58]. The pseudoautosomal regions of the X chromosome were not included in our study. Additional pre-filtering parameters were instituted based on strand-specific coverage and variant proximity to reduce the false positive rate in variant calling [51]. By comparing variant calls from the Unified Genotyper to nucleotide pileups from the aligned reads, this filter only retains variants with at least one alternate base call from each strand and only retains variants that are not present within ten nucleotides of another variant. This filter is applied in addition to the default *FisherStrand* covariate analysis

conducted during base recalibration by GATK. Variants were annotated for affected genes and coding changes using the ANNOVAR package [108].

2.4.4 Estimation of the Relatedness between Samples

The relatedness between samples was determined based on the fraction of total variant calls that are identical between each sample in the entire cohort [Identity ≈ (2 x number of variants identical between both samples)/(sum of variants of both samples)]. The closely related samples such as affected brothers and maternal cousins from the sample families generally shared the greatest identity, which confirms their expected relatedness (Figure 2-1). Relatedness was also validated visually using the Shared Segment filter step of the Affect Kindred/Cross-Cohort Filter (Figure 2-2).

2.4.5 Identification of Shared Segment that are Identical By Descent

Identification of regions of sharing between related samples was conducted using an automated 5 MB sliding window across all possible section of IBD. The regions defined by the sliding window were then manually refined. To determine if shared IBD segments align along known linkage intervals, SNPs that are in linkage were assessed for crossover along the boundaries of the shared segments. Retrieval of linked SNPs was conducted using PLINK v1.07 [plink --bfile hapmap3_r2_b36_fwd.consensus.qc.poly --blocks --noweb --chr X --from-bp 1 --to-bp 155270560 --missing-phenotype 1] on HapMap Phase 3 data (MAP and PED files) [62,109].

2.4.6 Determination of Population Stratification

Population stratification was determined by comparing samples to a cohort mean, and by comparing sample allele frequencies to population-specific allele frequencies derived from EVS data. For the cohort mean analysis, an average allele frequency for the cohort was obtained for each variant. A residual sum of squares was subsequently calculated for each sample compared to the cohort mean. The number of standard deviations from the residual is plotted on the Y-axis of Figure 2-4a. Comparison of sample allele frequencies to EVS was performed for both EVS European American (EA) and EVS African American (AA) population frequencies. For any given sample, an EA and an AA metric was obtained by averaging the population specific allele frequencies for all sample variants present in the EVS dataset. This EA metric for each sample is plotted on the X-axis of Figure 2-4. The AA metric for each sample is plotted on the Y-axis of Figure 2-4b.

2.4.7 Hypergeometric test for Known XLID Gene Enrichment

The full X chromosome gene set targeted by exome sequencing includes 975 coding genes, of which 103 are attributed to XLID in literature. Our combined filtering pipeline resulted in a large reduction in likely non-causal variants. When these variants are scored by function (nonsense, splicing, missense), we obtained a list of 89 prioritized genes, of which 24 are attributed to XLID in literature. A hypergeometric test was used to determined the probability of obtaining at least 24 (k) known XLID genes with 89 (n) random draws from a population of 975 (N) genes, of which only 103 (K) are known

XLID genes. The p-value for an over-representation of known XLID genes is the probability of randomly drawing *k* or more XLID genes with a total of *n* genes drawn from the total population of *K* in *N*, without replacement. (R code: [sum(dhyper(*k*:*n*,*K*,*N*-*K*,*n*))] or [1-phyper(*k*-1,*K*,*N*-*K*,*n*)]). This p-value was calculated to be less than 4E-6. As a negative control, we calculated the mutation frequency for each X chromosome gene, using our 162 male 1000G negative control cohort. It is necessary to take into consideration the mutation rate for each gene, as some XLID genes, such as DMD and MECP2 have higher mutation rates, which will increase the likelihood of enrichment. The probability of randomly selecting an XLID gene from the total X chromosome gene set, weighted for genic mutation rate, was calculated to be 0.103408.

2.4.8 Validation of Variants by Sanger Sequencing

Sanger sequencing for variant validation was conducted using the BigDye Terminator v3.1 Cycle Sequencing Kit on an ABI3100 automatic DNA analyzer (Applied Biosystems) following manufacturer's instructions. Analysis was done on standard sequence alignment software (CodonCode and MacVector) followed by manual investigations of the chromatograms.

2.4.9 "Non-Clinical" dbSNP, Male 1000 Genomes, and Male EVS as Variant Filters

A "Non-Clinical" dbSNP dataset was generated by the removal of variants that are present in the clinvar databases (CLINSIG = 4 [probable-pathogenic] or 5 [pathogenic]) from dbSNP (build 137), followed by the removal of variants with MAF <

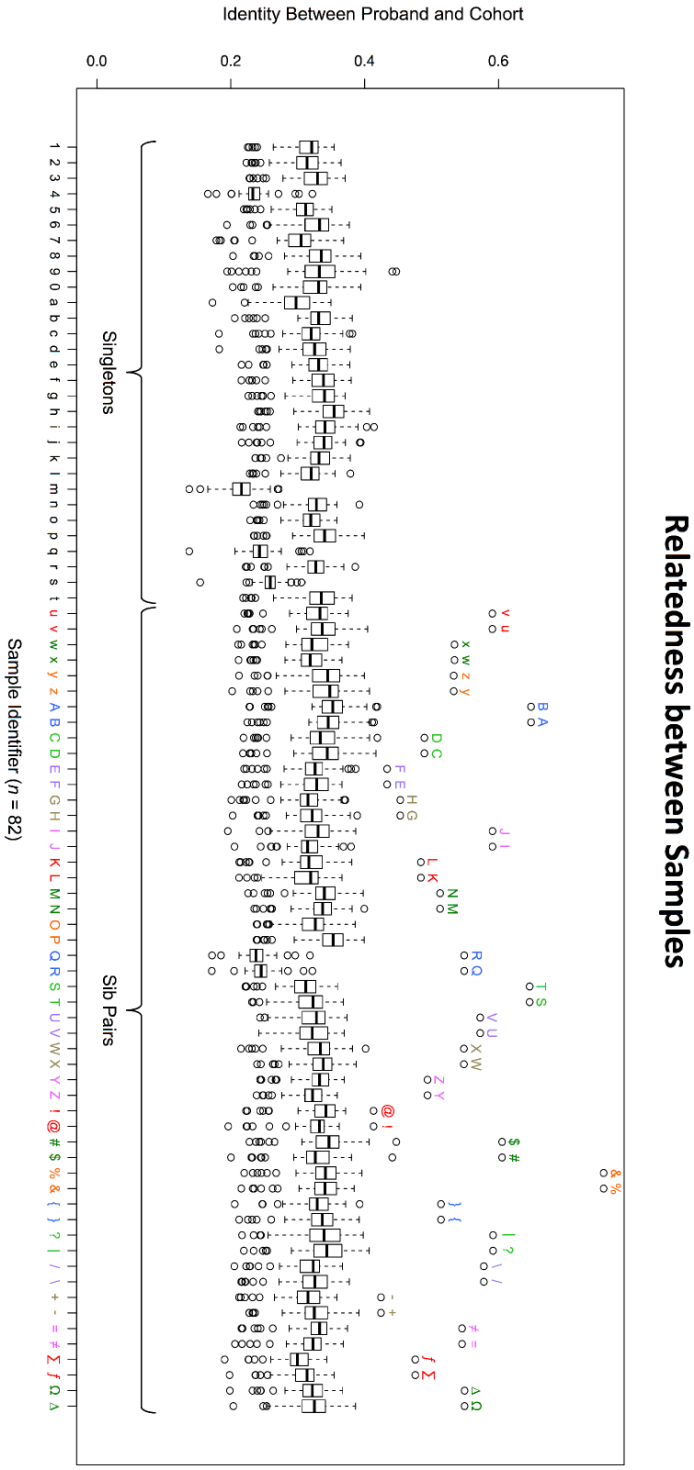
0.73%. For the 1000 Genomes project, variant data (SNPs and indels) from male samples only (n = 525) were extracted from the master variant output of the 1000 Genomes project (Integrated Phase 1, version 3: 20101123). SNP genotyping data (Omni Platform) was used to correlate variants calls from males in the 1000 Genomes. An additional variant call list, using default Unified Genotyper parameters, was conducted on 162 males samples from the 1000 Genomes alignment files. This alternate variant list was generously provided by the Chakravarti Lab (Johns Hopkins University). For the Exome Variant Server dataset, variants present in the male portion of the non-pseudoautosomal regions of the X chromosome were extracted based on variant annotations of hemizygous genotype.

2.4.10 Acknowledgements

I would like to thank Dr. Sarah Wheelan of Johns Hopkins University for critical reading of this manuscript. We thank Dr. Aravinda Chakravarti of Johns Hopkins University for providing valuable guidance and resource access. TSN and TT received training as graduate students of the Predoctoral Training Program in Human Genetics at Johns Hopkins University. We thank the patients and their families for participation in this project.

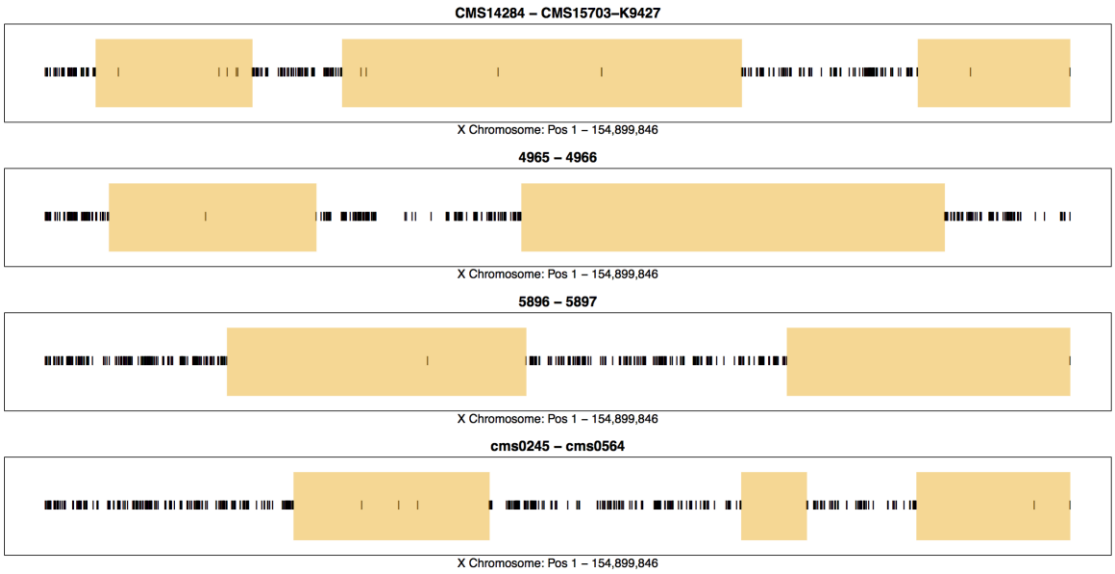
2.5 Figures: Chapter 2

Figure 2-1. Relatedness between study samples



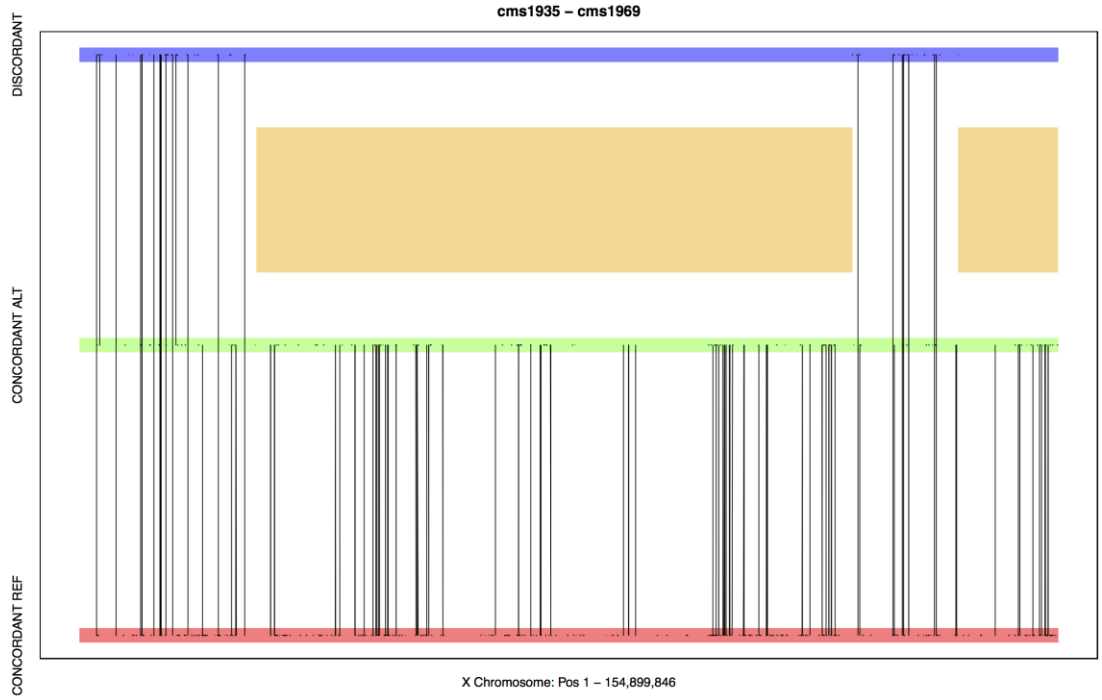
The estimated relatedness of samples is calculated by comparing the percentage of shared variants between two samples. The vertical axis shows the percentage of shared variants between two samples. The horizontal axis shows individual samples (n = 82) in this study cohort. Sporadic cases are samples that are not related to any other sample (left) while kindred pairs are two related affected males (right) (Table 1). The color-coded alphanumeric labels designate individual samples along the X-axis. There are 30 sporadic cases [1–9, 0, a-t; black labels] and 52 kinships (26 pairs) [u-z, A-Z, symbols; colored labels]. Box and whisker plots indicate overall identity between a proband and all other samples in the cohort. Identity ≈ (2 x # of variants identical between both samples) / (sum of variants of both samples). Sporadic cases generally share low identity with other samples. Paired kindred generally show the highest identity with each other. Paired kindred are juxtaposed with each other with the same color on the X-axis to simplify visualization of relationships. Outlier labels located above the hollow plots indicate the identifier for the sample that shares the highest identity, which is consistent with the known family relationship.

Figure 2-2. Shared IBD segments for selected representative sample pairs



Plot titles indicate sample identifiers (Sample Pair 1—Sample Pair 2). X-axis denotes position along the X chromosome. Far left is position 1 and far right is position 154,899,846, relative to the hg19 reference sequence. Black vertical bars indicate positions along the X chromosome at which a variant was called in one sample, but was not called in its paired sample (genotypic discordance between related samples). Orange blocks reflect regions lacking an abundance of discordant genotypes. These regions are shared Identical by Descent (inheritance) between the samples and contain the pathological variants of interest. All variants, both genotypically concordant (not shown) and discordant, that are located within the orange blocks are retained by the Shared Segment Filter. All possible pairwise relationships were assessed for segment sharing by an automated 5 MB sliding window and manual curation. Only samples with substantial sharing are plotted.

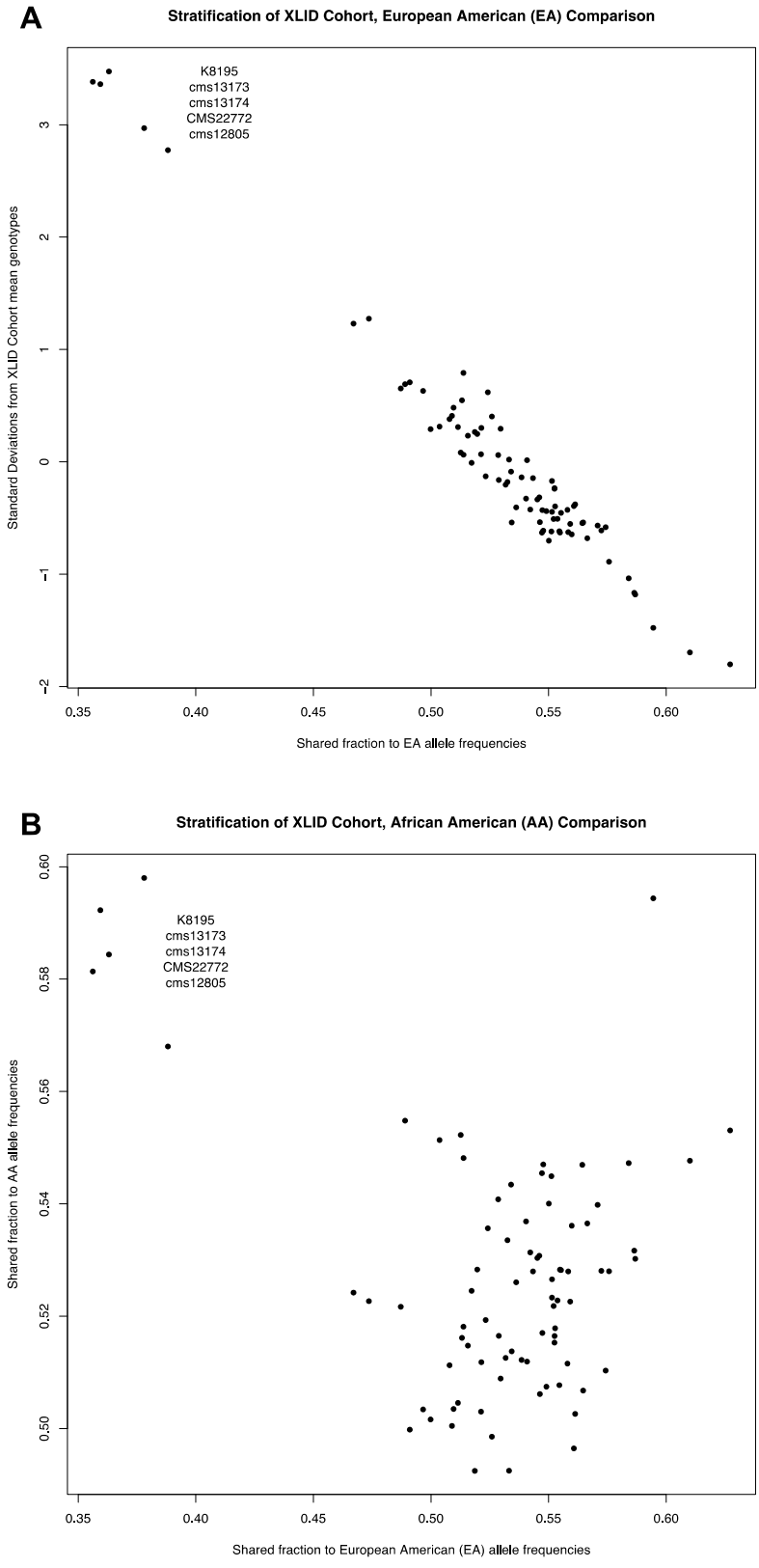
Figure 2-3. Correlation between shared segments of IBD and known linkage intervals; one sample presented



Plot titles indicate sample identifiers (Sample Pair 1—Sample Pair 2). X-axis denotes position along the X chromosome. Far left is position 1 and far right is position 154,899,846 relative to the hg19 reference sequence. Orange blocks reflect regions determined to be in IBD. These regions are shared by inheritance between the samples and contain the pathological variant of interest. Black dots are HapMap SNPs and are distributed across the Y-axis based on genotypic sharing between the sample pairs. SNP positions retaining the reference allele between both samples (Concordant Ref) are located at the bottom in the pink region. SNP positions retaining the alternate allele between both samples (Concordant Alt) are located at the middle in the green region. SNP positions that are genotypically discordant between the sample pairs (Discordant)

are located at the top in the blue region. Individual black dots reflect one SNP position. Vertical black lines connecting dots reflect SNPs that are known to be linked as determined using PLINK on HapMap Phase 3 data. Vertical lines rarely cross the orange IBD shared segments, suggesting that these segments are likely descending along known linkage intervals. However, only higher resolution, full chromosome genotyping (not exome sequence) can prove this conclusively.

Figure 2-4. Assessment of cohort population stratification



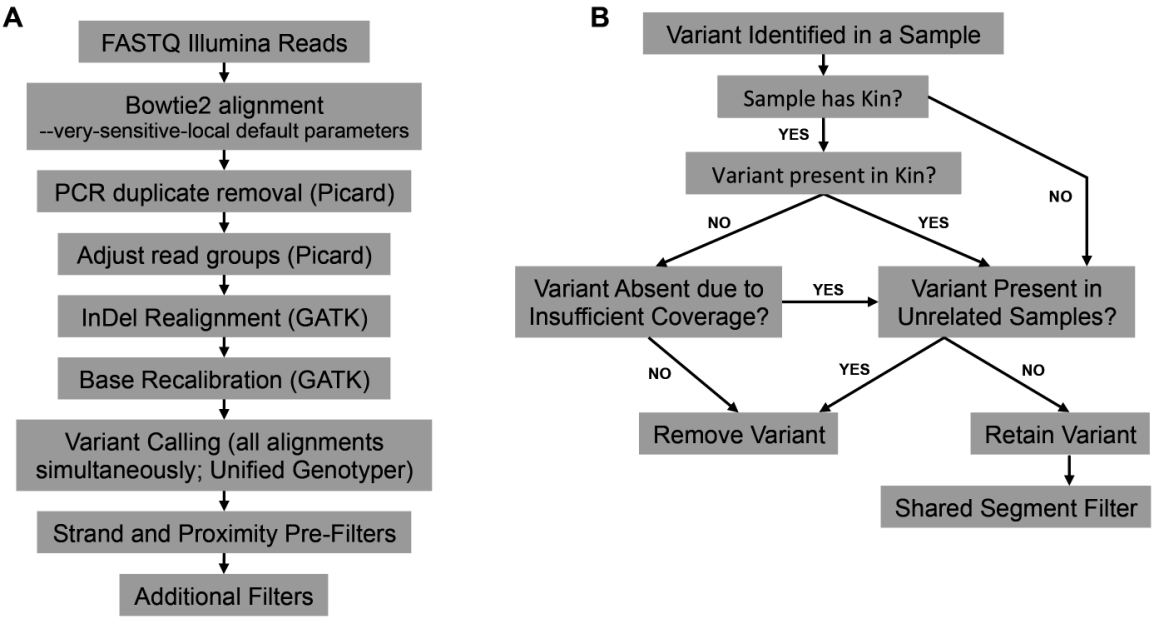
Population stratification was determined by comparing samples to a cohort mean, and by comparing sample allele frequencies to population-specific allele frequencies derived from EVS data. For the cohort mean analysis, an average allele frequency for the cohort was obtained for each variant. A residual sum of squares was subsequently calculated for each sample compared to the cohort mean. The number of standard deviations from the residual is plotted on the Y-axis of Panel A. Comparison of sample allele frequencies to EVS was performed for both EVS European American (EA) and EVS African American (AA) population frequencies. For any given sample, an EA and an AA metric was obtained by averaging the population specific allele frequencies for all sample variants present in the EVS dataset. This EA metric for each sample is plotted on the X-axis. The AA metric for each sample is plotted on the Y-axis of Panel B. The majority of samples cluster together with genotypes of primarily European ancestry. However, five samples show slight deviations from the main cluster. When SNP loads were compared to EVS data from the African American population, the sample cluster deviation was reproduced, indicating that these five samples have a small, but detectable contribution of African ancestry.

Figure 2-5. Shared segment filter and error reduction by strand/proximity pre-filter

The Shared Segment Filter (component of the Affected Kindred/Cross-Cohort Filter) retains chromosomal segments shared as Identical by Descent between two related samples in the XLID cohort. In this example, Panels A and B each reflect the same kindred pair, two brothers. The X-axis is position along the X chromosome exome. The Y-axis indicates the allelic status of a given variant for both siblings. Each point in the graph is a variant site for at least one sample. R|R allelic status indicates that the given point (genomic site) matches the reference sequence (hg19) in both samples (both samples are wildtype). A|A allelic status indicates the given point (variant site) is alternate to hg19 in both samples (both samples are hemizygous mutant). A|R allelic status indicates the given point matches reference in one sample and is alternate in the kindred sample (the samples are genotypically discordant). The orange blocks delineate chromosomal segments devoid of A|R points. All sequence in that segment is Identical by Descent between the two samples. The Shared Segment Filter retains variants (A|A)

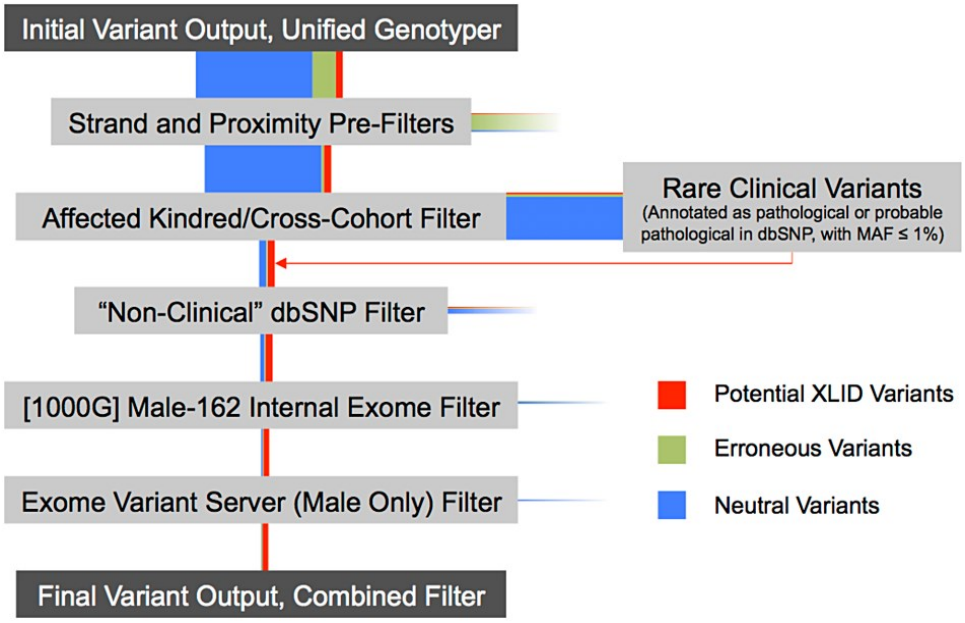
within the orange block. Panel A shows variant allele status in the Shared Segment Filter prior to the application of the strand- and proximity-based pre-filters. With the exception of the rare *de novo* mutation, there should be no discordant (A|R) variants within the orange block. Such variants are likely erroneous. Panel B shows the Shared Segment Filter after application of the strand- and proximity-based pre-filters. The A|R variants previously present in the orange block are eliminated, reflecting a reduction in erroneous variant calls as a result of these pre-filters.

Figure 2-6. Schematics of variant calling and affected kindred/cross-cohort analysis



Panel A: Illumina FASTQ sequenced read files are aligned to the human reference genome (hg19) using bowtie2, followed by removal of PCR duplicates, read group adjustment, indel realignment, and base recalibration. Variant calling is conducted using the Unified Genotyper. Variant calling is conducted in parallel on all alignments. Panel B: The Affected Kindred/Cross-Cohort Filter makes use of known relatedness. Unshared variants between related samples are removed. Shared variants between unrelated samples are removed. Shared variants between related samples are retained. The Affected Kindred/Cross-Cohort Filter accommodates for the possibility that the absence of a variant in a related sample may also be due to insufficient coverage or variant quality in the related sample. All retained variants are subsequently run through the Shared Segment Filter.

Figure 2-7. Schematic of variant reduction using a combined filter



The Combined Filter sequentially applies all the filters described in this study. Vertical colored bars reflect relative changes in the content of the variant pool after each filter step. Horizontal colored bars reflect rejected variants upon each filter step. The Strand and Proximity Pre-Filters are applied universally. Then the Affected Kindred/Cross-Cohort Filter (with Shared Segment Filter) is applied. The rejected variant pool in this step primarily eliminates neutral variants. Nonetheless, this rejected pool of variants is assessed for co-occurrence with rare dbSNP variants with known pathological function. Rejected variants that positively co-occur in the Rare Clinical Variants dataset are re-introduced (thin red arrow). Database-dependent filters are sequentially applied. Red bars reflect potential XLID variants that may be of functional interest. Green bars reflect variants that are likely sequencing errors. Blue bars reflect variants that are likely neutral in XLID etiology.

2.6 Tables: Chapter 2

Table 2-1. XLID cohort for X Chromosome exome sequencing

Relationship of Samples	Number (Pairs)
Affected Sporadic Cases	30
Affected Pairs	52 (26)
Brothers	44 (22)
Maternal Half-Brothers	4 (2)
Maternal Male First Cousins	2 (1)
Uncle-Nephew	2 (1)

All samples are diagnosed with an X-linked Intellectual Disorder. Criteria for X-linkage are described in Materials and Methods.

Table 2-2. Enrichment of potential pathological variants in X-Exome of XLID cohort with different variant filters

Application of Variant Filters	Non-Synonymous or Splicing Variants	Other Variants	Total Variants	% Original
Strand and Proximity Pre-filters Only [1]	221.8 ± 30.8	724.5 ± 137.0	946.3 ± 167.8	100.0%
+ Shared Segment Filter [2]	160.1 ± 65.0	511.3 ± 222.1	671.4 ± 287.1	71.0%
+ [1000G] Male-162 Internal Exome Filter [3]	62.5 ± 19.9	645.6 ± 129.2	708.1 ± 149.1	74.8%
+ Exome Variant Server (Male Only) Filter [4]	18.8 ± 4.6	545.9 ± 108.8	564.7 ± 113.4	59.7%
+ "Non-clinical" dbSNP Filter [5]	11.9 ± 5.4	48.8 ± 18.5	60.7 ± 23.9	6.4%
+ Affected Kindred/Cross-Cohort Filter [6]	7.5 ± 2.4	25.4 ± 2.2	32.9 ± 4.6	3.5%
All Filters [7]	2.1 ± 1.7	12.1 ± 10.0	14.2 ± 11.7	1.5%

Average number of variants remaining per sample after sequential or aggregate filtering steps.

[1] Strand and Proximity Pre-Filters are applied universally on top of all other filters. The percent of variants remaining after a filter is relative to the variant output after application of the Strand and Proximity Pre-Filters and is provided in column 5.

[2] Shared Segment Filter: for demonstration purposes, results of this filter are provided separately from the rest of the Affected Kindred/Cross-Cohort Filter.

[3] [1000G] Male-162 Internal Exome Filter: removes variants from the XLID cohort shared in common with 162 males from the 1000 Genomes.

[4] Exome Variant Server (Male Only) Filter: removes variants from the XLID cohort shared in common with variants of the male fraction of EVS.

[5] "Non-Clinical" dbSNP is redacted of known, probable, or potentially pathological variants in dbSNP Build 137.

[6] Affected Kindred/Cross-Cohort Filter: results exclude the Shared Segment Filter component (see Row 2).

[7] All filters, including re-introduction of known rare pathological variants (from dbSNP) that are inappropriately eliminated by the Affected Kindred/Cross-Cohort Filter.

Table 2-3. Calculation of rare variant minor allele frequency (MAF) cutoff

# of ♂ X Chrs (***m***)	# of ♀ X Chrs (***f***)	# of X Chrs Total
525	2 x 567	1659

# of mut. alleles in 1000 Genomes (***n***)	Prob. alleles only in ♀ carriers	Prob. (eq) alleles only in ♀ carriers
1	68.35%	
2	46.70%	$i=f-n+1$
3	31.88%	
4	21.76%	$\text{Prob} = \prod_{n}^{1} \frac{i}{i+\frac{m}{2}}$
5	14.84%	
6	10.11%	
7	6.89%	
8	4.69%	
9	3.19%	
10	2.17%	
11	1.48%	
12*	**1.00%**	**12 / 1659 = 0.73%**

* At 12 copies of a mutant allele in the 1000 Genomes dataset, the probability of seeing all 12 alleles in only female carriers is only 1%. At >12 copies of a mutant allele, the probability is less than 1%. 12 mutant allele copies is ~ 0.73% minor allele frequency. We can safely assume that potential pathological variants with a MAF < 0.73% could exist purely in a female carrier state. Therefore, such variants should be removed from dbSNP before implementation of a dbSNP-based filter.

Table 2-4. Ambiguous variant calls in the public 1000 Genomes variant dataset

Sex	Chromosome	Genotype	Heterozygous	Homozygous	Ambiguous
		Variant Call	1	2	Others
Males	X Chromosome		39.49%	51.55%	8.96%
Females	X Chromosome		90.54%	9.46%	0%
Males	Autosomes		92.34%	7.66%	0%
Females	Autosomes		92.34%	7.66%	0%

Variant call = 1: Percent of variant alleles present as one copy in a sample (heterozygous state). Variant call = 2: Percent of variant alleles present as two copies in a sample (homozygous state). Variant call = Others: Percent of variant alleles present in copies other than 1 or 2, including non-integer counts. All values are evaluated exclusively from coding sequence variants for the respective chromosomes and sexes. Only the male X chromosome dataset possesses ambiguous genotypes. All variants were obtained from the 1000 Genomes variant dataset, pre-separated by chromosome [Integrated Phase 1, version 3: 20101123].

Table 2-5. List of 89 potential XLID genes

Gene ID	Priority	Gene ID	Priority	Gene ID	Priority
ZC4H2*	5	GPC4	2	CLCN4	1
ATRX*	5	HSFX1,HSFX2	2	TLR8	1
DMD*	4	NXF4	2	PHKA2	1
UBE2A*	4	CSAG1	2	WBP5	1
MAGEC1	4	SAGE1	2	APLN	1
FLNA*	3	FMR1*	1	GPR101	1
PLXNA3	3	MED12*	1	MCF2	1
TAF1	3	MECP2*	1	HS6ST2	1
TRO	3	NHS*	1	ATP11C	1
HCFC1*	2	OPHN1*	1	ZNF185	1
HUWE1*	2	PGK1*	1	OGT	1
MAOA*	2	IDS*	1	AIFM1	1
ZNF711*	2	SLC16A2*	1	CFP	1
RBM10*	2	MID1*	1	ZNF630	1
PLP1*	2	NSDHL*	1	GUCY2F	1
OTC*	2	ARHGEF9*	1	PASD1	1
ZMYM3	2	CLIC2	1	CHM	1
RNF128	2	USP9X	1	UBA1	1
ZRSR2	2	SMARCA1	1	NKAP	1
EFNB1	2	MED14	1	ATP1B4	1
NHSL2	2	USP26	1	MAMLD1	1
ZMAT1	2	COL4A5	1	FAM47C	1
GABRE	2	FRMD7	1	RPS4X	1
PRICKLE3	2	SRPK3	1	SUV39H1	1
HAUS7	2	AMER1	1	NAP1L3	1
USP51	2	ALG13	1	FAM9A	1
BCORL1	2	MAGIX	1	FAM47A	1
GRIPAP1	2	SRPX	1	LOC100996648	1
MXRA5	2	LAS1L	1	MAGEA10	1
FHL1	2	KAL1	1		

The list presents the 89 genes that were found to be mutated in our XLID cohort after application of the Combined Filter. Genes were prioritized by presence of a missense, nonsense, or splicing mutation. Weight was given for nonsense and splicing mutations. Weight was given for mutations present in both siblings. An asterisk (*) after the Gene ID indicates that the gene has been previously associated with XLID. The current list of XLID genes as of this publication stands at 103. Our target-enrichment system targets 975 chromosome X genes. Of the 89 genes in the list, 24 have been previously associated to XLID. This reflects a statistically significant 2.55-fold enrichment for known or previously associated XLID genes ($p < 1E-5$, hypergeometric test).

