GENETIC ANALYSIS OF COMPLEX AND MENDELIAN DISEASES

Nicole Gabrielle Griffin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum of Genetics and Molecular Biology.

Chapel Hill
2014

Approved by:

Yun Li

Karen Mohlke

Kari North

Fernando Pardo-Manuel de Villena

Kirk Wilhelmsen

ABSTRACT

Nicole Gabrielle Griffin: Genetic Analysis of Complex and Mendelian Diseases
(under the direction of Kirk Wilhelmsen)

This work describes approaches for discovering genetic variants that contribute to the etiology of human diseases with complex and simple modes of inheritance through the use of linkage analysis, genome-wide association analysis, and massively parallel sequencing (MPS). The studies contained in this work illustrate both the capabilities and limitations of these approaches.

The two GWA studies in this work illustrated how reducing genetic and population heterogeneity could increase the ability to detect associations with genome-wide significance. The first, a GWA study of idiopathic Parkinson's disease (IPD), was able to detect an association signal that approached genome-wide significance across chromosome 12q12, including the *LRRK2* locus (average p-value=4.85 x 10^{-6}), which has been implicated in IPD by several linkage studies. The second, a pilot GWA study of dystonia, identified an association with genome-wide significance at *RNF213*. The second half of this work employed MPS approaches to investigate the genetics of familial presentations of disease. The first, a study of a family with an atypical presentation of frontotemporal dementia with amyotrophic lateral sclerosis, was unable to detect an obvious deleterious mutation despite sequencing the exomes of 10 individuals and the whole genome of 1 individual in this family. The exome sequencing data from this family were used to perform a multipoint linkage analysis, which potentially implicated chromosome 9q in this family. In this region, all affected family members shared a synonymous mutation in *CRB2*, a

gene in the gamma-secretase pathway. The final study featured a genome-wide linkage analysis of a pedigree affected with a microcoria myopathy and a combined whole genome and whole exome sequencing analysis of this pedigree and 7 unrelated individuals. The linkage analysis found a multipoint LOD score of 1.8 on Chromosome 5q35. Exome sequencing detected a missense mutation shared by the affected family members in *C5orf60*: c.97C>T (p.P33S) that was also found in an exome from an unrelated subject. Another missense mutation in *C5orf60*, c.64G>C (p.D22H), was present in the exomes from 5 of the unrelated subjects. These results suggest that mutations in *C5orf60* are a novel cause of microcoria and also corroborate the genetic heterogeneity of this condition.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ALS	amyotrophic lateral sclerosis
APM	affected pedigree member
ARMD	age-related macular degeneration
BAM	binary sequence alignment/map
BWA	Burrows-Wheeler aligner
CGH	comparative genomic hybridization
CK	creatine kinase
cM	centiMorgan
CNV	copy number variant
ESE	exonic splicing enhancer
ESP	Exome Sequencing Project
ESS	exonic splicing silencer
EST	expressed sequence tag
EVS	Exome Variant Server
FTD	frontotemporal dementia
FTD-ALS	frontotemporal dementia with amyotrophic lateral sclerosis
GATK	genome analysis toolkit

gDNA	genomic DNA
GWA	genome-wide association
GWAS	genome-wide association study
IBD	identical by descent
IBS	identical by state
IPD	idiopathic Parkinson’s disease
LD	linkage disequilibrium
LGMD	limb-girdle muscular dystrophy
LOAD	late-onset Alzheimer’s disease
LOD	logarithm of odds
MAF	minor allele frequency
MPS	massively parallel sequencing
NGS	next-generation sequencing
NPL	non-parametric linkage
PA-seq	poly-adenylation sequencing
PCR	polymerase chain reaction
Q-Q	quantile-quantile
RF	recombination frequency

RFLP	restriction fragment length polymorphism
SAM	sequence alignment/map
SF-A	San Francisco-A
SNP	single nucleotide polymorphism
VCF	variant call file
VNTR	variable number of tandem repeats
WES	whole exome sequencing
WGS	whole genome sequencing

Chapter 1

Introduction

An important goal of human genetics is the identification of sequence variants that produce traits of medical importance. Genetic variants that cause traits of interest occur with a wide range of frequencies and include changes as small as a single nucleotide to as large as a chromosomal duplication. To map variants to chromosome locations within the three billion base pairs that make up the human genome, researchers first used cytology to examine metaphase chromosomes for gross rearrangements. In most conditions, there are no visible rearrangements, such that cytology has been superseded by molecular biology and genetic techniques. Currently, the methods of choice for searching the genome include chromosome segregation analysis, which tests whether specific chromosome regions harbor sequence variants that transmit disease in families, or association analysis, which tests whether having specific variants is correlated with having a trait. With the development of technology for massively parallel molecular techniques and automation, it has become routine to systematically interrogate the entire genome by linkage and association analysis.

1.1 Linkage Analysis

Two loci are said to be linked if the hypothesis that they segregate independently is disproven. This typically occurs when two loci are on the same chromosome and are sufficiently

close that meiotic recombination between them does not occur frequently enough to make appear to segregate independently. The premise for linkage analysis of traits is that affected individuals in a pedigree are expected to have an increased probability of inheriting the same chromosome segments from common ancestors for the regions that have variants that lead to disease. For rare traits with simple modes of inheritance, there is a high probability that all the affected individuals in a family will inherit the causal sequence variants from the same common ancestors. The affected individuals in such a family are assumed to be identical by descent (IBD) for the causal variant. Because chromosomes are inherited from generation to generation with occasional meiotic recombinations, disease-causing variants will co-segregate with other markers on the same chromosome segment. Given enough affected individuals and pedigrees, linkage analysis can identify the region of the genome segregating with the disease variant. A typical chromosome region identified by linkage analysis will contain thousands of rare sequence variants that are effectively unique to the chromosome segment shared by the affected family members. Prior to the development of massively parallel sequencing strategies, linkage was followed by focused sequencing of coding sequences in the linked segment to look for mutations predicted to change the function of a gene product. The coding sequence of genes was the focus of investigations because it is often difficult to predict the effect of non-coding sequence changes and there was the prejudice that most trait producing sequence variants would affect the amino acids sequence of gene protein products.

Initially, the criteria for deciding that sequence changes in a gene were responsible for a rare trait were: 1) that the putative causal variants are not found in a large sampling of unaffected individuals; 2) that the putative causal variants segregate with the trait in pedigrees; and 3) that independent variants could be detected in the same gene in pedigrees with the same condition.

Candidate gene resequencing has now been replaced by massively parallel sequencing such that the absence of other likely causal mutations in a linkage interval can be used to increase the likelihood that the gene responsible for a trait has been identified.

To perform a linkage analysis, the minimum amount of information needed is a sufficiently large pedigree or collection of pedigrees with phenotype data and a set of genotyped markers. The type of analysis is determined by the characteristics of the disease inheritance and pedigree structure. For simple, Mendelian diseases where the mode of inheritance is well-understood, parametric linkage analysis is usually pursued; in this type of analysis, the parameters describe the frequency of and mode of inheritance of disease. For diseases with complex modes of inheritance, so-called non-parametric linkage analysis is often used. The less parameterized or fitted parameter methods typically require large collections of families with at least an affected relative pair. For the microcoria myopathy study and frontotemporal dementia with amyotrophic lateral sclerosis (FTD-ALS) study described in this dissertation, parametric linkage analysis was used because the mode of inheritance could be inferred from the respective pedigrees.

1.1.1 Design Considerations in Genetic Linkage Studies

To test for co-segregation of a chromosome segment and a trait, it is necessary to be able to identify chromosomes that are identical by descent. In 1980, Botstein et al. (1) proposed that sequence variations specific to chromosome position, usually called markers, could be used to construct genetic maps and determine whether chromosome segments are identical by descent. Any sequence variation can potentially be used as a marker and a series of markers across a region can be used to increase the confidence that chromosome segments are IBD. Since the

concept of using DNA sequence variations as markers emerged, numerous assay methods have been developed. Botstein et al. (1) first defined the use of restriction fragment length polymorphisms (RFLP). RFLP-based approaches used restriction enzymes from bacteria to cut DNA into fragments through the recognition of a specific sequence. Sequence variations that create or destroy the recognition sequence for a restriction enzyme can be assayed by monitoring the cleavage pattern of DNA. Most RFLP polymorphisms are dimorphic and thus often cannot distinguish between chromosomes with the same allele. Collections of markers on a chromosome segment can overcome the limited ability of a single marker to distinguish between chromosomes. When the cost of genotyping a marker was high, markers with many alleles were preferentially used. The most commonly used highly polymorphic markers detected sequences with variable number tandem repeats (VNTRs). In most cases, VNTRs were assayed by measuring the size of a fragment between restriction enzyme cleavage sites or polymerase chain reaction primer binding sites. The development of array-based approaches allowed for the simultaneous genotyping of large collections of single nucleotide polymorphisms (SNPs) in a single assay. Currently, SNP-based approaches allow for the genotyping millions of markers in a single assay. The human population is estimated to have ten million SNPs where the less frequent allele has a frequency greater than 5% in a commonly studied population.