*103 known XLID genes: *ABCD1*, *ACSL4*, *AFF2*, *AGTR2*, *AP1S2*, *ARHGEF6*, *ARHGEF9*, *ARX*, *ATP6AP2*, *ATP7A*, *ATRX*, *BCOR*, *BRWD3*, *CASK*, *CCDC22*, *CDKL5*, *CLIC2*, *CUL4B*, *DCX*, *DKC1*, *DLG3*, *DMD*, *FANCB*, *FGD1*, *FLNA*, *FMR1*, *FTSJ1*, *GDI1*, *GK*, *GPC3*, *GRIA3*, *HCCS*, *HCFC1*, *HDAC8*, *HPRT*, *HSD17B10*, *HUWE1*, *IDS*, *IGBP1*, *IKBKG*, *IL1RAPL1*, *IQSEC2*, *KDM5C*, *KIAA2022*, *KLF8*, *SHROOM4*, *L1CAM*, *LAMP2*, *MAGT1*, *MAOA*, *MBTPS2*, *MECP2*, *MED12*, *MID1*, *MTM1*, *NAA10*, *NDP*, *NDUFA1*, *NHS*, *NLGN3*, *NLGN4*, *NSDHL*, *NXF5*, *OCRL*, *OFD1*, *OPHN1*, *OTC*, *PAK3*, *PCDH19*, *PDHA1*, *PGK1*, *PHF6*, *PHF8*, *PLP1*, *PORCN*, *PQBP1*, *PRPS1*, *PTCHD1*, *RAB39B*, *RBM10*, *RPL10*, *RPS6KA3*, *SIZN1*, *SLC16A2*, *SLC6A8*, *SLC9A6*, *SMC1A*, *SMS*, *SOX3*, *SRPX2*, *SYN1*, *SYP*, *TIMM8A*, *TSPAN7*, *UBE2A*, *UPF3B*, *ZDHHC9*, *ZDHHC15*, *ZNF41*, *ZNF674*, *ZNF711*, *ZNF81*

Table 2-6. Identification of known and potentially novel genes for XLID using X Chromosome exome sequencing and affected kindred/cross-cohort analysis

Name	Abbrev.	Map	Known or Predicted Function	XLID Gene	No. of Mutations	Mutation Nomenclature	SIFT Prediction	PolyPhen-2 Prediction	Variant Segregates with Disease in Family	Primary Isoform	Diseases or Phenotype	Reference
C4H2 domain-containing zinc finger	ZC4H2	Xq11.2	Zinc finger transcription factor		3	IVS+5G>A p.L66H p.R190W	NA Damaging Damaging	NA Prob. Damaging Prob. Damaging	Yes Yes Yes	NM_001178033	Wieacker-Wolfe Syndrome	Hirata et al, 2013
Alpha thalassemia / mental retardation syndrome X-linked	ATRX	Xq21.1	ATP-dependent helicase; chromosome remodeling	Yes	2	p.R37X p.S1606N	Known Damaging Damaging	Known Damaging Prob. Damaging	Yes Yes	NM_000489	XLID with alpha thalassemia	Gibbons et al, 1995
Ubiquitin-conjugating enzyme E2A	UBE2A	Xq24	Ubiquitin-conjugating enzyme	Yes	2	p.P68R p.Q110E	Damaging Tolerated	Prob. Damaging Prob. Damaging	Yes Yes	NM_003336	XLID, Nascimento-type	Nascimento et al, 2006
Filamin A	FLNA	Xq28	Actin-binding protein; cytoskeletal reorganization	Yes	2	p.G1576R p.F2228L	Damaging low conf Tolerated	Prob. Damaging Prob. Damaging	Yes	NM_001110556	Multiple congenital malformation syndromes	Fox et al, 1998
Transcription Initiation TFIID Subunit 1	TAF1	Xq13.1	Initiation of transcription by RNA Polymerase II and cell cycle control		2	p.M21L p.Q1428P	Tolerated Damaging	Benign Prob. Damaging	Yes	NM_138923		
Methly CpG-Binding Protein 2	MECP2	Xq28	Chromatin-based transcriptional regulation	Yes	1	p.K268E	Damaging	Prob. Damaging	Yes	NM_001110792	Rett Syndrome, XLID	Amir et al, 1999;; Schule et al, 2008; Orrico et al, 2000
Host cell factor C1	HCFC1	Xq28	Cell cycle control	Yes	1	p.G342S	Damaging	Prob. Damaging	Yes	NM_005334	XLID with methylmalonic acidemia	Huang et al, 2012; Yu et al, 2013
Zinc finger protein 711	ZNF711	Xq21.1	Zinc finger transcription factor	Yes	1	p.N601S	Damaging	Prob. Damaging	Yes	NM_021998	X-linked intellectual disability	Tarpey et al, 2009
Rho Guanine Nucleotide Exchange Factor 9	ARHGEF9	Xq11.1-q11.2	Brain-specific regulation of glycine and GABA receptors clusters	Yes	1	p.R236W	Damaging	Prob. Damaging	Yes	NM_001173480	XLID, Epileptic encephalopathy	Shimojima et al, 2011; Marco et al, 2008
E3 Ubiquitin Ligase	HUWE1	Xp11.22	Degradation of proteins involved in apoptosis and DNA maintenance	Yes	1	p.R4187H	Tolerated	Prob. Damaging	Yes	NM_031407	XLID, Turner-type	Turner et al, 1994
Ephrin B1	EFNB1	Xq13.1	Ligand of Eph-related receptor tyrosine kinases		1	p.G290R	Tolerated	Prob. Damaging	Yes	NM_004429	Craniofrontonasal Syndrome	Wieland et al, 2004; Twigg et al, 2004, 2013
Plexin A3	PLXNA3	Xq28	Semaphorin receptor; cytoskeletal remodeling		1	p.V1304M	Tolerated	Benign	TBD	NM_017514		
Ring finger protein 128	RNF128	Xq22.3	E3 Ubiquitin protein ligase		1	p.R12H	Damaging	Prob. Damaging	TBD	NM_194463		
Prickle homolog 3	PRICKLE3	Xp11.23	LIM domain-containing protein		1	p.R175C	Damaging	Prob. Damaging	Yes	NM_006150		
Zinc finger, RNA-binding motif and serine/arginine rich 2	ZRSR2	Xp22.2	Essential splicing factor		1	p.R440Q	Tolerated	Benign	Yes	NM_005089		
Glutamate receptor interacting protein associated protein 1	GRIPAP1	Xp11.23	Interaction with AMPA receptor complex		1	p.R822Q	Tolerated	Prob. Damaging	Yes	NM_020137		
O-linked N-acetylglucosamine Transferase	OGT	Xq13.1	Post-translational glycosylation		1	p.L244F	Tolerated	Prob. Damaging	Yes	NM_181673		
SRSF (Ser/Arg Splicing Factors) Protein Kinase 3	SRPK3	Xq28	Homolog of SRPK1 with possible role in splicing regulation		1	p.H159D	Damaging	Poss. Damaging	Yes	NM_014370		

Chapter 3: Effective Detection of Rare Variants in Pooled DNA Samples Using Cross-pool Tailcurve Analysis

Sequencing targeted DNA regions in large samples is necessary to discover the full spectrum of rare variants. We report an effective Illumina sequencing strategy utilizing pooled samples with novel quality (Srfim) and filtering (SERVIC^{4}E) algorithms. We sequenced 24 exons in two cohorts of 480 samples each, identifying 47 coding variants, including 30 present once per cohort. Validation by Sanger sequencing revealed an excellent combination of sensitivity and specificity for variant detection in pooled samples of both cohorts as compared to publicly available algorithms [51].

3.1 Introduction

Next-generation sequencing and computational genomic tools permit rapid, deep sequencing for hundreds to thousands of samples [110-112]. Recently, rare variants of large effect have been recognized as conferring substantial risks for common diseases and complex traits in humans [113]. There is considerable interest in sequencing limited genomic regions such as sets of candidate genes and target regions identified by linkage and/or association studies. Sequencing large sample cohorts is essential to discover the full spectrum of genetic variants and provide sufficient power to detect differences in the allele frequencies between cases and controls. However, several technical and analytical challenges must be resolved to efficiently apply next-generation sequencing to large

samples in individual laboratories. First, it remains expensive to sequence a large number of samples despite a substantial cost reduction in available technologies. Second, for target regions of tens to hundreds of kilobases or less for a single DNA sample, the smallest functional unit of a next-generation sequencer (for example, a single lane of an Illumina Genomic Analyzer II (GAII) or HiSeq 2000 flow cell) generates a wasteful excess of coverage. Third, methods for individually indexing hundreds to thousands of samples are challenging to develop and limited in efficacy [114,115]. Fourth, generating sequence templates for target DNA regions in large numbers of samples is laborious and costly. Fifth, while pooling samples can reduce both labor and costs, it reduces sensitivity for the identification of rare variants using currently available next-generation sequencing strategies and bioinformatics tools [110,112].

We have optimized a flexible and efficient strategy that combines a PCR-based amplicon ligation method for template enrichment, sample pooling, and library indexing in conjunction with novel quality and filtering algorithms for identification of rare variants in large sample cohorts. For validation of this strategy, we present data from sequencing 12 indexed libraries of 40 samples each (total of 480 samples) using a single lane of a GAII Illumina Sequencer. We utilized an alternative base-calling algorithm, Srfim [116], and an automated filtering program, SERVIC^{4}E (Sensitive Rare Variant Identification by Cross-pool Cluster, Continuity, and tailCurve Evaluation), designed for sensitive and reliable detection of rare variants in pooled samples. We validated this strategy using Illumina sequencing data from an additional independent cohort of 480 samples. Compared to publicly available software, this strategy achieved an excellent

combination of sensitivity and specificity for rare variant detection in pooled samples through a substantial reduction of false positive and false negative variant calls that often confound next-generation sequencing. We anticipate that our pooling strategy and filtering algorithms can be easily adapted to other popular platforms of template enrichment, such as microarray capture and liquid hybridization [117,118].

3.2 Results

3.2.1 An optimized sample-pooling strategy

We utilized a PCR-based amplicon-ligation method because PCR remains the most reliable method of template enrichment for selected regions in a complex genome. This approach ensures low cost and maximal flexibility in study design compared to other techniques [118-120]. Additionally, PCR of pooled samples alleviates known technical issues associated with PCR multiplexing [121]. We sequenced 24 exon-containing regions (250 to 300 bp) of a gene on chromosome 3, *GRIP2* (encoding Glutamate Receptor Interacting Protein 2; [GenBank: AB051506]) in 480 unrelated individuals (Figure 3-1). The total targeted region is 6.7 kb per sample. We pooled 40 DNA samples at equal concentration into 12 pools, which was done conveniently by combining samples from the same columns of five 96-well plates. We separately amplified each of the 24 regions for each pool, then normalized and combined resulting PCR products at equal molar ratio. The 12 pools of amplicons were individually blunt-end ligated and randomly fragmented for construction of sequencing libraries, each with

a unique Illumina barcode [122]. These 12 indexed libraries were combined at equal molar concentrations and sequenced on one lane of a GAII (Illumina) using a 47-bp single-end module. We aimed for 30-fold coverage for each allele. Examples of amplicon ligation, distribution of fragmented products, and 12 indexed libraries are shown in Figure 3-2.

3.2.2 Data analysis and variant calling

Sequence reads were mapped by Bowtie using strict alignment parameters (-v 3: entire read must align with three or fewer mismatches) [123]. We chose strict alignment to focus on high quality reads. Variants were called using SAMtools (deprecated algorithms [pileup -A -N 80]; see Materials and methods) [124]. A total of 11.1 million reads that passed Illumina filtering and had identifiable barcodes were aligned to the human genome (hg19), generating approximately 520 megabases of data. The distribution of reads for each indexed library ranged from 641 k to 978 k and 80% of reads had a reported read score (Phred) greater than 25 (Figure 3-3a, b). The aggregate nucleotide content of all reads in the four channels across sequencing cycles was constant (Figure 3-3c), indicating a lack of global biases in the data. There was little variability in total coverage per amplicon pool, and sufficient coverage was achieved to make variant calling possible from all amplicon pools (Figure 3-4). Our data indicated that 98% of exonic positions had an expected minimum coverage of 15× per allele (approximately 1,200× minimum coverage per position) and 94% had an expected minimum coverage of 30× (approximately 2,400× minimum coverage per position). Overall average expected

allelic coverage was 68×. No exonic positions had zero coverage. To filter potential false positive variants from SAMtools, we included only high-quality variant calls by retaining variants with consensus quality (cq) and SNP quality (sq) scores in 95% of the score distributions (cq ≥ 196, sq ≥ 213; Figure 3-5a). This initially generated 388 variant calls across the 12 pools. A fraction of these variant calls (n = 39) were limited to single pools, indicating potential rare variants.

3.2.3 Tailcurve analysis

Initial validations by Sanger sequencing indicated that approximately 25% or more of these variant calls were false positives. Sequencing errors contribute to false positive calls and are particularly problematic for pooled samples where rare variant frequencies approach the error rate. To determine the effect of cycle-dependent errors on variant calls [116], we analyzed the proportions of each nucleotide called at each of the 47 sequencing cycles in each variant. We refer to this analysis as a tailcurve analysis due to the characteristic profile of these proportion curves in many false-positive variant calls (Figure 3-6; Figure 3-7). This analysis indicated that many false positive calls arise from cycle-dependent errors during later sequencing cycles (Figure 3-6d). The default base-calling algorithm (BUSTARD) and the quality values it generates make existing variant detection software prone to false positive calls because of these technical biases. Examples of tailcurves reflecting base composition by cycle at specific genetic loci for wild type, common SNP, rare variant, and false positive calls are shown in Figure 3-6.

3.2.4 Quality assessment and base calling using Srfim

To overcome this problem, we utilized Srfim, a quality assessment and base-calling algorithm based on a statistical model of fluorescence intensity measurements that captures the technical effects leading to base-calling biases [116]. Srfim explicitly models cycle-dependent effects to create read-specific estimates that yield a probability of nucleotide identity for each position along the read. The algorithm identifies nucleotides with highest probability as the final base call, and uses these probabilities to define highly discriminatory quality metrics. Srfim increased the total number of mapped reads by 1% (to 11.2 million), reflecting improved base-calling and quality metrics, and reduced the number of variant calls by 20% (308 variants across 12 pools; 33 variant calls present in only a single pool).

3.2.5 Cross-pool filtering using SERVIC^{4}E

Further validation by Sanger sequencing indicated the persistence of a few false positive calls from this dataset. Analysis of these variant calls allowed us to define statistics that capture regularities in the base calls and quality values at false positive positions compared to true variant positions. We developed SERVIC^{4}E, an automated filtering algorithm designed for high sensitivity and reliable detection of rare variants using these statistics.

Our filtering methods are based on four statistics derived from the coverage and qualities of variant calls at each position and pool: (1) continuity, defined as the number of cycles in which the variant nucleotide is called (ranges from 1 to 47); (2) weighted

allele frequency, defined as the ratio of the sum of Phred quality scores of the variant base call to the sum of Phred quality scores of all base calls; (3) average quality, defined as the average quality of all base calls for a variant; and (4) tailcurve ratio, a metric that captures strand-specific tailcurve profiles that are characteristic of falsely called variants. SERVIC^{4}E employs filters based on these four statistics to remove potential false-positive variant calls. Additionally, SERVIC^{4}E searches for patterns of close-proximity variant calls, a hallmark of errors that have been observed across different sequenced libraries and sequencing chemistries (Figure 3-8), and uses these patterns to further filter out remaining false positive variants. In the next few paragraphs we provide rationales for our filtering statistics, and then define the various filters employed.

The motivation for using continuity and weighted allele frequency is based on the observation that a true variant is generally called evenly across all cycles, leading to a continuous representation of the variant nucleotide along the 47 cycles, and is captured by a high continuity score. However, continuity is coverage-dependent and should only be reliable when the variant nucleotide has sufficient sequencing quality. For this reason, continuity is assessed in the context of the variant's weighted allele frequency. Examples of continuity versus weighted allele frequency curves for common and rare variants are shown in Figure 3-9. Using these two statistics, SERVIC^{4}E can use those pools lacking the variant allele (negative pools) as a baseline to isolate those pools that possess the variant allele (positive pools).

SERVIC^{4}E uses a clustering analysis of continuity and weighted allele frequency to filter variant calls between pools. We use k-medioid clustering and decide the number

of clusters using average silhouette width [125]. For common variants, negative pools tend to cluster and are filtered out while all other pools are retained as positives (Figure 3-9a, b). Rare variant pools, due to their lower allele frequency, will have a narrower range in continuity and weighted allele frequency. Negative pools will appear to cluster less, while positive pools cluster more. SERVIC^{4}E will retain as positive only the cluster with highest continuity and weighted allele frequency (Figure 3-9c, d).

The second filter used by SERVIC^{4}E is based on the average quality of the variant base calls at each position. One can expect that the average quality score is not static, and can differ substantially between different sequencing libraries and even different base-calling algorithms. As such, the average quality cutoff is best determined by the aggregate data for an individual project (Figure 3-10). Based on the distribution of average qualities analyzed, SERVIC^{4}E again uses cluster analysis to separate and retain the highest quality variants from the rest of the data. Alternatively, if the automated clustering method is deemed unsatisfactory for a particular set of data, a more refined average quality cutoff score can be manually provided to SERVIC^{4}E, which will override the default clustering method. For our datasets, we used automated clustering to retain variants with high average quality.

The third filtering step used by SERVIC^{4}E captures persistent cycle-dependent errors in variant tailcurves that are not eliminated by Srfim. Cycle-specific nucleotide proportions (tailcurves) from calls in the first half of sequencing cycles are compared to the proportions from calls in the second half of sequencing cycles. The ratio of nucleotide proportions between both halves of cycles is calculated separately for plus and minus

strands, thereby providing the tailcurve ratio added sensitivity to strand biases. By default, variant calls are filtered out if the tailcurve ratio differs more than ten-fold; we do not anticipate that this default will need adjustment with future sequencing applications, as it is already fairly generous, chiefly eliminating variant pools with clearly erroneous tailcurve ratios. This default was used for all our datasets.

The combination of filtering by average quality and tailcurve structure eliminates a large number of false variant calls. Figure 3-11 demonstrates the effect of these filtering steps applied sequentially on two sets of base call data.

In addition to these filtering steps, SERVIC^{4}E employs limited error modeling. The pattern of errors observed in many libraries may be dependent on the sequence context of the reads, the preparation of the library being sequenced, the sequencing chemistry used, or a combination of these three factors. We have observed that certain erroneous variant calls tend to aggregate in proximity. These clusters of errors can sometimes occur in the same positions across multiple pools. These observations appeared in two independent datasets in our studies. Importantly, many of the false positive calls that escaped our tailcurve and quality filtering fell within these clusters of errors. To overcome this problem, SERVIC^{4}E conducts error filtering by analyzing mismatch rates in proximity to a variant position of interest and then determining the pattern of error across multiple pools. This pattern is defined as the most frequently occurring combination of pools with high mismatch rates at multiple positions within the isolated regions. The similarity between a variant call of interest and the local pattern or error across pools can then be used to eliminate that variant call (Figure 3-8). The

consequences of these sequential filtering steps on variant output are outlined in Table 3-1 for both cohorts tested in this study.

Finally, SERVIC^{4}E provides a trim parameter that masks a defined length of sequence from the extremes of target regions from variant calling. This allows for SERVIC^{4}E to ignore spurious variant calling that may occur in primer regions as a result of the concatenation of amplicons. By default, this parameter is set to 0; for our datasets, we used a trim value of 25, which is the approximate length of our primers.

3.2.6 Reliable detection of rare variants in pooled samples

Using SERVIC^{4}E, we identified 68 unique variants (total of 333 among 12 pools), of which 34 were exonic variants in our first dataset of 480 samples (Table 3-2). For validation, we performed Sanger sequencing for all exonic variants in individual samples in at least one pool. A total of 4,050 medium/high-quality Sanger traces were generated, targeting approximately 3,380 individual amplicons. Total coverage in the entire study by Sanger sequencing was approximately 930 kb (approximately 7.3% of total coverage obtained by high-throughput sequencing). Sanger sequencing confirmed 31 of the 34 variants. Fifteen rare exonic variants were identified as heterozygous in a single sample in the entire cohort.

3.2.7 A comparison with available variant calling algorithms

We compared our variant calling method to publicly available algorithms, including SAMtools, SNPSeeker, CRISP, and Syzygy [110,112,124,126]. Because some

variants are present and validated in multiple pools and each pool is considered as an independent discovery step, we determined the detection sensitivity and specificity on a variant pool basis. Results are shown in Table 3-3.

To call variants with SAMtools [124], we used the deprecated Maq algorithms (SAMtools pileup -A -N 80), as the regular SAMtools algorithms failed to identify all but the most common variants. As a filtering cutoff we retained only the top 95th percentile of variants by consensus quality and SNP quality score (cq $\geq$ 196 and sq $\geq$ 213 for standard Illumina base calls, Figure 3-5a; cq $\geq$ 161 and sq $\geq$ 184 for Srfim base calls, Figure 3-5b).

SNPSeeker [110] uses large deviation theory to identify rare variants. It reduces the effect of sequencing errors by generating an error model based on internal negative controls. We used exons 6 and 7 as the negative controls in our analysis (total length = 523 bp) as both unfiltered SAMtools analysis and subsequent Sanger validation indicated a complete absence of variants in both exons across all 12 pools. Only Illumina base calls were used in this comparison because of a compatibility issue with the current version of Srfim. The authors of SNPSeeker recently developed a newer variant caller called SPLINTER [127], which requires both negative and positive control DNA to be added to the sequencing library. SPLINTER was not tested due to the lack of a positive control in our libraries.

CRISP [126] conducts variant calling using multiple criteria, including the distribution of reads and pool sizes. Most importantly, it analyzes variants across multiple

pools, a strategy also employed by SERVIC^{4}E. CRISP was run on both Illumina base calls and Srfim base calls using default parameters.

Syzygy [112] uses likelihood computation to determine the probability of a non-reference allele at each position for a given number of alleles in each pool, in this case 80 alleles. Additionally, Syzygy conducts error modeling by analyzing strand consistency (correlation of mismatches between the plus and minus strands), error rates for dinucleotide and trinucleotide sequences, coverage consistency, and cycle positions for mismatches in the read [128]. Syzygy was run on both Illumina and Srfim base calls, using the number of alleles in each pool (80) and known dbSNP positions as primary input parameters.

SERVIC^{4}E was run using a trim value of 25 and a total allele number of 80. All other parameters were run at default. The focus of our library preparation and analysis strategy is to identify rare variants in large sample cohorts, which necessitates variant calling software with very high sensitivity. At the same time, specificity must remain high, primarily to ease the burden during validation of potential variants. In addition to calculating sensitivity and specificity, we calculated the Matthews correlation coefficient (MCC; see Materials and methods) for each method (Table 3-3) in order to provide a more balanced comparison between the nine methods.

For validation of our dataset, we focused primarily on changes in the exonic regions of our amplicons. Any intronic changes that were collaterally sequenced successfully were also included in our final analysis (Table 3-3). Sixty-one exonic positions were called as having a variant allele in at least one pool by one or more of the

nine combinations of algorithms tested. We generated Sanger validation data in at least one pool for 49 of the 61 positions identified. Genotypes for validated samples are indicated in Table 3-4.

SNPSeeker (with Illumina base calls) performed with the highest specificity (97.3%), but with the worst sensitivity (62.2%), identifying less than half of the 15 valid rare exonic variants (Table 3-3). This is likely due to an inability of this algorithm to discriminate variants with very low allele frequencies in a pool; 84% of SNPSeeker's true positive calls have an allele frequency $\geq$ 1/40, while only 13% of the false negative calls have a frequency $\geq$ 1/40 (Table 3-2 and Table 3-5). SNPSeeker's MCC score was low (61.8%), due in large part to its very low false positive rate.

SAMtools alone with Illumina base calls achieved a 92.2% sensitivity, identifying all 15 rare exonic variants; however, these results were adulterated with the highest number of false positives, resulting in the worst specificity (56.2%) and MCC score (52.8%) among the nine methods (Table 3-3). Incorporation of Srfim base calls cut the number of false positives by 60% (from 32 to 13) without a sizable reduction in the number of true positive calls (from 83 to 80). Fourteen of the fifteen valid rare exonic variants were successfully identified, which while not perfect, is an acceptably high sensitivity (Table 3-3). Srfim made noticeable improvements to individual base quality assessment as reflected in a substantial reduction in low quality variant calls (Figure 3-5) by reducing the contribution of low quality base calls to the average quality distribution (Figure 3-10b) and by reducing the tailcurve effect that leads to many false positives (Figure 3-11). Most low quality variant calls that were eliminated when transitioning to

Srfim were not valid; nonetheless, three low quality valid variant calls were similarly affected by Srfim, and their loss resulted in a slight reduction in the true positive rate.

CRISP using Illumina base calls achieved a sensitivity slightly lower than SAMtools (87.8% versus 92.2%). Additionally, CRISP identified only 13 of the 15 valid rare exonic variants. Though this is lower than SAMtools, it is a large improvement over SNPSeeker; for the purposes set forth in our protocol, the > 75% sensitivity for extremely rare variants achieved by CRISP (using either base-calling method) is acceptable (Table 3-3).

Syzygy achieved the second highest sensitivity (94.4%) using Illumina base calls, but specificity remained low (67.1%). Fourteen of the fifteen rare exonic variants were successfully identified. CRISP and Syzygy achieved relatively average MCC values (50.5% and 65.0%, respectively), reflecting better performance than SAMtools with Illumina base calls.

SERVIC^{4}E using Illumina base calls achieved the highest sensitivity (97.8%) and identified all 15 valid rare exonic variants. Both sensitivity and specificity were improved over SAMtools, CRISP, and Syzygy (Table 3-3), reflected in the highest MCC score of all the tested methods (84.2%). Taken together, the combination of SERVIC^{4}E with either base-calling algorithm provides the highest combination of sensitivity and specificity in the dataset from pooled samples.

As previously mentioned, Srfim greatly improved variant calling in SAMtools, as is reflected in the 19% increase in SAMtools' MCC value (from 52.8% to 71.4%). CRISP, Syzygy, and SERVIC^{4}E benefited little from using Srfim base calls: the MCC

value for CRISP improved by only 6% (from 50.5% to 56.5%), Syzygy diminished by 4.6% (from 65.0% to 60.4%), and SERVIC^{4}E diminished by 6.5% (from 84.2% to 77.7%). Importantly, use of Srfim base calls with Syzygy diminished its capacity to detect rare variants by a third. These three programs are innately designed to distinguish low frequency variants from errors using many different approaches. As such, it can be inferred from our results that any initial adjustments to raw base calls and quality scores by the current version of Srfim will do little to improve that innate capacity. In contrast, SAMtools, which is not specifically built for rare variant detection and would therefore have more difficulty distinguishing such variants from errors, benefits greatly from the corrective pre-processing provided by Srfim.

In addition to performance metrics like sensitivity and specificity, we analyzed annotated SNP rates, transition-transversion rates, and synonymous-non-synonymous rates of the nine algorithms on a variant-pool basis (Table 3-6).

The variant pools with the greatest discrepancies between the various detection methods tended to have an estimated allele frequency within the pool that is less than the minimum that should be expected (1/80; Table 3-2**,** Table 3-5, and Table 3-7). Such deviations are inevitable, even with normalization steps, given the number of samples being pooled. This underscores the importance of having careful, extensive normalization of samples to minimize these deviations as much as possible, and the importance of using variant detection methods that are not heavily reliant on allele frequency as a filtering parameter or are otherwise confounded by extremely low allele frequencies.

3.2.8 Validation using data from an independent cohort of samples

To further assess the strength of our method and analysis software, we sequenced the same 24 *GRIP2* exons in a second cohort of 480 unrelated individuals. The same protocol for the first cohort was followed, with minor differences. Firstly, we pooled 20 DNA samples at equal concentration into 24 pools. The first 12 pools were sequenced in one lane of a GAII and the last 12 pools were sequenced in a separate lane (Figure 3-12). Additionally, the libraries were sequenced using the 100-bp paired-end module, and sequencing was conducted using a newer version of Illumina's sequencing chemistry. These 24 libraries occupied approximately 5% of the total sequencing capacity of the two lanes. The remaining capacity was occupied by unrelated libraries that lacked reads originating from the *GRIP2* locus.

To map reads from this dataset, we initially used Bowtie's strict alignment parameters (-v 3), as we had done with our first dataset, but this resulted in a substantial loss of coverage in the perimeters of target regions. This is likely due to reads that cross the junctions between our randomly concatenated amplicons; such reads, which have sequence from two distant amplicons, appear to have extensive mismatching that would result in their removal. This effect became pronounced when using long read lengths (100 bp), but was not noticeable when using the shorter reads in our first dataset (Figure 3-13). This effect should not be an issue when using hybridization enrichment, where ligation of fragments is not needed.

In order to improve our coverage, we used Bowtie's default parameter, which aligns the first 28 bases of each read, allowing no more than two mismatches. To focus

on *GRIP2* alignments, we provided a fasta reference of 60 kb covering the *GRIP2* locus. A total of 6.4 million reads (5.6% of all reads) aligned to our reference template of the *GRIP2* locus. The depth of coverage for each amplicon pool is shown in Figure 3-14. For exonic positions, the average allelic coverage was 60.8×, and the minimum coverage was 10×; 99.9% of exonic positions were covered at least 15× per allele, and 98.5% were covered at least 30× per allele.

We did not apply Srfim base calls to our variant calling as Srfim has not yet been fully adapted to the newer sequencing chemistry used with this cohort. For variant calling, we tested Syzygy and SERVIC^{4}E, the two most sensitive software identified in our first dataset when using only the standard Illumina base calls (Table 3-3). Syzygy was provided with a template-adjusted dbSNP file and a total allele number of 40 as input parameters. All other parameters were run at default. Syzygy made a total of 474 variant calls across 24 pools (74 unique variant calls). Of the 74 unique calls made, 36 were exonic changes. SERVIC^{4}E was run using a trim value of 25 and a total allele number of 40. All other parameters were run at default. SERVIC^{4}E made a total of 378 variant calls across 24 pools (68 unique variant calls). Of the 68 unique calls made, 33 were exonic changes. Between Syzygy and SERVIC^{4}E, a total of 42 unique exonic sequence variant calls were made (Table 3-8 and Table 3-9).

For validation of these results, we again targeted variants within exons for Sanger sequencing. Sanger data were successfully obtained from individual samples in at least one pool for 41 of the 42 exonic variants. Genotypes for validated samples are indicated in Table 3-10. Results are summarized in Table 3-11 and include any intronic variant

pools that were collaterally Sanger sequenced successfully. Of the 41 exonic variants checked, 29 were valid. Sixteen were identified as occurring only once in the entire cohort of 480 individuals. Syzygy achieved a high sensitivity of 85.5% but a fairly low specificity of 59.4%. Of the 16 valid rare exonic variants, 13 (81.25%) were identified. The MCC score was low (45.9%), primarily as a result of the low specificity (Table 3-11). SERVIC^{4}E achieved a higher sensitivity of 96.4% and a higher specificity of 93.8%. All 16 valid rare exonic variants were identified and a high MCC score (89.9%) was obtained. The combined analysis of the first and second cohorts identified 47 valid coding variants, of which 30 were present only once in each cohort.

3.3 Discussion

We have developed a strategy for targeted deep sequencing in large sample cohorts to reliably detect rare sequence variants. This strategy is highly flexible in study design and well suited to focused resequencing of candidate genes and genomic regions from tens to hundreds of kilobases. It is cost-effective due to substantial cost reductions provided by sample pooling prior to target enrichment and by the efficient utilization of next-generation sequencing capacity using indexed libraries. Though we utilized a PCR method for target enrichment in this study, other popular enrichment methods, such as microarray capture and liquid hybridization [117-119], can be easily adapted for this strategy.

Careful normalization is needed during sample pooling, PCR amplification, and library indexing, as variations at these steps will influence detection sensitivity and

specificity. While genotyping positive pools will be needed for validation of individual variants, only a limited number of pools require sequence confirmation as this strategy is intended for discovery of rare variants.

SERVIC^{4}E is highly sensitive to the identification or rare variants with minimal contamination by false positives. It consistently outperformed several publicly available analysis algorithms, generating an excellent combination of sensitivity and specificity across base-calling methods, sample pool sizes, and Illumina sequencing chemistries in this study. As sequencing chemistry continues to improve, we anticipate that our combined sample pooling, library indexing, and variant calling strategy should be even more robust in identifying rare variants with allele frequencies of 0.1 to 5%, which are within the range of the majority of rare deleterious variants in human diseases.

3.4 Materials and Methods

3.4.1 Sample pooling and PCR amplification

De-identified genomic DNA samples from unrelated patients with intellectual disability and autism, and normal controls were obtained from Autism Genetics Research Exchange (AGRE), Greenwood Genomic Center, SC, and other DNA repositories [129]. An informed consent was obtained from each enrolled family at the respective institutions. The Institutional Review Board at the Johns Hopkins Medical Institutions approved this study.

DNA concentration from each cohort of 480 samples in 5 × 96-well plates was measured using a Quant-iT™ PicoGreen® dsDNA Kit (Invitrogen, Carlsbad, CA, USA) in a Gemini XS Microplate Spectrofluorometer. These samples were normalized and mixed at equal molar ratio into 12 pools of 40 samples each (first cohort) or 24 pools of 20 samples each (second cohort). For convenience, first cohort samples from the same column of each 5 × 96-well plate were pooled into a single well (Figure 3-1). The same principle was applied to the second cohort, with the first two and a half plates combined into the first 12 pools, and the last two and a half plates combined into the last 12 pools (Figure 3-12). PCR primers for individual amplicons were designed using the Primer3 program. PCR reaction conditions were optimized to result in a single band of the expected size. Phusion Hot Start High-Fidelity DNA Polymerase (Finnzymes, Thermo Fisher Scientific, Waltham, MA, USA) and limited amplification cycles (n = 25) were used to minimize random errors introduced during PCR amplification. PCR reactions were carried out in a 20- μl system containing 50 ng of DNA, 200 μM of dNTP, 1× reaction buffer, 0.2 μM of primers, and 0.5 units of Phusion Hot Start High-Fidelity Polymerase in a thermocycler with an initial denaturation at 98°C for 30 seconds followed by 25 cycles of 98°C for 10 seconds, 58 to 66°C for 10 seconds, and 72°C for 30 seconds. The annealing temperature was optimized for individual primer pairs. Successful PCR amplification for individual samples was then verified by agarose gel electrophoresis. The concentration for individual PCR products was measured using the Quant-iT™ PicoGreen® dsDNA Kit (Invitrogen) on Gemini XS Microplate Spectrofluorometer, and converted to molarity. PCR amplicons intended for the same

indexed library were combined at equal molar ratio, purified using QIAGEN (Hilden, Germany) QIAquick PCR Purification Kit, and concentrated using Microcon YM-30 columns (Millipore, Billerica, MA, USA).

3.4.2 Amplicon ligation and fragmentation

The pooled amplicons were ligated using a Quick Blunting and Quick Ligation Kit (NEB, Ipswich, MA, USA) following the manufacturer's instructions. For blunting, a 25- µl reaction system was set up as follows: 1× blunting buffer, 2 to 5 µg of pooled PCR amplicons, 2.5 µl of 1 mM dNTP mix, and 1 µl of enzyme mix including T4 DNA polymerase (NEB #M0203) with 3' → 5' exonuclease activity and 5' → 3' polymerase activity and T4 polynucleotide kinase (NEB #M0201) for phosphorylation of the 5' ends of blunt-ended DNA. The reaction was incubated at 25°C for 30 minutes and then the enzymes were inactivated at 70°C for 10 minutes. The blunting reaction products were purified using a MinElute PCR purification column (QIAGEN) and then concentrated using a Microcon YM-30 column (Millipore) to 5 µl volume in distilled water. For ligation, 5 µl of 2× Quick-ligation buffer was mixed with 5 µl of purified DNA. Quick T4 DNA ligase (1 µl; NEB) was added to the reaction mixture, which was incubated at 25°C for 5 minutes and then chilled on ice. The reaction product (0.5 µl) was checked for successful ligation using 1.5% agarose gel electrophoresis. The ligation products were then purified using a MinElute PCR purification column (QIAGEN). Random fragmentation of the ligated amplicons was achieved using either one of the two methods: (1) nebulization in 750 µl of nebulization buffer at 45 psi for 4 minutes on ice following a

standard protocol (Agilent); or (2) using a NEBNext dsDNA Fragmentase Kit following the manufacturer's instructions (NEB). One-twentieth of the product was analyzed for successful fragmentation to a desired range using 2% agarose gel electrophoresis.

3.4.3 Library construction and Illumina sequencing

The Multiplexing Sample Preparation Oligonucleotide Kit (Illumina PE-400-1001) was used to generate 1 × 12 (first cohort) and 2 × 12 (second cohort) individually indexed libraries following the manufacturer's instructions. The indexed libraries were quantified individually and pooled at equal molar quantity. The concentration of the final pooled library was determined using a Bioanalyzer (Agilent). All 12 pooled libraries from the first cohort were run in one lane of a flow cell on an Illumina Genomic Analyzer II (GAII). The first 12 pooled libraries from the second cohort were run in one lane of a GAII, while the last 12 pooled libraries were run in another lane in the same flow cell. Illumina sequencing was done at the UCLA DNA Sequence Core and Genetic Resource Core Facility at the Johns Hopkins University.

3.4.4 Sequence data analysis

Raw intensity files and fastq-formatted reads were provided for both cohort datasets. Output had been calibrated with control lane PhiX DNA to calculate matrix and phasing for base calling. A custom script was used on first cohort sequence data to identify the 12 Illumina barcodes from the minimum edit distance to the barcode and assign a read to that pool if the distance index was unique (demultiplexing). Second

cohort sequence data were provided to us already demultiplexed. Read mapping was done independently on each pool using BOWTIE (options: -v 3 for first cohort, default for second cohort). As reference templates, hg19 was used for the first cohort and a 60-kb fragment of the *GRIP2* regions was used for the second cohort (*GRIP2* region-chr3:14527000-14587000).

Variant calling using SAMtools was done independently on each pool using SAMtools' deprecated algorithms (options: pileup -vc -A -N 80). Variants identified were first filtered by eliminating non-*GRIP2* variants, and then filtered by consensus quality and SNP quality scores (cq ≥ 196 and sq ≥ 213 for Illumina base calls; cq ≥ 161 and sq ≥ 184 for Srfim base calls). Deprecated (Maq) algorithms were used, as the current SAMtools variant-calling algorithms failed to call all but the most common SNPs. Quality cutoff is based on the 95th percentile of scores in the quality distributions observed amongst all reported SAMtools variants in the *GRIP2* alignment region, after excluding variants with the maximal quality score of 235). Reads were base-called using Srfim using default filtering and quality parameters.

SERVIC^{4}E was given the location of sorted alignment (BAM) files. Though alignment files are maintained separately for each pool, the locations of each file are given all together. A trim value was set at 25. This trims 25 bases away from the ends of aligned amplicons, so that variant calling is focused away from primer regions. Use of shorter primers during library preparation allows for a smaller trim value. Hybridization enrichment will always result in a trim value of zero, regardless of what trim value is actually set. The total number of alleles in each pool was also provided as input (80

alleles for the first cohort; 40 alleles for the second cohort). SERVIC^{4}E (release 1) does not call insertions or deletions.

SNPSeeker was run on first cohort data using author recommended parameters. Reads (Illumina base calls) were converted to SCARF format. Srfim base calls could not be used due to an unknown formatting issue after SCARF conversion. Alignment was conducted against *GRIP2* template sequences. Exons 6 and 7 reference sequences were merged so that their alignments could be used as a negative control to develop an error model. All 47 cycles were used in the alignment, allowing for up to three mismatches. Alignments were tagged and concatenated, and an error model generated using all 47 cycles, allowing for up to three mismatches, and using no pseudocounts. The original independent alignment files (pre-concatenation) were used for variant detection. As per recommendation by the authors, the first third of cycles was used for variant detection (15 cycles). A P-value cutoff of 0.05 was used. Lower cutoffs generated worse results when checked against our validation database.

CRISP was run using default parameters. A CRISP-specific pileup file was generated using the author-provided sam_to_pileup.py script and not generated using the pileup function in SAMtools. A separate pileup was generated for each pool for both alignments from Illumina base calls and alignment from Srfim base calls. A BED file was provided to focus pileup at *GRIP2* loci. CRISP analysis for variant detection was conducted using all 47 cycles and a minimum base quality of 10 (default). All other parameters were also kept at default.

Syzygy [112,128] was run on both cohorts using 80 and 40 as the total number of alleles, respectively. A dbSNP file was provided for known chromosome 3 variants. A TGF file was provided to focus variant calling at *GRIP2* target regions. Hg19 was used as the reference sequence for the first cohort, while the same abridged *GRIP2* sequence that was used by SERVIC^{4}E was also used by Syzygy for the second cohort. All other parameters were run at default.

Reads used for analysis, both Illumina and Srfim base calls, are available through the public data repository at the NCBI (accession number SRP007694). Srfim is available as an R package, while SERVIC^{4}E is available as a set of R scripts. Both are available for download online [130].

3.4.5 Validation by Sanger sequencing

Sanger sequencing of positive pools for variant validation was conducted using the BigDye Terminator v3.1 Cycle Sequencing Kit on an ABI3100 automatic DNA analyzer (Applied Biosystems, Foster City, CA, USA) following the manufacturer's instructions.

Sanger sequencing was done on each sample within a pool separately (40 traces per pool with the first cohort, 20 traces per pool for the second cohort). Only traces with low quality or ambiguous calls were sequenced bidirectionally. In the event that a positive sample was verified at least once in the pool, further sequencing of that pool was halted. Sequencing primers were the same primers used in target enrichment to build the libraries for next-generation sequencing.

Standard sequence alignment software (CodonCode, MacVector) followed by manual investigations of the chromatograms was used to identify any variants that might have been missed by all nine combinations of programs.

3.4.6 Calculation of Matthews Correlation Coefficient

The MCC is intended as a measure of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs), without being influenced by potential extreme sizes by one or more of the groups. An MCC = 1 indicates perfect correlation between predicted results (variants identified by next-generation sequencing and various combinations of base-calling and variant-calling algorithms) and the observed results (validation by Sanger sequencing). An MCC = 0 indicates that the algorithm is no better than random. An MCC = -1 indicates an inverse correlation. MCC = (TP × TN-FP × FN)/SQRT [(TP + FP) × (TP + FN) × (TN + FP) × (TN + FN)]. Sensitivity (true positive rate, recall): TP/(TP + FN). Specificity (true negative rate): TN/(FP + TN). Positive predictive value (precision): TP/(TP + FP). Negative predictive value: TN/(TN + FN). Accuracy: (TP + TN)/(TP + TN + FP + FN). False positive rate (fall-out): 1-True negative rate. False discovery rate: FP/(FP + TP).

3.4.7 Abbreviations

bp: base pair; cq: consensus quality score generated by SAMtools pileup; GAII: Genome Analyzer II (Illumina Sequencing Machine); *GRIP2*: glutamate-receptor interacting protein 2; MCC: Matthews correlation coefficient; PCR: polymerase chain

reaction; SERVIC^{4}E: Sensitive Rare Variant Identification by Cross-pool Cluster: Continuity: and tailCurve Evaluation; SNP: single nucleotide polymorphism; sq: SNP quality score generated by SAMtools pileup.

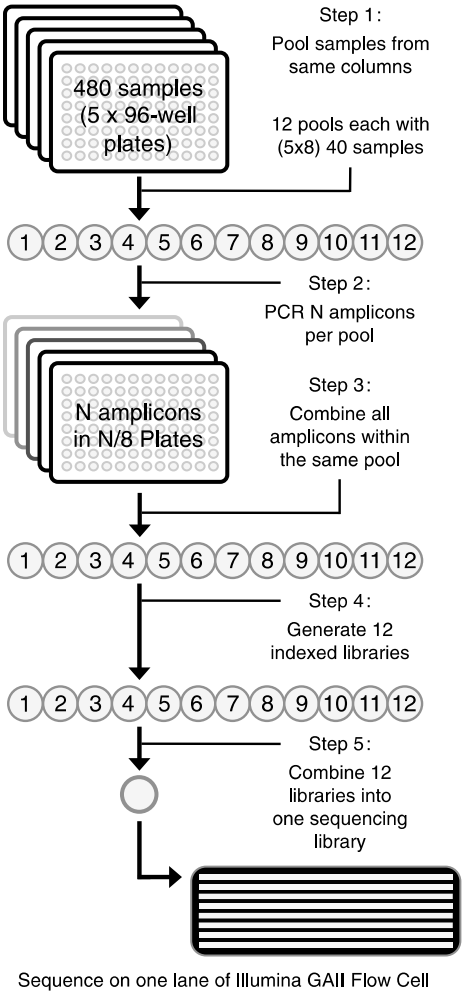
3.4.8 Authors' contributions and acknowledgements

AA and TSN generated Illumina libraries and conducted sequence validation. TSN, HCB, and MAT performed sequence data analysis. RI, SW, and TW conceived of the study, participated in its design, and supervised its execution. TSN, AA, HCB and TW wrote the manuscript. All authors read and approved the final manuscript.

We thank Dr David Valle of Johns Hopkins University for critical reading of this manuscript, Autism Genetics Research Exchange (AGRE) and Greenwood Genetic Center (GGC) for DNA samples used in this study, and Dr Manuel Rivas of the Massachusetts Institute of Technology and the University of Oxford with assistance in running Syzygy. This work was supported in part by a pilot grant (#2487) from Autism Speaks (to TW), a research grant from the Brain Science Institute at Johns Hopkins University, R01HD52680 from NICHD (to TW), and R01HG005220 from NHGRI (to RI). TSN is a student of the Predoctoral Training Program in Human Genetics at Johns Hopkins University.

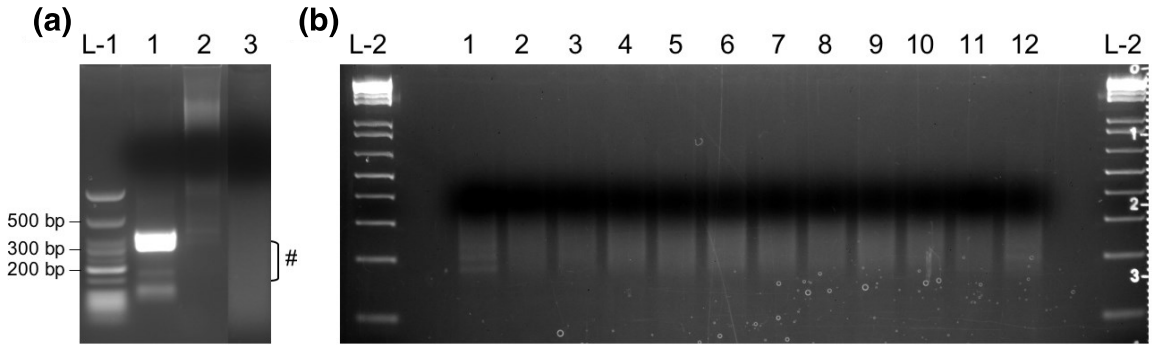
3.5 Figures: Chapter 3

Figure 3-1. Schematic diagram of the sequencing strategy



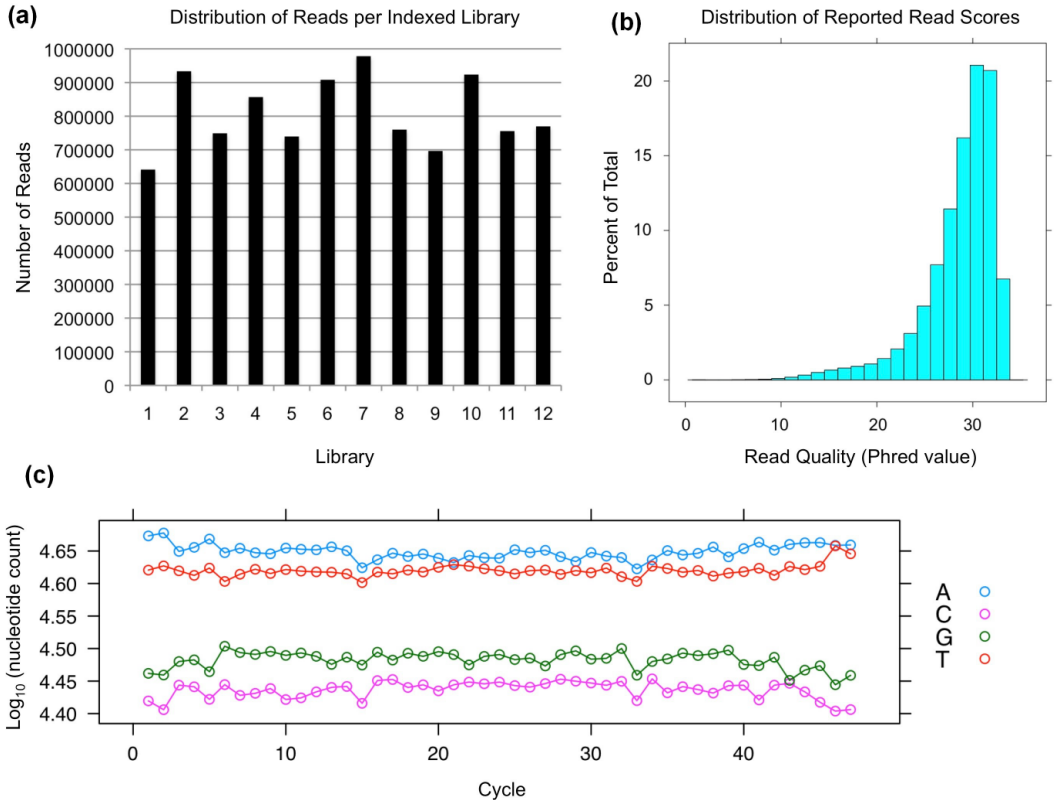
Sample pools of 40 samples × 12 pools were generated from a cohort of 480 individuals for PCR amplification of individual exons. After blunt-ended ligation and random fragmentation, PCR amplicons from individual sample pools were used to generate indexed sequence libraries. The 12 indexed libraries were combined in equal molar amounts and sequenced in one lane of a flow cell using an Illumina GAII.

Figure 3-2. Amplicon ligation, fragmentation, and indexed Illumina libraries



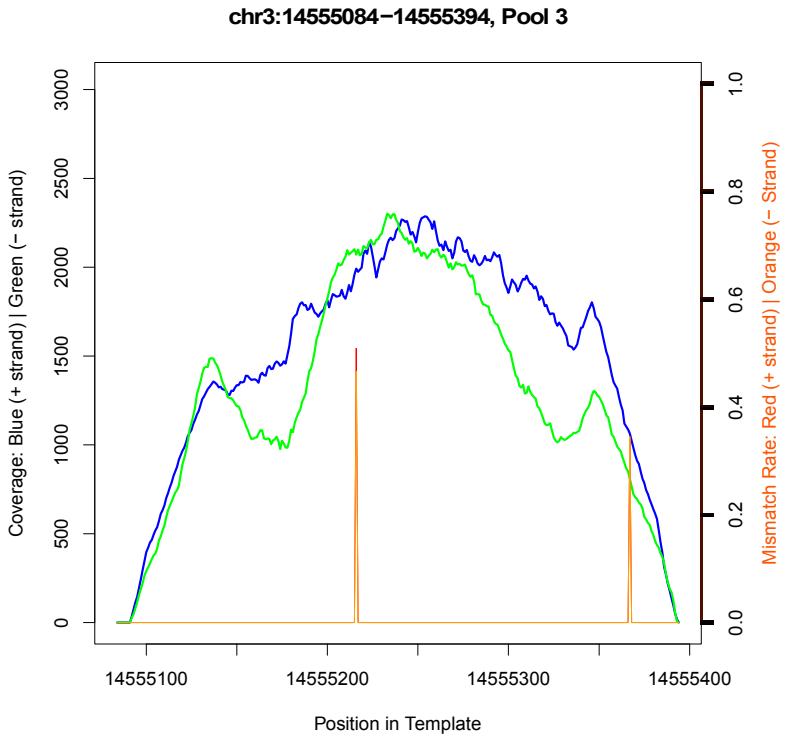
(a) Amplicon ligation and fragmentation: L-1, low molecular weight marker; lane 1, PCR amplicons before ligation; lane 2, PCR amplicons after ligation; lane 3, random fragmentation using Fragmentase (NEB). #The bracket indicates fragments of desired length. (b) Indexed Illumina libraries: L-2, 1-kb ladder; lanes 1 to 12, size distribution of 12 indexed Illumina libraries.

Figure 3-3. Quality assessment of the Illumina sequence data



(a) Number of reads with barcodes that passed Illumina filtering and aligned to the reference templates using Bowtie from individually indexed libraries (n = 12). Range, 641 k to 978 k reads; mean ± standard deviation, 809 k ± 107 k. (b) Percentage of total (unaligned) reads that fall into a mean Phred quality interval. Note > 80% of the reads have mean Phred quality scores ≥25. (c) Nucleotide content as a function of sequencing cycles (n = 47). Note that the nucleotide proportions closely match the expected proportions as determined from the templates.

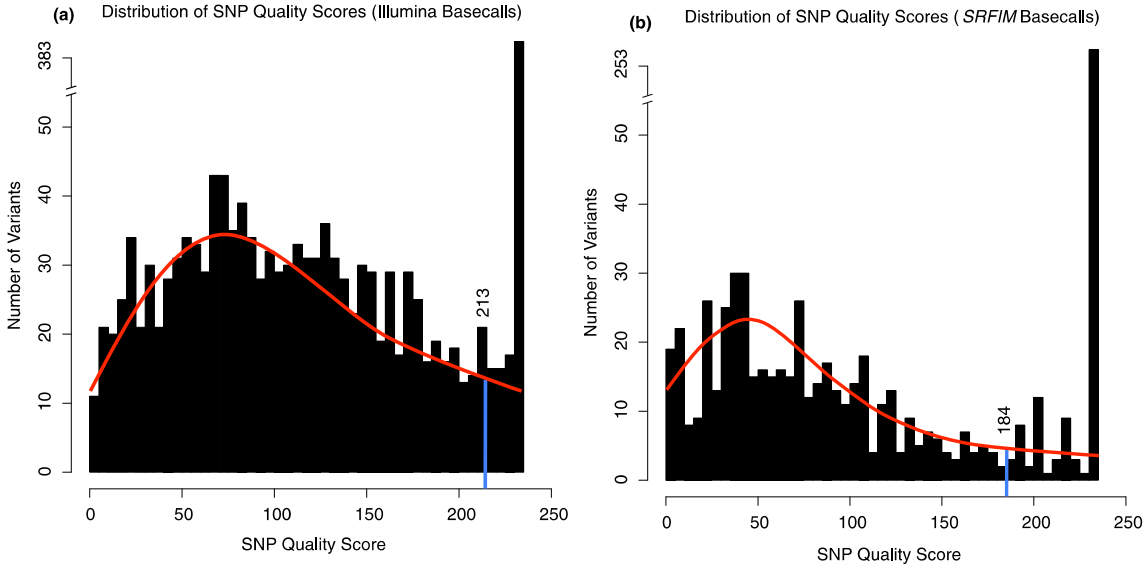
Figure 3-4. Depth of coverage of a selected representative amplicon-pool derived from first cohort sequencing data



Blue line depicts absolute coverage for plus-strand aligned reads. Green line depicts coverage of minus-strand aligned reads. Light red line indicates presumptive mismatch rate determined from plus-strand aligned reads. Light orange line indicates presumptive mismatch rate determined from minus-strand aligned reads. Ratio of mismatch rate between plus and minus strands is later incorporated into the tailcurve factor used in filtering by SERVIC^{4}E. Depth of coverage for all amplicon-pools is available for download as Additional File 1 from the publication [51] or from the following URL.

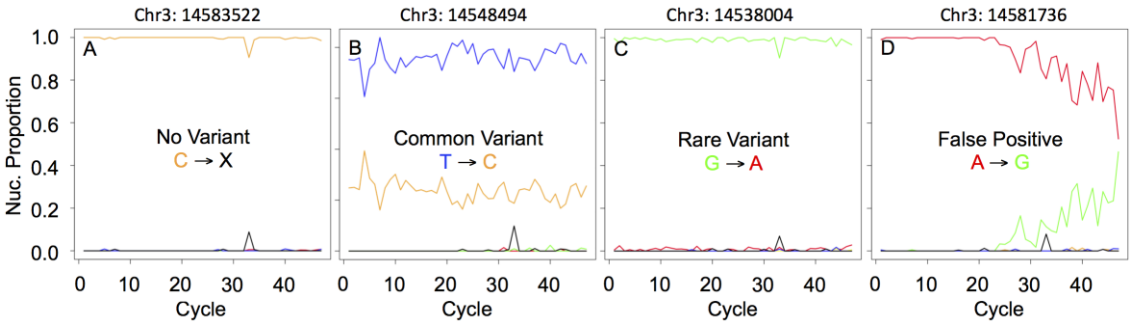
[http://genomebiology.com/content/supplementary/gb-2011-12-9-r93-s1.pdf]

Figure 3-5. Distribution of quality scores from SAMtools pileup



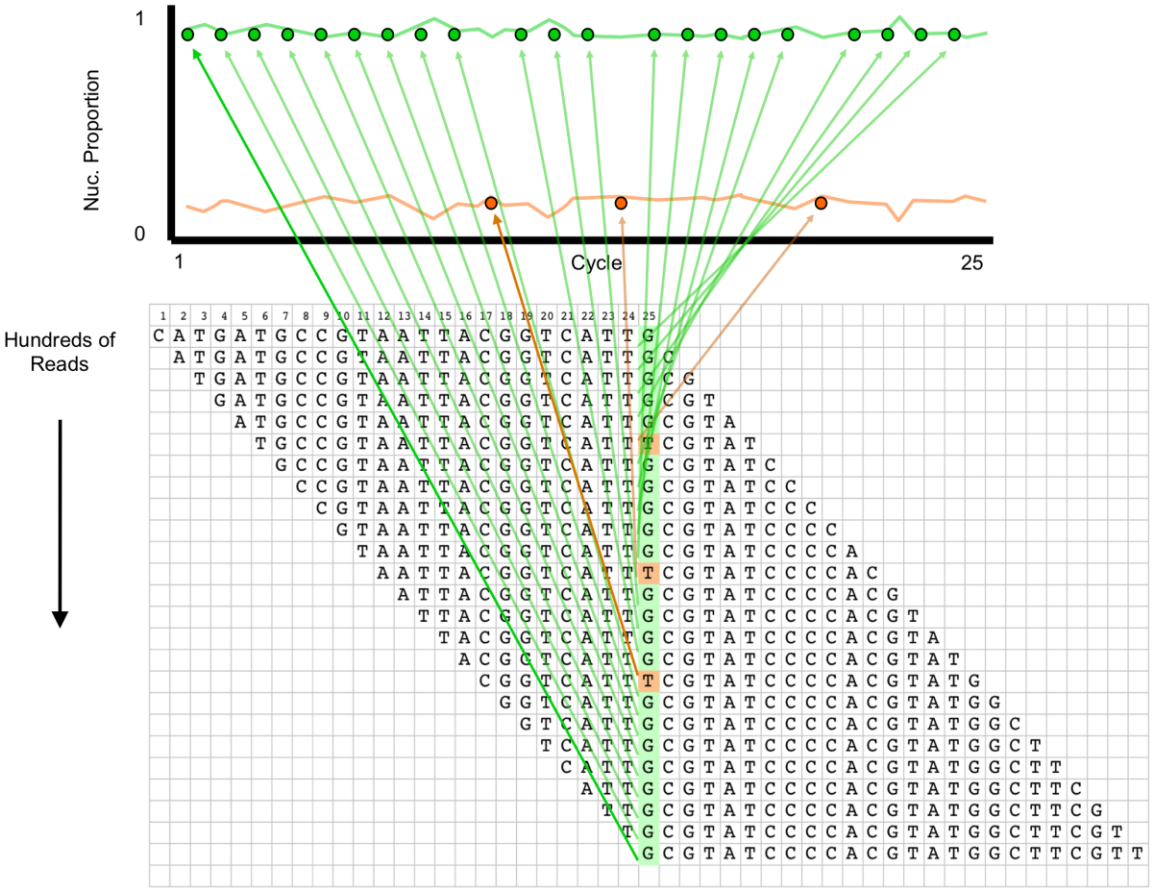
Filtering was conducted at the 95th percentile of the consensus and SNP quality distributions reported by SAMtools; only the distribution of SNP quality values is depicted here. The blue bar is the 95th percentile score cutoff, discounting variants with max score. (a) SNP quality scores derived from Illumina base calls. (b) SNP quality scores derived from Srfim base calls.

Figure 3-6. Representative base reads and tailcurves for common and rare variants and error calls



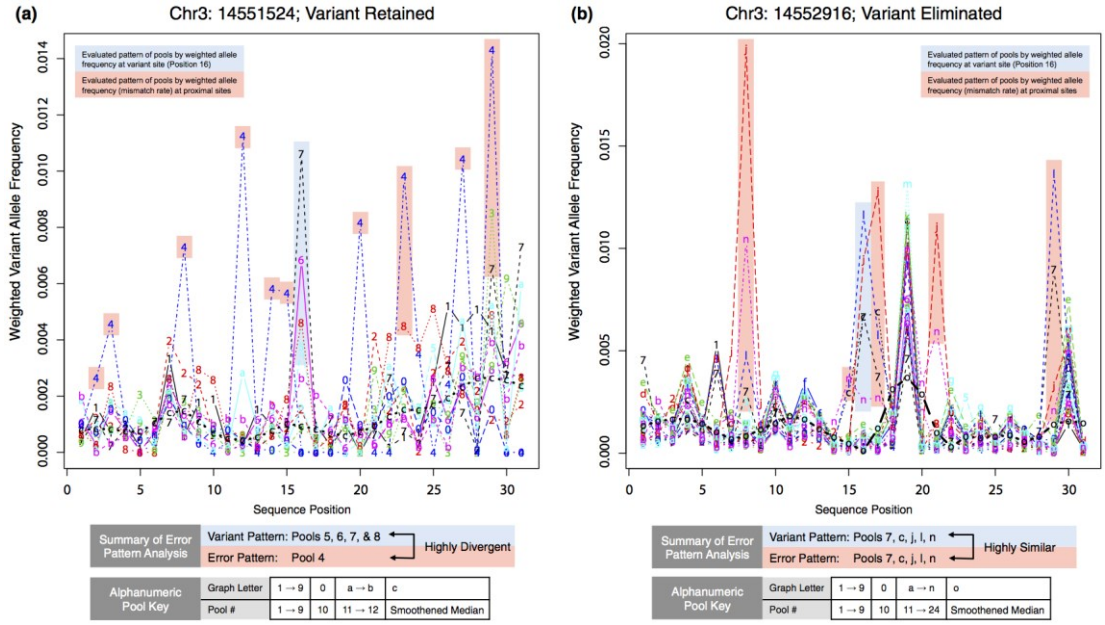
(a) Position with no variant. (b) Position with a common variant. (c) Position with a rare variant. (d) Position with a false positive call.

Figure 3-7. Description of tailcurve (nucleotide proportion at individual cycles along the sequence read)



With perfect random fragmentation, a given position and its associated base calls (consensus and variant) should be represented at multiple sequencing cycles. With high coverage, a particular base call will be present for that position at all or most cycles. Example: for a sequencing module of 25 cycles with several hundred (24 shown) overlapping reads covering the highlighted position, all the cycles are represented by 'G', with variant reads producing the 'T' at a handful of cycles (potential variant).

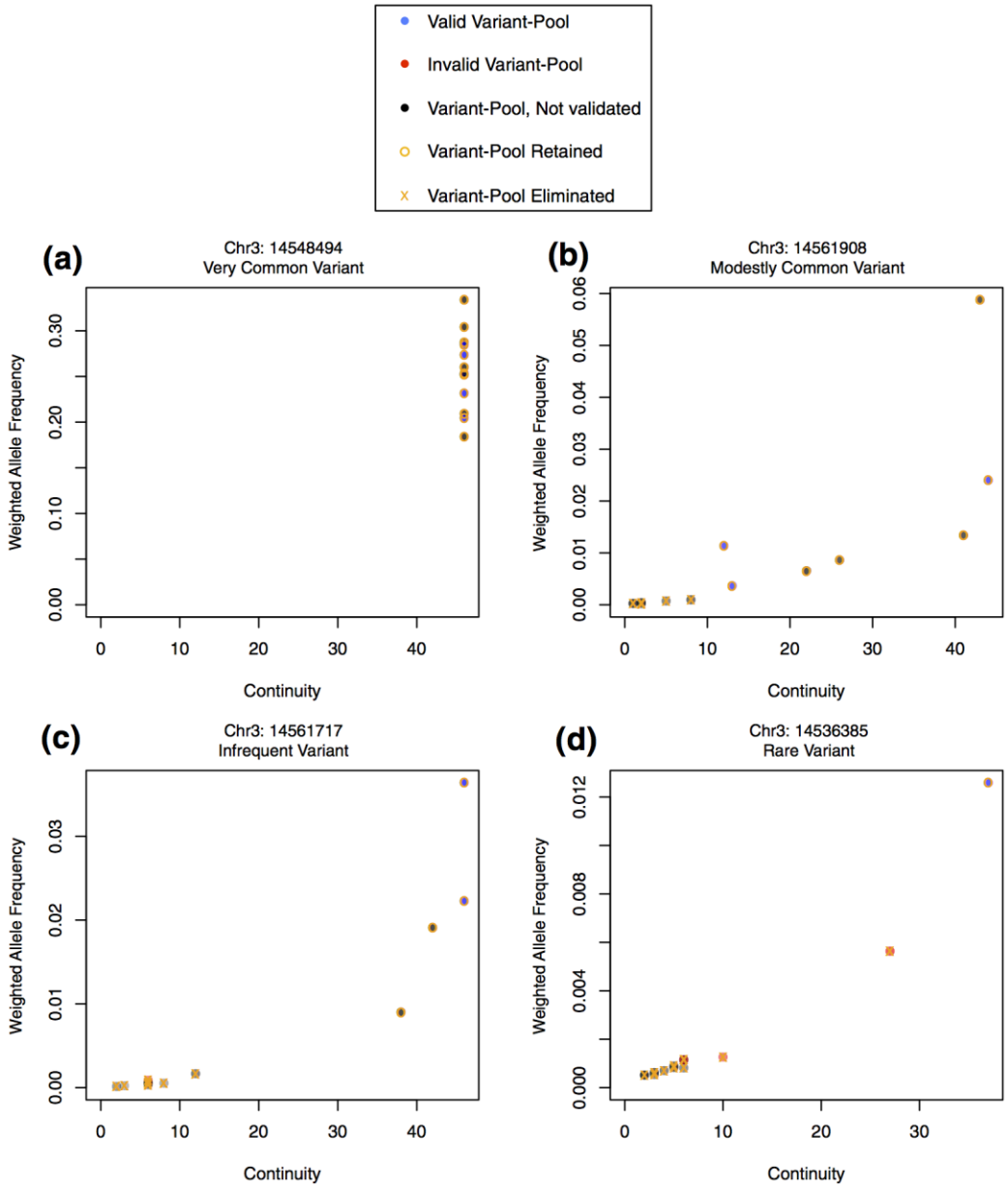
Figure 3-8. Local pool patterns for error analysis



X-axes denote position in a local sequence. Position 16 is the variant site being analyzed, positions 1 to 15 are immediately upstream and positions 17 to 31 are downstream. Y-axes denote the weighted allele frequency of the most prominent non-reference allele at each position (mismatch rate). Individual pools are denoted by a unique line pattern, color, and number/letter. Light shading indicates the pool pattern that is most recognizable by SERVIC^{4}E for each position. (a) Local weighted allele frequencies for each pool at position 14,551,524 ± 15 in chromosome 3 from the first cohort. The evaluated pattern of pools at the variant position involves pools 5, 6, 7, and 8, while the evaluated pattern at proximal positions involves pool 4. The dissimilarity between patterns results in retention of chr3:14551524 as a variant site. (b) Local weighted allele frequencies for each pool at position 14,552,916 ± 15 in chromosome 3 from the second cohort. The evaluated pattern of pools at the variant position involves pools 7, 13 (c), 20

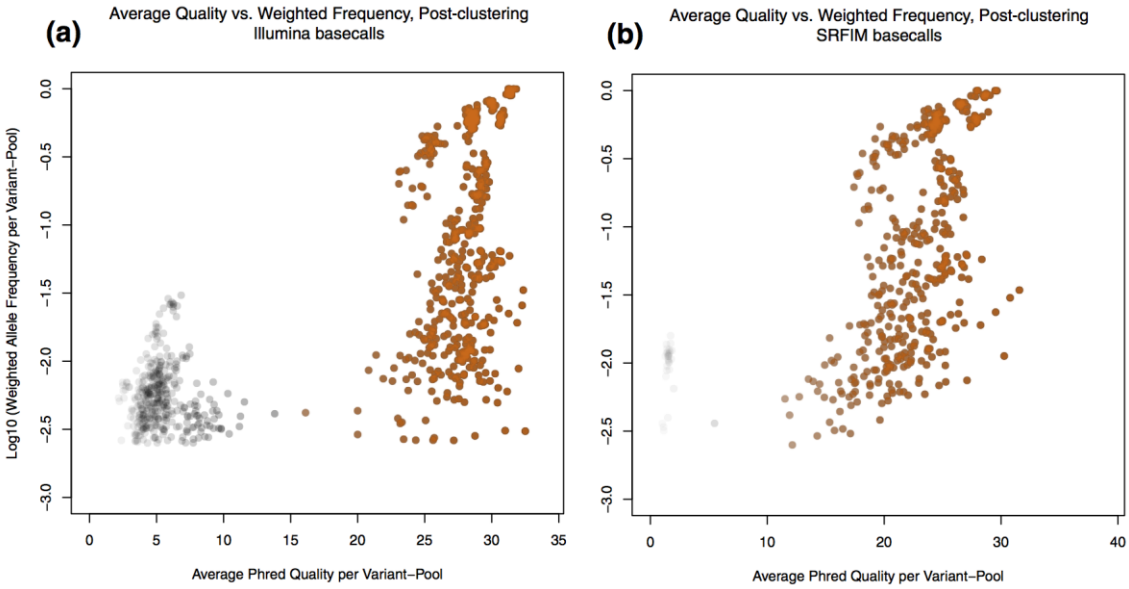
(j), 22 (l), and 24 (n), and the evaluated pattern at proximal positions involves the same pools. The similarity between patterns results in elimination of chr3:14552916 as a variant site.

Figure 3-9. Continuity vs. weighted allele frequency curves for selected variants



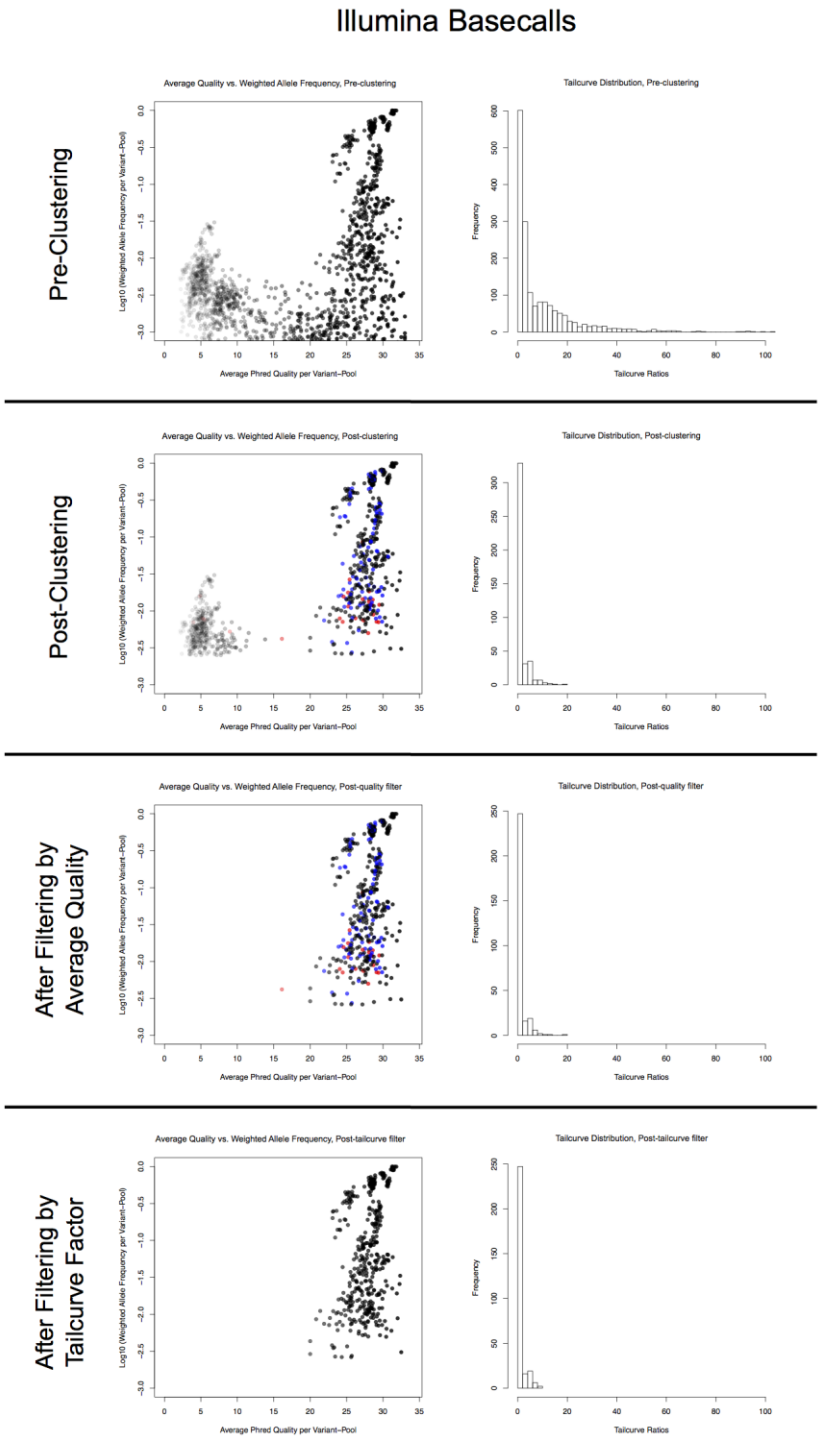
(a) Very common variant present in all 12 pools. (b) Modestly common variant present in the majority of pools. (c) Infrequent variant present in a minority of pools. (d) Rare variant present in only one pool. Gold circles indicate variant pools retained by cluster analysis, while a gold 'x' indicates a variant pool that has been eliminated.