Prior to collecting genotype data for linkage analysis, it is prudent to estimate the power to detect linkage. Frequently for pedigrees with simple modes of inheritance where sufficient genotype data will be obtained to confidently determine whether family members are identical by descent, power can be estimated by counting the number of informative meioses in the pedigree. For families with complex modes of inheritance, it is usually necessary to simulate genotype data using what is referred to as a gene-dropping approach (2). In this approach, the genotypes are

randomly assigned to the individuals for whom no ancestral data is available- the founders. Random segregation of the founders' chromosomes is used to "drop," or assign, the genotypes of the remaining members of the pedigree. The genotypes of the individuals that will not be genotyped are then removed. When pedigrees have already been ascertained the observed phenotypes and pedigree structures are used. Otherwise phenotypes are simulated using the parameters deduced from the observed mode of inheritance of the trait. The key parameters that need to be estimated are the number of trait loci, their allele frequencies, mode of inheritance and the effect size. For traits with simple modes of inheritance it is often assumed that there is a single trait locus with estimated allele frequencies and genotype penetrances. The genotype penetrance is the estimated fraction of the time that an individual with a trait genotype will express the trait. Linkage analysis is performed for the simulated data. The process is repeated over and over to estimate the null distribution for a linkage signal. The distribution can be compared to the distribution obtained where genotypes are simulated assuming linkage. Following this assessment of power, linkage between markers and a trait can be evaluated by several different tests. The Elston-Stewart and Lander-Green algorithms are the two standard methods for calculating the genotype likelihoods in pedigrees. Briefly, the Elston-Stewart algorithm evaluates the likelihood of the pedigree data using the probability of the founder genotypes, the probability of the phenotype given the genotype (i.e., the penetrance), and the probability of a child's genotype given the parental genotypes (i.e., the transmission probability) (3). Because the offsprings' genotypes can be collapsed or "peeled" onto the parent, this algorithm is better suited to handling large pedigrees; i.e., the complexity increases linearly as with the size of the pedigree and exponentially with the number of loci. In contrast, the complexity of the Lander-Green algorithm increases exponentially with the number of pedigree

founders and linearly with the number of loci. The Lander-Green algorithm incorporates a Hidden Markov Model (HMM) to determine the likelihood of the pedigree data given the inheritance vectors for each marker, the genotype probabilities for each marker, and the transition probabilities from one marker to the next (4).

Tests for Genetic Linkage

Parametric linkage analysis

Parametric linkage analysis tests for segregation between a trait locus and a set of markers with known positions through the use of a specific trait model. Parametric models typically include the trait allele frequencies and genotype penetrances. Parametric linkage analysis can allow for age and gender specific penetrances that are either continuous or discontinuous functions. For traits with simple modes of inheritance, often called Mendelian diseases, the penetrance of disease associated genotypes are high and the penetrance of non-disease associated genotypes are very low. By making simple assumptions, causal allele frequencies can be estimated based on the frequency of the trait in the population. Rare traits are predicted to be caused by rare alleles. Penetrances are inferred from the fraction of at risk individual in a pedigree that manifest the trait given the inferred mode of inheritance and genotype frequencies.

Often by annotating the status (i.e., affected, carrier, unaffected, or unknown) of each individual in a pedigree, the particular mode of inheritance of a disease locus often becomes clear. For example, in an autosomal dominant mode of inheritance with complete penetrance, all individuals in a pedigree harboring one copy of the trait allele will express the trait; there will be male-to-male transmission, and approximately half of the offspring of an affected individual will

also be affected. Other modes of inheritance include autosomal recessive, X-linked, Y-linked, and mitochondrial. The observed mode of inheritance is used to develop the model used in parametric linkage analysis.

Once the trait model is developed, the probability that a known genotyped marker is segregating with disease can be calculated. The logarithm-of-odds (LOD) score compares the probability that the pattern of genotypes would be seen if the marker locus and trait locus were linked to the probability that the pattern of genotypes would be seen if the marker locus and trait locus were not linked.

Non-parametric linkage analysis

Non-parametric linkage (NPL) analyses were developed to map loci of traits with complex inheritance. NPL studies analyze the amount of IBD sharing among affected individuals in pedigrees. The simplest form of NPL analysis is the affected sib-pair test, which requires genotype information from nuclear families with two affected siblings. The Haseman-Elston regression performs this sib-pair test by regressing the square of number of alleles shared IBD at a given locus against the square of difference in the quantitative phenotypes of the sib-pair (5). Variance components analysis, another form of NPL analysis, allows for the inclusion of components that might influence a trait, such as age or gender, to isolate the additive and dominance genetic effects for relatives with any type of familial relationship (6).

1.1.2 Limitations of Genetic Linkage Studies

For linkage analysis to be successful there must be sufficiently dense coverage of the genome with genotyped markers and sufficiently large and accurate pedigree information. For

parametric linkage analysis the disease variant must have sufficient penetrance to deduce the mode of inheritance for parametric linkage analysis, and there must be a sufficient number of informative meioses to uniquely specify the region of the genome that is segregating with the variant.

Because the number of crossovers that typically occur on a single chromosome is small, the first major limitation of linkage analysis is that a linkage study typically does not provide the resolution necessary to determine the causal variant for a disease. By chance, it is possible for the critical recombination events to occur within close proximity, but most studies with the minimum power to detect linkage can narrow the genome to regions that are ten to twenty cM in size. Candidate gene resequencing was first used to search for causal variants after linkage was detected, but the advent of massively parallel sequencing technologies also allows for the identification of such variants.

The other primary limitation of linkage analysis is that this type of analysis can only resolve the genetics of certain diseases. Diseases with complex inheritance are less easily resolved. While linkage analyses succeeded in mapping the genes for thousands of diseases, new approaches were needed to map the genes for the majority of human traits, which have complex modes of inheritance.

1.2. Genome-wide Association Studies

With the ability to genotype hundreds of thousands of markers at once came a new kind of genetic analysis: genome-wide association (GWA) analysis. GWA studies focus on common genomic variation which entail the genotyping of many markers with minor allele frequencies (MAF) typically greater than 5%. The underlying hypothesis is that common variation can be

responsible for common diseases. While linkage studies seek regions of the genome that are IBD in affected individuals, GWA studies seek to find alleles that are IBS. A significant association signal indicates that an allele occurs more frequently in individuals with the disease (i.e., the cases) than in individuals in the general population (i.e., the controls). Because markers located within proximity of each other on the same chromosome segment are more likely to be inherited together, certain combinations of alleles occur more frequently than would be expected simply by their frequencies in the population. This phenomenon is known as linkage disequilibrium (LD). In turn, because of this LD, association testing can identify direct associations at a given SNP and indirect associations with SNPs that were not genotyped but that are in LD with the genotyped SNP. The genotyped SNP is said to "tag" the SNPs with which it is in LD. Furthermore, the effect sizes of the variants implicated by GWA and linkage studies differ substantially. Most of the SNPs implicated in human disease by GWA analysis have small effect sizes. GWA analysis can pinpoint thousands of markers with small effect sizes as contributing to a phenotype, which can give a more complete view of that phenotype's underlying biology. GWA studies have paved the way for the genetic analysis of common, complex diseases.

The first published GWA study investigated age-related macular degeneration (ARMD) (7). ARMD is one of the most common causes of blindness or vision impairment in the elderly. While linkage had previously implicated a locus on chromosome 1q32, these researchers used data for 116,204 SNPs from 96 cases and 50 controls to find a significant association signal in this region for two SNPs in the gene for complement factor H (*CFH*). Another early success in the field of GWA studies was the association of the E4 allele of apolipoprotein-E (*APOE*) with the risk for late onset Alzheimer's disease (LOAD). Prior to the advent of GWA studies, the Roses' lab first provided evidence of linkage of LOAD to chromosome 19, and a candidate gene

study implicated the E4 allele (8-9). Following these discoveries, Grupe et al. (10) published a GWA study of 1,808 LOAD cases and 2,062 controls, who were genotyped for 17,343 SNPs, and found significant association signals for SNPs near *APOE*. This association has been confirmed by the majority of subsequent GWA studies of LOAD (11).