Figure 3-10. Average quality vs. weighted allele frequency for variant pools after filtering by clustering

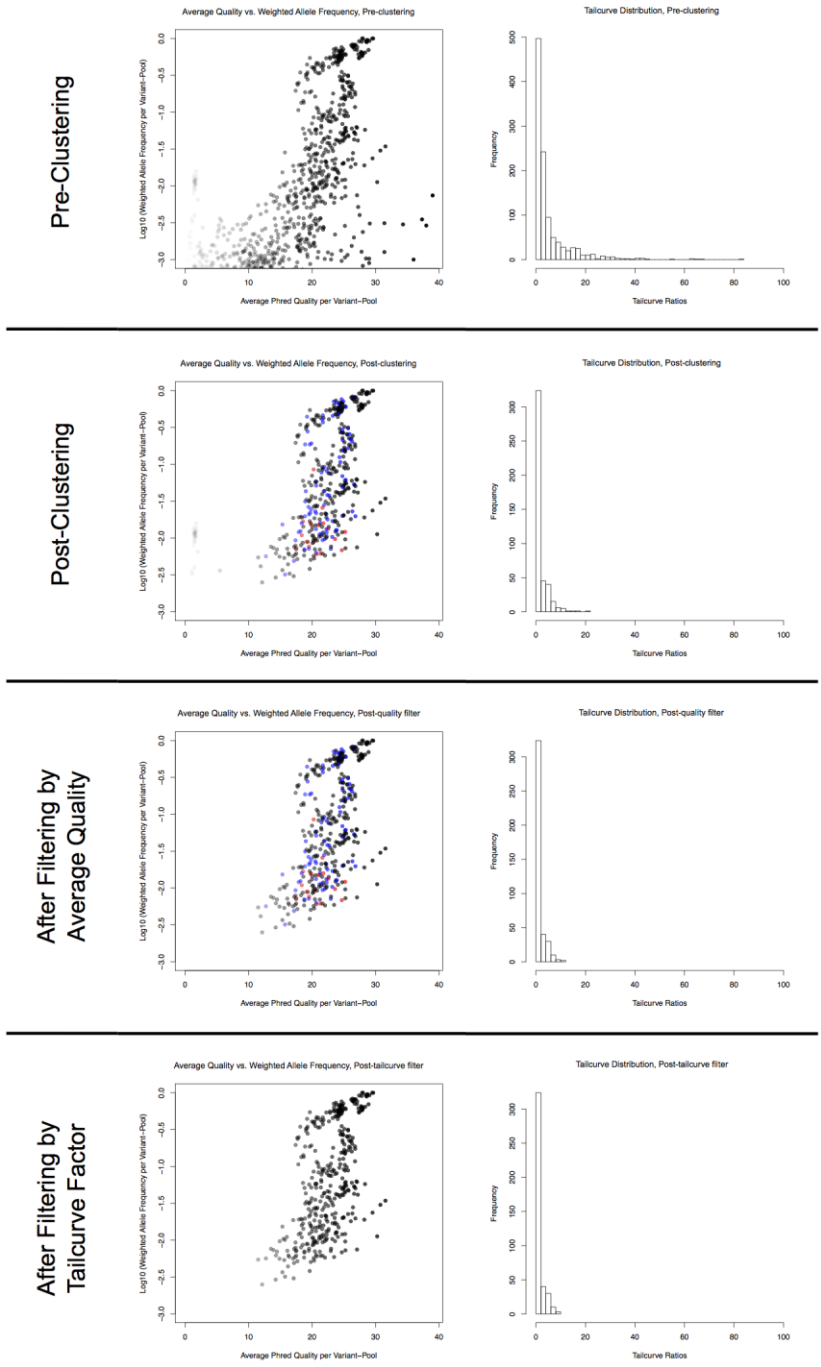


The X-axis is average Phred sequencing quality score and the Y-axis is weighted allele frequency (ratio of the sum of Phred quality scores for the variant allele at a position to the sum of all Phred quality scores at that position) in log10 scale. Characteristic distribution shapes make it possible to cluster and retain only high quality variants (orange points). (a) Illumina base calls. (b) Srfim base calls.

Figure 3-11. Diagrammatic output of first three filtering steps using SERVIC^{4}E on first cohort data

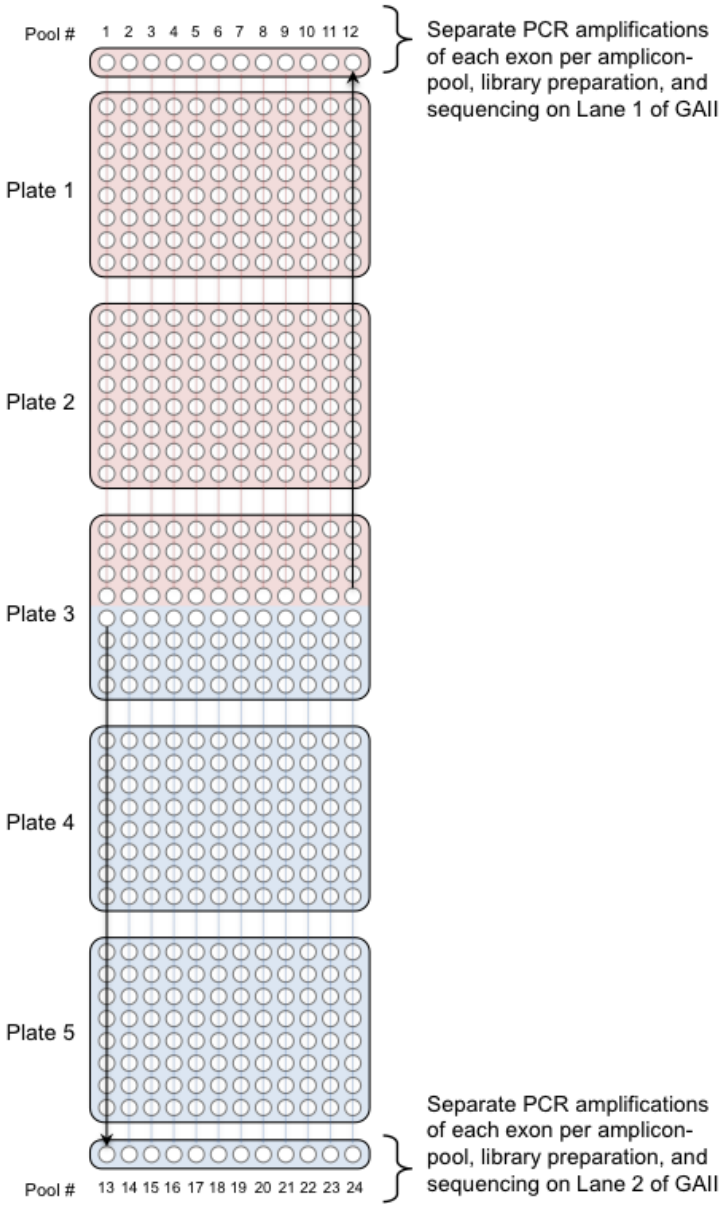


SRFIM Basecalls



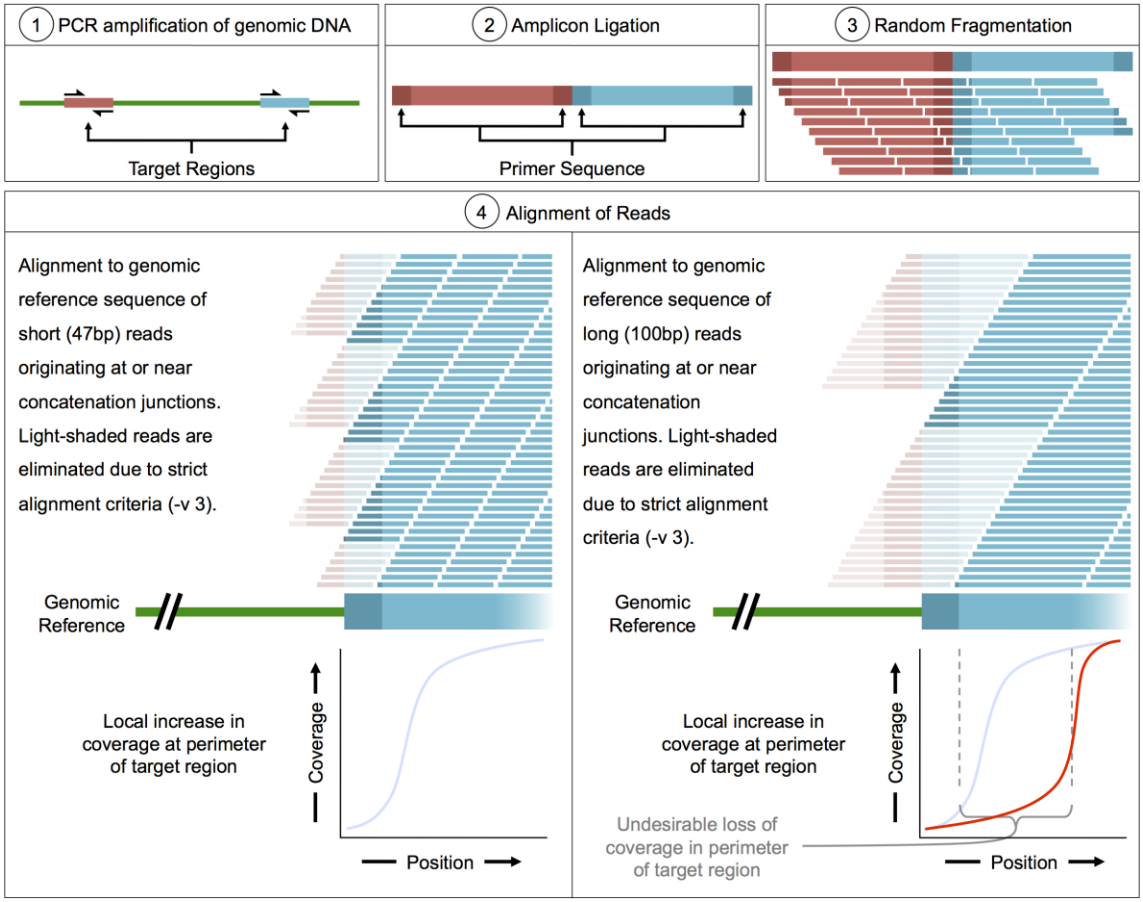
The top panel set uses Illumina base calls. The bottom panel set uses Srfim base calls. Individual filtering steps progress while moving down each panel. Colored dots incorporate validation data for visualization purposes; blue dots are valid variant pools and red dots are invalid variant pools. Within each panel, the graphs on the left are Average quality versus Weighted allele frequency distributions. X-axis is average Phred quality for each variant-pool. Y-axis is log10 of weighted allele frequency. Histograms on the right depict the frequency of evaluated tailcurve ratios across bins of length = 2.

Figure 3-12. Pooling strategy for second cohort samples



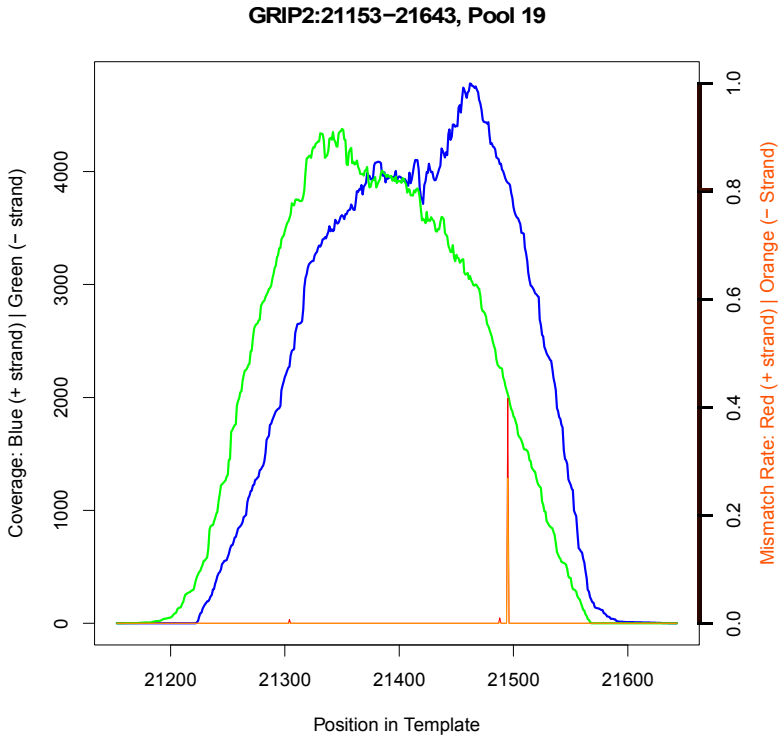
Example: Normalized DNA samples from column 12 of plates 1 and 2 as well as samples from plate 3, column 12, rows A, B, C, and D are pooled together to form pool 12. Normalized DNA samples from column 1 of plates 4 and 5 as well as samples from plate 3, column 1, rows E, F, G, and H are pooled together to form pool 13.

Figure 3-13. Effect of strict alignment on coverage from concatenated amplicons



Panel 1 indicates targets for amplification (primers denoted by black half-arrows). Color-coding for each unique target region is retained in all panels. Panel 2 depicts ligation (concatenation) of amplicons. Only two amplicons are depicted; in practice many amplicons ligate together in a row. Darker shaded regions are from primer sequence. Panel 3 depicts random fragmentation to generate 150- to 200-bp segments for sequencing. Panel 4 depicts subsequent strict alignment of short (left) and long (right) reads to genomic reference sequence.

Figure 3-14. Depth of coverage of a selected representative amplicon-pool derived from second cohort sequencing data



Blue line depicts absolute coverage for plus-strand aligned reads. Green line depicts coverage of minus-strand aligned reads. Light red line indicates presumptive mismatch rate determined from plus-strand aligned reads. Light orange line indicates presumptive mismatch rate determined from minus-strand aligned reads. Ratio of mismatch rate between plus and minus strands is later incorporated into the tailcurve factor used in filtering by SERVIC^{4}E. Depth of coverage for all amplicon-pools is available for download as Additional File 11 from the publication [51] or from the following URL.

[http://genomebiology.com/content/supplementary/gb-2011-12-9-r93-s11.pdf]

3.6 Tables: Chapter 3

Table 3-1. Effect of sequential filtering by SERVIC^{4}E on variant output

	Dataset 1		Dataset 2
	Illumina	**Srfim**	**Illumina**
Number of variant-pools after filtering			
Prior to cluster analysis	1,656	1,056	7,272
After cluster analysis	929	905	5,123
After average quality filtering	437	342	422
After tailcurve filtering	341	340	412
After error-modeling	333	334	378

Reported values indicate the total number of variant positions (across all pools) that remain after each filtering step. Dataset 1: sequencing output of *GRIP2* exons in a first cohort of 480 samples. Dataset 2: sequencing output of *GRIP2* exons in a second independent cohort of 480 samples.

Table 3-2. Partial list of variant calls from first cohort analyses

Position	Pool	REF	VAR	Illumina SAMtools	Srfim Samtools	SNPSeeker	Illumina CRISP	Srfim CRSIP	Illumina SERVIC4E	Srfim SERVIC4E	Illumina Syzygy	Srfim Syzygy	Rarity	P	Valid	Exon	Dist
14535190	2	T	Y	+	+				+		+	+		+	-	25	240
14535190	4	T	Y	+	+		+	+	+	+	+	+		+	+	25	240
14535190	8	T	Y								+	+		+		25	240
14535190	9	T	Y	+	+		+	+	+		+	+		+	+	25	240
14535190	11	T	Y						+		+	+		+	-	25	240
14535265	7	C	M	+										+		25	165
14535341	2	G	S	+	+				+						+	25	89
14535341	5	G	S	+	+	+	+	+	+	+	+	+			+	25	89
14535341	9	G	S	+	+		+	+	+	+	+	+			+	25	89
14535341	11	G	S	+	+				+	+	+	+				25	89
14535341	12	G	S	+	+	+	+	+	+	+	+	+			+	25	89
14536319	3	C	Y									+				24	272
14536319	4	C	Y	+	+		+	+	+	+	+	+	+		+	24	272
14536385	1	C	Y	+	+		+	+	+	+	+		+	+	+	24	206
14536385	2	C	Y	+	+					+	+			+	-	24	206
14536396	4	C	Y	+										+	-	24	195
14536508	1	A	R							+		+			-	24	83
14536508	2	A	R		+					+	+	+			-	24	83

All positions are given in reference to chromosome 3 of hg19. For each program, a '+' value indicates that a variant call was made by that program for that variant position and pool. Column 'P' indicates the position is in exonic sequence (not intronic). Column 'Valid' indicates validation results for each variant-pool tested; '+' indicates a valid call and '-'indicates an invalid call. Column 'Dist' indicates the position of the variant call in each amplicon. The full list of variant calls from analyses of the first cohort is available for download as Additional File 4 from the publication [51] or from the following URL.

[http://genomebiology.com/content/supplementary/gb-2011-12-9-r93-s4.xls]

Table 3-3. Validation analysis of variant calling from first cohort samples

	Illumina					Srfim			
	SNPSeeker	SAMTools*	CRISP	Syzygy	*SERVIC^{4}E*	SAMTools#	CRISP	Syzygy	*SERVIC^{4}E*
Variant identification and validation									
True positive	56	83	79	85	88	80	78	80	84
True negative	72	41	44	49	62	60	50	51	61
False positive	2	32	29	24	11	13	23	22	12
False negative	34	7	11	5	2	10	12	10	6
Statistical analysis (%)									
Sensitivity	62.22	92.22	87.78	94.44	97.78	88.89	86.67	88.89	93.33
Specificity	97.26	56.16	60.27	67.12	84.93	82.19	68.49	69.86	83.56
PPV	96.55	72.17	73.15	77.98	88.89	86.02	77.23	78.43	87.50
NPV	67.62	85.42	80.00	90.74	96.88	85.71	80.65	83.61	91.04
FPR	2.74	43.84	39.73	32.88	15.07	17.81	31.51	30.14	16.44
FDR	3.45	27.83	26.85	22.02	11.11	13.98	22.77	21.57	12.50
Accuracy	77.91	76.07	75.46	82.21	92.02	85.89	78.53	80.37	88.96
MCC	61.78	52.79	50.54	65.05	84.22	71.41	56.50	60.37	**77.72**
Rare exonic variant detection and validation									
Detected total variants (n = 15)	7	15	13	14	15	14	12	10	15
Detection rate (%)	46.67	100	86.67	93.33	100	93.33	80.00	66.67	100

Descriptions of calculations used in statistical data analysis are provided in Materials and methods. FDR, false discovery rate; FPR, false positive rate; MCC, Matthews correlation coefficient; NPV, negative predictive value; PPV, positive predictive value. SNPSeeker: variant called uses the first 15 cycles (author recommended). P-value cutoff of 0.05 gave the best results. SAMTools* pileup -A -N 80: filtered for variants with consensus quality score ≥ 194 and SNP quality scores ≥ 213. CRISP: all 47 cycles used in alignment. Minimum base quality set to a default of 10. Syzygy: default parameters used. SAMTools# pileup -A -N 80: filtered for variants with consensus quality score ≥ 161 and SNP quality scores ≥ 184. SERVIC^{4}E: default parameters used.

Table 3-4. Partial list of genotyping results from individual first cohort samples

	Position (chr3)	14547182	14547198	14547331	14548277	14548457	14548474	14548494
	Reference Allele	G	G	C	G	G	G	A
	Variant Allele	R	R	Y	R	R	R	R
Row	Column (pool)	Patient Identifier						
D	05	Plate1-D-05			-	-	-	+
E	05	Plate1-E-05			-	-	-	-
F	05	Plate1-F-05			-	-	-	+
G	05	Plate1-G-05			-	-	-	+
H	05	Plate1-H-05			-	-	-	-
A	06	Plate1-A-06			-	-	-	-
B	06	Plate1-B-06			-	-	-	+
C	06	Plate1-C-06			-	-	-	-
D	06	Plate1-D-06			-	-	-	-
E	06	Plate1-E-06			-	-	-	-
F	06	Plate1-F-06			-	-	-	+
G	06	Plate1-G-06			-	-	-	-
H	06	Plate1-H-06			-	-	-	++
A	07	Plate1-A-07	-					
B	07	Plate1-B-07	-					
C	07	Plate1-C-07	-					
D	07	Plate1-D-07	-					
E	07	Plate1-E-07	-					
F	07	Plate1-F-07	-					
G	07	Plate1-G-07	-					
H	07	Plate1-H-07	-					
A	08	Plate1-A-08			-	-		-
B	08	Plate1-B-08			-	-	-	+
C	08	Plate1-C-08			-	-	-	+
D	08	Plate1-D-08			-	-	-	-
E	08	Plate1-E-08			-	-	-	

For all samples validated by Sanger sequencing, homozygous wild types are indicated by '-', heterozygotes are indicated by '+', and homozygous mutants are indicated by '++'. The full table of genotyping results of the first cohort is available for download as Additional File 5 from the publication [51] or from the following URL.

[http://genomebiology.com/content/supplementary/gb-2011-12-9-r93-s5.xls]

Table 3-5. Partial list of variant call output of SERVIC^{4}E on the first cohort using Illumina sequencing output

CHR	POS	POOL	REF	VAR	AVE_QUAL	CONTINUITY	VAR_COV	TOT_COV	FRQ	TAIL
chr3	14535190	2	T	Y	24.06	22	34	3652	0.0093	1.94
chr3	14535190	4	T	Y	23.89	34	76	4073	0.0187	1.53
chr3	14535190	9	T	Y	24.29	29	48	2584	0.0186	1.91
chr3	14535190	11	T	Y	24.46	24	26	3167	0.0082	1.80
chr3	14535341	2	G	S	25.71	16	24	7937	0.0030	1.68
chr3	14535341	5	G	S	27.54	42	108	7044	0.0153	1.30
chr3	14535341	9	G	S	27.25	38	83	6393	0.0130	1.28
chr3	14535341	11	G	S	28.66	31	62	7116	0.0087	2.14
chr3	14535341	12	G	S	27.68	46	231	7979	0.0290	1.23
chr3	14536385	1	C	Y	25.35	37	52	3836	0.0136	1.36
chr3	14536508	5	A	R	28.72	28	57	5589	0.0102	6.32
chr3	14536508	7	A	R	27.93	28	61	7474	0.0082	7.30
chr3	14536319	4	C	Y	28.09	36	57	3054	0.0187	2.50
chr3	14538047	2	C	Y	27.96	18	28	5618	0.0050	1.40
chr3	14538004	5	C	Y	28.76	36	83	4083	0.0203	1.15
chr3	14538081	3	C	Y	23.00	4	4	815	0.0049	6.79
chr3	14538081	6	C	Y	30.16	41	114	5786	0.0197	1.19
chr3	14538081	7	C	Y	30.03	45	455	7603	0.0598	1.11
chr3	14538081	8	C	Y	28.60	28	43	4166	0.0103	2.20

CHR: reference sequence chromosome. POS: position in reference sequence chromosome. POOL: pool in which the variant is called. REF: reference nucleotide. VAR: variant nucleotide (IUPAC merge). AVE_QUAL: average quality of variant nucleotide (average of base quality scores). VAR_COV: coverage of variant nucleotide. TOT_COV: total coverage of position. TAIL: tailcurve factor. The full list of variant call output of SERVIC^{4}E on the first cohort is available for download as Additional File 6 from the publication [51] or from the following URL.

[http://genomebiology.com/content/supplementary/gb-2011-12-9-r93-s6.txt]

Table 3-6. Comparison of annotated SNPs, transition-transversion ratios, and synonymous-non-synonymous ratios

Base-calling Methods	Illumina					SRFIM			
Variant Detection Algorithm	SNPSeeker	SAMTools	CRISP	Syzygy	SERVIC4E	SAMTools	CRISP	Syzygy	SERVIC4E
Annotated Variants - # of Variant-Pools -									
Annotated	191	235	235	247	246	232	234	244	242
Not Annotated	30	153	150	132	87	76	128	132	92
% Unique Variant-Pools Annotated	*86.43%*	*60.57%*	*61.04%*	*65.17%*	*73.87%*	*75.32%*	*64.64%*	*64.89%*	*72.46%*
Annotated Variants not called	23	19	19	19	19	20	19	20	19
Transitions vs. Transversions - # of Variant-Pools -									
Transitions	205	313	306	332	297	278	298	308	298
Transversions	16	75	79	47	36	30	64	68	36
Transition:Transversion Ratio	12.81	4.17	3.87	7.06	8.25	9.27	4.66	4.53	8.28
Synonymous vs. Non_Synonymous Changes - # of Variant-Pools -									
Synonymous	99	110	105	100	102	100	104	99	102
Non-Synonymous	40	103	69	63	64	69	69	57	64
Synonymous:Non-synonymous Ratio	2.48	1.07	1.52	1.59	1.59	1.45	1.51	1.74	1.59

Calculated metrics for annotation rates, transition-transversion rates, and synonymous-non-synonymous rates are for first cohort data only.

Table 3-7. Partial list of variant call output of SERVIC^{4}E on the first cohort using Srfim base calls

CHR	POS	POOL	REF	VAR	AVE_QUAL	CONTINUITY	VAR_COV	TOT_COV	FRQ	TAIL
chr3	14535190	4	T	Y	15.33	35	87	3878	0.0224	2.14
chr3	14535341	5	G	S	21.29	45	118	6741	0.0175	1.17
chr3	14535341	9	G	S	17.90	40	94	6087	0.0154	1.08
chr3	14535341	11	G	S	23.63	29	62	6659	0.0093	1.86
chr3	14535341	12	G	S	21.90	45	228	7755	0.0294	1.11
chr3	14536385	1	C	Y	17.40	29	43	3712	0.0116	1.50
chr3	14536385	2	C	Y	21.33	26	33	5340	0.0062	1.79
chr3	14536508	1	A	R	16.21	17	28	4424	0.0063	2.79
chr3	14536508	2	A	R	17.33	28	51	6707	0.0076	5.04
chr3	14536508	3	A	R	14.88	22	33	3872	0.0085	2.86
chr3	14536508	5	A	R	20.56	28	61	5387	0.0113	3.73
chr3	14536508	6	A	R	17.52	26	48	6285	0.0076	3.55
chr3	14536508	7	A	R	17.47	29	66	7237	0.0091	4.97
chr3	14536508	8	A	R	14.31	24	48	5035	0.0095	2.43
chr3	14536508	10	A	R	14.48	23	58	5508	0.0105	2.66
chr3	14536508	11	A	R	15.33	19	30	5322	0.0056	7.29
chr3	14536319	4	C	Y	21.80	31	49	2937	0.0167	2.72
chr3	14538047	2	C	Y	20.71	23	35	5501	0.0064	1.67
chr3	14538004	5	C	Y	22.25	40	88	4240	0.0208	1.10

CHR: reference sequence chromosome. POS: position in reference sequence chromosome. POOL: pool in which the variant is called. REF: reference nucleotide. VAR: variant nucleotide (IUPAC merge). AVE_QUAL: average quality of variant nucleotide (average of base quality scores). VAR_COV: coverage of variant nucleotide. TOT_COV: total coverage of position. TAIL: tailcurve factor. The full list of variant call output of SERVIC^{4}E on the first cohort is available for download as Additional File 8 from the publication [51] or from the following URL.

[http://genomebiology.com/content/supplementary/gb-2011-12-9-r93-s8.txt]

Table 3-8. Partial list of variant calls from second cohort analyses

Position	Pool	REF	VAR	Illumina SERVIC4E	Srfim SERVIC4E	Rarity	P	Valid	Exon	Dist
14535190	14	T	Y	+	+				25	240
14535277	4	C	Y	+	+	+	+	+	25	153
14536385	14	C	Y	+	+	+	+	+	24	206
14536473	14	G	R	+	+	+	+	+	24	118
14536508	5	A	R		+				24	83
14536508	13	A	R	+	+			+	24	83
14536508	14	A	R	+	+			+	24	83
14536508	17	A	R	+	+				24	83
14536508	18	A	R	+					24	83
14536508	19	A	R	+	+				24	83
14536508	21	A	R	+	+				24	83
14536508	22	A	R	+	+				24	83
14536508	24	A	R	+	+				24	83
14536544	22	C	Y	+	+				24	47
14536544	24	C	Y		+				24	47
14537948	3	G	R	+					23	205

All positions are given in reference to chromosome 3 of hg19. For each program, a '+' value indicates that a variant call was made by that program for that variant position and pool. Column 'P' indicates the position is in exonic sequence (not intronic). Column 'Valid' indicates validation results for each variant-pool tested; '+' indicates a valid call and '-'indicates an invalid call. Column 'Dist' indicates the position of the variant call in each amplicon. The full list of variant calls from analyses of the second cohort is available for download as Additional File 12 from the publication [51] or from the following URL.

[http://genomebiology.com/content/supplementary/gb-2011-12-9-r93-s12.xls]

Table 3-9. Partial list of variant call output of SERVIC^{4}E on the second cohort using Illumina base calls

CHR	POS	POOL	REF	VAR	AVE_QUAL	CONTINUITY	VAR_COV	TOT_COV	FRQ	TAIL
chr3	14535277	4	C	Y	26.67	99	892	2492	0.3579	1.04
chr3	14535190	14	T	Y	13.67	56	125	3706	0.0337	2.96
chr3	14536508	13	A	R	22.00	29	52	3735	0.0139	1.41
chr3	14536508	14	A	R	24.58	47	135	3895	0.0347	3.09
chr3	14536508	17	A	R	22.08	44	120	4882	0.0246	3.39
chr3	14536508	18	A	R	15.63	29	48	3861	0.0124	2.47
chr3	14536508	19	A	R	19.99	41	95	4821	0.0197	3.75
chr3	14536508	21	A	R	23.70	36	86	4575	0.0188	5.12
chr3	14536508	22	A	R	20.24	35	71	4279	0.0166	6.11
chr3	14536508	24	A	R	15.63	27	64	3403	0.0188	4.03
chr3	14536385	14	C	Y	23.26	68	172	5377	0.0320	1.30
chr3	14536473	14	G	R	27.24	72	143	5522	0.0259	1.41
chr3	14536544	22	C	Y	27.63	42	80	2246	0.0356	1.72
chr3	14537948	3	G	R	18.17	25	30	1971	0.0152	1.84
chr3	14538081	13	C	Y	15.49	36	65	2927	0.0222	1.56
chr3	14538081	14	C	Y	19.22	46	76	3543	0.0215	1.13
chr3	14538081	18	C	Y	18.27	54	102	2932	0.0348	1.35
chr3	14538081	19	C	Y	24.75	85	327	3384	0.0966	2.40
chr3	14538081	24	C	Y	23.40	69	168	2238	0.0751	2.44

CHR: reference sequence chromosome. POS: position in reference sequence chromosome. POOL: pool in which the variant is called. REF: reference nucleotide. VAR: variant nucleotide (IUPAC merge). AVE_QUAL: average quality of variant nucleotide (average of base quality scores). VAR_COV: coverage of variant nucleotide. TOT_COV: total coverage of position. TAIL: tailcurve factor. The full list of variant call output of SERVIC^{4}E on the second cohort is available for download as Additional File 13 from the publication [51] or from the following URL.

[http://genomebiology.com/content/supplementary/gb-2011-12-9-r93-s13.txt]

Table 3-10. Partial list of genotyping results for individual second cohort samples

Exon	Pool	POS	Pool:	1	2	3	4	5	6	7	8	9	10
21													
	12												
		14547145		-	-	-	-	-	-	-	-		-
		14547182		-	-	-	-	-	+	-	-		++
		14547198		-	-	+	-	-	-	-	-		-
		14547270		-	-	-	-	-	-	-	+		-
	17												
		14547145		-	-	-	-	-	-	-	-	-	-
		14547182		-	-	-	-	-	-	-	-	-	-
		14547198		-	-	-	-	-	-	-	-	-	-
		14547270		-	-	-	-	-	-	-	-	-	-
	21												
		14547145		-	-	-	-	-	-	-	-	-	-
		14547182		-	-	-	+	-	-	+	-	-	-
		14547198		-	-	-	-	-	-	-	-	-	-
		14547270		-	-	-	-	-	-	-	-	-	-
20													
	17												
		14548315						-	-	-	-	-	-
		14548392						-	-	-	-	-	-
		14548425						-	-	-	-	-	-
		14548457						-	-	-	-	-	-
	18												
		14548315						-	-	-	-	-	-
		14548392						-		-	-	-	-
		14548425						-		-	-	-	-
		14548457						-		-	-	-	-
	19												
		14548315								-	-	-	-

For all samples validated by Sanger sequencing, homozygous wild types are indicated by '-', heterozygotes are indicated by '+', and homozygous mutants are indicated by '++'. The full table of genotyping results of the second cohort is available for download as Additional File 14 from the publication [51] or from the following URL.

[http://genomebiology.com/content/supplementary/gb-2011-12-9-r93-s14.xls]

Table 3-11. Validation analysis of variant calling from second cohort samples

	Illumina	
	Syzygy	*SERVIC^{4}E*
Variant identification and validation		
True positive	47	53
True negative	38	60
False positive	26	4
False negative	8	2
Statistical analysis (%)		
Sensitivity	85.45	96.36
Specificity	59.38	93.75
PPV	64.38	92.98
NPV	82.61	96.77
FPR	40.63	6.25
FDR	35.62	7.02
Accuracy	71.43	94.96
MCC	45.90	89.93
Rare exonic variant detection and validation		
Detected total variants (n = 16)	13	16
Positive detection rate (%)	81.25	100

Descriptions of calculations used in statistical data analysis are provided in Materials and methods. FDR, false discovery rate; FPR, false positive rate; MCC, Matthews correlation coefficient; NPV, negative predictive value; PPV, positive predictive value For both algorithms, an allele count of 40 was used. Syzygy: default parameters used. SERVIC^{4}E: trim value of 25 used. Default used for all other parameters.

Chapter 4: Glutamate Signaling Defects and the Role of *GRIP1/2* in Autism

Autism spectrum disorders are clinically and genetically heterogeneous, and likely involve hundreds of risk genes. Though the underlying causes remain unknown, there are many indications that glutamate signaling is involved. Glutamate signaling is the primary excitatory neurotransmission system in the brain, and is involved in a wide range of neural pathways. Glutamate signaling involves hundreds of key components in the pre-synaptic and post-synaptic areas of the neuron; two of these key components are the Glutamate Receptor Interacting Proteins 1 and 2, also known as *GRIP1* and *GRIP2*. In the previous chapter, an effective method for the detection of rare variants in a large cohort using next-generation sequencing was discussed, with the gene *GRIP2* as the primary focus. Within that study, two cohorts, each of 480 samples, were sequenced. The first cohort is composed of 480 males diagnosed with autism spectrum disorder, and the second cohort is an ethnically matched control. A number of rare *GRIP2* variants were identified as a result of that study. The following chapter will provide functional evidence on how rare variants in *GRIP2*, and more broadly on glutamate signaling, can influence autism phenotype.

4.1 Introduction

Autism is a common neurodevelopmental disorder with a prevalence of at least one in 100 children, and possibly even higher [131]. It is characterized by deficits in social interaction, verbal communication, and repetitive or stereotypical behaviors. These core traits are part of a broad set of phenotypes that present variably in any given individual with autism. Such broad clinical traits also include neuropsychological symptoms, such as deficits in executive function like memory or planning, neuroanatomical abnormalities, such as reduced activation of the prefrontal cortex during certain tasks, immunological abnormalities, such as increases in pro-inflammatory factors, or morphological abnormalities, such as macrocephaly and microcephaly [132]. The broad presentation and variability of expressivity underlies a complex genetic etiology, likely involving hundreds to thousands of genes, with multiple genes altered in each individual.

Multiple lines of evidence from GWAS, linkage, cytogenetic, molecular, and sequencing studies indicate that many genes in the glutamate signaling pathway are likely involved in autism etiology [133, 134]. Indeed, there is substantial overlap between genes implicated by GWAS and linkage studies and known genes for glutamate receptors, glutamate transporters, and glutamate receptor interacting proteins (Figure 4-1). As such, it has been hypothesized that an imbalance in glutamatergic signaling, the primary form of excitatory neurotransmission in the brain, in concert with possible dysregulation of GABAergic signaling, the primary form of inhibitory neurotransmission, can produce neural abnormalities that generate autism phenotype [134].

4.1.1 Evidence of an imbalance in glutamatergic and GABAergic signaling in autism

Key to this hypothesis has been the recognition that the proper construction and maintenance of neural synapses requires a balancing act between excitatory and inhibitory neurotransmission [135]. This balance allows for the appropriate strengthening and retention, or weakening and pruning, of synaptic connections. This ability to selectively control synaptic strength governs the most important functional feature of the brain: synaptic plasticity. Synaptic plasticity provides the brain with the ability to encode new information, in turn producing our capacity for behavior, communication, and intelligence. As such, an imbalance in excitatory and inhibitory signaling, through aberrant modulation of specific neural circuits, could influence human behavior, communication, and intelligence in a manner consistent with autism [136].