Since these initial studies, shifts in the scale and scope of GWA studies have occurred. One of the first large-scale GWA studies was pursued by the Wellcome Trust Case Control Consortium (12) and involved 14,000 cases for 7 diseases and 3,000 shared controls. GWA studies can now entail tens or hundreds of thousands of individuals and approximately one million SNP markers. The largest GWA studies have focused on traits such as height and type 2 diabetes. There are several considerations in designing a study of this scope.

1.2.1. Design Considerations in Genome-wide Association Studies

Power

The first step in performing a GWA study involves the selection of the phenotype of interest and the collection of DNA samples from a population. The design of a GWA study will affect the power to detect associations. Increases in power lead to decreases in type II error. Power depends on the number of samples, the frequency of the trait in the population, the ratio of cases to controls, the effect size of the association, and the disequilibrium between typed polymorphisms and the causal polymorphism. Power is becoming less of an issue as consortia are formed to tackle the genetics of some of the most common diseases; these study populations are often on the order of tens or hundreds of thousands of subjects.

Population stratification

The subjects in a GWA study are often assumed to be unrelated. Failure to correct for kinship can inflate measures of association. It is possible to directly measure the relationship (kinship co-efficient) between study participants with a large collection of typed markers and correct measures of association for their relationship (13). However, it is also important that the subjects be derived from the same population. A more pernicious source of false association occurs when there is population substructure such that there is a correlation between phenotype and subpopulation membership. In the event that a study population is actually comprised of two or more sub-populations with significantly different minor allele frequencies and differences in disease frequency, a false positive association signal may arise.

The most commonly used method for detecting stratification is genomic control (14). The genomic control statistic, λ , indicates the presence of the potential inflation of the association signals by dividing the average of the (Chi-squared) test statistics for approximately 50 SNPs by the median of this distribution. Several methods have been developed for identifying the sub-populations within a study when stratification is present. These include structure-association based approaches, such as STRUCTURE (15), and principal component analysis-based methods, such as Eigensoft (16, 17).

Adjusting for multiple testing

While increasing the power of a study can lessen the likelihood of type II error, the potential for type I error is an equally pressing concern in GWA study design. Type II error (also called a false negative) results when the null hypothesis is incorrectly accepted, while type I error

(also called a false positive) results when the null hypothesis is incorrectly rejected. Each genotyped marker that is in equilibrium with other typed markers represents the testing of a different hypothesis. One statistical test used commonly in GWA studies is the Chi-squared test, which compares the allele counts or genotype frequencies in the cases and controls. For the purposes of the two GWA studies described in this dissertation, the Fisher's exact test was used to compare the distribution of the genotypes for each SNP between the cases and controls because this test accounts better than the Chi-squared test for low allele/genotype counts. Thus, the p-values obtained from GWA studies have to be adjusted to account for hundreds of thousands of hypotheses. The most conservative of these adjustments is the Bonferroni Correction, which tests each hypothesis at a significance level of α/n, where α is the probability of incorrectly rejecting the null hypothesis and n is the number of hypotheses, such that the type I error rate for the group of hypotheses as a whole is equal to α. The generally accepted p-value cutoff for genome-wide significance is 5×10^{-8} (18). The Bonferroni correction is overly conservative when markers are not independent of each other. To illustrate the point, imagine that a large gene has hundreds of polymorphisms, but only two haplotypes. By testing whether a single SNP in the gene is associated effectively tests whether any of the SNPs in the gene is associated with the trait. To correct the threshold value for significance by the number of equivalent tests is obviously overly conservative.

Even when adjusting for multiple comparisons, false positive results can still occur. The false discovery rate (FDR) describes the expected proportion of false positive results; Benjamini and Hochberg first described a method for controlling the FDR (19). One method for evaluating the appropriateness of adjustments for multiple comparisons is quantile-quantile (Q-Q) plotting.

A Q-Q plot compares the expected distribution of p-values under the null hypothesis with the observed distribution of p-values.

Imputation

It is estimated that there are ten million common (MAF > 0.01) SNPs in the human genome. Due to LD, it is possible to infer the genotypes of most common SNPs if the genotypes of approximately one million SNPs are known. Computational methods for imputing genotypes rely on haplotype phasing.

Individuals within populations can be shown to be related by increased allele sharing that reflects population histories. Homogeneous populations tend to have unique LD structures, facilitating imputation, and the number of possible haplotypes for a gene is generally much lower than the number of possible haplotypes present in the population.

When phasing haplotypes in a group of unrelated individuals, the inclusion of genotype information from more individuals results in a better estimation of the haplotypes (20). The majority of computational phasing approaches take into account the fact that most haplotypes in a population are similar to each other because of constraints on recombination and mutation over short distances; this observation forms the basis of approximate coalescent models and allows for hidden Markov model-based approaches to haplotype phasing. Once a set of haplotypes has been generated, these can be used to impute the genotypes for additional SNPs using the known genotypes and the knowledge of LD. PHASE (21), BEAGLE (22), MaCH (23), and IMPUTE2 (24) are commonly used phasing methods; the latter three can also be used for genotype imputation based on phased genotypes.

Thus, through the use of genome-wide SNP arrays, more densely typed reference populations and imputation, it is possible to capture a substantial amount of genetic variation through a GWA approach.

1.2.2. Limitations of Genome-wide Association Studies

GWA studies have identified thousands of genetic markers that are associated with human disease. However, these studies do not seem to explain the majority of the heritability of a phenotype. For example, in 2010, Lango et al. published a GWA study of height with 183,727 subjects of European descent and found significant associations for 180 loci (25). Because of its very large sample size, this study should have had sufficient power to detect associations. However, in the aggregate, these 180 loci explained only 10% of the variation in height. Because height is predicted to be 80-90% heritable, the findings of the study suggest that GWA studies are insensitive to much of the genetic variation that can explain a phenotype. This problem is known as the "case of the missing heritability."

Several explanations have been offered for this missing heritability. These explanations exemplify the limitations of GWA studies and are discussed below.

Genetic and clinical heterogeneity

Complex human diseases can be characterized by both genetic and clinical heterogeneity. Genetic heterogeneity can contribute to the problem of missing heritability because a given variant may only contribute to a phenotype in a portion of the cases. If this is the case, the association signal in the region containing this variant may not achieve genome-wide significance. Genetic heterogeneity can be divided into two types: locus and allelic

heterogeneity. The former occurs when variants in different genes result in the same phenotype. Locus heterogeneity is easier to define for loci with high genotypic relative risks, such as those that cause disease with simple modes of inheritance (e.g., Charcot-Marie-Tooth and spinocerebellar degeneration). There are many examples of apparently similar conditions that are due to mutations in different genes. Similarly often more than one mutation in a gene, i.e., allelic heterogeneity, can result in the same condition. For alleles and loci with smaller genotypic relative risks, the effects of genetic heterogeneity have a more nuanced effect on measured heritability.

The occurrence of either locus or allelic heterogeneity can obscure an association signal in a GWA study by diluting the allele frequencies of SNPs. Genetic heterogeneity can be reduced by selecting a study group from a homogeneous population.

In addition to genetic heterogeneity, clinical heterogeneity is problematic because while subjects might be classified as having the same disease, this classification may encompass a variety of distinct syndromes with similar symptoms. Thus, diseases that may be regarded as "common" may not be that common after all. Researchers have proposed focusing on intermediate phenotypes instead, such as using blood glucose levels as a proxy for diabetes, to lessen the possibility for phenotypic heterogeneity.

The limitations of SNPs

While SNPs are the most common type of polymorphism, many of the SNPs implicated by GWA studies do not cause a change in protein-coding sequence. Many SNPs associated with disease have been found outside of protein-coding sequences in introns or in intergenic regions, where their effects are unclear. It is also possible that an associated SNP can affect a gene that is

thousands of base pairs away. Furthermore, there are many types of variants beyond SNPs that are not tagged well in GWA studies.

For example, researchers have also investigated the role of copy number variants (CNVs) in common disease. CNVs can constitute small or large insertions or deletions, as well as duplications, of DNA. While CNVs can be shared by individuals in a population, they can also occur as *de novo* mutations. Through the application of comparative genomic hybridization (CGH), researchers have identified *de novo* CNVs as playing a major role in autism spectrum disorders (26).

Thus, the "missing heritability" may be explained in part by the inadequacy of SNPs in representing genetic variation throughout the genome.

Gene-gene interactions

In the case of complex diseases for which many genes have been implicated, the missing heritability may also arise from gene-gene interactions. That is, the etiology of a disease may be explained by a combination of alleles in different genes. However, it is computationally expensive and statistically burdensome to model all of the potential gene-gene interactions that could contribute to a phenotype. Moreover, these interactions are not captured by conventional GWA methods.

Gene-environment interactions

Gene-environment interactions may also contribute to the problem of the missing heterability. A gene-environment interaction arises when the contribution of a variant to a phenotype is influenced by an environmental exposure such that the combined effect of the

genotype and environmental exposure is greater (or less) than the sum of these factors' individual effects. One possible consequence of a gene-environment interaction is that a variant only contributes to the risk for a phenotype in the portion of the population that has been introduced to a certain exposure; this interaction could potentially obscure an association signal. For example, researchers have investigated the extent to which physical activity may interact with common variants in *FTO* contribute to type II diabetes (27, 28). As is the case for gene-gene interactions, it would be computationally difficult and statistically burdensome to model these interactions in the context of a GWA study.

Transgenerational epigenetic effects

While SNP arrays can capture variation at the level of DNA sequence, the regulation of gene expression depends on more than just the primary sequence of nucleotides. Within nuclei, DNA is packaged as chromatin, which is wrapped around proteins known as histones to form nucleosomes. Histone modifications, such as methylation and acetylation, can either activate or suppress the transcription of genes into mRNA. These modifications, which collectively form the epigenome, are overlooked by GWA analysis because they do not alter the primary DNA sequence and may explain part of the missing heritability if they are passed from one generation to the next and contribute to the variation in a phenotype.

1.2.3 Summary

GWA studies, in concert with the International HapMap project, have thoroughly explored the common disease-common variant hypothesis and have described a vast amount of the common variation in the human genome. The approach has also explained a substantial proportion of the heritability of common diseases. However, just as the ability to genotype

common variants caused a shift from linkage analysis to GWA analysis, the ability to rapidly sequence the entire genome is causing a shift from GWA studies to sequencing studies to explain the genetics of disease.

1.3. Massively Parallel Sequencing Studies

On April 14, 2003, the Human Genome Project announced its completion. It had taken $3 billion and 13 years to sequence the 3 billion base pairs in a single haploid genome. Ten years later, the diploid genome could be sequenced for approximately $5000 in a matter of days. In the near future, researchers hope to sequence an individual genome for less than $1000.

This advancement was made possible the by the development of next-generation, massively parallel sequencing (MPS) technologies. (Sanger sequencing is regarded as the first generation of sequencing technology). As MPS technologies have become more affordable, researchers have increasingly made use of sequencing to identify variants in their subjects. MPS studies represent a significant shift in the scope of genetic analysis; these technologies allow the extension of linkage and association analysis to the scope of the full genome. However, because MPS approaches can identify nearly all the variants in an individual genome, it becomes more difficult to determine which variants are "important." The pursuit of "private" variants that contribute to a disease with MPS studies signals a departure from the common variant-common disease model and a transition into the era of personalized medicine. To successfully identify the mutations that explain a particular phenotype, the scope of the study, sequencing platform, approach to variant calling, and variant interpretation are all important design considerations.

1.3.1. Design Considerations in Massively Parallel Sequencing Studies

Sequencing Target

While it is possible to sequence the whole genome, many researchers employ an approach that exclusively targets, enriches, and sequences the exome, i.e., the protein coding sequences of the genome. As described above, the focus on coding sequences can be partly explained by the difficulty that comes with trying to explain the effect of variants in non-coding sequences. While the exome comprises only 1 to 2% of the genome, it is predicted to contain 85% of the variants that explain single-gene, Mendelian diseases. Substantial efforts have been made to develop efficient strategies for targeting these regions. Furthermore, sequencing only the exome can ensure better coverage of the sequenced regions and reduces the computational resources needed to call and analyze the variants. This approach was first successfully used by Ng et al. (29) to determine the causal variant for Miller syndrome using data from 4 individuals from 3 separate families, and exome sequencing has since been used to identify causal variants for many Mendelian diseases.

Sequencing the exome, alone, however, may not be sufficient to find the variants that contribute to a phenotype. Even in the case of Mendelian diseases, it is possible that the causal variant resides outside of the exome or is not adequately targeted by current exome enrichment strategies. Furthermore, the analysis of sequenced variants may exclude variants that do not cause a change in the predicted amino acid sequence. Thus, researchers should use the characteristics of their disease of interest to determine whether whole genome sequencing or exome sequencing should be used. For Mendelian diseases with a known mode of inheritance

and for which the causal variant is suspected to have a high penetrance, exome sequencing can be more effective and less costly than whole genome sequencing.

Sequencing Platform

Several technologies have been developed in the field of next-generation sequencing. MPS is high-throughput and massively parallel, meaning that thousands of bases can be sequenced simultaneously. The sequencing studies of microcoria myopathy and FTD-ALS described in this dissertation made use of the Illumina (Solexa) sequencing platform. The underlying principle of this technology is sequence-by-synthesis. The Illumina platform makes use of flow cells; the surfaces of these flow cells are covered with a lawn of special primers to which genomic DNA fragments are ligated. The fragments are sequenced using a process called cyclic reversible termination. Briefly, the attached DNA fragments are amplified with fluorescent dye-labeled nucleotides to form clusters of DNA fragments, which are imaged, allowing the calling of nucleotides. Other sequencing platforms include those developed by Roche (454), Helicose Bioscences, Pacific BioSciences, and Life Technologies (SOLiD and Ion Torrent).

Sequencing alignment and variant calling

Once the sequence reads have been produced, the next step involves aligning the reads to a reference genome. Many software tools have been developed to perform this task. Most of these methods rely on indexing the reads, the reference genome, or both. The earliest aligning tools used hashing to create these indexes. However, this form of indexing is memory-intensive, which prompted the development of diverse algorithms to minimize the needed computational

resources. Several of these algorithms make use of the Burrows-Wheeler transformation to compress the sequence information and perform the alignment.

A second crucial computational task involves the calling of genotypes once the sequence reads have been aligned. This process takes into account, at the bare minimum, the number of reads for a given nucleotide and the quality of these reads at each position in the genome. Probabilistic methods for variant calling take into account other information, such as LD structure and allele frequencies, to make genotype calls (30).

Variant interpretation

Once the variants in an individual have been called, the next task is to determine which of these contribute to the phenotype of interest. Software tools, such as Polyphen, can predict whether a variant has the potential to cause a change in protein sequence or structure. However, variants might be excluded from analysis if it is unclear how they affect the product of a gene.

1.3.2. Limitations of Massively Parallel Sequencing Studies

Read length

The majority of sequencing platforms produce reads that are in the range of 100 nucleotides in length. The short length of these reads can result in mis-alignment or unplaced reads, leading to a loss of data. Read length will become less of an issue as the technology is able to synthesize longer reads.

Resolving copy number variants and repetitive regions

Because MPS relies on reference genomes, this method is not as adept at resolving mutations that more than a few nucleotides in length. In the case of cancer, for example, large chromosomal rearrangements are not uncommon, and these rearrangements are difficult to capture with standard genome alignment tools. Despite these challenges, MPS technologies will continue to advance the field of genetics and enable the development of personalized medicine.

1.4 Chapter Overviews

The studies contained within this manuscript highlight the use of the approaches described above. Chapters 2 and 3 describe two GWA studies that were performed in the Ashkenazi population. Chapter 2, a study of idiopathic Parkinson's disease, illustrates how genetic heterogeneity can obscure association signals. Chapter 3 presents a pilot GWA study of dystonia, a movement disorder. Chapters 4 and 5 focus on the use of MPS technologies. Chapter 4 describes the use of exome sequencing to analyze a familial presentation of frontotemporal dementia with amyotrophic lateral sclerosis (FTD-ALS). Chapter 5 describes a genome-wide linkage analysis of a pedigree affected with microcoria myopathy and a combined whole genome and whole exome sequencing analysis of this pedigree and 7 unrelated individuals. Finally, Chapter 6 provides conclusions and future directions for the studies as well as insights for the future of human genetic analysis.

1.5 REFERENCES

1. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980; 32: 314–331.

2. MacCluer JW, Vandeburg JL, Read B and Ryder OA. Pedigree analysis by computer simulation. Zoo Biol. 1986; 5:149-160.

3. Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Hum Hered. 1971; 21: 523–542.

4. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Nat Acad Sci. 1987; 84(8): 2363–2367.

5. Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet. 1972; 2:3–19.

6. Lange K, Westlake J, Spence MA. Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet. 1976; 39:485–491.

7. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308(5720): 385-389.

8. Pericak-Vance MA, Bebout JL, Gaskell PC Jr, et al. Linkage studies in familial Alzheimer disease: Evidence for chromosome 19 linkage. Am J Hum Genet 1991; 48(6): 1034–50.

9. Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993; 261(5123): 921–3.

10. Grupe A, Abraham R, Li Y, et al. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum Mol Genet. 2007;16.8:865-873.

11. Bertram L, Tanzi RE. Genome-wide association studies in Alzheimer's disease. Hum Mol Genet. 2009;18.R2: R137-R145.

12. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 2010; 42:348-54.

13. Burton PR, Clayton DG, Cardon LR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447(7145): 661-678.

14. Devlin B, Roeder K. Genomic control for association studies, Biometrics. 1999; 55(4):997–1004.

15. Pritchard J K, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945-959.

16. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2.12: e190.

17. Price, Alkes L., et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8): 904-909.

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

19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B. 1995; 57: 289–300.

20. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics. 2011;12(10): 703-714.

21. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006; 78(4): 629-644.

22. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American journal of human genetics, 2007; 81(5), 1084.

23. Li Y, Willer CJ, Ding J, Scheet P and Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34:816-834.

24. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6): e1000529.

25. Allen HL, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010: 467(7317):832-838.

26. Sebat J, Lakshmi B, Malhotra D, et al. Strong association of de novo copy number mutations with autism. Science. 2007; 316(5823):445-449.

27. Andreasen CH, Stender-Petersen KL, Mogensen MS, et al. Low physical activity accentuates the effect of the *FTO* rs9939609 polymorphism on body fat accumulation. Diabetes. 2008;57:95–101.

28. Rampersaud E, Mitchell BD, Pollin TI, et al. Physical activity and the association of common *FTO* gene variants with body mass index and obesity.Arch Intern Med 2008;168:1791–1797.

29. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2009;42(1): 30-35.

30. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Reviews Genet. 2011:12(6): 443-451.

Chapter 2

A Genome-wide Association Study of Idiopathic Parkinson's Disease

2.1 Introduction

As the one of the most common neurodegenerative diseases, idiopathic Parkinson's Disease (IPD) is a debilitating disease presenting with chronic, progressive difficulty with motor function. While the symptoms of the disease can be managed, its progression cannot. IPD is estimated to affect 1% to 2% of the population over 65 and can occur in families or sporadically. Despite the identification of some environmental factors and numerous genetic studies, the etiology of this disorder is largely unexplained in greater than 95% of cases. The study of the disease is further complicated by genetic and phenotypic heterogeneity; this phenotypic heterogeneity is evidenced by the presence of both early-onset and late-onset forms of the disease. The disease can also be monogenic or polygenic. Thus, substantial effort is needed to determine the genetic causes of this disease.

Family-based linkage studies, which have the ability to identify rare variants with high penetrance, constituted some of the first attempts to identify genes that cause IPD. Polymeropoulos et al. (1, 2) mapped the first IPD susceptibility locus to chromosome 4q21-23 in an Italian kinship, which they later deduced to be a mutation in the gene for alpha-synuclein (*SNCA).* Since this study, the following loci have been identified as linked to the disease: *PARK3* (3), *PARK5/UCHL1* (4), *PARK8/LRRK2* (5), *PARK10* (6,7), *PARK11* (8), *PARK2/SOD2* (9-12), *PARK7/DJ1* (13,14), *PARK15/FBXO7* (15), and *PARK12* (8). Among the genes identified, leucine-rich repeat kinase 2 (*LRRK2*, 12q12) was first linked to IPD using a family-based linkage

study in a Japanese family (5). Subsequently, a moderate penetrant (G2019S) mutation in *LRRK2* was identified in several populations that would not have been detected by conventional linkage analysis; the G2019S mutation is estimated to have a 30% lifetime penetrance (16).

While these linkage studies could elucidate the etiology of IPD in select families, new methods were needed to begin the task of identifying the causes of most cases of IPD. Approximately 15% of IPD cases are familial, while the rest are sporadic. With the advent of genome-wide single nucleotide polymorphism (SNP) genotyping platforms, researchers began looking for common variants that contribute to IPD. Such genome-wide association studies (GWAS) have the ability to identify common variants that contribute to the risk for a disease, and it was hoped that because IPD is a common disease that common variants could explain the etiology of the disease. The first IPD GWAS, performed by the Linked Efforts to Accelerate Parkinson's Solutions (LEAPS), studied 443 discordant sibling pairs with 332 case-unrelated control pairs and identified 11 SNPs as likely to be associated with IPD (17). The second major GWAS, performed by the National Institute of Neurological Diseases and Stroke (NINDS), tested 547 unrelated individuals and was not able to replicate the LEAPS findings (18). Because of its detection in many different linkage studies, GWAS researchers had expected to find a significant association at *LRRK2*. Neither the LEAPS study nor the NINDS study was able to detect polymorphisms near *LRRK2* as being associated with IPD. A meta-analysis of 3,458 cases and 3,719 controls from 10 different populations conducted by the Genetic Epidemiology of Parkinson’s Disease (GEOPD) was also unable to replicate the findings of these studies (19). The conflicting results from these studies raised skepticism about the ability to find susceptibility loci for sporadic IPD under the common disease-common variant model implicit to GWAS.

This inability to discover associated loci could be explained by the fact that these studies did not satisfy certain conditions. For loci to have significant association signals in a GWAS, there needs to be sufficient allelic and locus homogeneity such that the genotyped markers are in LD with the causal allele(s). In the case of a polygenic disorder such as IPD, this condition is difficult to achieve. Also, the effect of an associated variant has to be modest in size, and the variant has to have a relatively high frequency. The study population needs to be homogeneous; population substructure, the presence of two or more groups with different genetic backgrounds in a study, can cause false positive signals that obscure true associations and be a source of locus heterogeneity. Because of their small sample sizes and heterogeneous populations, previous IPD GWAS were under-powered to detect variants that contribute to IPD.

In order to increase power, it will be necessary to increase sample size and reduce locus heterogeneity. GWA studies of other very common complex, multifactorial diseases, including type 2 diabetes and coronary artery disease, now involve tens to hundreds of thousands of samples genotyped at over a million markers to search for causative mutations. An alternate approach is to increase power by focusing on specific populations with a shared genetic background; the underlying logic is that individuals with a disease in the same population will have inherited disease-causing variants from common ancestors.

This claim is supported by two GWA studies performed in different populations; the first of these was a study conducted by Satake et al. in 2,000 cases and 18,000 controls in a Japanese population in which the researchers were able to detect a significant association on a genome-wide level for *LRRK2*, which is a very promising finding (20). The second study was conducted by Simon-Sanchez et al. in a European population of 3,000 cases and 4,000 controls, and the association signal at *LRRK2* was not significant (21). Because these studies incorporated large

sample sizes, they had substantial power to detect associations. Both of these studies were able to identify associated loci with genome-wide significance; Satake et al. found a novel risk locus, which they named *PARK16*, as well as SNPs in *BST1*, *SNCA*, and *LRRK2* to be associated with IPD in their Japanese study population, while Simon-Sanchez et al. found SNPs in *SNCA*, *MAPT*, and *PARK16* to be significantly associated in their European study population. The largest IPD GWAS to date, which was conducted by Do et al. in individuals of European descent, consisted of 3,426 cases and 29,624 controls (22). This study identified two novel associations with genome-wide significance at *SCARB2* and *SREBF1/RAI1* and confirmed associations at *LRRK2, GBA*, *SNCA*¸*MAPT*, *GAK*, and the *HLA* region. Thus, these studies indicate that genetic heterogeneity within the study population and the inability of common SNPs to "tag" the G2019S mutation, and not simply a lack of power, prevented previous IPD GWAS from finding an association signal at *LRRK2*.

In the European population, studies have suggested that G2019S-positive individuals with late onset IPD share a common haplotype (23). The most recent estimates suggest that this mutation arose from a common ancestry some 2500 years ago (24). This claim is supported by the observation that G2019S mutations are far more common in the Ashkenazi and Middle Eastern/North African populations than in the general European populations (25).

The Ashkenazi population is unique in that it features haplotypes that are European and North African in origin (26). In this population, G2019S mutations are associated with 10 and 28%, respectively, of sporadic and family history positive patients. The best estimate of the cumulative risk of IPD for G2019S carriers is 36% at 59 years, 59% at 69 years, and 80% at 79 years (27). By examining this population separately in a GWAS for causative polymorphisms,

this study will have increased power to detect therapeutic targets for IPD and contribute to the search for more effective treatments for this debilitating disease.