Supportive evidence for this model of autism includes molecular observations from model organisms and from humans. Mice that have been pharmacologically rendered into a hypoglutamatergic state show cognitive impairment consistent with both autism and schizophrenia [137]. Additionally, abnormal levels of key genes in the glutamate signaling pathway, including AMPA-type glutamate receptor 1 and glutamate transporters, have been reported from post-mortem autism brain samples [138]. In addition to glutamate receptor 1, glutamate receptor 6 and 8 have also been implicated by multiple linkage and association studies [139-141].

Analyses of autism post-mortem brain samples has revealed abnormal levels of Glutamic Acid Decarboxylases 63 and 67 in both parietal and cerebellar cortices,

particularly in inhibitory cerebellar Purkinje cells and interneurons [142-144]. Glutamic acid decarboxylases are required for the generation of GABA, the primary neurotransmitter in inhibitory neurotransmission. GABA receptors, including the GABA receptor subunits *GABRB1* and *GABRA4*, have also been associated with autism by allelic association analysis [145]. Furthermore, significantly altered levels of *GABRA1*, *GABRA3*, and *GABRB3* have been observed in multiple areas of autism post-mortem brains [146]. These studies in aggregate provide support that modification of this neurotransmitter pathway can play a role in autism through an excitatory/inhibitory imbalance.

In addition to the neurotransmitter receptors, transporters, and enzymes involved in neurotransmitter synthesis, key regulators and interaction partners of the glutamatergic and GABAergic systems have also been implicated. *SHANK3*, a scaffolding protein in the post-synaptic dendrite, is an important regulator of glutamatergic activity, through its binding and organization of multiple glutamate signaling partners, including the neuroligin genes, and it is a necessary component in the maturation of dendritic spines. Mutations in *SHANK3* have been strongly associated with autism, particularly to impairment in social communication [147]. The neuroligin genes are synaptic cell-adhesion molecules that are important in establishing synaptogenesis via their pre-synaptic ligands, the ß-neurexins. Rare, private mutations identified in the X-linked neuroligins, *NLGN3* and *NLGN4*, have been associated with autism, as have mutations in *NLGN1* and *NLGN2* [148,149]. Importantly, a mutation identified in *NLGN1* in an autism

patient has been shown functionally to inhibit normal activity of excitatory synapses [149].

Taken together, these multiple lines of evidence provide strong support that components of the excitatory glutamatergic and inhibitory GABAergic signaling pathways should be an important focus in studying autism etiology.

4.1.2 Glutamate Receptor Interacting Proteins as candidate autism genes

Several converging lines of evidence recommend the Glutamate Receptor Interacting Proteins (*GRIP*) as candidate genes in autism etiology. *GRIP1*, located at 12q14-23 has positive association through multiple genetic markers with autism, including positive linkage to D12S338 at 12q23.3, positive association to D12S395 at 12q24.23, positive association with SNPs around *AVPR1a* at 12q14-15, and positive genome-wide linkage to rs1445442 at 12q14.13 [150-153]. *GRIP2*, located at 3p25.1, is similarly supported by genetic evidence, including positive linkage to D3S3680 at 3p25.2, positive association to SNPs in *ATP2B2* at 3p25, positive association to D3S3594 at 3p25.2, and positive association to SNPs near *OXTR* at 3p25.3 [154,155,151,156].

GRIP1 was originally determined to bind to AMPA receptor subunits and to co-localize with AMPA receptors at excitatory synapses [157]. AMPA receptors are a subtype of glutamate receptors that can be selectively activated by the agonist AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazoleproprionic acid). This subtype of glutamate receptors is referred to as ionotropic, due to its ability to immediately channel the influx

of calcium ions into the cell. AMPA receptors are heterotetramers. The most frequently studied subunits in the context of *GRIP* are *GluA2* and *GluA3*.

NMDA-type glutamate receptors are also ionotropic receptors, but differentiated from AMPA-type receptors by their ability to be activated by the agonist NMDA. Metabotropic glutamate receptors conversely do not act as ion channels, but rather activate downstream effectors upon their own activation by extracellular glutamate or agonist.

GRIP1 is widely expressed in multiple brain regions, particularly in the cerebral cortex and hippocampus, and is enriched in post-synaptic densities [158,159]. Similar observations have been made with *GRIP2*, including binding to AMPA receptor subunits, particularly *GluA2/3*, but not NMDA-type receptors. While *GRIP1* has a generally much broader tissue expression pattern, *GRIP2* is nonetheless also widely expressed across the CNS [160]. Both *GRIP*s are also found in inhibitory GABAergic nerve terminals, and show extensive sub-cellular colocalization, though *GRIP2* has been observed more strongly in pyramidal neurons, while *GRIP1* is strongest in nonpyramidal neurons [158,161]. Interestingly, proximal dendritic spines can be enriched for one *GRIP* type over the other [162].

Though the exact mechanisms are still to be determined, it is clear that the specific functions of *GRIP1* and *GRIP2* in AMPA receptor trafficking, clustering, internalization, and recycling are important in establishing synaptic plasticity. Post-synaptic dendritic activation by glutamate neurotransmitter can induce calcium influx, by ionotropic NMDA- and AMPA-type glutamate receptors, or by metabotropic glutamate

receptors that can activate *PLC* (phospholipase C), releasing intracellular stores of calcium [163]. Intracellular calcium can in turn activate the glutamate scaffolding protein *PICK1* via the signal transduction enzyme *PKC* (protein kinase C). *PKC* will phosphorylate serine-880 on AMPA receptor subunits (*GluA2/3*). This phosphorylation prevents binding between *GluA2* and *GRIP*, but improves accessibility of *GluA2* to *PICK1*, resulting in *GluA2* (and overall AMPA receptor) internalization, with a general reduction in membrane levels of AMPA receptor [164, 165]. Internalized *GluA2* can then be stabilized internally or recycled back to the surface membrane by *GRIP* [166,167].

In addition to binding *GluA2/3* AMPA receptor subunits, *GRIP1* and *GRIP2* can bind to *liprin-alpha-1*, and to *ephrinB1* and *ephrinB2*, the ligands of *Eph* receptors. Interaction between *GRIP* and *liprin-alpha-1* is necessary for proper targeting of AMPA receptors to the surface membrane, using the microtubule-associated protein *GIT1* to mediate such trafficking [168,169]. Interaction between *GRIP* and *liprin-alpha-1* has been shown as necessary for the proper induction of long-term depression (LTD) in hippocampal neurons, which is an important component in regulating synaptic strength and plasticity [170]. Additionally, *GRIP* has been shown to mediate trafficking of *ephrinB1/2* to the surface membrane [171]. Reverse signaling of *ephrinB* ligand by stimulation with its extracellular *Eph* receptor can induce microdomain clustering of *ephrinB*, in turn forming subsurface clusters of *GRIP1/2*. This process is mediated by the phosphorylation of a serine on *ephrinB* that increases binding strength between *ephrinB* and *GRIP1/2* [172,173]. This strong binding between surface *ephrinB* and subsurface *GRIP1/2* stabilizes AMPA receptor clusters at the membrane, supported by the

observation that *ephrinB* knockout results in constitutive internalization of AMPA receptors [173].

This complex system of multiple interaction partners is necessary within the excitatory glutamatergic pathway to induce LTD, which is a well-established neural model by which new information is encoded through modulation of the strength of synaptic connections. This pathway allows the neuron to change the levels of glutamate receptors on the synaptic surface through internalization and recycling activities, induced as a response to initial glutamate receptor activation by pre-synaptic neurons. *GRIP1/2* double knockout cerebellar Purkinje cells exhibit a loss in LTD capability [174]. This is likely a result of the inability of *GRIP1/2* double knockout neurons to properly recycle AMPA receptors back to the synaptic membrane in response to activation by glutamate receptor agonists [167]. These observations underlie the importance of *GRIP1* and *GRIP2* in the glutamate signaling pathway and in synaptic plasticity.

4.1.3 Structure of GRIP1 and GRIP2

Topologically, the *GRIP*s are composed of seven repeating PDZ domains, which allow for binding to a broad set of interaction partners (Figure 4-2). The first three PDZ domains are known to interact with components of the exocyst complex [167]. PDZ4 and PDZ5 bind to AMPA receptor subunits as well as to other *GRIP* proteins [129]. PDZ6 binds to *liprin-alpha-1*, *ephrinB1*, and *ephrinB2* [168,172]. The region between PDZ6 and PDZ7 can bind to *KIF5*, a kinesin that may mediate some of *GRIP*s trafficking

capabilities. Lastly, PDZ7 can bind to *GRASP1*, which also appears to have an important role in proper targeting of AMPA receptors to the synaptic membrane [168].

The *GRIP1* and *GRIP2* proteins can be functionally and structurally separated into three separate domains, reflective both of the unique functions of each domain and the consistency of amino acid conservation (Figure 4-3, Panel A). The regions of PDZ123 can be viewed as one conservation domain, PDZ456 as another, PDZ7 as a domain on its own, and the linker region between PDZ6 and PDZ7 as a region undergoing frequent evolutionary changes and poorly conserved. These regions will be important during discussion of mutation burden analysis further in the chapter.

4.1.4 Autism-associated GRIP1 gain-of-function variants increase AMPA receptor recycling

In Mejias et al., a statistically significant burden of functional rare variants was identified in *GRIP1* from a large cohort of autism males [129]. Five *GRIP1* autism-associated variants clustered in the PDZ456 region, known to bind AMPA receptor subunits. Three of these variants were determined to increase binding to *GluA2* and *GluA3* by Yeast-Two-Hybrid and co-immunoprecipitation assays, indicating a gain-of function effect.

To determine the consequence of these gain-of-function rare variants in neurons, a recycling assay was performed in mouse hippocampal neurons expressing a pH-sensitive fluorescent *GluA2* and *GRIP1* with or without an autism-associated variant. An increased rate of recycling of AMPA receptor subunits was observed for all three variants upon

NMDA receptor stimulation. No change was observed for the internalization of AMPA receptors. This observation is in direct contrast to the decrease in recycling rate observed in *Grip1/2* deficient hippocampal neurons, further indicating that these variants present a gain-of-function effect. In addition to faster recycling, higher overall surface expression of *GluA2* was observed for at least two of the variants [129].

Given that neurons derived from *Grip1/2* double knockout mice show the opposite effect, these double knockout mice were subjected to a range of tests, to determine if *Grip* has a role in modulating behavior. These mice were found to have an increase in sociability to unfamiliar mice, a trait that is in contrast to what would be expected in autism. Given the *Grip1/2* double knockout more closely resembles a loss-of-function effect, it would suggest that the *GRIP1* gain-of-function mutations may have the opposite effect on mouse social behavior, though this has yet to be determined.

4.1.5 GRIP2 as an autism-susceptibility gene

In this chapter, I will discuss work performed in identifying rare mutations in *GRIP2* in a cohort of autism males. This is the same cohort of autism samples and ethnically matched controls that was used in Mejias et al. to identify *GRIP1* autism-associated variants [129]. The methodology to identify rare variants in these two cohorts using next-generation sequencing was discussed in the previous chapter [51].

Throughout this chapter, I will provide genetic evidence that *GRIP2* variants, particularly in the PDZ456 region, are significantly associated with autism. Clinical evidence will be presented that shows a correlation between *GRIP2* variants and deficits

in social behavior and communication in affected individuals. Functional evidence will also be provided, showing that *GRIP2* variants produce biologically detectable changes. In conjunction with behavioral analyses performed in *Grip2* knockout mice, which display deficits in social behavior, *GRIP2* will be discussed as a putative gene for autism susceptibility.

4.2 Results

4.2.1 Sequencing of an autism cohort and ethnically matched control

We sequenced the exons of the gene *GRIP2* from 480 individuals diagnosed with a strict autism spectrum disorder. DNA for these samples was obtained from the Autism Genetic Research Exchange (AGRE) and the South Carolina Autism Project (SCAP). Additionally, sequencing was performed on 480 ethnically matched controls with apparent normal cognitive functions and behaviors, as determined by the Greenwood Genetics Center, South Carolina (Table 4-1).

GRIP2 possesses 23 exons. The coding portions of these exons (excluding untranslated regions) were enriched using a sample pooling and PCR amplification strategy. The methodologies for sample pooling, enrichment, library preparation, sequencing, and data analysis for variant calling are described in detail in the previous chapter [51].

Sequencing identified a total of 78 unique variants. 43 of these are non-coding variants, of which one is a splicing variant and was confirmed by Sanger sequencing. The

remaining 35 variants are coding variants, of which Sanger sequencing confirmed 34. Nine of these variants are extremely rare, existing in only one sample from all 960 samples sequenced between the two cohorts. Of the 35 coding variants, 16 are synonymous and were all confirmed. 18 are missense, with 17 confirmed by Sanger sequencing. One is nonsense and was also confirmed. 20 of the 78 variants are relatively common and exist in the general population at a frequency greater than 1%, as determined from dbSNP.

4.2.2 GRIP2 PDZ456 is significantly burdened with autism-associated variants

A total of 28 non-unique non-synonymous coding variants were identified (Figure 4-3, Panel B). To determine if *GRIP2* possesses a statistically significant burden of mutation in autism cases, we performed a two-tailed Fisher's exact test across the entire gene. However, no significant difference was observed between cases and controls. We then performed the Fisher's test across each of the three functionally conserved regions of *GRIP2* described previously. The conserved region containing PDZ123 and the conserved region containing PDZ7 with the PDZ6-7 linker also does not show a significant burden of mutation ($p = 0.306$, $p = 0.773$). However, the conserved region containing PDZ456 shows a significant burden of mutation for autism cases vs. controls (Figure 4-3, Panel C; $p < 0.03$). This result mirrors what was obtained when analyzing *GRIP1* variants using the same cohort sets in Mejias et al., where a statistically significant burden of mutation was observed in *GRIP1* PDZ456 only [129].

This result may have major implications for future gene-disease association studies. Given that burden is highest within a specific functionally conserved domain for both genes, and not across the entire gene, one can make the case that analysis of mutation burden or variant association should not just be done at the level of an entire gene. This is in direct contrast to current methods of association or rare variant analysis, which often groups variants by gene level annotation [59]. Instead, it may be more appropriate to perform such analyses at the level of a functional domain or region of conservation. This makes sense from a molecular level, as individual genes are composed of multiple functional units working both independently and in tandem. A particular region of a gene may be more relevant to a particular disease, and that certainly holds true for genes expressed in the synapse, where different domains in the same gene can function in separate but parallel biological pathways. By assessing for mutation burden across the entire gene, the mutation load may become diluted by variants with unrelated function, and fail to achieve statistical significance, thereby masking true significance within a more defined region of that gene.

4.2.3 GRIP2 autism-associated variants correlate with more severe disease

To quickly assess if the variants identified in our study are truly relevant to disease in our autism cohort, we performed a genotype-phenotype correlation (Table 4-2). All samples in our cohort with a *GRIP1* or *GRIP2* mutation identified by sequencing were assessed for family members, and those family members were genotyped for the mutation. This is an important step, as genotyping family members allows us to

determine if any of the mutations are *de novo* or fail to properly segregate with disease. One mutation, the nonsense change W1013X, was the only *de novo* occurrence identified. Non-paternity and non-maternity were excluded in this case.

To determine a genotype-phenotype correlation, it is useful to assess any changes in phenotype that segregate between genotypically discordant individuals in the same family. This can be quite difficult for autism families. Diagnosis of autism is more nuanced than the simple binary presence of disease. Instead, multiple endophenotypes are analyzed and quantitatively scored. Family members not diagnosed with autism receive such detailed scoring very rarely. As such, it is difficult to phenotypically compare affected individuals to their unaffected family members. In order to perform a genotype-phenotype correlation we instead restricted our analysis to siblings who have both been diagnosed with autism (and therefore quantitatively assessed), but are genotypically discordant for the mutation in the proband identified during sequencing.

Restricting our analysis to such cases provided five autism families out of the 480 sequenced. Genotype-phenotype correlation and pedigrees for these families are provided in Table 4-2. Four of these five sibling sets have a proband that is positive for a missense mutation identified in PDZ456 (Table 4-2, Rows 1-4). Additionally, two sibling sets fitting the genotypic discordance criteria for *GRIP1* variants are also shown (Table 4-2, Rows 6-7).

An important pattern emerges in the analysis of these five families. For all samples, quantitative scores of reciprocal social interaction (SOCT_CS), verbal/nonverbal communication (COMVT_CS/COMNVT_CS), and repetitive behavior

(BEHT_CS) are provided. Higher scores indicate a more severe disease phenotype. A nonverbal score is always more severe than a verbal score. In each sibling set, the sibling with a *GRIP1/2* mutation always has a more severe quantitative score for reciprocal social interaction and verbal/nonverbal communication than the sibling that does not have that mutation (Table 4-2, red arrows).

Family 111 has a son and daughter, both affected with autism. The affected son is heterozygous for the A575T change, while his affected sister is wildtype. His social score is nine points more severe and his verbal score is six points more severe than his sister (Table 4-2, Row 1). In family 772, the affected son is heterozygous for the N610S change, while his affected sister is wildtype. His social score is 11 points more severe than his sister, and he has a nonverbal score, which is more severe than his sister's verbal score (Table 4-2, Row 2). In family 388, one affected son is heterozygous for the G749D change, while his affected brother is wildtype. His social score is six points more severe than his wildtype brother, and he has a nonverbal score, which is more severe than his brother's verbal score (Table 4-2, Row 3).

In family 656, one affected son and one affected daughter are both heterozygous for the E773K change. Another affected son and affected daughter are both wildtype. The heterozygous son has a social score 12 points more severe and a verbal score four points more severe than his wildtype brother, and he has a social score nine points more severe and a verbal score five points more severe than his wildtype sister. The heterozygous daughter has a social score seven points more severe and a verbal score four points more

severe than her wildtype sister, and she has a social score ten points more severe and a verbal score three points more severe than her wildtype brother (Table 4-2, Row 4).

This pattern repeats itself for the A130T change in *GRIP2*, and the M794R and A625T changes in *GRIP1* (Table 4-2, Rows 5-7). Importantly, with the exception of the A130T change, all of these missense changes occur within the PDZ456 regions of *GRIP1/2*. Taken cumulatively, the increased severity of autism phenotype, particularly for reciprocal social interaction, between genotypically discordant autistic siblings is statistically significantly different (Figure 4-4; $p <$ E-3). These results, while not conclusive evidence that *GRIP2* is an autism susceptibility gene, nonetheless provide support that *GRIP2* autism-associated variants, and in particular PDZ456 variants, are worthwhile candidates for functional study. Given the statistically significantly increased likelihood of more severe disease with *GRIP1/2* mutations, one possible explanation for the role of *GRIP1/2* in autism is that they may act as modifier genes, increasing risk severity on a permissive genetic background, but are not necessarily driving disease on their own.

4.2.4 GRIP2 PDZ456 autism-associated variants modify binding strength to GRIP2 interaction partners

Ten missense changes were identified in the PDZ456 region of *GRIP2* for both autism cases and unaffected controls. Autism-specific missense changes include R517Q, N610S, V664M, G749D, E773K, and R843C (*GRIP2* autism-associated variants). Missense changes found only in unaffected controls include P543Q and R711C (*GRIP2*

control variants). Changes found both in cases and controls include T540M and A575T (*GRIP2* common variants). The changes specific to PDZ4 are R517Q, T540M, P543Q, A575T, and N610S. The changes specific to PDZ5 are V664M and R711C. Importantly, the V664M change rests just adjacent to the ELGI binding site in PDZ5 for *GluA2/3*. The changes specific to PDZ6 are G749D, E773K, and R843C (Figure 4-3, Panel B). The location of each missense change is important, as it may influence which binding partner affinities are most affected by a particular variant. PDZ4 and PDZ5 are particularly important in binding *GluA2/3* and other *GRIP1/2* proteins. PDZ6 is important in binding *liprin-alpha-1* and *ephrinB1/2* (Figure 4-3).

To determine if these mutations could modify *GRIP2* binding ability to its interaction partners, a high-throughput Yeast-Two-Hybrid (Y2H) assay was performed. The Y2H is a gene-based colorimetric assay that allows for quantitative measuring of binding between two interacting proteins. Binding strength influences the ratio of bound protein (dimers) to free protein in the cell. Stronger binding yields a higher population of bait-prey dimers, while weaker binding yields a higher population of unbound bait and prey. Through the Y2H, we determined that each of the ten mutations affects binding strength to multiple *GRIP2* interaction partners. Only statistically significantly different changes in binding strength ($p < 0.05$) are considered.

For interaction between *GRIP2* and *GluA2*, binding strength is most attenuated with the V664M missense mutation, which lies just adjacent to the ELGI peptide sequence in PDZ5 that is bound by the c-terminal domain of *GluA2/3*. This effect diminishes with distance from the binding site, as observed by the increased binding

strengths of R711C and G749D. Most of the missense changes do not change binding strength by more than 20%. Only the autism-associated variant V664M and the control variant R711C attenuate binding substantially (>20%), while the two common variants, T540M and A575T, strengthen binding substantially (Figure 4-5).

Similar observations were made between *GRIP2* variants and *GluA3*. Few of the variants change binding strength more than 20%, with the exceptions of the autism-associated variants V664M and E773K, which significantly attenuate binding. Both common variants T540M and A575T again increase binding. At a less stringent effect size of 10% or greater (instead of 20%), three and five autism-associated variants out of six attenuate binding to *GluA2* and *GluA3* respectively. This is opposed to the control variants, for which only one of two (R711C) attenuates binding to *GluA2/3*. Based on these results, there is a general attenuation of binding strength between *GRIP2* and *GluA2/3* in the presence of several autism-associated variants, which is best described as a loss-of-function effect (Figure 4-5).

PDZ6 of *GRIP2* binds to *ephrinB1/2*. It is the variants closest to or within PDZ6 that show the greatest changes in binding strength between *GRIP2* and the *ephrinB*s. Missense changes R517Q through N610S, and R711C, show little effect on binding to *ephrinB1*. Though control variants P543Q and R711C produce little change on binding to *ephrinB1*, they do produce a substantial change with *ephrinB2*, with P543Q decreasing and R711C increasing binding (Figure 4-6). The presence of only two control variants is not sufficient to infer biological importance.

On the other hand, autism-associated variants V664M, E773K, and R843C (VER) increase binding, while G749D (G) attenuates binding to *ephrinB1*. Conversely for *ephrinB2*, V664M, E773K, and R843C (VER) attenuate binding, while G749D (G) increases binding. It is an interesting pattern that where an autism-associated variant increases binding with one *ephrinB*, it decreases binding with the other (Figure 4-6). This reciprocal phenomenon may underpin the unique differences in function of *ephrinB1* versus *ephrinB2* in neuronal development and synaptic function.

PDZ6 also binds to *liprin-alpha-1*. Just as is observed with *ephrinB1/2*, variants closest to or within PDZ6 show the greatest change in binding strength between *GRIP2* and *liprin-alpha-1*. Uniquely, all the variants, independent of cohort origin, display increased binding to *liprin-alpha-1*, with the exception of the control variant P543Q, which decreases binding. Of particular note are the autism-associated variants G749D and E773K, which increase binding 4x and 3x of wildtype levels respectively. The control variant R711C also produces a large increase in binding strength at ~2.5x (Figure 4-7). As noted earlier, interaction between *GRIP2* and *liprin-alpha-1* is necessary for proper targeting of AMPA receptors to synaptic membranes and induction of long-term depression (LTD). Small increases in binding between *GRIP2* and *liprin-alpha-1* may not have substantial effects on AMPA receptor function, but it is possible that the more substantial mutations could negatively affect LTD and synaptic plasticity.

Four of six autism-associated variants, R517Q, G749D, E773K, and R843C, show significantly increased binding strength between *GRIP2* and *GRIP1*, while only one, N610S, shows attenuation, and the last, V664M, shows no difference. The control and

common variants, P543Q and A575T, show significant attenuation of binding to *GRIP1* (Figure 4-8). There is evidence, based on *Grip1/2* double knockout studies in mice, that suggests both *GRIP*s act in a cooperative manner [167,174]. However, behavioral analysis of *Grip1/2* double knockout versus *Grip2* single knockout mice indicates that *Grip1* and *Grip2* can have opposite effects on behavior, and that they may have opposing function [129]. The true nature of both *GRIP*s and how they act together is likely more complex.

The changes in binding strength between *GRIP2*-*GRIP2* homodimers are more complex. Four autism-associated variants demonstrate increased binding strength, while two attenuate binding. The two control variants attenuate binding, while one common variant, T540M, increases binding (Figure 4-8). Given the lack of a clear pattern, it is difficult to deduce how *GRIP2*-*GRIP2* binding is important in the synapse, or how these specific mutations may influence *GRIP2* activity in general. Further functional work must be done, particularly in the context of the full scaffolding protein complex. Simple analysis of one-on-one interaction between *GRIP2* homodimers and other interaction partners may produce misleading results, as variant effect my change in the context of which proteins are included in the complex during analysis.

Though there are not enough common or control variants to identify any unique patterns, when limiting the scope to just the autism-associated *GRIP2* variants, certain unique patterns do emerge, which may be relevant to biological function. These patterns provide a roadmap for future functional studies. As outlined in Figure 4-9, there is a general loss-of-function pattern observed through reduced binding between *GRIP2* and

GluA2 and *GluA3*, in the presence of autism-associated variants (Figure 4-9, Row 1). Four variants nears PDZ6 induce reciprocal changes in binding strength to *ephrinB1* and *ephrinB2* (Figure 4-9, Row 2). All six autism-associated variants produce a gain-of-function pattern by inducing increased binding between *GRIP2* and *liprin-alpha-1*, with the greatest changes observed in variants around PDZ6, the known binding domain to *liprin-alpha-1* (Figure 4-9,Row 3). There is a pattern of increased binding for autism-associated variants between *GRIP2* and *GRIP1*. However, the effects of these variants on *GRIP2* to *GRIP2* interaction are more difficult to decipher (Figure 4-9, Row 4).

The results of this Yeast-Two-Hybrid assay are informative, but do not immediately reveal a molecular mechanism of disease. Where possible, I have provided hypotheses (see Discussion) of how some of the autism-associated variants may influence activity between *GRIP2* and its binding partners, and how it may be relevant to synaptic function. Nonetheless, the most effective way to determine how these variants are functioning in disease is to directly, functionally test them in neurons.

One thing that can be noted about the observations made from this assay is that all these variants have a functional effect, and they all produce a different pattern of changes when analyzed across all the different interaction partners of *GRIP2*. This diversity of functional effect could indicate that multiple different pathways may need to be affected in a single individual in order to produce sufficient neurological change. If this is indeed the case, it will have important implications for therapeutic design.

4.2.5 GRIP2 knockout mice demonstrate deficits in social behavior

The previous experiments have established a statistically significantly increased burden of autism-associated variants to the PDZ456 region of *GRIP2*, that the variants correlate phenotypically with more severe behavioral and communication deficits, and that these variants produce a detectable biological effect by modulating binding capacity between *GRIP2* and its various interaction partners.

While these multiple lines of evidence suggest that *GRIP2* is worthy of further study as an autism susceptibility gene, they so far have not given decisive evidence that *GRIP2* is relevant in general to behavior consistent with autism phenotype. In order to determine if *GRIP2* is relevant in behavior, a series of tests using *Grip2* knockout mice was performed.

Grip2 conventional knockout mice were obtained from the lab of Rick Huganir. In all tests, ten male *Grip2* knockout mice were tested against ten male age-matched C57BL/6 control mice. In the social behavior tests, all reference mice are male age-matched C57BL/6 background.

An Open Field Test was first performed. Several metrics are calculated, including the amount of time spent rearing, the amount of time spent in the periphery or in the central field, the amount of time in ambulation (crude movement), and time spent in still activity (fine movement, such as grooming). *Grip2* knockout mice spent significantly less time in the central field for both gross ambulatory and fine movement ($p < 0.01$), although ambulation in the peripheral field was not changed relative to control mice. Total activity was also significantly reduced for knockout mice (Figure 4-10; $p < 0.01$).

The reduction in movement in the central field may indicate an anxiety-like phenotype. The reduction in total activity may reflect issues in anxiety, strength, and/or motor function. Additionally, total time spent rearing was significantly reduced (Figure 4-11; $p < E\text{-}6$). Rearing is an exploratory behavior, and a reduction in such behavior may again be indicative of an issue with anxiety, strength or motor function.

An Elevated Plus Maze was performed next, to further determine if these mice demonstrate an anxiety-like phenotype. The Elevated Plus Maze consists of an elevated platform composed of two open arms and two closed arms. Mice that spend more time in the closed arm likely have issues with fear and/or anxiety [175]. *Grip2* knockout mice spent significantly less time in the open arms relative to control mice, further supporting a possible anxiety disorder (Figure 4-12; $p < 0.03$).

As these observations could also be explained by defects in physical strength or motor coordination, both the Grip Strength Test and the Rotarod Test were performed. With the Grip Strength Test, the test mouse is place on grate to which it can cling. The grate is attached to a force meter (force measured in grams). The mouse is pulled from the grate, and the maximal force at which the mouse can no longer continue gripping the grate is measured. No statistically significant difference was observed between *Grip2* knockout mice and controls, indicating that strength deficits are an unlikely cause for the reduced ambulation and general activity observed (Figure 4-13).

With the Rotarod Test, mice are placed on a progressively accelerating rotating rod. At a maximal rotation rate, the mice can no longer stay on the rod and fall off. The amount of time spent on the rod is recorded. The Rotarod Test measures both strength

and motor coordination. Because mice improve on the Rotarod over time, the test is performed in triplicate, three days in a row. This improvement is reflective of the development of motor memory. As such, an inability to improve over time can also suggest difficulties in memory. The performance of *Grip2* knockout mice on the Rotarod Test is not statistically significantly different compared to control mice, indicating that *Grip2* knockout mice do not have deficits in muscle strength, motor coordination, or motor memory (Figure 4-13).

Additionally, an Olfaction Test was performed, in which a test mouse was placed in a clean cage with extra bedding. After a period of acclimation, the mouse is temporarily removed, a small block of cheese is hidden under the bedding, and the test mouse is returned to the cage. The amount of time to discover the block of cheese is recorded. Mice that take more time to find the block may have issues with their sense of smell. This is an important concern, as the sense of smell is one of the primary mechanisms of interaction and communication between mice. For social behavior tests performed later, what may appear to be a social deficit may in fact be an olfactory deficit. As such, deficits in olfaction must be ruled out.

An initial analysis indicated that *Grip2* knockout mice take a substantially, but not statistically significantly, longer time to discover the buried block of cheese, which would suggest an olfaction deficit (Figure 4-14). However, upon closer examination, it was discovered that *Grip2* knockout mice spend a significant amount of time immobile after replacement to the cage, in spite of prior acclimation. As such, crude time for finding the hidden cheese was divided into delay time (duration between replacement of the mouse

and its first steps with hind feet), and search time (duration between initial movement of hind feet and discovery of the block of cheese). Based on this analysis, the delay time is statistically significantly higher (6-fold; $p < 0.05$) for *Grip2* knockout mice compared to controls. This deficit in rapid exploratory behavior provides further support for a possible anxiety-like disorder. Search time is not significantly different between *Grip2* knockout and control, with the sample sizes used. As such, a deficit in olfaction could not be observed (Figure-14).

The first behavioral test performed was the Male Dyad Social Interaction Test. Two mice are separated in an open field by a divider. One mouse is a test mouse (*Grip2* knockout or control) and the other is a reference mouse (always WT) that is unfamiliar to the test mouse. After a five minute acclimation, the divider is removed and the mice are free to interact for ten minutes. Social behaviors of sniffing and following are recorded, as well as time spent self-grooming. *Grip2* knockout and control mice spent the same amount of time sniffing reference mice, but *Grip2* knockout mice spent a statistically significantly shorter amount of time following reference mice ($p < 0.05$), indicating a possible deficit in social interest. Additionally, *Grip2* knockout mice spent a statistically significantly longer time grooming, indicating a possible issue with anxiety or repetitive/compulsive behavior (Figure 4-15; $p < 0.05$).

Sociability Tests and Social Novelty Tests were performed next. For the Sociability Test, a test mouse (*Grip2* knockout or control) is acclimated to an open field for five minutes. At opposite corners of the open field are two mesh cages through which mice can interact but not touch. After acclimation, an unfamiliar reference mouse (WT) is

placed under the first mesh cage. The time spent by the test mouse interacting with the reference mouse is recorded over a five minute window. The Social Novelty Test continues from the Sociability Test. After the five minute window completes, a second unfamiliar reference mouse is placed under the second mesh cage. Over a second five minute window, time spent by the test mouse interacting with the first (familiar mouse) or second (novel mouse) mesh cage is recorded. With the Sociability Test, normal mice should spend more time with the first mesh cage than the second. With the Social Novelty Test, normal mice should spend more time interacting with the second mesh cage than the first.

For the Sociability Test, no significant difference was observed between *Grip2* knockout mice and control mice. As expected, both groups of mice spent more time with the first mesh cage with its unfamiliar reference mouse than with the empty second mesh cage (Figure 4-16). For the Social Novelty Test, both *Grip2* knockout and control mice spent roughly the same amount of time with the first mesh cage, with its now familiar mouse. As expected, control mice spent more time with the second mesh cage with its unfamiliar mouse than with the first. However, *Grip2* knockout mice did not spend more time with the second cage, indicating a deficit in preference for social novelty (Figure 4-16; $p < 5E\text{-}3$).

Based on these results, *Grip2* knockout mice do not appear to have deficits in strength, motor function, or olfaction. However, they are generally less active, possibly due to an anxiety-like defect. Additionally, they present with a repetitive or compulsive behavior as inferred from excessive grooming. And lastly, *Grip2* knockout mice display

deficits in social interaction and preference for social novelty. These results indicate that *Grip2* has an important role in regulating behavior in mice and loss of normal *Grip2* function produces behavioral deficits reminiscent of autism phenotypes, including defects in social interaction and repetitive behaviors. These observations suggest that *GRIP2* may similarly play an important role in human autism behavior, and that specific mutation of *GRIP2* could contribute to autism etiology.

4.3 Discussion

In order to establish that a particular gene is relevant for a genetic disorder, it is necessary to build a logical path. The gene must be burdened with an excess of functionally pathologic variants compared to controls, though pathogenicity may not be immediately apparent. Those mutations must perturb a particular biological pathway in appropriate tissues and developmental time points. This altered pathway must then be shown to influence physiology and behavior consistent with the disease phenotype. Genetic evidence leads to functional evidence, which leads to clinical evidence, thereby establishing a molecular mechanism of disease. Throughout this chapter, I have provided some genetic, functional, behavioral, and clinical evidence to establish *GRIP2* as a putative autism susceptibility gene, likely through perturbation of the glutamate signaling pathway.

Firstly, adult *Grip2* knockout mice display reduced social interactions and reduced preference for social novelty, as well as anxiety-like and repetitive behaviors, compared to normal age- and strain-matched controls, indicating *GRIP2* is a relevant

gene for studying autism-like behaviors. Secondly, there is an increased load of non-synonymous coding variants in *GRIP2*-PDZ456 between autistic cases and normal controls, as obtained through targeted next-generation sequencing. Thirdly, affected siblings carrying PDZ456 variants show a more severe deficit in reciprocal social interactions compared to affected siblings not carrying those mutations. Though this lends support to a possible role of *GRIP2* in autistic behavior, it does not necessarily indicate that *GRIP2* is driving disease; rather, it may simply be acting as a modifier under a risk background.

Lastly, autism-associated *GRIP2*-PDZ456 variants exhibit altered interaction with multiple binding partners, as determined by Yeast-Two-Hybrid (Y2H) analysis. These *GRIP2*-PDZ456 mutations have functional effects that may be relevant to disease. The mechanisms by which these mutations influence disease are not immediately apparent, but it is possible to hypothesize mechanisms for further functional study based on patterns of changes observed by Y2H.

GRIP2 autism-associated variants demonstrate a pattern of loss-of-function with binding to *GluA2/3* (Figure 4-9, Row 1). Given that *Grip2* knockout in mice (a loss-of-function model) produces a social deficit, loss-of-function mutations influencing interaction between *GRIP2* and *GluA2/3* may duplicate this behavioral change. Additionally, it has already been shown that gain-of-function *GRIP1* mutations increase binding to *GluA2/3*, accelerate AMPA receptor recycling rates, and increase AMPA receptor density at the synaptic surface, and *GRIP1/2* double knockout in mice (a loss-of-function model) generates improved social behaviors [129].

From this information, one possible mechanism for *GRIP2* mutations in glutamate signaling is that *GRIP2* functions counter to *GRIP1*. *GRIP2* loss-of-function mutations to *GluA2/3* could independently increase AMPA receptor recycling rates and synaptic surface density. Alternatively, if *GRIP2* is in competition with *GRIP1*, these mutations may make *GluA2/3* more accessible to *GRIP1*. This would mimic the activity of *GRIP1* gain-of-function variants, thereby producing increases in AMPA receptor recycling and surface density. These changes in the AMPA receptor pathway would interfere with the ability to modulate synaptic strength. This hypothesis depends on *GRIP1* and *GRIP2* having some competitive relationship; however at this time, the only evidence regarding the relationship between *GRIP1* and *GRIP2* indicates a cooperative one, complicating this hypothesis [167,174]. It is possible for both *GRIP*s to function cooperatively and competitively under different conditions, suggested by documentation of their independent and intersecting expression patterns [158-162]. To fully answer this question requires further study of *GRIP1/2*.

Autism-associated variants produce a unique pattern of binding changes with *ephrinB1/2*. The mutations V664M, E773K, and R843C (VER) increase *GRIP2* binding to *ephrinB1*, but decrease binding to *ephrinB2*. G749D (G) decreases *GRIP2* binding to *ephrinB1*, but increases binding to *ephrinB2* (Figure 4-9, Row 2). This reciprocal pattern may be a manifestation of the functional differences between the *ephrinB*s and their unique roles in neuronal and synaptic function across different brain regions. For example, *ephrinB1* has been shown to be highly expressed in the developing neocortex, with a specific expression pattern distinct from *ephrinB2* [176]. Conversely, specific

expression of *ephrinB2*, and not *ephrinB1*, is important in determining distinct mesolimbic and mesostriatal dopaminergic pathways in the developing midbrain [177]. Differential temporal and spatial expression of *ephrinB1* to embryonic primary olfactory neurons and *ephrinB2* to olfactory ensheathing cells during different stages of development of the olfactory bulb is important in establishing the proper structure and integration of primary and second-order neurons [178]. The unique expression patterns of the *ephrinB*s are necessary in establishing specific brain regions and the pathways that connect these regions. The reciprocal pattern of autism-associated *GRIP2* variants on binding to *ephrinB1/2* could play different roles in different regions, but cooperatively exacerbate proper neuronal activity when those different brains regions communicate.

A general increase in binding between *GRIP2* autism-associated variants and *liprin-alpha-1* was observed (Figure 4-9, Row 3). *Liprin-alpha-1* has been shown as necessary for the proper targeting of AMPA receptors to the synapse through *GRIP*, and necessary for production of LTD. The gain-of-function mutations observed between *GRIP2* and *liprin-alpha-1* may likely influence proper targeting of AMPA receptors. For example, given that *liprin-alpha-1* knockout results in loss of surface AMPA receptor density, these gain-of-function mutations may increase surface density, reminiscent of the higher AMPA receptor surface density observed with *GRIP1* gain-of-function mutations in Mejias et al. [168,169,129]. Future experiments should focus on co-localizations of AMPA receptor subunits, *GRIP2*, *GRIP1*, and *liprin-alpha-1*.

There is a pattern of increased binding (gain-of-function) between *GRIP2* autism-associated variants and *GRIP1*, while the control or common variants have no effect or

loss-of-function effect. The relationship between *GRIP1* and *GRIP2* in the same cell is still an active area of study. As indicated previously, the loss-of-function variants between *GRIP2* and *GluA2/3* can mechanistically explain perturbation of glutamate signaling in a manner consistent with *GRIP1* gain-of-function variants, if *GRIP1* and *GRIP2* are in competition with *GluA2/3* binding.

However, there is evidence that *GRIP1* and *GRIP2* may have a cooperative or compensatory relationship based on observation made from *Grip1* and/or *Grip2* knockout in mouse Purkinje neurons. Simple knockout of *Grip2* does not change receptor recycling rate, but double knockout of *Grip1/2* does successfully slow receptor recycling [167]. No observation of receptor recycling in single *Grip1* knockout neurons has been made yet. As such, *Grip2* does not seem to be sufficient alone in influencing AMPA receptor recycling. *Grip1* knockout appears to be necessary under *Grip2* knockout to slow recycling, providing support for a compensatory, and not competitive relationship. However, this can only be established by determining if *Grip1* knockout alone is sufficient to modulate receptor recycling rates.

At the same time, while *Grip1* knockout has been shown to completely abolish LTD induction in Purkinje neurons, *Grip2* knockout produces only a partial loss of LTD induction [174]. If there is compensation or cooperativity between the *GRIPs*, it is not fully compensatory or cooperative, and there is room for each *GRIP* to possess unique, and possibly divergent function.

One explanation for how the autism-associated variants may be functioning is that they are improving cooperativity between *GRIP1* and *GRIP2*. The increased binding

between *GRIP2* and *GRIP1* in the presence of these variants could allow *GRIP2* to unload itself of bound AMPA receptor, and pass on this AMPA receptor to the more tightly bound *GRIP1*. This could translate into improved interaction between *GRIP1* and AMPA receptors, consistent with activity observed with the *GRIP1* gain-of-function autism-associated variants previously studied in this same cohort set [129].

If in fact *GRIP2* and *GRIP1* have a competitive relationship, one alternate explanation is that these *GRIP2* autism-associated gain-of-function variants may increase competitive action between *GRIP1* and *GRIP2*. Stronger binding may sterically preclude *GRIP2* from binding AMPA receptors. In this way, a gain-of-function effect between *GRIP2* and *GRIP1* may generate a loss-of-function effect between *GRIP2* and AMPA receptor subunits. This loss-of-function effect could mirror or exaggerate the reduced binding observed between some *GRIP2* autism-associated variants and *GluA2/*3. This hypothetical model would be particularly exaggerated in the case of E773K, a PDZ6 autism-associated variant. E773K generates a >60% reduction in binding to *GluA3* and a 20% increase in binding to *GRIP1*. Both these changes could substantially deplete AMPA receptors of bound *GRIP2*, making them more accessible for *GRIP1* binding. Stronger *GRIP1-GluA3* interaction would then accelerate AMPA receptor recycling rates and increase AMPA receptor synaptic density [129].

Of course, our understanding of the relationship between the *GRIP*s is still developing. These two alternate explanations do not have to be mutually exclusive, and both may be occurring under different conditions.

Though individually each mutation has broadly different effects with different *GRIP2* interaction partners, it is possible to build a testable hypothesis regarding how some effects may be relevant to autism. There is an autism-like behavior in the *Grip2* knockout mice, which is a loss-of-function effect, and *GRIP2* autism-associated variants produce a loss-of-function effect between *GRIP2* and AMPA receptor subunits. Given these observations, I hypothesize the following: loss-of-function mutations in *GRIP2* reduce interaction between *GRIP2* and AMPA receptors, resulting in abnormal rates of receptor recycling to the synaptic surface. This will aberrantly alter that synapse's ability to produce long-term potentiation (LTP) or long-term depression (LTD). Without proper regulation of LTP or LTD, neurons cannot selectively strengthen or weaken specific synapses in response to glutamate signaling from pre-synaptic neurons. Without control over synaptic strength, the brain loses synaptic plasticity and with it, the ability to encode new information. Selective loss of plasticity in specific brain regions would then results in intellectual and behavioral features observed in Autism Spectrum Disorders.

Building evidence to support each of these steps in disease mechanism requires conducting a range of specific experiments that are currently ongoing. Firstly, neuronal assays are being performed with *GRIP2* variants, particularly V664M, G749D, and E773K, looking for changes in neuronal morphology, sub-cellular distribution of *GRIP2* binding partners, and changes to synaptic responses to glutamate signaling, such as AMPA receptor recycling assays. Additionally, in order to determine what brain regions may be most important, histological studies and immunochemistry are being performed on fixed post-mortem brain from *Grip2* knockout mice. In order to establish that these

specific mutations can influence behavior, a *Grip2* knock-in mouse model for V664M has been generated, and behavioral analysis of this model is pending.

Lastly, based on the results of these experiments, it may be possible to identify useful points in the glutamate signaling pathway to target pharmacologically, in order to modify the behavioral patterns of these transgenic mice. We are currently testing glutamate pathway agonists on *Grip1* and *Grip2* knockout mice to determine changes in behavior. These experiments will provide valuable knowledge on appropriate approaches for targeted treatment of autism.

4.4 Materials and Methods

4.4.1 Yeast-Two-Hybrid

The Yeast-Two-Hybrid (Y2H) protocol uses an endogenously-expressing ß-galactosidase yeast strain, such as Y190. This ß-galactosidase can only be expressed upon transcriptional activation by binding of an upstream activation sequence (UAS), immediately 5' of the ß-galactosidase gene. The yeast cell is transformed with two plasmids, a bait plasmid and a prey plasmid. These two plasmids separately express two genes whose gene products interact with each other. The gene on the bait plasmid is in fusion with a GAL4 DNA-binding domain (DBD); the gene on the prey plasmid is in fusion with a GAL4 transcriptional-activation domain (TAD). The UAS is bound by the bait through the DBD, and the bait is bound by the prey, due to the physical interaction

between the two bait and prey genes. The prey, with its TAD, can then recruit and activate RNA Polymerase II to transcribe the downstream ß-galactosidase gene.

The amount of ß-galactosidase enzyme in the cell is directly proportional to the strength of binding between the bait and prey. The amount of ß-galactosidase can be assayed colorimetrically using enzymatically reactive dyes, such as CPRG (chlorophenol red-ß-D-galactopyranoside) or ONPG (*o*-nitrophenyl-ß-D-galactopyranoside) [179]. Using this system, mutations in the peptide sequence of the bait or prey can be quantitatively assayed for changes in binding strength between interaction partners by observing the change in color of the dye.

Two yeast vectors were generated for bait and prey proteins. For the bait, PDZ456 of *GRIP2* was inserted into the multiple cloning site of the pPC97 vector, thereby placing *GRIP2*-PDZ456 in fusion with a GAL4 DNA-binding domain sequence already present on the vector. This vector, containing the wildtype form of *GRIP2*-PDZ456 derived from rat (*Rattus norvegicus*) peptide sequence (highly similar to mouse *GRIP2* sequence), was subjected to site-directed mutagenesis using the QuikChange system through Agilent Technologies. Site-directed mutagenesis produced individual *GRIP2*-PDZ456-pPC97 clones containing one of the *GRIP2* mutations identified by sequencing: R517Q, T540M, P543Q, A575T, N610S, V664M, R711C, G749D, E773K, and R843C. Additionally, a separate clone was generated using site-directed mutagenesis to produced a double mutant vector containing both R577A & K578A (KR.AA) changes. This KR.AA mutant of *GRIP2*-PDZ456 prevents binding of PDZ456 to both *GluA2* and *GluA3*, thereby acting as a negative control [129]. Individual clones were selected and sequenced using the

Sanger method to validate the presence of mutation and to exclude clones with off-target mutations.

For the prey, the pPC86 vector was used, which contains the GAL4 transcriptional activation domain (TAD). Cloned into the multiple cloning site in fusion with the TAD are the c-terminal 50 amino acids of *GluA2* or *GluA3*, PDZ456 of *GRIP1* (all obtained from Mejias et al.), PDZ456 of *GRIP2*, the c-terminal domains of *ephrinB1* or *ephrinB2*, or full-length *liprin-alpha-1* (all obtained from the Huganir Lab). A mutant form of pPC86-*GluA3* was obtained containing a deletion of the last four amino acids of the c-terminal domain of *GluA3* (WTCΔ4). This mutant cannot bind *GRIP2*-PDZ456, thereby acting as a negative control. Mutant forms of *ephrinB1* and *ephrinB2* c-terminal domains were obtained with deletion of the last three amino acids (WTCΔ3; NC). These mutants cannot bind *GRIP2*-PDZ456, thereby acting as negative controls. A mutant form was also obtained for *liprin-alpha-1* containing mutation of the terminal seven amino acids (-TVRTYSC) known to bind PDZ456 [168]. This mutant also cannot bind *GRIP2*-PDZ456, thereby acting as a negative control (all negative controls obtained from the Huganir Lab). For interaction between *GRIP2*-PDZ456-pPC97 (WT or mutants) with *GRIP1/2*-PDZ456-pPC86, empty vectors of pPC86 were used as negative controls.

Combinations of bait plasmids (*GRIP2*-PDZ456-pPC97, WT or mutant) and prey plasmids (WT *GluA2*, *GluA3*, *ephrinB1*, *ephrinB2*, *liprin-alpha-1*, *GRIP1*, *GRIP2*, or appropriate negative controls) were co-transformed into yeast cells of the Y190 strain, using a lithium acetate method [179]. Positive clones were selected from colonies grown

on selective plates (-Trp, -Leu) and grown in selective media (-Trp, -Leu), with six independent clones for each co-transformation.

ß-galactosidase assays were performed using a modified high-throughput protocol, described as follows:

High-Throughput Yeast-Two-Hybrid Protocol

Day 1:

1. Grow-up 500 µL of yeast over-night. Start at 2-4 pm. Use SDA (-Trp, -Leu) media. Use autoclaved 2 mL deep 96-well plates with 4mm glass beads at the bottom.

Day 2:

1. Harvest culture in the morning.
2. Transfer 100 µL of culture to 384-well OD (optical density) plate.
3. Take OD (optical density/absorbance) at 600 nm
    a. Subtract blank from each reading
    b. If X is adjusted reading, then take y=500*0.3/X µL of over-night culture and add to a new 2mL deep 96-well plate
    c. Add an additional 500 - y (500 minus y) µL of SDA media to each well
4. Grow new culture to log-phase for 3 hrs. Set aside additional growth wells to do time curve. Check OD every half hour until OD = 0.6-0.8. Save log OD.
5. Add 50 µL of culture to a well of a 96-well plate. Do this in quadruplicate (200 µL transferred total).

6. Centrifuge plates at max for 25 minutes at 4°. Plates can be stacked, but make sure they are completely balanced and that the bottoms of the plates are clean.
7. Turn plates upside down on a paper towel, and briefly centrifuge the supernatant out (just a few seconds, do not let rpm exceed 500). Quickly pull the plates out are turn them right side up.
8. Add 33.3 µL of Buffer 1 (see below) to each well; briefly centrifuge contents down, and vortex at 2000 rpm for 30 seconds.
9. Centrifuge plates at max for 25 minutes at 4°. Plates can be stacked, but make sure they are completely balanced and that the bottoms of the plates are clean.
10. Turn plates upside down on a paper towel, and briefly centrifuge the supernatant out (just a few seconds, do not let rpm exceed 500). Quickly pull the plates out are turn them right side up.
11. Add 10 µL of Buffer 1 to each well. Vortex at 2000 rpm for 5 minutes.
12. Freeze wells in liquid nitrogen and then thaw at 47°. Repeat an additional 2 times. Wrap reactions in aluminum and store at -80°. Freeze/thaw cycles break the yeast cells.
13. Prepare sterile Buffer 2 (see below) with CPRG (chlorophenol red-ß-D-galactopyranoside) concentration of 6.244 mM (26.558 mg of CPRG per 7 mL of Buffer 1).
14. Conduct time curve analysis checking reaction times by 30-second intervals.
    a. It is recommended that each time point be done in triplicate.

b. Start reaction by mixing 25 µL of Buffer 2 to each well of one of the freeze/thaw plates (one of the quadruplicate plates; the remaining three plates will be used for the actual Y2H)

c. It is best to try and start all reactions at the same time, and then mix together using plate vortexer.

d. Stop reaction using 35 µL of 6mM $ZnCl_2$ when the darkest accelerating reaction well completes its linear phase of the ß-galactosidase reaction

e. Record the time at which reaction was stopped. A fraction of this time will be used for the actual Y2H reactions, in order to ensure all reactions are stopped during the linear phase.

15. Take OD of ß-gal reaction (time curve analysis of step 14) using 578 nm wavelength.

    a. Transfer reactions using manual multi-channel pipet to 384-well plate and remove air bubbles as quickly as possible.

    b. Take optical density (578 nm).

    c. Subtract blank values for ß-gal ODs and log-growth phase ODs.

    d. Divide adjusted ß-gal values by adjusted log-growth phase OD values (final value).

    e. Choose a time point where the adjusted ß-gal values are between 0.25-1.8 and the final values are in the linear phase.

16. Do the actual reactions.

a. Start reaction with 25 µL of Buffer 2. Try to start all reactions at the same time. Mix with plate vortexer.
b. Stop reaction with 35 µL of 6mM of $ZnCl_2$ after specified time. Try to stop all reactions at the same time.
c. Transfer reactions to 384-well plate using manual multi-channel pipet. Remove air bubbles as quickly as possible.

17. Take OD of ß-gal reaction using 578 nm wavelength.
    a. Subtract blank values for ß-gal ODs and log ODs.
    b. Divide adjusted ß-gal values by adjusted log OD values (final value).

The protocol for making Buffers 1 and 2 are as follows:

Buffer 1 to prepare a 100 mL solution. Dissolve the following components in 75 mL of deionized water. Adjust pH to 7.25-7.30, then bring the volume to 100 mL. Filter sterilize. Store at 4ºC for up to 3 months.

HEPES	2.38 grams
NaCl	0.9 grams
L-Aspartate (hemi-Mg salt)	0.065 grams
BSA	1.0 grams
Tween 20	50.0 µL

Buffer 2 to prepare a 7 mL solution. Dissolve 26.558 mg of CPRG in 7 mL of Buffer 1, giving a final concentration of CPRG as 6.25 mM.

4.4.2 Generating a Grip2 Knockout Mouse

Grip2 knockout mice were obtained from the Huganir Lab. *Grip2* knockout mice were generated by targeting a PGK-neo cassette using EcoRI and BspEI sites to exons of PDZ1 of *Grip2* derived from a genomic phage library [180]. Linearized constructs were electroporated into murine embryonic stem cells and selected for resistance to G418. Homologous recombination was tested for by PCR and Southern Blot analysis. Chimeric mice were backcrossed to C57BL/6 strain for ten generations. Absence of *Grip2* was confirmed by Western Blot (Figure 4-17).

4.4.3 Behavioral Analysis of Grip2 Knockout Mice

Open Field Test. An Open Field Test was performed to assess general motor activity, movement patterns, stereotypical movement, and exploratory activity (such as rearing). Time spent in periphery versus the central field is evaluated for anxiety. Four open field chambers were used simultaneously, with each chamber 16 x 16 inches wide. Evaluations were conducted over a 30-minute window, with mouse movement tracked using the SDI Photobeam Activation System (San Diego Instruments). Time was measured for overall activity, rearing movement, fine movement, crude movement, presence in central field, and presence in periphery.

Elevated Plus Maze. An Elevated Plus Maze was performed to test anxiety by comparing time spent in the open arms versus the closed arms of an elevated plus-shaped platform. Timed recording were made with a camcorder, using a five minute recording window for each mouse. The elevated plus maze platform (SDI) is made of stainless steel and consists of two closed arms of 19.5 inches in length x 4 inches in width x 15 inches in height, and two open arms of 19.5 inches in length x 4 inches in width. The four arms are connected by a 4 x 4 inch platform, onto which each mouse was placed at the beginning of each recording.

Rotarod Test. A Rotarod Test was performed to test motor function, coordination, and overall strength. Tests were conducted using the Rotamex-V (Columbus Instruments). Mice were tested on the rotarod for no longer than 5 minutes per trial. Each mouse was tested in triplicate. The entire protocol was repeated for two additional days. Starting rotation speed was 5 RPM, and would accelerate by 1 RPM every 5 seconds. Amount of time spent on the rotarod before the mouse falls was recorded by photobeam.

Social Interaction Test. The Social Interaction Test was performed to examine exploratory behavior to a novel, unfamiliar mouse, analyzing metrics such as time spent conducting aggressive behavior, social behavior, like sniffing, following, or allo-grooming, or non-social behavior, such as self-grooming. A 16 x 16 inch square plastic chamber is divided diagonally by a high separator, partitioning the chamber into two triangular fields. A test mouse (*Grip2* knockout or WT control) is placed in one field, and a reference mouse (always WT C57BL/6) is placed in the other field. The mice are allowed to habituate to the field for five minutes. They cannot interact with each other

through the separator. After five minutes, the separator is removed, opening up the whole square chamber for exploration and interaction between the mice. Their interaction is video recorded for 10 minutes. The video is manually analyzed for time spent sniffing, following, and self-grooming. Test mice did not display aggressive behavior (attacks, biting, or tail flicks) or allo-grooming behavior.

Sociability and Social Novelty Test. The Sociability Test was performed to determine how sociable the test mice are to novel, unfamiliar mice. Two circular mesh cages are placed at opposite corners of a 16 x 16 inch plastic chamber. Both mesh cages are empty. A test mouse (*Grip2* knockout or WT control) is placed in the chamber and allowed to explore and habituate for five minutes. After five minutes, video recording is started and a reference mouse (WT C57BL/6) is placed under the first mesh cage. The test mouse and the reference mouse have never met before, including the previously conducted Social Interaction Test. The test mouse and reference mouse are allowed to interact through the mesh cage (no physical contact) for five minutes. Social interaction is recorded manually by analyzing the video for amount of time spent by the test mouse sniffing either the first or second mesh cage (placing the nose in close juxtaposition to either mesh cage). Normal mice will spend more time with the first mesh cage with the unfamiliar reference mouse than with the second, empty mesh cage.

After five minutes, a second reference mouse is placed in the second mesh cage. This marks the Social Novelty Test, which analyzes a mouse's preference for social novelty in the presence of a familiar social experience. The reference mouse in the first mesh cage is now a familiar mouse, and the reference mouse in the second mesh cage is

now the unfamiliar mouse, as it and the test mouse have never met before, including the previously conducted Social Interaction Test. The test mouse is allowed to interact with both reference mice for five minutes. The video is again manually analyzed for time spent by the test mouse interacting with either mesh cage. Normal mice will spend more time interacting with the second mesh cage with its unfamiliar mouse than with the first cage with its familiar mouse.

Statistical Analysis of Behavioral Studies. Statistical analysis and derivation of significance values was performed under all behavioral tests using a two-tailed Student's t-test, assuming unequal variance.

4.5 Figures: Chapter 4

Figure 4-1. Chromosomal locations for Glutamate Receptors, Transporters, and Interacting Proteins overlapping with autism susceptibility loci that are identified in published chromosomal, linkage, and/or association studies

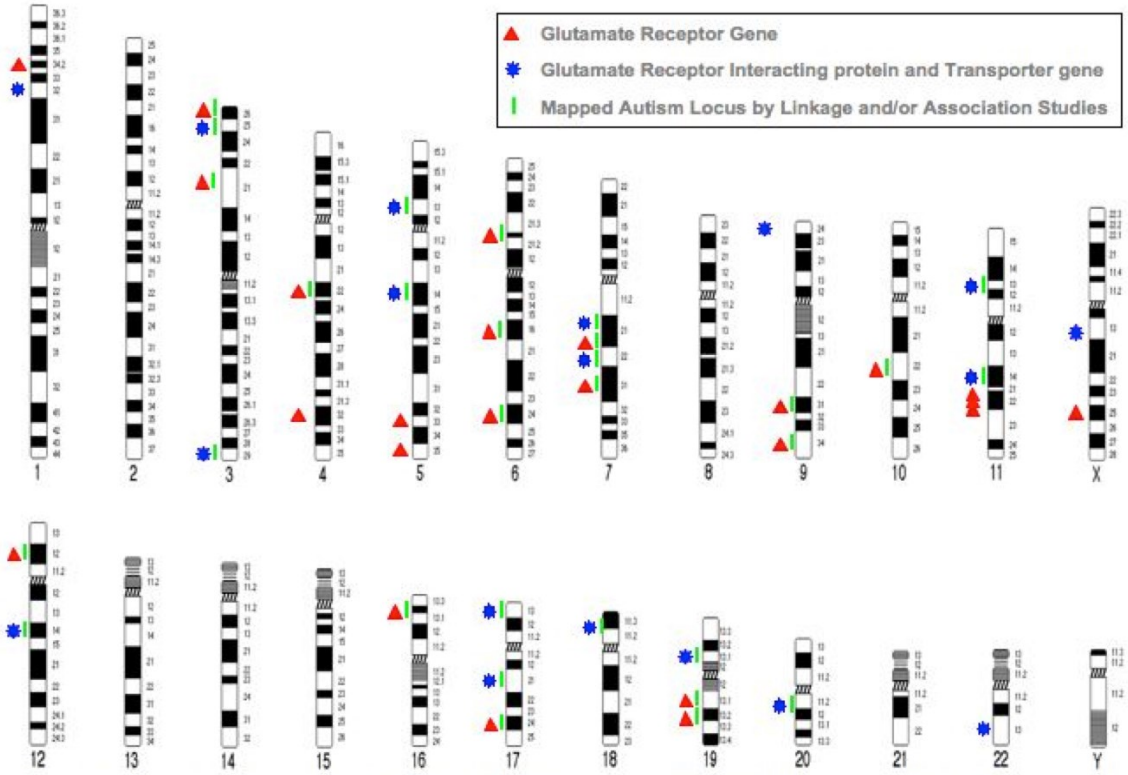
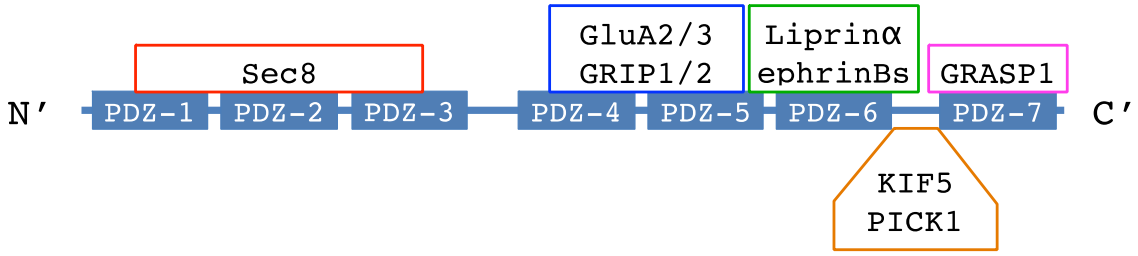


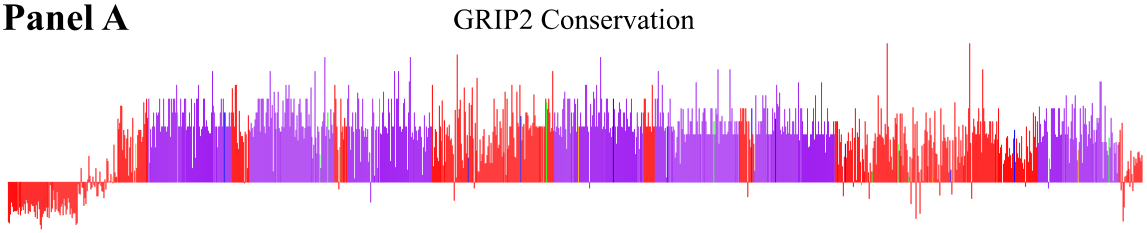
Figure 4-2. *GRIP1/2* PDZ domains and their respective interaction partners



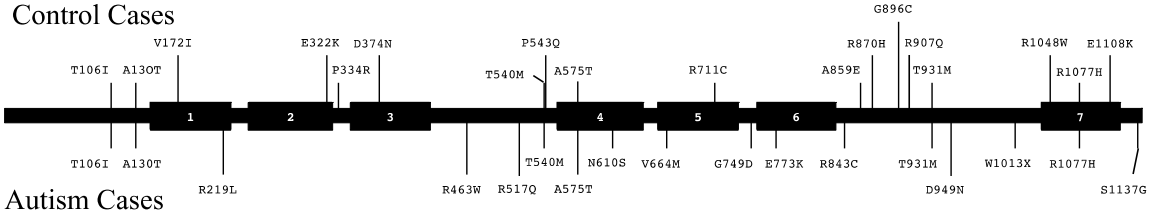
GRIP1 and *GRIP2* interact with many proteins involved in glutamatergic and GABAergic neurotransmission. Some key interaction partners are described above. PDZ123 binds members of the exocyst complex, such as Sec8. PDZ45 binds to *GluA2/3* and other *GRIP1/2* proteins. PDZ6 binds to *ephrinB1/2* and *liprin-alpha-1*. PDZ7 binds to *GRASP1*. The linker region between PDZ6 and PDZ7 binds to *KIF5*, a kinesin motor protein, and *PICK1*, which is important in AMPA receptor internalization and recycling. The diagram does not include all known interaction partners.

Figure 4-3. *GRIP2* conservation and topology

Panel A



Panel B



Panel C

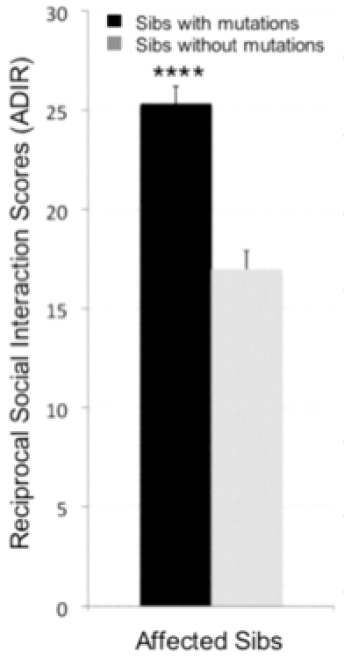
Conserved Region A **Conserved Region B** **Conserved Region C**

	Full Length GRIP2		Conserved Region A	Conserved Region B	Conserved Region C
	Total Alleles	Cumulative NSCV	Cumulative NSCV	Cumulative NSCV	Cumulative NSCV
Autism	960	28	5	18	5
Control	960	20	7	6	7
Fisher Exact (two tailed)		$p = 0.306$	$p = 0.773$	$\mathbf{p = 0.022}$	$p = 0.773$

Panel A shows relative conservation of amino acids across *GRIP2*, using a 42-way multi-species alignment. Each vertical line is an amino acid residue, with vertical height reflecting relative conservation (amino acid similarity using a BLOSUM62 matrix). Taller lines indicate higher conservation. Purple regions are PDZ domains. Blocks of conservation around multiple PDZ domains are reflective of their shared functionality. As such, PDZ123 can be considered one conserved region (A), and PDZ456 can be considered another conserved region (B). Panel B provides the positions of nonsynonymous coding changes identified through high-throughput sequencing of *GRIP2*. Variants identified in control cases are provided above the *GRIP2* topology map

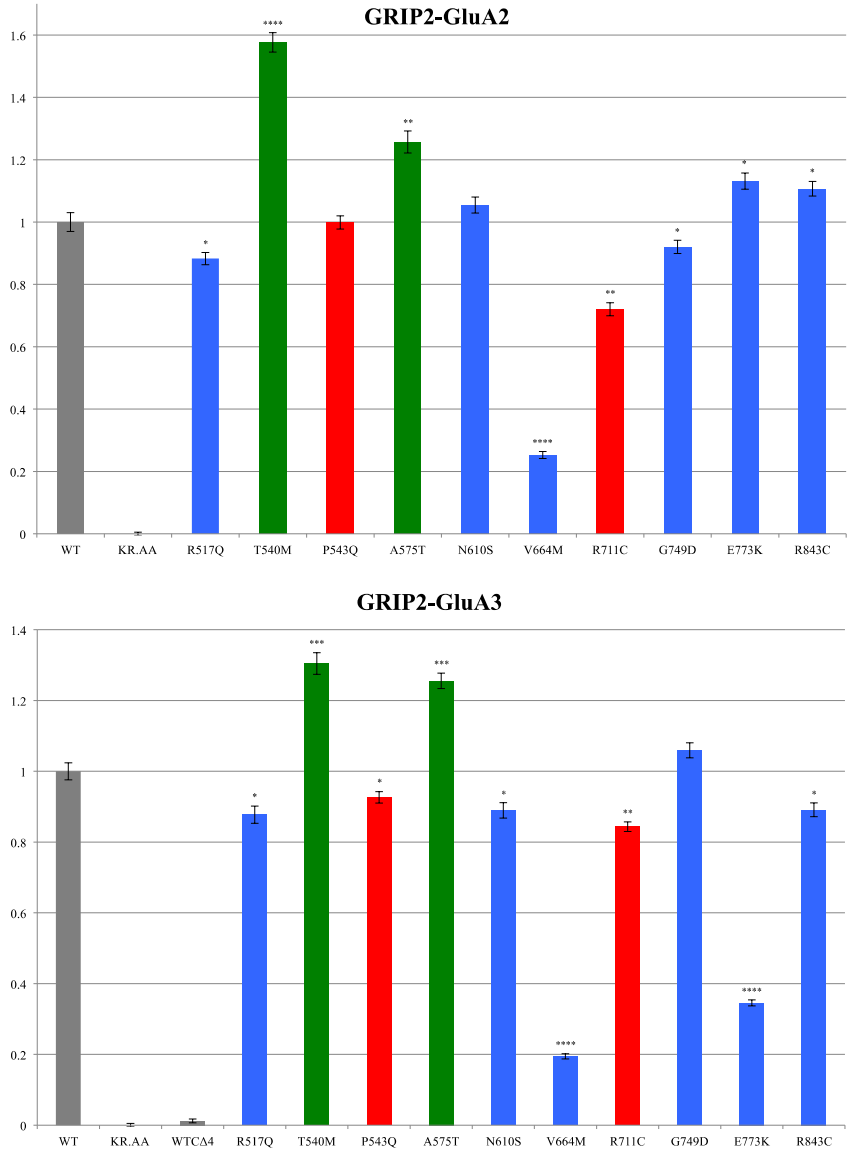
(in black), and variants identified in autism cases are provide below the map. Panel C tabulates the number of nonsynonymous coding variants (NSCV) in each conserved region. Calculation of statistical mutation burden is conducted using Fisher's exact test. Only Conserved Region B (PDZ456) has a statistically significantly increased burden of mutation for autism cases.

Figure 4-4. *GRIP1/2* PDZ456 variants correlate with more severe social deficits



Autism cases and affected siblings who are genotypically discordant are quantified for cumulative social deficits by summation of the reciprocal social interaction scores (see Table 4-2). Autism cases with a *GRIP1/2* coding mutation are significantly more likely to have a higher, and more severe social score than their affected sibling without a *GRIP1/2* mutation ($p <$ E-3).

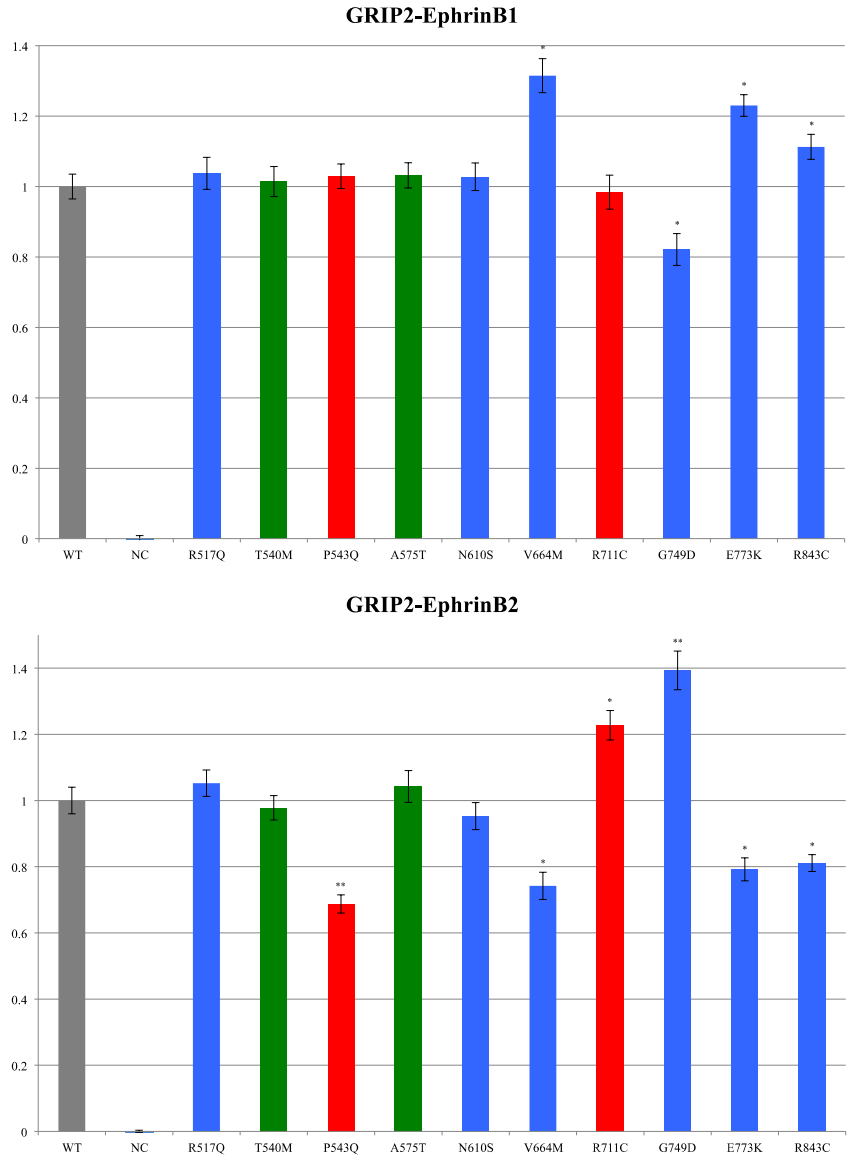
Figure 4-5. *GRIP2* variants change binding strength to *GluA2/3* in Y2H assay



PDZ456 domains from WT and mutant *GRIP2* are tested against *GluA2/3*. R577A-K578A (KR.AA) is a *GRIP2* double mutation acting as negative control against both *GluA2* and *GluA3*. WTCΔ4 is a four residue terminal deletion of *GluA3* acting as a second negative control. Relative β-galactosidase activities for individual mutants are normalized with WT and negative controls, and are presented as mean ± sem in triplicate

studies. WT and negative control levels are presented in grey. Autism-associated variants are in blue. Common variants are in green. Control variants are in red. Student's T test was performed for comparison of two means between mutants and WT. *, $p < 0.05$; **, $p < 0.01$; ***, $p <$ E-8; ****, $p <$ E-20.