Thus, a small, prospective GWAS in the Ashkenazi, followed by a larger replication GWAS, was performed to look for common variants that could explain the genetics of IPD in this population. While the initial GWAS was unable to detect an association signal near *LRRK2*, the replication GWAS was able to detect a strong association signal across this region that approached genome-wide significance. On chromosome 12q12, several SNPs in the genes *CNTN1* and *SLC2A13* also approached genome-wide significance.

2.2 Methods

Samples

For the initial phase, a GWAS was conducted in an Ashkenazi population from New York consisting of 25 *LRRK2*-G2019S positive individuals, 96 IPD cases, and 96 controls. The replication phase of the study consisted of 166 IPD cases and 1,436 controls. The cases included 50 individuals that were known *LRRK2* G2019S mutation carriers. The controls included 105 dystonia patients, 407 schizophrenia patients, 273 Crohn's disease patients, and 651 healthy controls.

Only individuals with 4 Ashkenazi grandparents were included in the study. The IPD cases were confirmed by Dr. Susan Bressman, a movement disorder specialist, and were obtained in collaboration with Dr. Laurie Ozelius of the Mount Sinai School of Medicine. Family history information was collected from each patient, and none of the subjects included in the analysis was known to be related or had a measured pair-wise kinship coefficient greater than 0.1 with another participant (i.e., less related than first cousins) as measured with the PLINK

software package (28). In the replication study, 5 cases and 32 controls were excluded based on relatedness.

Genotyping

For the first phase of the study, the samples were genotyped on the Affymetrix GeneChip Human Mapping 500K SNP array. The genotype calls were made using the Birdseed V2 algorithm included in the Affymetrix Genotyping Console Software package. For the replication phase of the study, all samples were genotyped on the Affymetrix Genome-Wide Human SNP Array 6.0 platform according to the manufacturer's protocol, and the genotype calls were made using the Birdseed V2 algorithm.

Quality control of SNP Genotyping and Association Analysis

Standard quality control measures were used to filter the genotyping calls with the PLINK software. Only SNPs with a minor allele frequencies > 0.05 and genotyping rates > 90% that did not violate Hardy-Weinberg equilibrium were included in the association analysis. After removing these SNPs, 357,319 markers and 634,338 markers remained in the initial study and the replication study, respectively. The p-values for the association of each of the SNPs in the initial study were calculated by comparing the distribution of the genotypes between the cases and controls with a Fisher's 2x3 exact test. The association analysis for the replication study was performed for the genotyped and imputed SNPs (see below) for the replication study using the mach2dat tool in the MaCH software package. An association was considered to have genome-wide significance if the p-value was less than 2.0×10^{-8}; this threshold was determined using Bonferroni's correction (p = 0.05/2,500,000 markers).

To reduce type I error that resulted from aggressive genotyping calling by the Birdseed algorithm, the fluorescent probe intensities for each allele for each genotype were plotted for SNPs with p-values that were less than 1.0×10^{-5}. These scatter plots were used to determine if the genotype calls had been made inappropriately.

Imputation

For the replication phase of the study, additional genotypes were imputed using MACH (MArkov Chain Haplotyping) (29, 30). A set of reference haplotypes was created using the HapMap2 CEU population after first ensuring that the alleles from the current study matched the HapMap alleles for SNPs that were found in both data sets (i.e., that they were genotyped on the same strand). The genotypes from the HapMap data were phased to produce a set of reference haplotypes with the following parameters: --rounds 20, --states 300, --phase, --interim 5, --sample 5, --compact.

Then, this set of reference haplotypes was used to impute additional markers with the following options in MaCH: --rounds 5, --states 200. For genotype inference, only SNPs genotyped on the Affymetrix platform with genotyping rates > 90% and minor allele frequencies > 5% that did not violate Hardy-Weinberg equilibrium ($p< 0.000001$) were included in the input files.

Analysis of Population Stratification

Population stratification is known to confound the results of association studies. To assess population stratification, the genomic control inflation factor, λ, was calculated by selecting a subset of 50 SNPs from the replication GWAS data that were in linkage

disequilibrium with each other. The median of the observed chi-squared distribution of the association values for these SNPs was divided by the expected median of this distribution to obtain a value of 1.04, which was indicative of negligible population stratification.

To further rule out population stratification, principal components analysis was performed in EIGENSTRAT (31) by comparing the study population with the CEU and YRI HapMap populations. These two populations were selected because evidence suggests that the Ashkenazi population is characterized by the presence of haplotypes that are both European and North African in origin. The first two principal components were plotted, and the population clustered together, reflecting the coancestry of the study population.

Plots

The Manhattan plots and Q-Q plots were produced using the "qqman" package in the R software package. The plot of the association signal for *LRRK2* was produced in Locus Zoom (32) using data from the 1000 Genomes project European population (hg19).

2.3 Results

Initial GWAS

An initial GWAS was conducted in an Ashkenazi population from New York consisting of 25 *LRRK2*-G2019S positive individuals, 96 IPD cases, and 96 controls. Only individuals with 4 Ashkenazi grandparents were included in the study, and the absence of population admixture was confirmed by genetic analysis with the Eigenstrat software package. IPD cases were confirmed by Dr. Susan Bressman, a movement disorder specialist, and were obtained in collaboration with Dr. Laurie Ozelius of the Mount Sinai School of Medicine. Family history

information was collected from each patient, and none of the subjects included in the analysis was known to be related or had a measured pair-wise kinship coefficient greater than 0.1 with another participant (i.e., less related than first cousins) as measured by PLINK. The samples were genotyped on the Affymetrix GeneChip Human Mapping 500K platform, and the genotypes were called with the Birdseed V2 algorithm included in the Affymetrix Genotyping Console Software package. SNPs were filtered based on the MAF, missingness of genotypes, and violation of Hardy-Weinberg equilibrium; after this, 357,319 markers remained. The significance of the associations was calculated by comparing the distribution of the genotypes between the cases and controls with Fisher's exact test.

An examination of the distribution of the association statistics indicated the possibility of Type I errors. There were far more significant associations than what would have been expected, which was confirmed by a Q-Q plot (Figure 2.1). The high rate of positive associations was due to the aggressiveness of the genotype calling algorithm. Visual analysis of the genotyping cluster plots allowed for the exclusion of SNPs that were not called correctly.

By reducing the possibility of false positive associations resulting from aggressive genotype calling through visual genotype plot analysis, the list of loci potentially associated with IPD was refined. The most significantly associated SNPs are listed in Table 2.1. A significant association signal was not detected across *LRRK2*, most likely because the Affymetrix 500K platform does not tag this mutation well and because the signal was obscured by the cases without *LRRK2* mutations (Figure 2.2). *LRRK2* was significantly associated in this sample only when considering the *LRRK2*-positive cases vs. the controls (Figure 2.3), which illustrates why previous IPD GWAS have been unable to find a significant association in the region of *LRRK2* (17, 18). Of the associations that were identified, the association at rs11167267 on chromosome

20q11.22 was the most significant with a p-value of 9.21 X 10^{-7}, which is greater than the threshold for genome-wide significance. This variant is located in *C20ORF173*. Because of the small sample size of this initial GWAS, a replication GWAS was performed.

Replication GWAS

For the replication phase of the study, the study group initially consisted of 166 IPD cases and 1,436 controls. The cases included 50 individuals that were known *LRRK2* G2019S mutation carriers. The controls included 105 dystonia patients, 407 schizophrenia patients, 273 Crohn's disease patients, and 651 healthy individuals. After filtering out individuals for relatedness, 161 cases and 1,404 controls remained. All samples were genotyped on the Affymetrix 6.0 genotyping platform, which allows for the genotyping of nearly one million SNPs. After filtering for missingness and MAF as described above, 634,338 SNPs remained. To provide greater coverage of the genome, imputation was performed with MaCH to infer the genotypes for additional markers using the HapMap CEU population. When including the imputed SNPs, the dataset consisted of 2,557,252 SNPs. The association statistics were calculated using the mach2dat tool. The p-values for the most significantly associated SNPs are listed in Table 2.2, and the results are displayed as a Manhattan plot in Figure 2.4.

Two SNPs on chromosome 12 were found to have association signals with genome-wide significance: rs10506095 ($p= 2.49x10^{-10}$) and rs7316771($p=1.19x10^{-9}$). Both are located in an intergenic region on chromosome 12p11.21, but these markers flanked *FGD4*. One SNP on chromosome 17q21.2 was also found to have a genome-wide significant association signal: rs3887424 ($p=4.29x10^{-8}$). This SNP was located in an intronic region of *ACLY*.