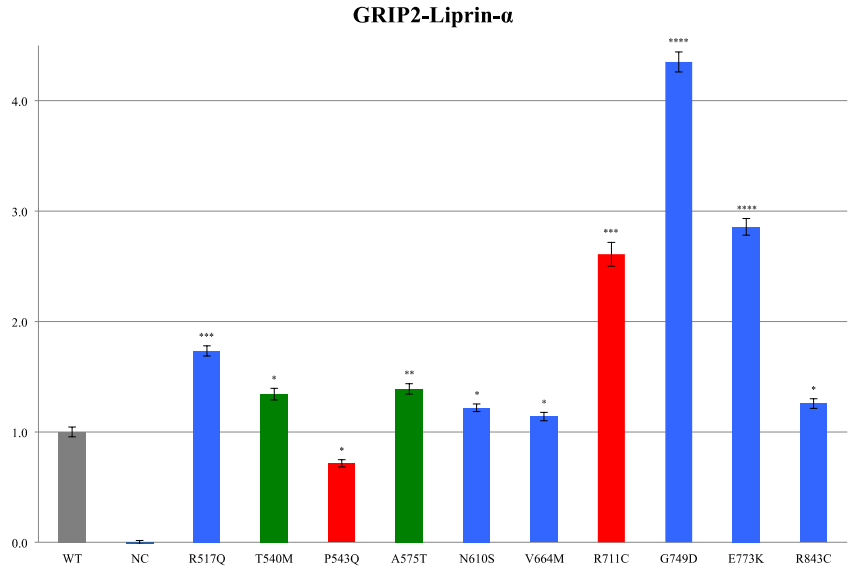
Figure 4-6. *GRIP2* variants change binding strength to *ephrinB1/2* in Y2H assay



PDZ456 domains from WT and mutant *GRIP2* are tested against *ephrinB1/2*. NC is the negative control, using co-transformation of the R577A-K578A (KR.AA) *GRIP2* double mutant with a WTCΔ3 mutant for both *ephrinB1* and *ephrinB2*. WTCΔ3 is a three residue terminal deletion of *ephrinB1/2*, preventing binding to *GRIP2*. Relative β-galactosidase activities for individual mutants are normalized with WT and negative

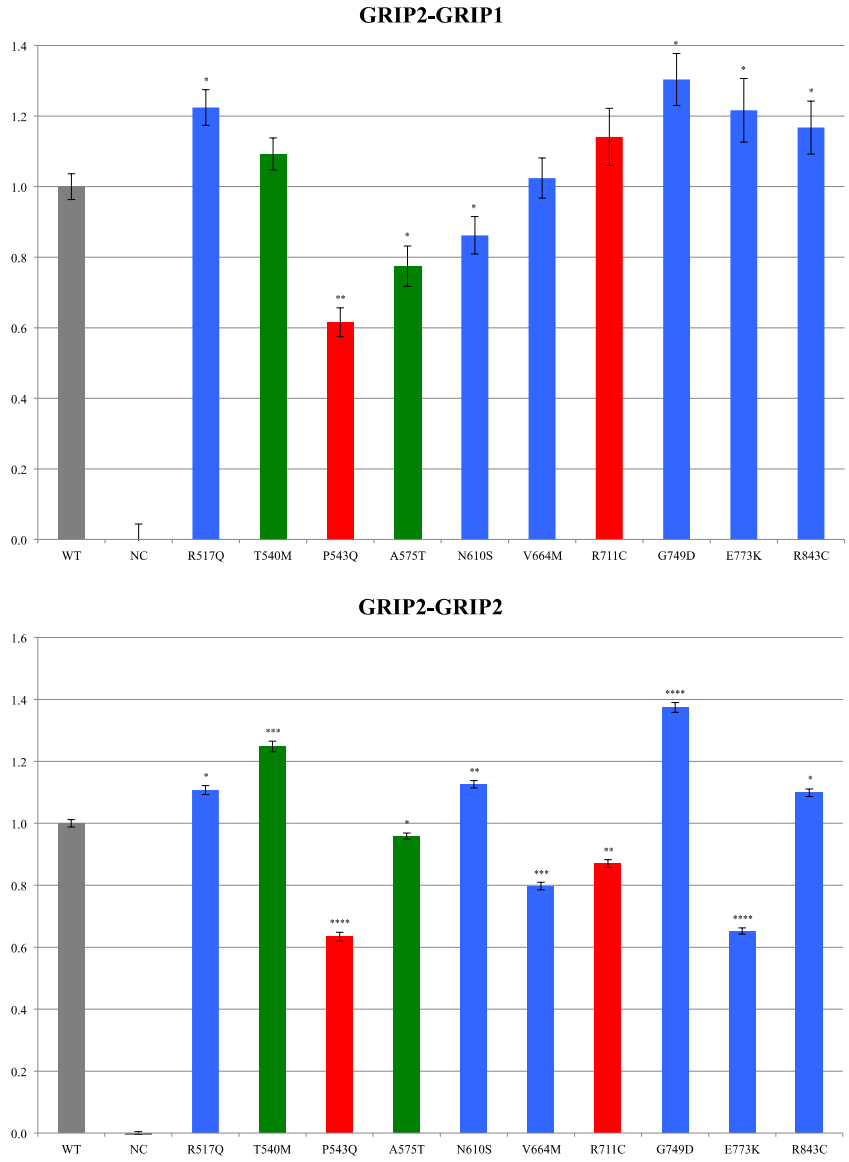
controls, and are presented as mean ± sem in triplicate studies. WT and negative control levels are presented in grey. Autism-associated variants are in blue. Common variants are in green. Control variants are in red. Student's T test was performed for comparison of two means between mutants and WT. *, $p < 0.05$; **, $p < 0.01$.

Figure 4-7. *GRIP2* variants change binding strength to *liprin-alpha-1* in Y2H assay



PDZ456 domains from WT and mutant *GRIP2* are tested against *liprin-alpha-1*. NC is the negative control, using co-transformation of the R577A-K578A (KR.AA) *GRIP2* double mutant, and a *liprin-alpha-1* mutant containing a c-terminal heptapeptide mutation that precludes binding to *GRIP2*. Relative β-galactosidase activities for individual mutants are normalized with WT and negative controls, and are presented as mean ± sem in triplicate studies. WT and negative control levels are presented in grey. Autism-associated variants are in blue. Common variants are in green. Control variants are in red. Student's T test was performed for comparison of two means between mutants and WT. *, $p < 0.05$; **, $p < 0.01$; ***, $p <$ E-8; ****, $p <$ E-20.

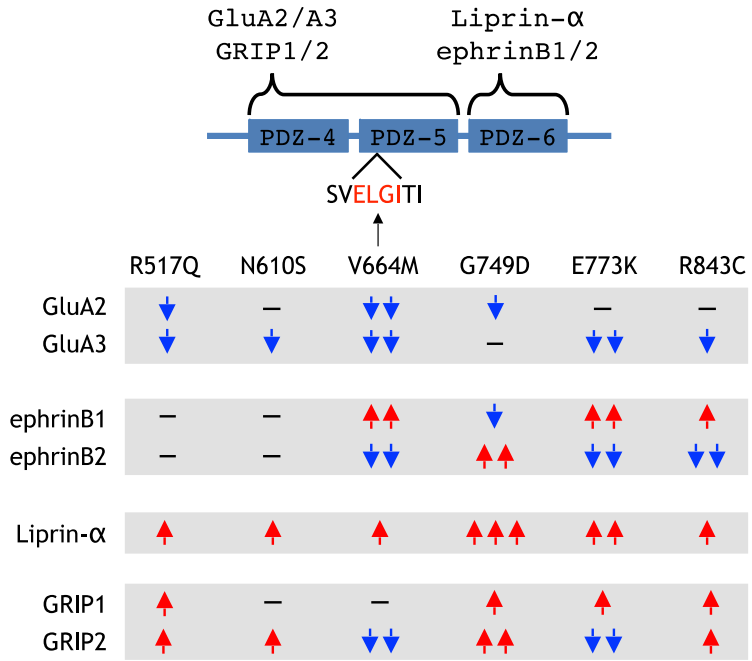
Figure 4-8. *GRIP2* variants change binding strength to *GRIP1/2* in Y2H assay



PDZ456 domains from WT and mutant *GRIP2* are tested against *GRIP1* and *GRIP2*. NC is the negative control, using co-transformation of the R577A-K578A (KR.AA) *GRIP2* double mutant, and empty prey vector (pPC86). Relative β-galactosidase activities for individual mutants are normalized with WT and negative controls, and are presented as mean ± sem in triplicate studies. WT and negative control levels are presented in grey.

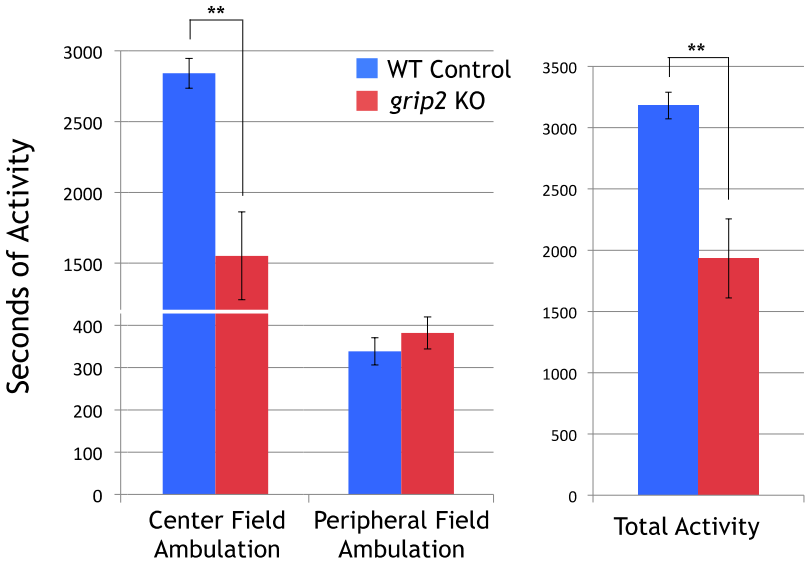
Autism-associated variants are in blue. Common variants are in green. Control variants are in red. Student's T test was performed for comparison of two means between mutants and WT. *, $p < 0.05$; **, $p < 0.01$; ***, $p <$ E-8; ****, $p <$ E-20.

Figure 4-9. *GRIP2* autism-associated variants produce consistent patterns of changes with *GRIP2* interaction partners



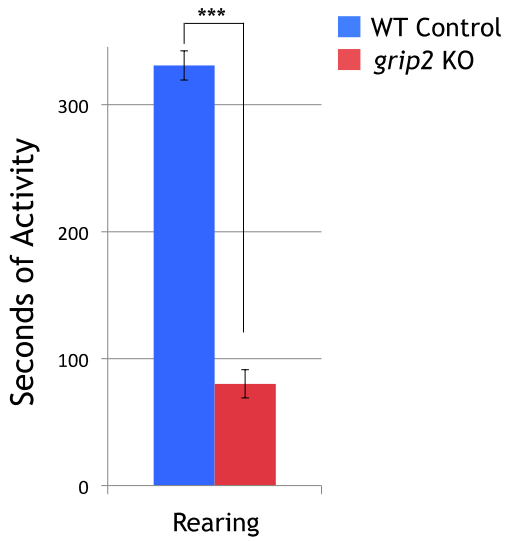
At top is shown the PDZ domains and their respective interaction partners for PDZ456. The ELGI peptide sequence of PDZ5 is the known binding site of *GluA2/3*. The valine just prior to ELGI is the valine of the V664M missense mutation that results in almost complete attenuation of binding between *GRIP2*-PDZ456 and *GluA2/3*. Autism-associated variants show a pattern of reduced binding to *GluA2/3*, reciprocal binding patterns for *ephrinB1/2*, increased binding pattern to *liprin-alpha-1*, and a generally increased binding to *GRIP1*. Importantly, all six mutations have some significant effect on at least three interaction partners.

Figure 4-10. *Grip2* knockout mice display reduced activity in the Open Field Test



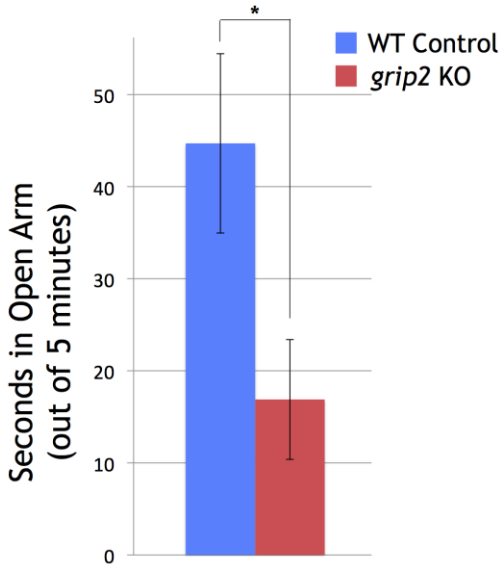
Time in activity is recorded on the Y-axis. Blue bars indicate WT control mice, and red bars indicated *Grip2* knockout mice. Average time in activity is presented as mean ± sem (n = 10 for both mouse cohorts). *Grip2* knockout mice display reduced crude movement activity (ambulation) in the central field of the open field test compared to controls (left graph; $p < 0.01$). The same amount of ambulation time is spent in the periphery. Total activity is provided in the graph on the right, for which *Grip2* knockout mice are significantly less active. A decrease in central field movement and total activity may indicate issues with anxiety, strength, or motor function.

Figure 4-11. *Grip2* knockout mice spend less time rearing in the Open Field Test



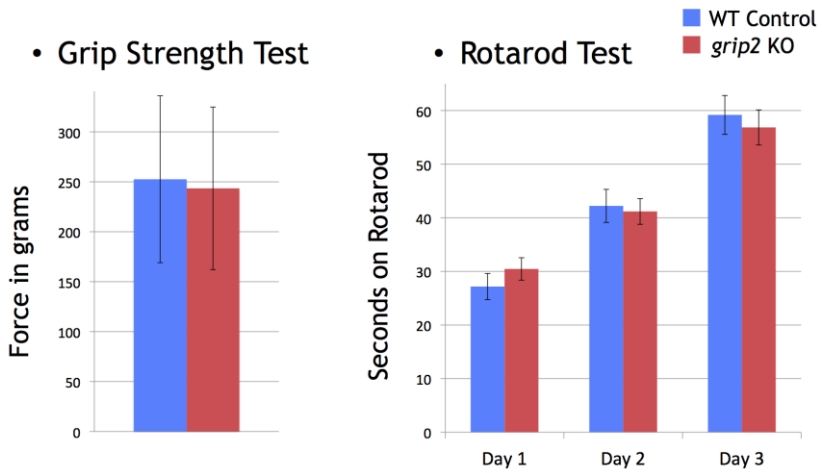
The amount of time spent rearing, an exploratory behavior, is significantly reduced ($p <$ E-6) for *Grip2* knockout mice (red bars) compared to WT controls (blue bars). Average time in activity is presented as mean ± sem ($n = 10$ for both mouse cohorts). A decrease in rearing may further support an issue in anxiety, strength, or motor function.

Figure 4-12. *Grip2* knockout mice display anxiety traits in the Elevated Plus Maze



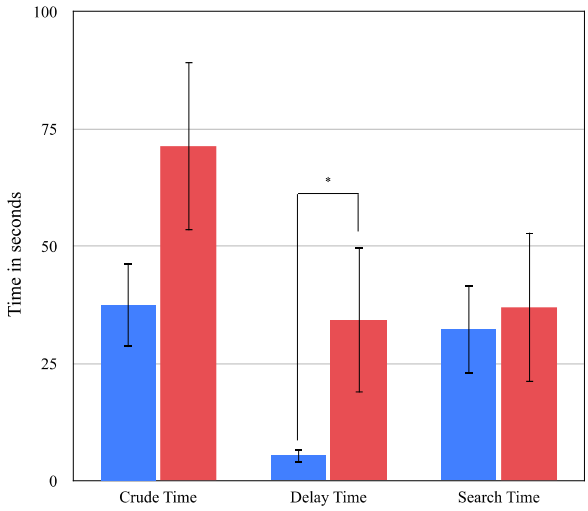
The Elevated Plus Maze provides a measure of anxiety levels in mice, by comparing the amount of time spent in an open arm (exposed) instead of a closed arm (sheltered). Average time in the open arm is presented as mean ± sem (n = 10 for both mouse cohorts). *Grip2* knockout mice (red bars) spend significantly less time in the open arm ($p < 0.05$) during a five minute test window, compared to WT control mice (blue bars). This decrease in time in the open arms may reflect an issue with anxiety for *Grip2* knockout mice.

Figure 4-13. *Grip2* knockout mice do not display deficits in strength or motor function



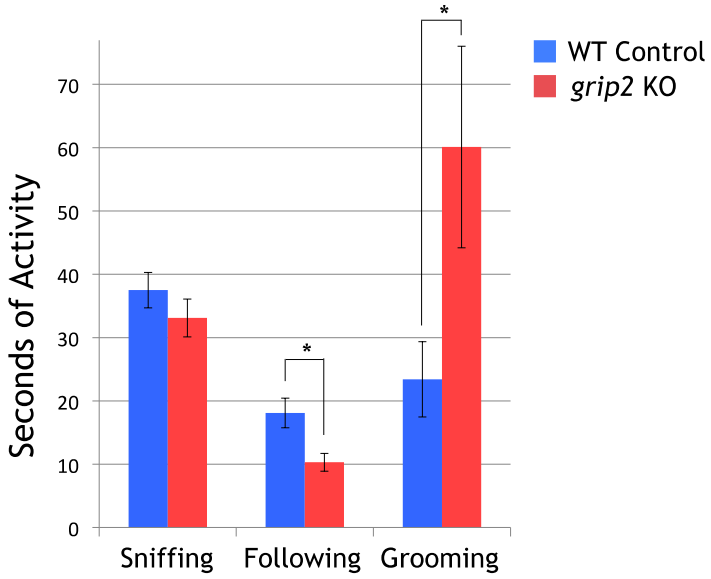
On the left graph is presented the results of the Grip Strength Test. Average strength is presented as mean ± sem in grams (n = 10 for both mouse cohorts). *Grip2* knockout mice (red bars) do not display a significant deficiency in strength compared to WT controls (blue bars), given the sample sizes used. On the right graph is presented the results of the Rotarod Test, performed over three consecutive days. Average time on the Rotarod before falling is presented as mean ± sem (n = 10 for both mouse cohorts). *Grip2* knockout mice do not demonstrate a significant difference in motor coordination, motor strength, or motor memory, compared to WT controls, given the sample sizes used. As expected, both cohorts of mice improve (longer time spent on Rotarod) over time.

Figure 4-14. *Grip2* knockout mice do not display an olfaction deficit



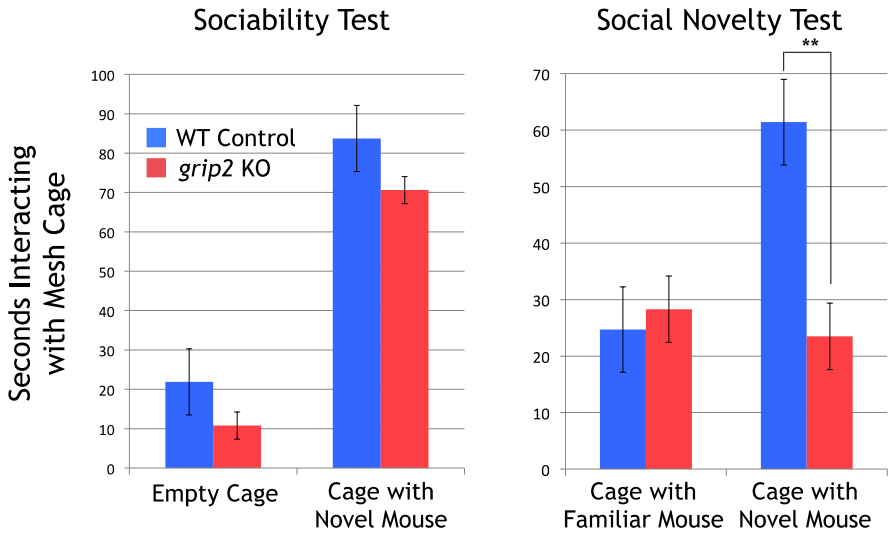
Recorded is the amount of time taken by *Grip2* knockout mice (red bars) and WT controls (blue bars) to find a buried block of cheese, as a test of olfactory sense. Average time in activity is presented as mean ± sem (n = 10 for both mouse cohorts). Crude Time on the left is time taken from initial drop of the mouse into the cage to the moment when the block of cheese is found. *Grip2* knockout mice spend a substantial (but not statistically significant) amount of Crude Time finding the block of cheese. Middle and right graphs show Crude Time broken into Delay Time and Search Time. Delay Time is the time taken from initial drop into the cage to the moment the mouse first moves its hind feet. Search Time is the time taken from movement of hind feet to the moment the block of cheese is found. *Grip2* knockout mice have a significantly longer delay time ($p <$ 0.05) compared to WT controls, possible reflective of an anxiety issue. There is no significant difference in Search Time, given the cohort sizes used, likely indicating that there is no deficit in olfaction.

Figure 4-15. *Grip2* knockout mice have impaired social and grooming behavior in the Social Interaction Test



Presented in the amount of time spent by *Grip2* knockout mice (red bars) and WT controls (blue bars) engaging in social behaviors in the Social Interaction Test. Average time in activity is presented as mean ± sem (n = 10 for both mouse cohorts). No significant difference is observed for time spent sniffing (left bars) a wildtype reference mouse. *Grip2* knockout mice spent significantly less time following (middle bars) reference mice compared to WT controls ($p < 0.05$), indicating an anomaly in social behavior. *Grip2* knockout mice spend significantly more time self-grooming (right bars) compared to WT controls ($p < 0.05$), indicating a possible concern with repetitive or compulsive behavior. None of the mice displayed aggressive behavior or allo-grooming.

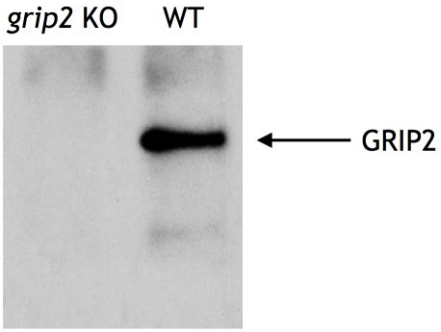
Figure 4-16. *Grip2* knockout mice display a reduced preference for social novelty in the Sociability and Social Novelty Tests



Provided on the left graph are the results of the Sociability Test. Average time spent interacting with a particular mesh cage is presented as mean ± sem (n = 10 for both mouse cohorts). As expected, both *Grip2* knockout and WT control mice spend more time interacting with the mesh cage with a novel mouse than with the mesh cage that is empty. There are no statistical differences between *Grip2* knockout mice and WT controls.

Provided on the right graph are results of the Social Novelty Test. As expected, both *Grip2* knockout and WT control mice spend little time interacting with the original mesh cage with its more familiar mouse than the cage with the more novel mouse. However, *Grip2* knockout mice spend far less time interacting with the novel mouse in the second mesh cage, compared to WT controls ($p < 0.01$). As such, *Grip2* knockout mice display a reduced preference for social novelty.

Figure 4-17. Western blot confirms loss of *Grip2* protein in *Grip2* knockout mouse



Grip2 protein was undetectable in *Grip2* knockout mouse brain using a *Grip2*-specific antibody to the *Grip2* c-terminus. Protein was detectable in strain- and age-matched WT control brain.

4.6 Tables: Chapter 4

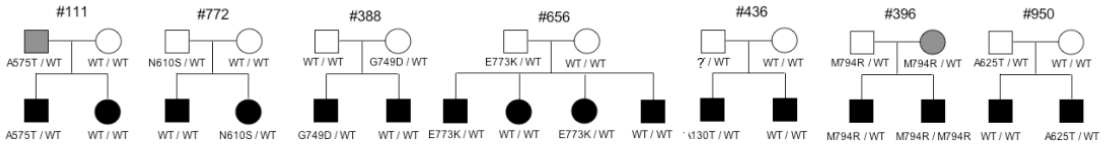
Table 4-1. Race/ethnicity of cohorts of patients with autism and matched controls

	Autism*		Control[T]	
Race and ethnicity	Number	Percentage	Number	Percentage
White	336	70.0	366	76.2
Black or African American	31	6.5	30	6.3
Asian	6	1.3	5	1.0
More Than One Race	16	3.3	0	0.0
Hawaiian/Pacific Islander	1	0.0	0	0.0
Unknown	90	18.9	79	16.5
Total	480	100.0	480	100.0

*Autism Genetics Research Exchange and South Carolina Autism Project.
[†]Individuals with apparently normal cognitive function and behaviors from the Greenwood Genetics Center, SC.

Table 4-2. Genotype-Phenotype Correlations of *GRIP1/2* variants between discordant autism siblings

Index Family	Autism Phenotype	Gene	Genetic Variants: AA change	Genetic Variants: Genotype	Social Interaction Score (SOCT_CS)	Communication Score: Verbal (COMVT_CS)	Communication Score: Nonverbal (COMNVT_CS)	Behavioral Score (BEHT_CS)
111	Affected son	GRIP2	A575T	het	20	17		6
	Affected daughter			wt	11	11		3
772	Affected son	GRIP2	N610S	het	27		14	2
	Affected daughter			wt	16	13		4
388	Affected son	GRIP2	G749D	het	27		14	7
	Affected son			wt	21	12		6
656	Affected son	GRIP2	E773K	het	28	18		5
	Affected son			wt	16	14		6
	Affected daughter			het	26	17		5
	Affected daughter			wt	19	13		4
436	Affected son	GRIP2	A130T	het	24	19		6
	Affected son			wt	11	15		3
396	Affected son	GRIP1	M794R	homo	26		13	8
	Affected son			het	22	16		10
950	Affected son	GRIP1	A625T	het	24	18		6
	Affected son			wt	20	16		4



Provided are quantitative phenotypes for affected sib-pairs with discordant genotypes for *GRIP* variants in proband families. Cumulative scores are summations of standard ADI-R tests. Higher scores indicate more severe phenotype. A nonverbal score suggests a more severe communication deficit. WT is wildtype. HET is heterozygous. Red arrows show a consistent directional pattern of more severe disease in siblings with a *GRIP* variant. Family pedigrees are provided at the bottom. Filled symbols indicate children with autism. Open symbols are not diagnosed with autism. A grey square for pedigree #111 indicates an eccentric, "loner" father with speech delay requiring therapy, but not diagnosed with autism.

Chapter 5: Concluding Remarks

5.1 The Spectrum of Complexity, Revisited

The results obtained throughout the different studies in this dissertation provide additional support for the spectrum of complexity previously introduced. It is well known that different genes that have been implicated in single-gene disorders are also relevant to disease in more complex disorders, even if the inheritance patterns and clinical presentations differ substantially. This overlap of genetic etiology should not come as a surprise. Innumerable parallel and intersecting biological pathways are involved in normal brain function. As such, a collection of mutated genes can produce roughly similar neurological and psychiatric phenotypes (genetic heterogeneity), with different mutations and genetic backgrounds influencing the degree of phenotype severity (variable expressivity). Individually each gene can in turn influence multiple separate disease phenotypes (pleiotropy).

This pleiotropy is likely influenced by the contribution of a single gene to multiple parallel pathways. This is a result both of the structure of biological pathways, but is also a result of the structure of genes. As described by François Jacob, "Nature is a tinkerer, not an inventor" [181]. This is particularly true for genes, which are composed of multiple independent functional domains that can link parallel biological pathways together to form more sophisticated systems. The *GRIP1/2* proteins are an excellent example of this. One set of domains, PDZ123, provides function in cellular trafficking across multiple cell types, from neurons to skin cells [180]. Another set of domains,

PDZ456, provides functionality to glutamatergic neurotransmission. A linker region provides trafficking mobility. PDZ7 provides functionality in GABAergic neurotransmission. This phenomenon is a result of the manner in which genes have evolved, through a process of tinkering with interchangeable and mutable parts to construct greater complexity.

As such, it is important to analyze disease variant association in the context of domains, and not simply at the level of an entire gene. As was demonstrated with the *GRIP1/2* study, mutation burden and genotype-phenotype correlation provided no positive results when looking across the entire gene. However, focusing on one highly conserved domain set provided valuable and significant association, and laid groundwork for developing a focused hypothesis on how *GRIP1/2* may function in disease.

Returning to the concept of multiple overlapping biological pathways influencing a spectrum of disease complexity, the studies in this dissertation provide additional support of how different genes in vastly different roles can impinge on the same pathways and produce different phenotypic outcomes. Insight into how this process works should be valuable in bridging therapeutics used in one disease, and transferring its utility to another disease.

In the study of X-linked Intellectual Disability (XLID), *ZC4H2*, a zinc-finger transcription factor, was identified at the top of our prioritized list, with two missense changes and one splicing change. Both missense changes are predicted to be damaging by SIFT and PolyPhen-2. The pedigrees for the probands possessing the two missense changes are provide in Figure 5-1, and show a clear pattern of X-linked inheritance. Two

of the variants identified, the R190W missense change and the splicing change, were submitted to a collaborator (Hwang et al.) for analysis in zebrafish. *Zc4h2* zebrafish knockouts were generated using TALEN technology. Developing zebrafish were observed to have a loss of GABAergic interneurons as determined by a loss of *Gad1* staining. When WT human or zebrafish *zc4h2* mRNA was reintroduced into these knockout fish, it resulted in a complete recovery of the phenotype. However, when mutant *zc4h2* mRNA containing either of the two mutations was introduced into these knockout fish, it only resulted in a partial recovery of the phenotype [data not shown]. These multiple points of evidence suggest that *ZC4H2* is a strong candidate gene for XLID, with a mechanism of disease possibly involving the GABAergic neurotransmission pathway [Hwang et al., unpublished].

As previously mentioned, *GRIP1* and *GRIP2* are important in GABA signaling and are robustly expressed in inhibitory GABAergic interneurons [158,161]. *ZC4H2* and *GRIP1/2* share little in common in gene structure or function. The former is a transcription factor expressed early in development, and the later are scaffolding proteins expressed from early development throughout adulthood. Yet both appear relevant to normal function of GABAergic neurons and both are important in different neurological diseases. Though there is little evidence that *GRIP1/2* are relevant to intellectual disability, or that *ZC4H2* is relevant to autism, this biological convergence acts as further support that, at least at the pathway level, there is important overlap between otherwise clinically distinguishable diseases. Building our understanding of signaling pathways,

such as glutamatergic and GABAergic neurotransmission, will be relevant not only to autism, but to many neurodevelopmental conditions.

5.2 The Necessity of Functional Experimentation

The results from studying *ZC4H2* XLID-associated mutations also underscores the importance of conducting functional studies on as many potential disease-causing variants as possible. While *ZC4H2* was obtained as a potential XLID gene by statistical analysis, how it functions in disease can only be understood by placing the mutation into a biological system and recording its effect. Though this seems an obvious requirement in the study of disease, it is important to underscore this point, because it is easy to forget that functional experimentation of individual variants for all relevant disease genes is in fact a daunting task with little workaround.

This is made all the more important by the results of studying *GRIP2* autism-associated variants. Many of the variants identified in *GRIP2* sequencing were excluded from further functional analysis due to insufficient statistical association or high allelic frequency in the population. Though association and variant frequency are appropriate metrics by which to prioritize variants, it would be careless to completely exclude such variants from further analysis. As was demonstrated in the analysis of *GRIP2*-PDZ456, common and control variants also can have functional effect. In fact, in certain instances the common and control variants produce biological effects counter to what was observed with autism-associated variants, suggesting a possible protective effect, which is important in our understanding of disease mechanism. That *GRIP2* common or control

variants outside of PDZ456 may have important functional relevance to autism is a distinct possibility. As such, the general exclusion of such variants independent of disease, variant frequency, or cohort origin, from further analysis should only ever be temporary.

5.3 Significance for Therapeutics

Functional work is performed to establish a molecular mechanism and identify targets for pharmacologic intervention. However, with so many variants performing different functions, can just one or a few drugs be sufficient to effectively treat disease? Alternatively, with hundreds to thousands of possible disease genes and divergent pathways in intellectual disability and autism, would we need hundreds to thousands of therapeutic options?

If we were to assume that *GRIP2* is truly involved in autism etiology, *GRIP2* autism-associated variants then provide an excellent theoretical paradigm on the complexity of treating such disease variants, given their individually unique influences on different interaction partners. Because each variant does something different, it would be difficult to design individualized therapies for each variant. However, as there are consistent patterns of variant effect, such as general loss-of-function between *GRIP2* and *GluA2/3* and general gain-of-function between *GRIP2* and *liprin-alpha-1* and *GRIP1*, it may be possible to design drugs that individually target each of those interaction pathways, and then combine drugs as appropriate for each patient.

However, the case of *GRIP2* autism-associated variants and *ephrinB1/2* may complicate this scenario. The reciprocal nature of these variants in their influence on binding to *ephrinB1/2* demonstrates how potential risk variants produce a biological nuance that may be difficult to negate therapeutically. As described previously, the differences between *ephrinB1* and *ephrinB2* are substantial, in their interaction partners and ligands, their effects on forward and reverse signaling, and their functions in defining separate brain regions and communication projections [176-178]. Any treatment strategy directly targeting the *ephrinB*s must discriminate between the two. Any treatment strategy circumventing the *ephrinBs* must exert its influence with consideration of imbalances that may be present between different brain regions.

5.4 Ethical Considerations

A number of ethical consideration must be addressed, both in regard to the nature of disease variants and the consequences of effective treatment.

For single-gene disorders like XLID, pathologic variants are likely pathologic under all conditions, with the exception of rare cases of heterozygous advantage [182,183]. For complex disorders, the pathogenicity of a variant is dependent on genetic background and environmental factors. It is convenient in lay terms to identify a variant, common or rare, that is important for a disease such as autism, and permanently and publicly tag that variant as a cause for disease. However, just as some Mendelian disorders are intertwined with cases of heterozygous advantage, it is just as possible that variants for complex disorders may have non-disease functions. Under non-risk genetic

or environmental backgrounds, it is possible that these "disease" variants may provide an advantage in intellectual or behavioral function. Alternatively, they may simply provide for more nuanced differences in intelligence, interest, and behavior that makes each individual person unique. Non-disease effects may in fact explain much of the difficulty in establishing disease association for common variants, and may even be partially influencing the gradual increase in autism incidence over time [131].

This has strong ethical implications, particularly as next-generation sequencing becomes more accessible. For any given individual, a particular gene may be identified as mutated. If that gene has been linked to a genetic disease, and even more seriously, if the variant in that gene has been linked to disease, it is an easy mental shortcut for a scientist, clinician, or patient to act as though disease risk is elevated. If proper consideration of genetic background and environment takes place, disease risk may in fact be negligible.

Much of this dissertation has been devoted to understanding *GRIP1/2* as autism susceptibility genes. It is just as plausible that *GRIP1/2* variants, and even the autism-associated variants, under different conditions can contribute to a wider pleiotropy of phenotype, from other neuropsychiatric disorders like schizophrenia or intellectual disability, to general personality differences. Under the right conditions, *GRIP1/2* variation could even make one susceptible to creative genius.

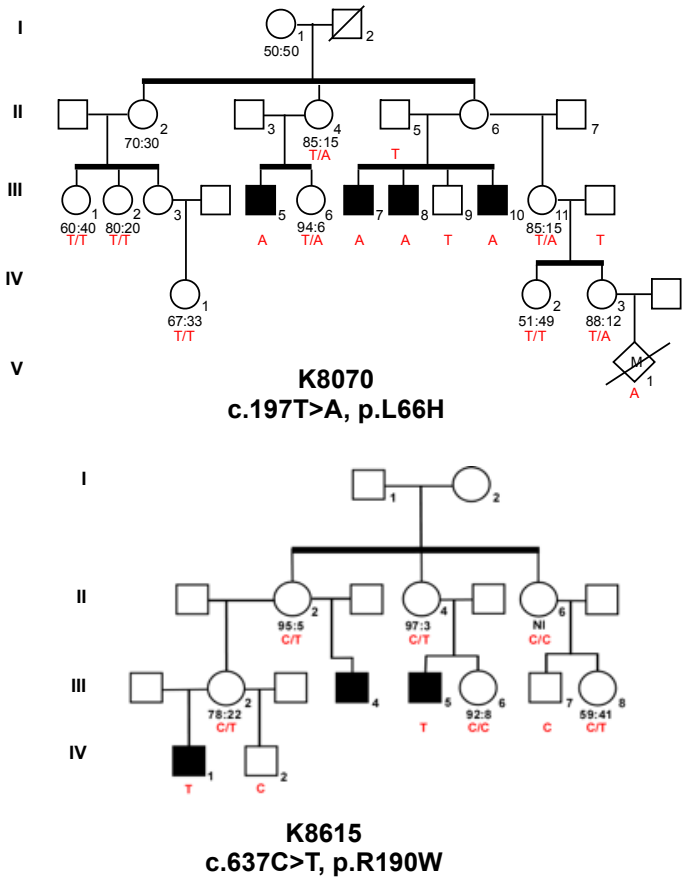
Any discussion of the contribution of genetic variation to complex disease must be delivered in such a way as to ensure that clinicians and patients are not taking unnecessary medical risks. Importantly, any discussion of the contribution of genetic

variation to psychiatric disease must be delivered in such a way as to not further stigmatize mental health.

A second ethical concern, if only a bit premature, is in regard to the consequences of successful treatment of genetic disorders, both Mendelian and complex. These disorders necessarily have a reproductive negative selection placed against them, thereby preventing the number of affected individuals from rising too high in following generations. Successful treatment of genetic disorders, should it become available, reduces the negative selection placed on deleterious/disease risk variants. These variants are much more likely to be passed on to the next generation. The next generation, with a higher burden of disease variants, will be more dependent upon treatment. From a commercial standpoint, this presents a moral hazard by which treatment of disease guarantees more patients to treat in the future. From a medical standpoint, this produces over the course of many generations a diminishing return on overall public health, which could have serious consequences for the population at large. As such, serious consideration should be given to identifying treatment routes that allow affected individuals to have full, healthy, reproductive lives, while somehow maintaining negative selection on true disease variants.

5.5 Figures: Chapter 5

Figure 5-1. *ZC4H2* mutations segregate with disease



Provided are pedigrees for two families, each with a proband relative identified as having a *ZC4H2* missense mutation. Full genotyping of the variant was subsequently conducted in all family members with available genomic DNA samples. In the first family (top), four affected males inherit the disease allele A from their female heterozygous carrier mothers. The same appears to apply to at least two affected males in the second family (bottom) for which genotyping was possible.

References

1. Mendel G (1866) Versuche über Pflanzen-Hybriden. Verh. Naturforsch 4: 3-47
2. López-Muñoz F, Boya J, Alamo C (2006) Neuron theory, the cornerstone of neuroscience, on the centenary of the Nobel Prize award to Santiago Ramón y Cajal. Brain Res Bull 70(4-6): 391-405
3. Garrod AE (1902) The incidence of Alkaptonuria: A Study in chemical individuality. Lancet 160(4137): 1616-1620
4. Hamosh A, Scott AF, Amberger JS, et al. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucl Acids Res 33(suppl 1): D514-D517
5. Stevenson RE, Schwartz CE (2009) X-linked intellectual disability: Unique vulnerability of the male genome. Dev Disabil Res Revs 15: 361-368
6. American Psychiatric Association (2000) Diagnostic and statistical manual of mental disorders: DSM-IV. Washington, DC: Author.
7. Shapiro SL, Sheppard, Jr. GL, Dreifuss FE, et al. (1966) X-linked recessive inheritance of a syndrome of mental retardation with hyperuricemia. Exp Biol Med (Maywood) 122: 609-611
8. Tsui LC, Buchwald M, Barker D, et al. (1985) Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. Science 230(4729): 1054-1057
9. Ng SB, Turner EH, Robertson PD, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276
10. Bao R, Huang L, Andrade J, et al. (2014) Review of current methods, applications,

and data management for the bioinformatics analysis of whole exome sequencing. Cancer Inform 13(suppl 2): 67-82

11. Jacquemont S, Coe BP, Hersch M, et al. (2014) A higher mutation burden in females supports a "female protective model" in neurodevelopmental disorders. Am J Hum Genet 94(3): 415-425

12. Lichtenstein P, Carlström E, Råstam M, et al. (2010) The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. Am J Psychiatry 167: 1357-1363

13. Hallmayer J, Cleveland S, Torres A, et al. (2011) Genetic heritability and shared environmental factors among twin pairs with autism. Arch Gen Psychiatry 68(11): 1095-1102

14. Gaugler T, Klei L, Sanders SJ, et al. (2014) Most genetic risk for autism resides with common variation. Nat Genet 46: 881-885

15. Kendler KS, Diehl SR (1993) The genetics of schizophrenia: a current, genetic-epidemiologic perspective. Schizophr Bull 19(2): 261-285

16. Cannon TD, Kaprio J, Lönnqvist J, et al. (1998) The genetic epidemiology of schizophrenia in a Finnish twin cohort. Arch Gen Psychiatry 55(1): 67-74

17. Kendler KS, Gatz M, Gardner CO, et al. (2006) A Swedish nation twin study of lifetime major depression. Am J Psychiatry 163: 109-114

18. Smoller JW, Finn CT (2003) Family, twin, and adoption studies of bipolar disorder. Am J Med Genet 123C: 48-58

19. Edvardsen J, Torgersen S, Røysamb E, et al. (2007) Heritability of bipolar

spectrum disorders: unity or heterogeneity? J Affect Disord 106(3): 229-240

20. McDuffie A, Thurman AJ, Hagerman RJ, et al. (2014) Symptoms of autism in males with Fragile X Syndrome: A comparison to nonsyndromic ASD using current ADI-R scores. J Autism Dev Disord doi: 10.1007/s10803-013- 2013-6 [Epub ahead of print]

21. Tsai LY (1992) Is Rett syndrome a subtype of pervasive developmental disorders? J Autism Dev Disord 22(4): 551-561

22. Verkerk AJ, Pieretti M, Sutcliffe JS, et al. (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in Fragile X Syndrome. Cell 65(5): 905-914

23. Laing S, Partington M, Robinson H, et al. (1991) Clinical screening score for the fragile X (Martin-Bell) syndrome. Am J Med Genet 38: 256–259

24. Darnell JC, Van Driesche SJ, Zhang C, et al. (2011) FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell 146(2): 247-261

25. Harrison JE, Bolton PF (1997) Annotation: Tuberous Sclerosis. J Child Psychol Psychiatry 38: 603–614

26. Amir RE, Van den Veyver IB, Wan M, et al. (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet 23: 185-188

27. Couvert P, Bienvenu T, Aquaviva C, et al. (2001) MECP2 is highly mutated in X-linked mental retardation. Hum Mol Genet 10(9): 941-946

28. Nagarajan RP, Hogart AR, Gwye Y, et al. (2006) Reduced MeCP2 expression is frequent in autism frontal cortex and correlates with aberrant MECP2 promoter methylation. Epigenetics 1(4):e1-11

29. Watson P, Black G, Ramsden S, et al. (2001) Angelman syndrome phenotype associated with mutations in MECP2, a gene encoding a methyl CpG binding protein. J Med Genet 38:224-228

30. Kim S, Chahrour M, Ben-Shachar S, et al. (2013) Ube3a/E6AP is involved in a subset of MECP2 functions. Biochem Biophys Res Commun 437(1): 67-73

31. Sanders SJ, Ercan-Sencicek AG, Hus V, et al. (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

32. McCarthy SE, Makarov V, Kirov G, et al. (2009) Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet 41: 1223-1227

33. Iyenger SK, Elston RC (2007) The genetic basis of complex traits: rare variants or "common gene, common disease"? Methods Mol Bio 376: 71-84

34. Duerr RH, Taylor KD, Brant SR, et al. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314(5804): 1461-1463

35. Sillén A, Andrade J, Lilius L, et al. (2008) Expanded high-resolution genetic study of 109 Swedish families with Alzheimer's disease. Eur J Hum Genet 16(2): 202-208

36. Anney R, Klei L, Pinto D, et al. (2012) Individual common variants exert weak

effects on the risk for autism spectrum disorders. Hum Mol Genet 21(21): 4781-4792

37. Panoutsopoulou K, Tachmazidou I, Zeggini E (2013) In search of low-frequency and rare variants affecting complex traits. Hum Mol Genet 22(R1): R16-21

38. Levy D, Ronemus M, Yamrom B, et al. (2011) Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. Neuron 70(5): 886-897

39. Sanders SJ, Murtha MT, Gupta AR, et al. (2012) *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485(7397): 237-241

40. Iossifov I, O'Roak BJ, Sanders SJ, et al. (2014) The contribution of *de novo* coding mutations to autism spectrum disorder. Nature 515(7526): 216-221

41. Griffiths AJF, Miller JH, Suzuki DT, et al. (2000) An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman and Company.

42. Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. PLoS Comput Biol 8(12): e1002822

43. Hindorff LA, Sethupathy P, Junkins HA, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci 106(23): 9362-9327

44. Wang K, Zhang H, Ma D, et al. (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 459(7246): 528-533

45. Weiss LA, Arking DE; Gene Discovery Project of Johns Hopkins & the Autism Consortium, Daly MJ, Chakravarti A (2009) A genome-wide linkage and

association scan reveals novel loci for autism. Nature 461(7265): 802-808

46. Correia C, Oliveira G, Vicente AM (2014) Protein interaction networks reveal novel autism risk genes within GWAS statistical noise. PLoS One 9(11): e112399

47. Cukier HN, Dueker ND, Slifer SH, et al. (2014) Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders. Mol Autism 5(1): 1

48. Butler MG, Rafi SK, Hossain W, et al. (2015) Whole exome sequencing in females with autism implicates novel and candidate genes. Int J Mol Sci 16(1): 1312-1335

49. Buermans HP, den Dunnen JT (2014) Next generation sequencing technology: Advances and applications. Biochim Biophys Acta 1842(10): 1932-1941

50. Bentley DR, Balasubramanian S, Swerdlow HP, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218): 53-59

51. Niranjan TS, Adamczyk A, Bravo HC, et al. (2011) Effective detection of rare variants in pooled DNA samples using Cross-pool tailcurve analysis. Genome Biol 12(9): R93

52. Bashiardes S, Veile R, Helms C, et al. (2005) Direct genomic selection. Nat Methods 2(1): 63-69

53. Kiialainen A, Karlberg O, Ahlford A, et al. (2011) Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery. PLoS One 6(2): e16486

54. Chilamakuri CS, Lorenz S, Madoui MA, et al. (2014) Performance comparison of

four exome capture systems for deep sequencing. BMC Genomics 15: 449

55. Samorodnitsky E, Datta J, Jewell BM, et al. (2015) Comparison of custom capture for targeted next-generation DNA sequencing. J Mol Diagn 17(1): 64-75

56. Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11(5): 473-483

57. Liu X, Han S, Wang Z, et al. (2013) Variant callers for next-generation sequencing data: a comparison study. PLoS One 8(9): e75619

58. McKenna A, Hanna M, Banks E, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297-1303

59. Kennedy B, Kronenberg Z, Hu H, et al. (2014) Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data. Curr Protoc Hum Genet 81: 6.14.1–6.14.25

60. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4: 1073-1081

61. Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 76: 7.20:7.20.1–7.20.41

62. Purcell S, Neale B, Todd-Brown K, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3): 559-575

63. Sham PC, Purcell SM (2014) Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet 15(5):335-346

64. Hu H, Roach JC, Coon H, et al. (2014) A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat Biotechnol 32(7): 663-669

65. Maragh SD (2014) RBM24 is essential for normal early embryonic development. Diss. Johns Hopkins University, Baltimore, MD

66. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915): 520-562

67. Bućan M, Abel T (2002) The mouse: genetics meets behaviour. Nat Rev Genet 3(2): 114-123

68. Niranjan TS, Skinner C, May M, et al. (2015) Affected kindred analysis of human X chromosome exomes to identify novel X-linked intellectual disability genes. PLoS One 10(2): e0116454

69. Lubs HA, Stevenson RE, Schwartz CE (2012) Fragile X and X-linked intellectual disability: four decades of discovery. Am J Hum Genet 90: 579-590

70. Gecz J, Shoubridge C, Corbett M (2009) The genetic landscape of intellectual disability arising from chromosome X. Trends Genet 25: 308-316

71. Ropers HH (2010) Genetics of early onset cognitive impairment. Annu Rev Genomics Hum Genet 11: 161-187

72. Raychaudhuri S (2011) Mapping rare and common causal alleles for complex

human diseases. Cell 147: 57-69

73. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46

74. Mamanova L, Coffey AJ, Scott CE, et al. (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7: 111-118

75. Bamshad MJ, Ng SB, Bigham AW, et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12: 745-755

76. Biesecker LG (2010) Exome sequencing makes medical genomics a reality. Nat Genet 42: 13-14

77. Gilissen C, Hoischen A, Brunner HG, et al. (2011) Unlocking Mendelian disease using exome sequencing. Genome Biol 12(9): 228

78. Gonzaga-Jauregui C, Lupski JR, Gibbs RA (2012) Human genome sequencing in health and disease. Annu Rev Med 63: 35-61

79. Kleefstra T, Kramer JM, Neveling K, et al. (2012) Disruption of an EHMT1-associated chromatin-modification module causes intellectual disability. Am J Hum Genet 91: 73-82

80. Krawitz PM, Murakami Y, Hecht J, et al. (2012) Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. Am J Hum Genet 91: 146-151

81. Ng SB, Buckingham KJ, Lee C, et al. (2010) Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet 42: 30-35

82. Ng SB, Bigham AW, Buckingham KJ, et al. (2010) Exome sequencing identifies

MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42: 790-793

83. O'Sullivan J, Bitu CC, Daly SB, et al. (2011) Whole-exome sequencing identifies FAM20A mutations as a cause of amelogenesis imperfecta and gingival hyperplasia syndrome. Am J Hum Genet 88: 616-620

84. Caramins M, Colebatch JG, Bainbridge MN, et al. (2013) Exome sequencing identification of a GJB1 missense mutation in a kindred with X-linked spinocerebellar ataxia (SCA-X1). Hum Mol Genet 22(21): 4329-4338

85. Zhang SQ, Jiang T, Li M, et al. (2012) Exome sequencing identifies MVK mutations in disseminated superficial actinic porokeratosis. Nat Genet 44: 1156-1160

86. Heinzen EL, Depondt C, Cavalleri GL, et al. (2012) Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. Am J Hum Genet 91: 293-302

87. Wang DG, Fan J, Siao C, et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280: 1077-1082

88. Bainbridge MN, Wang M, Wu Y, et al. (2011) Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher densities. Genome Biol 12: R68

89. Sherry ST, Ward M-H, Kholodov M, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311

90. Genomes Project C, Abecasis GR, Auton A, et al. (2012) An integrated map of

genetic variation from 1,092 human genomes. Nature 491: 56-65

91. Piton A, Redin C, Mandel JL (2013) XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. Am J Hum Genet 93(2): 368-383

92. Hirata H, Nanda I, van Riesen A, et al. (2013) ZC4H2 mutations are associated with arthrogryposis multiplex congenita and intellectual disability through impairment of central and peripheral synaptic plasticity. Am J Hum Genet 92: 681-695

93. Simmons AD, Püschel AW, McPherson JD, et al. (1998) Molecular cloning and mapping of human Semaphorin F from the Cri-du-chat candidate Interval. Biochem Biophy Res Comm 242: 685-691

94. Athanasakis E, Licastro D, Faletra F, et al. (2013) Next generation sequencing in nonsyndromic intellectual disability: From a negative molecular karyotype to a possible causative mutation detection. Am J Med Genet Part A 164A: 170-176

95. Ye B, Liao D, Zhang X, et al. (2000) GRASP-1: A neuronal RasGEF associated with the AMPA Receptor/GRIP Complex. Cell 26: 603-617

96. Hamdan FF, Gauthier J, Araki Y, et al. (2011) Excess of *de novo* deleterious mutations in genes associated with glutamatergic systems in nonsyndromic intellectual disability. Am J Hum Genet 88: 306-316

97. Muddashetty RS, Kelic S, Gross C, et al. (2007) Dysregulated metabotropic glutamate receptor-dependent translation of AMPA receptor and postsynaptic density-95 mRNAs at synapses in a mouse model of fragile X syndrome. J

Neurosci 27(20): 5338-5348

98. Verheij C, Bakker CE, de Graaff E, et al. (1993) Characterization and localization of the FMR-1 gene product associated with fragile X syndrome. Nature 363: 722-724

99. Edens AC, Lyons MJ, Duron RM, et al. (2011) Autism in two females with duplications involving Xp11.22-p11.23. Dev Med Child Neurol 53: 463-466

100. Chung BH, Drmic I, Marshall CR, et al. (2011) Phenotypic spectrum associated with duplication of Xp11.22-p11.23 includes Autism Spectrum Disorder. Eur J Med Genet 54: e516-520

101. Bush JO, Soriano P (2009) Ephrin-B1 regulates axon guidance by reverse signaling through a PDZ-dependent mechanism. Genes Dev 23: 1586-1599

102. Twigg SRF, Babbs C, van den Elzen MEP, et al. (2013) Cellular interference in craniofrontonasal syndrome: males mosaic for mutations in the X-linked EFNB1 gene are more severely affected than true hemizygotes. Hum Molec Genet 22: 1654-1662

103. Arvanitis DN, Behar A, Drougard A, et al. (2014) Cortical abnormalities and non-spatial learning deficits in a mouse model of craniofrontonasal syndrome. PLoS ONE 9(2): e88325

104. Vella P, Scelfo A, Jammula S, et al. (2013) TET proteins connect the O-linked N-acetylglucosamine transferase OGT to chromatin in embryonic stem cells. Mol Cell 49(4): 645-656

105. Huang L, Jolly LA, Willis-Owen S, et al. (2012) A noncoding, regulator mutation

implicates HCFC1 in nonsyndromic intellectual disability. Am J Hum Genet 91(4): 694-702

106. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359

107. DePristo MA, Banks E, Poplin R, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498

108. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164

109. The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52-58

110. Druley TE, Vallania FL, Wegner DJ, et al. (2009) Quantification of rare allelic variants from pooled genomic DNA. Nat Methods 6: 263-265

111. Out AA, van Minderhout IJ, Goeman JJ, et al. (2009) Deep sequencing to reveal new variants in pooled DNA samples. Hum Mutat 30: 1703-1712

112. Calvo SE, Tucker EJ, Compton AG, et al. (2010) High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. Nat Genet 42: 851-858

113. McClellan J, King M-C (2010) Genetic heterogeneity in human disease. Cell 141: 210-217

114. Hamady M, Walker JJ, Harris JK, et al. (2008) Error-correcting barcoded primers

for pyrosequencing hundreds of samples in multiplex. Nat Methods 5: 235-237

115. Craig DW, Pearson JV, Szelinger S, et al. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods 5: 887-893

116. Bravo HC, Irizarry RA (2009) Model-based quality assessment and base-calling for second-generation sequencing data. Biometrics 66: 665-674

117. Thomas AJ, Molla MN, Muzny DM, et al. (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4: 903-905

118. Bainbridge MN, Wang M, Burgess DL, et al. (2010) Whole exome capture in solution with 3 Gbp of data. Genome Biol 11: R62

119. Okou DT, Steinberg KM, Middle C, et al. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4: 907-909

120. Porreca GJ, Zhang K, Li JB, et al. (2007) Multiplex amplification of large sets of human exons. Nat Methods 4: 931-936

121. Markoulatos P, Siafakas N, Moncany M (2002) Multiplex polymerase chain reaction: a practical approach. J Clin Lab Anal 16: 47-51

122. Quail MA, Kozarewa I, Smith F, et al. (2008) A large genome center's improvements to the Illumina sequencing system. Nat Methods 5: 1005-1010

123. Langmead B, Trapnell C, Pop M, et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25

124. Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence Alignment/Map (SAM) format and SAMtools. Bioinformatics 25: 2078-2079

125. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and

validation of cluster analysis. J Comput Appl Math 20: 53-65

126. Bansal V (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics 26: i318-i324

127. Vallania FLM, Druley TE, Ramos E, Wang J, et al. (2010) High-throughput discovery of rare insertions and deletions in large cohorts. Genome Res 20: 1711-1718

128. Rivas M, Daly M: Syzygy Documentation Release 0.9. [http://www.broadinstitute.org/software/syzygy/sites/default/files/Syzygy.pdf]

129. Mejias R, Adamczyk A, Anggono V, Niranjan T, et al. (2011) Gain-of-function glutamate receptor interacting protein 1 variants alter GluA2 recycling and surface distribution in patients with autism. Proc Natl Acad Sci USA 108: 4920-4925

130. Niranjan T, Bravo HC (2011) Download Links for Srfim and SERVIC^{4}E scripts. [http://www.cbcb.umd.edu/~hcorrada/secgen]

131. Autism and Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators, Centers for Disease Control and Prevention (2014) Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010. MMWR Surveill Summ 63(2): 1-21

132. Persico AM, Sacco R (2014) Endophenotypes in Autism Spectrum Disorders. Comprehensive Guide to Autism, Springer New York 2014: 77-95

133. Li X, Zou H, Brown T (2012) Genes associated with autism spectrum disorder. Brain Res Bull 88(6): 543-552

134. Carlsson ML (1998) Hypothesis: is infantile autism a hypoglutamatergic disorder? Relevance of glutamate - serotonin interactions for pharmacotherapy. J Neural Transm 105(4-5): 525-535

135. Cline H (2005) Synaptogenesis: a balancing act between excitation and inhibition. Curr Biol 15(6):R203-205

136. Rubenstein JL, Merzenich MM (2003) Model of autism: increased ratio of excitation/inhibition in key neural systems. Genes Brain Behav 2(5):255-267

137. Nilsson M, Waters S, Waters N, et al. (2001) A behavioural pattern analysis of hypoglutamatergic mice--effects of four different antipsychotic agents. J Neural Transm 108(10):1181-1196

138. Purcell AE, Jeon OH, Zimmerman AW, et al. (2001) Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. Neurology 57(9):1618-1628

139. Jamain S, Betancur C, Quach H, et al. (2002) Linkage and association of the glutamate receptor 6 gene with autism. Mol Psychiatry 7(3):302-310

140. Shuang M, Liu J, Jia MX, et al. (2004) Family-based association study between autism and glutamate receptor 6 gene in Chinese Han trios. Am J Med Genet B Neuropsychiatr Genet 131B(1):48-50

141. Serajee FJ, Zhong H, Nabi R, Huq AH (2003) The metabotropic glutamate receptor 8 gene at 7q31: partial duplication and possible association with autism. J Med Genet 40(4):e42

142. Fatemi SH, Halt AR, Stary JM, et al. (2002) Glutamic acid decarboxylase 65 and

67 kDa proteins are reduced in autistic parietal and cerebellar cortices. Biol Psychiatry 52(8):805-810

143. Yip J, Soghomonian JJ, Blatt GJ (2007) Decreased GAD67 mRNA levels in cerebellar Purkinje cells in autism: pathophysiological implications. Acta Neuropathol 113(5):559-568

144. Yip J1, Soghomonian JJ, Blatt GJ (2008) Increased GAD67 mRNA expression in cerebellar interneurons in autism: implications for Purkinje cell dysfunction. J Neurosci Res 86(3):525-530

145. Collins AL, Ma D, Whitehead PL, et al. (2006) Investigation of autism and GABA receptor subunit genes in multiple ethnic groups. Neurogenetics 7(3):167-174

146. Fatemi SH, Reutiman TJ, Folsom TD, Thuras PD (2009) GABA(A) receptor downregulation in brains of subjects with autism. J Autism Dev Disord 39(2):223-230

147. Durand CM, Betancur C, Boeckers TM, et al. (2007) Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. Nat Genet 39(1):25-27

148. Jamain S, Quach H, Betancur C, et al. (2003) Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. Nat Genet 34(1):27-29

149. Chubykin AA, Atasoy D, Etherton MR, et al. (2007) Activity-dependent validation of excitatory versus inhibitory synapses by neuroligin-1 versus neuroligin-2. Neuron 54(6):919-931

150. International Molecular Genetic Study of Autism Consortium (1998) A Full Genome Screen for autism with Evidence for Linkage to a Region on Chromosome 7q. Hum Mol Genet 7(3): 571-578

151. Lauritsen MB, Als TD, Dahl HA, et al. (2006) A genome-wide search for alleles and haplotypes associated with autism and related pervasive developmental disorders on the Faroe Islands. Mol Psychiatry 11(1): 37-46

152. Tansey KE, Hill MJ, Cochrane LE, et al. (2011) Functionality of promoter microsatellites of arginine vasopressin receptor 1A (AVPR1A): implications for autism. Mol Autism 2(1): 3

153. Ma DQ, Cuccaro ML, Jaworski JM, et al. (2007) Dissecting the locus heterogeneity of autism: significant linkage to chromosome 12q14. Mol Psychiatry 12(4): 376-384

154. Shao Y, Wolpert CM, Raiford KL, et al. (2002) Genomic screen and follow-up analysis for autistic disorder. Am J Med Genet 114(1): 99-105

155. Carayol J, Sacco R, Tores F, et al. (2011) Converging evidence for an association of ATP2B2 allelic variants with autism in male subjects. Biol Psychiatry 70(9): 880-887

156. LoParo D, Waldman ID (2014) The oxytocin receptor gene (OXTR) is associated with autism spectrum disorder: a meta-analysis. Mol Psychiatry [Epub ahead of print]

157. Dong H, O'Brien RJ, Fung ET, et al. (1997) GRIP: a synaptic PDZ domain-containing protein that interacts with AMPA receptors. Nature 386(6622): 279-284

158. Dong H, Zhang P, Song I, et al. (1999) Characterization of the glutamate receptor-interacting proteins GRIP1 and GRIP2. J Neurosci 19(16): 6930-6941

159. Wyszynski M, Kim E, Yang FC, Sheng M (1998) Biochemical and immunocytochemical characterization of GRIP, a putative AMPA receptor anchoring protein, in rat brain. Neuropharmacology 37(10-11): 1335-1344

160. Srivastava S, Osten P, Vilim FS, et al. (1998) Novel anchorage of GluR2/3 to the postsynaptic density by the AMPA receptor-binding protein ABP. Neuron 21(3): 581-591

161. Wyszynski M, Valtschanoff JG, Naisbitt S, et al. (1999) Association of AMPA receptors with a subset of glutamate receptor-interacting protein *in vivo*. J Neurosci 19(15): 6528-6537

162. Burette A, Khatri L, Wyszynski M, et al. (2001) Differential cellular and subcellular localization of ampa receptor-binding protein and glutamate receptor-interacting protein. J Neurosci 21(2): 495-503

163. Masgrau R, Servitja JM, Young KW, et al. Characterization of the metabotropic glutamate receptors mediating phospholipase C activation and calcium release in cerebellar granule cells: calcium-dependence of the phospholipase C response. Eur J Neurosci 13(2): 248-256

164. Matsuda S, Mikawa S, Hirai H (1999) Phosphorylation of serine-880 in GluR2 by protein kinase C prevents its C terminus from binding with glutamate receptor-interacting protein. J Neurochem 73(4): 1765-1768

165. Perez JL, Khatri L, Chang C, et al. (2001) PICK1 targets activated protein kinase

Calpha to AMPA receptor clusters in spines of hippocampal neurons and reduces surface levels of the AMPA-type glutamate receptor subunit 2. J Neurosci 21(15): 5417-5428

166. Braithwaite SP, Xia H, Malenka RC (2002) Differential roles for NSF and GRIP/ABP in AMPA receptor cycling. Proc Natl Acad Sci USA 99(10): 7096-7101

167. Mao L, Takamiya K, Thomas G, et al. (2010) GRIP1 and 2 regulate activity-dependent AMPA receptor recycling via exocyst complex interactions. Proc Natl Acad Sci USA 107(44): 19038-19043

168. Wyszynski M, Kim E, Dunah AW, et al. (2002) Interaction between GRIP and liprin-alpha/SYD2 is required for AMPA receptor targeting. Neuron 34(1): 39-52

169. Ko J, Kim S, Valtschanoff JG, et al. (2003) Interaction between liprin-alpha and GIT1 is required for AMPA receptor targeting. J Neurosci 23(5): 1667-1677

170. Dickinson BA, Jo J, Seok H, et al. (2009) A novel mechanism of hippocampal LTD involving muscarinic receptor-triggered interactions between AMPARs, GRIP and liprin-alpha. Mol Brain 2: 18

171. Hoogenraad CC, Milstein AD, Ethell IM, et al. GRIP1 controls dendrite morphogenesis by regulating EphB receptor trafficking. Nat Neurosci 8(7): 906-915

172. Brückner K, Pablo Labrador J, Scheiffele P, et al. (1999) EphrinB ligands recruit GRIP family PDZ adaptor proteins into raft membrane microdomains. Neuron 22(3): 511-524

173. Essmann CL, Martinez E, Geiger JC, et al. (2008) Serine phosphorylation of ephrinB2 regulates trafficking of synaptic AMPA receptors. Nat Neurosci 11(9): 1035-1043

174. Takamiya K, Mao L, Huganir RL, Linden DJ (2008) The glutamate receptor-interacting protein family of GluR2-binding proteins is required for long-term synaptic depression expression in cerebellar Purkinje cells. J Neurosci 28(22): 5752-5755

175. Pellow S, Chopin P, File SE, Briley M (1985) Validation of open:closed arm entries in an elevated plus-maze as a measure of anxiety in the rat. J Neurosci Methods 14(3): 149-167

176. Stuckmann I, Weigmann A, Shevchenko A, et al. (2001) Ephrin B1 is expressed on neuroepithelial cells in correlation with neocortical neurogenesis. J Neurosci 21(8): 2726-2737

177. Yue Y, Widmer DA, Halladay AK, et al. (1999) Specification of distinct dopaminergic neural pathways: roles of the Eph family receptor EphB1 and ligand ephrin-B2. J Neurosci 19(6): 2090-2101

178. St John JA, Key B (2001) EphB2 and two of its ligands have dynamic protein expression patterns in the developing olfactory system. Brain Res Dev Brain Res 126(1): 43-56

179. Clontech Laboratories, Inc. (2009) Yeast Protocols Handbook. Protocol No. PT3024-1, Version No. PR973283

180. Takamiya K, Kostourou V, Adams S, et al. (2004) A direct functional link between

the multi-PDZ domain protein GRIP1 and the Fraser syndrome protein Fras1. Nat Genet 36(2): 172-177

181. Jacob F (1977) Evolution and tinkering. Science 196(4295): 1161-1166

182. Poolman EM, Galvani AP (2007) Evaluating candidate agents of selective pressure for cystic fibrosis. J R Soc Interface 4(12): 91-98

183. Chen ZJ (2010) Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci 15(2): 57-71

Tejasvi S Niranjan

3201 Saint Paul St., Unit 222 Baltimore, Maryland 21218
Phone: 862.432.8653 E-Mail: Tejasvi.Niranjan@jhu.edu

Education

- *PhD* (expected completion: May 2015)
  Molecular Biology and Human Genetics

  The Johns Hopkins University School of Medicine
  McKusick-Nathans Institute of Genetic Medicine
  Pre-Doctoral Trainee in Human Genetics
  Laboratory of Tao Wang, MD PhD

- Dissertation: Gene Discovery and Glutamate Signaling Defect in Intellectual Disability and Autism

- *BS* in Cellular and Molecular Biology

  The Johns Hopkins University
  Krieger School of Arts and Sciences
  Minor in Writing

Experience

- Pre-Doctoral Trainee in Human Genetics — 2009 – present
  - Laboratory of Tao Wang, MD PhD
  - The Johns Hopkins University School of Medicine
  - Gene discovery for psychiatric disorders: high-throughput sequencing with rigorous molecular functional analysis to identify disease-causing mutations
    - X-Linked Intellectual Disability: pre-prediction functional variant enrichment [Publication 1], with molecular follow-up
    - Autism: assessment of functional mutation burden in the synaptic genes GRIP2 [Publication 2] and GRIP1 [Publication 3], as well as statistical analysis of global synaptic gene mutation burdens in a large cohort (n = 800) of discordant Autism siblings
- Teaching Assistant
  - Introduction to Genetics: Hoyt & Cunningham. Dept. of Biology, JHU — 2012
  - Evolution of the Concept of the Gene: Valle, Smith, & Kazazian. IGM, JHU — 2013
- Volunteer Mentor and Team Lead at Incentive Mentoring Program — 2008 - present
  - Engaging at-risk and underperforming Baltimore high school students to break socioeconomic barriers to academic and personal advancement
  - Serving as a social and academic mentor
  - Serving as a Team Lead for post-high school (post-graduate) projects, including college/technical school application, standardized test preparation, and professional development

Experience (*continued*)

- Co-founder and Graduate Mentor 2007 - 2008
  - JHU iGEM Team (International Genetically Engineered Machines)
  - Over one hundred international bioscience teams participate in the iGEM competition at MIT to present their work in developing a genetically engineered machine (organism) for scientific and artistic advancement.
- Research Assistant 2006 - 2008
  - Laboratory of Jef Boeke, PhD
  - The Johns Hopkins University School of Medicine
  - Transposon Insertion Profiling ChIP (TIP-ChIP) Project
    - Development of microarray technology for genome-wide mapping of human LINE-1 retrotransposons
    - May lead to discovery of human phenotype variation caused by LINE1 insertions [Publication 4]
  - Synthetic Yeast Project
    - Pilot development to build from scratch, an entire synthetic version of the yeast genome (*Saccharomyces cerevisiae*)
    - Will allow rapid re-engineering of the genome for functional studies (e.g. finding the most minimal eukaryotic genomic structures)

Skills

- Basic Programming: JAVA, Perl, Shell script (Bash, Awk, Sed, etc.), C/C++, Python
- Web Programming: HTML, JavaScript, CGI, AJAX, PHP
- Database Literacy: SQL/SQLite/MySQL, FileMaker
- Computational Biostatistics: R and JAVA
  - High-throughput sequencing analysis: target capture design, alignment, error-correction, variant discovery, variant filtering, and functional variant prediction
  - Microarray data analysis: probe design, standardization, intensity pattern recognition, and statistical analysis (expression profiling and arrayCGH)
- Bioinformatics:
  - High-throughput automated primer design, integrated index sequence searching, and restriction enzyme coverage optimization
  - Proficiency in BLAT, BLAST, multiple sequence alignment, and genome browser track building

Skills (*continued*)

- Molecular Biology: PCR & gel electrophoresis, Sanger sequencing, restriction mapping, vector design and construction, Western blotting, bacterial and yeast transformations, Yeast-2-Hybrid, protein co-immunoprecipitation, beta-galactosidase colorimetric assay
- Cellular Biology: rat/mouse brain dissection, neuronal transfection and electroporation, mammalian cell culturing (including neuronal culturing), immunohistochemistry, fluorescent confocal microscopy

Publications

1. Niranjan T, Skinner C, May M, *et al.* (2015) **Affected Kindred Analysis of Human X Chromosome Exomes to Identify Novel X-linked Intellectual Disability Genes**. *PloS One* 10(2): e0116454
2. Niranjan T, Adamczyk A, Bravo H, Taub M, *et al.* (2011) **Effective Detection of Rare Variants in Pooled DNA Samples Using Cross-Pool Tailcurve Analysis**. *Genome Biol* 12: R93
3. Mejías R, Adamczyk A, Anggono V, Niranjan T, *et al.* (2011) **Gain-of-function Glutamate Receptor Interacting Protein 1 Variants Alter GluA2 Recycling and Surface Distribution in Patients with Autism**. *PNAS* 108(12): 4920-4925
4. Huang C, Schneider A, Lu Y, Niranjan T, *et al.* (2010) **Mobile Interspersed Repeats are Major Structural Variants in the Human Genome**. *Cell* 141(7): 1171-1182

Posters and Presentations

1. Niranjan T, *et al.* (2013) **Disease Gene Discovery: Bridging the Gap from Mendelian Disorder to Complex Disease**. Presentation. *7th Annual Young Investigators Symposium on Genomics*, Johns Hopkins Center for Computation Genomics
2. Niranjan T, *et al.* (2013) **Affected Sib Analysis of Human X-Exome Data to Identify Novel X-linked Intellectual Disability Genes**. Poster. *ASHG*, Boston, MA
3. Niranjan T, *et al.* (2012) **Cumulative Mutation Load in PDZ Domains 4, 5, & 6 of Glutamate Receptor Interacting Protein 2 in Autism**. Poster. *ASHG*, San Francisco, CA
4. Niranjan T, *et al.* (2012) **Identification of Novel X-Linked Intellectual Disability Genes by Human X Chromosome Exome Sequence**. Poster. *ASHG*, San Francisco, CA
5. Niranjan T, *et al.* (2012) **Isolation of Low Frequency Variants from Sequencing Errors within Next-Generation Sequencing Data**. Presentation. *Computational Genomics Seminar Series*, Johns Hopkins Center for Computational Genomics

Posters and Presentations (*continued*)

6. Niranjan T, *et al.* (2011) **Amplicon Ligation of Pooled DNA Samples for Effective Detection of Rare Variants in a Large Cohort**. Poster. *ASHG*, Washington, DC 2010 & *Beyond the Genome*

Affiliations/Memberships

- American Society of Human Genetics 2010 – present
- Incentive Mentoring Program 2008 - present