Another noteworthy finding was the presence of a genome-wide suggestive association signal across 23 SNPs on chromosome 12q12, which included the *LRRK2* locus, as well as the *SLC2A13* and *CNTN1* loci (Figure 2.5); these loci are located upstream and downstream of *LRRK2*, respectively. This finding was somewhat expected because of the inclusion of known *LRRK2* mutation carriers in the study. The average p-value for this signal was 4.85×10^{-6}; this signal most likely did not reach genome-wide significance because of the inclusion of subjects without known *LRRK2* mutations. This finding suggests that even within this population, there is genetic heterogeneity in the etiology of IPD. Furthermore, these results confirm those of a previous GWAS of IPD performed in a Japanese population that found significant associations for SNPs that were upstream of *LRRK2* (20).

2.4 Discussion

The results of this study provide further evidence that IPD is genetically heterogeneous. The inclusion of 50 known *LRRK2* G2019S mutation carriers in the replication study group enabled the detection of an association signal across 12q12. The strong association signal suggests that IPD patients with a *LRRK2* mutation share a distinct genetic etiology that may have arisen from a common ancestor. *LRRK2* was also recently associated with Crohn's disease, which, like IPD, is also more common in the Ashkenazi than in the general European population. Recent studies have found that this gene is expressed in immune cells and could participate in processes such as inflammation (33). In IPD, the penetrance of *LRRK2* mutations is relatively low, which indicates that additional mutations must be identified. Future research into this gene's pathways will provide a better understanding of its role in the etiology of both IPD and Crohn's disease.

Because the Ashekenazi population is characterized by shared coancestry, the number of potential haplotypes in the population was lower than would be expected, which made it more likely that causal variants were shared IBD from a common ancestor. Thus, future IPD GWAS could increase their power to detect associations, not only by increasing the number of subjects, but also by focusing on specific populations.

With the advent of massively parallel sequencing (MPS) technologies, however, it is far more likely that future studies will focus on whole genome/exome sequencing to identify variants that contribute to the etiology of IPD. For the past few years, researchers have debated how much of the genetic variation in a disease can be explained by the common variants sought by GWAS. In the case of IPD, the MPS approaches can complement the findings of GWAS by identifying variants in known IPD susceptibility genes, such as *MAPT*, *SNCA,* and *LRRK2*. Future research should yield insight into the biological pathways that cause this disease and identify potential therapeutic targets.

Contributions

Within this work, I performed the Birdseed genotyping, quality control of genotyping, development of the genotype correction tool, association analysis, imputation, population stratification analysis, general data analysis, and drafting of the manuscript. Dr. Kirk Wilhelmsen provided oversight and mentorship; he also developed software tools used for the quality control of genotyping calls and imputation. Dr. Amy Webb assisted with the design of the initial GWA study. Dr. Scott Chasse assisted with the Affymetrix genotyping. Dr. Susan Bressman and Dr. Laurie Ozelius provided the samples and some of the Affymetrix genotyping files.

2.5 REFERENCES

1. Polymeropoulos MH, Higgins JJ, Golbe LI, et al. Mapping of a gene for Parkinson's disease to chromosome 4q21-q23. Science. 1996;274(5290):1197-9.

2. Polymeropoulos MH, Lavedan C, Leroy E, et al. Mutation in the a-synuclein gene identified in families with Parkinson's disease. Science. 1997; 276:2045-2047.

3. Gasser T, Müller-Myhsok B, Wszolek ZK, et al. A susceptibility locus for Parkinson's disease maps to chromosome 2p13. Nat Genet. 1998; 18(3):262-5.

4. Liu Y, Fallon L, Lashuel HA, Liu Z, Lansbury PT. The UCH-L1 gene encodes two opposing enzymatic activities that affect alpha-synuclein degradation and Parkinson's disease susceptibility. Cell. 2002;111: 209-218.

5. Funayama M, Hasegawa K, Kowa H, Saito M, Tsuji S, Obata F. A new locus for Parkinson's disease (PARK8) maps to chromosome 12p11.2-q13.1. Ann Neurol. 2002; 51(3):296-301.

6. Li YJ, Deng J, Mayhew GM, Grimsley JW, Huo X, Vance JM.. Investigation of the PARK10 gene in Parkinson disease. Ann Hum Genet. 2007;71: 639-647.

7. Hicks AA, Pétursson H, Jónsson T, et al. A susceptibility gene for late-onset idiopathic Parkinson's disease.Ann Neurol. 2002;52(5):549-55.

8. Pankratz N, Nichols WC, Uniacke SK, Genome screen to identify susceptibility genes for Parkinson disease in a sample without parkin mutations. Am J Hum Genet. 2002; 71(1):124-35.

9. Matsumine H, Yamamura Y, Hattori N, Kobayashi T, Kitada T, Yoritaka A, Mizuno Y. A microdeletion of D6S305 in a family of autosomal recessive juvenile parkinsonism (PARK2). Genomics. 1998; 49(1):143-6.

10. Jones AC, Yamamura Y, Almasy L, et al. Autosomal recessive juvenile parkinsonism maps to 6q25.2-q27 in four ethnic groups: detailed genetic mapping of the linked region.Am J Hum Genet. 1998;63(1):80-7.

11. Tassin J, Dürr A, de Broucker T, et al. Chromosome 6-linked autosomal recessive early-onset Parkinsonism: linkage in European and Algerian families, extension of the clinical spectrum, and evidence of a small homozygous deletion in one family. The French Parkinson's Disease Genetics Study Group, and the European Consortium on Genetic Susceptibility in Parkinson's Disease. Am J Hum Genet. 1998; 63(1):88-94.

12. Saito M, Matsumine H, Tanaka H, et al. Refinement of the gene locus for autosomal recessive juvenile parkinsonism (AR-JP) on chromosome 6q25.2-27 and identification of markers exhibiting linkage disequilibrium. J Hum Genet. 1998; 43(1):22-31.

13. van Duijn CM, Dekker MC, Bonifati V, et al. Park7, a novel locus for autosomal recessive early-onset parkinsonism, on chromosome 1p36.Am J Hum Genet. 2001; 69(3):629-34.

14. Bonifati V, Dekker MC, Vanacore N, et al. Autosomal recessive early onset parkinsonism is linked to three loci: PARK2, PARK6, and PARK7. Neurol Sci. 2002;23(Suppl 2):S59-60.

15. Shojaee S, Sina F, Farboodi N, et al. A clinic-based screening of mutations in exons 31, 34, 35, 41, and 48 of LRRK2 in Iranian Parkinson's disease patients. Mov Disord. 2009;24(7):1023-7.

16. Ozelius LJ, Senthil G, Saunders-Pullman R, LRRK2 G2019S as a cause of Parkinson's disease in Ashkenazi Jews. NEJM. 2006;354(4): 424-425.

17. Maraganore DM, de Andrade M, Lesnick TG, et al. High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet. 2005; 77: 685-693.

18. Fung HC, Scholz S, Matarin M, et al. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2006;5: 911-916.

19. Evangelou E, Maraganore DM, Annesi G, et al. Non-replication of association for six polymorphisms from meta-analysis of genome-wide association studies of Parkinson's disease: large-scale collaborative study. Am J Med Genet B Neuropsychiatr Genet. 2010; 153B: 220-8.

20. Satake W, Nakabayashi Y, Mizuta I, et al. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nat Genet. 2009;41(12):1303-7.

21. Simón-Sánchez J, Schulte C, Bras JM, et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat Genet. 2009;41(12):1308-12.

22. Do CB, Tung JY, Dorfman E, et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet. 2011; *7*(6): e1002141.

23. Kachergus J, Mata IF, Hulihan M, et al. Identification of a novel LRRK2 mutation linked to autosomal dominant parkinsonism: evidence of a common founder across European populations.Am J Hum Genet. 2005;76(4):672-80.

24. Bar-Shira A, Hutter CM, Giladi N, Zabetian CP, Orr-Urtreger A. Ashkenazi Parkinson's disease patients with the LRRK2 G2019S mutation share a common founder dating from the second to fifth centuries. Neurogenetics. 2009; 4:355-8.

25. Lesage S, Leutenegger AL, Ibanez P, Janin S, Lohmann E, Dürr A, Brice A; French Parkinson's Disease Genetics Study Group. LRRK2 haplotype analyses in European and

North African families with Parkinson disease: a common founder for the G2019S mutation dating from the 13th century. Am J Hum Genet. 2005;77(2):330-2.

26. Hammer MF, Redd AJ, Wood ET, et al. Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. Proc Natl Acad Sci USA 97. 2000;(12):6769-74.

27. Healy DG, Falchi M, O'Sullivan SS, et al. Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. Lancet Neurol. 2008;7(7):583-90.

28. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 2007; 81(3): 559–575.

29. Li Y, Willer CJ, Ding J, Scheet P and Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34:816-834.

30. Li Y, Willer CJ, Sanna S and Abecasis GR. Genotype Imputation. Annu Rev Genomics Hum Genet. 2009; 10:387-406.

31. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904-909.

32. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18): 2336.2337.

33. Hakimi M, Selvanantham T, Swinton E, et al. Parkinson's disease-linked LRRK2 is expressed in circulating and tissue immune cells and upregulated following recognition of microbial structures. J Neural Transmission. 2011;118(5) : 795-808.

Figure 2.1: Quantile-quantile plot of the observed (blue) vs. the expected (green) distribution of p-values for the initial IPD GWAS

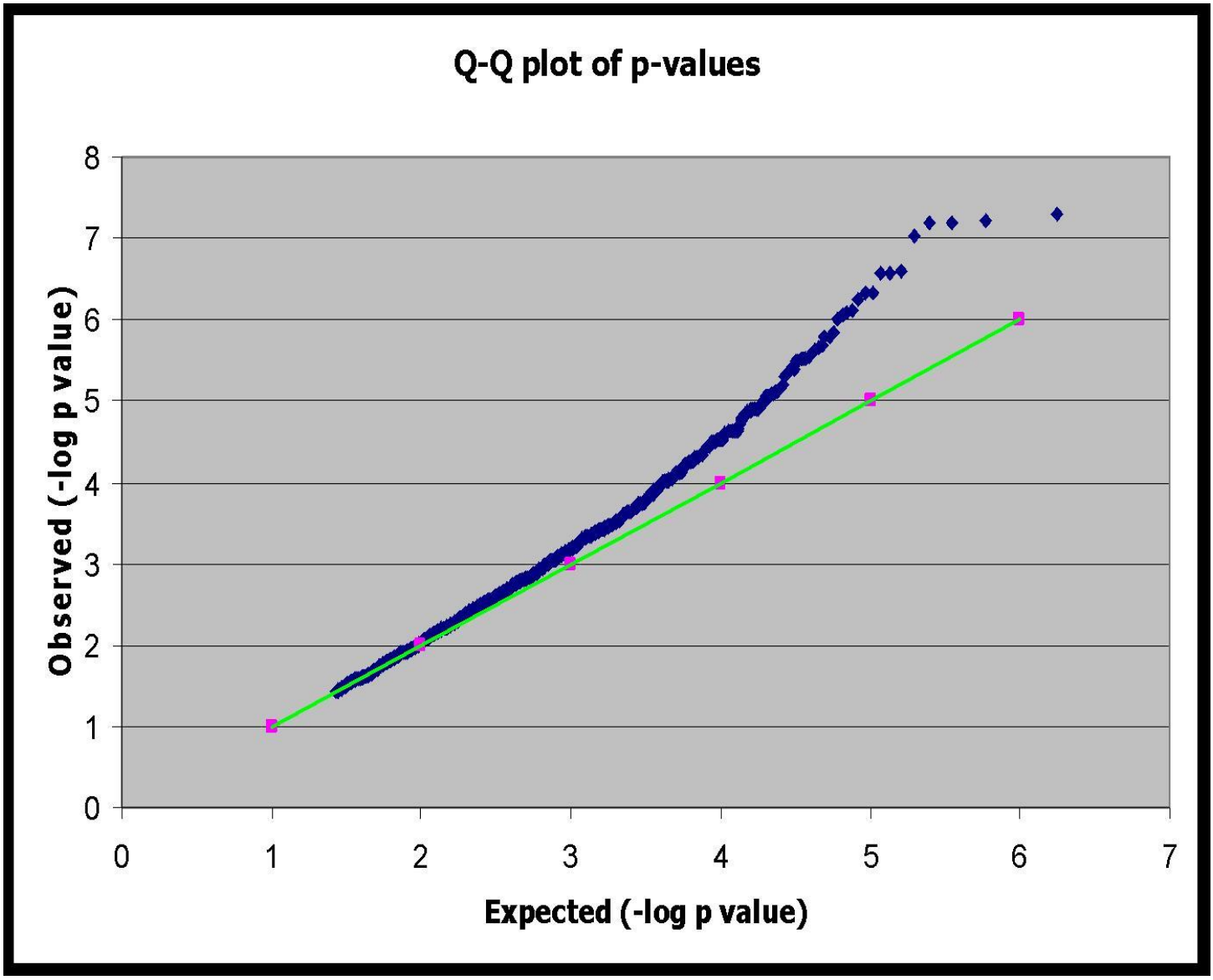


Figure 2.2: Manhattan plot of the $-\log_{10}$(p-values) of the initial IPD GWAS.

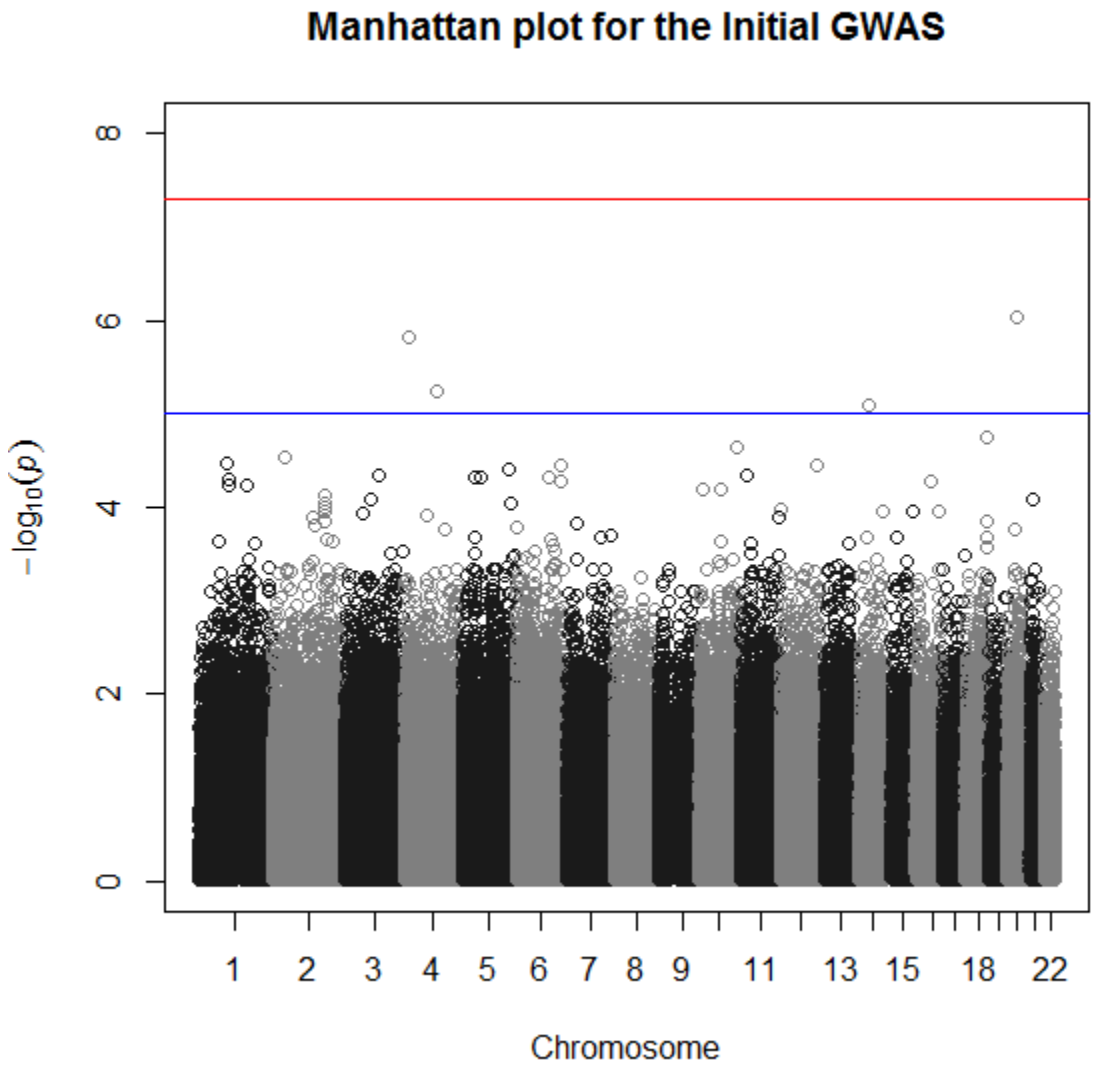


Figure 2.3: Manhattan plot of the association signal across *LRRK2* for the initial GWAS

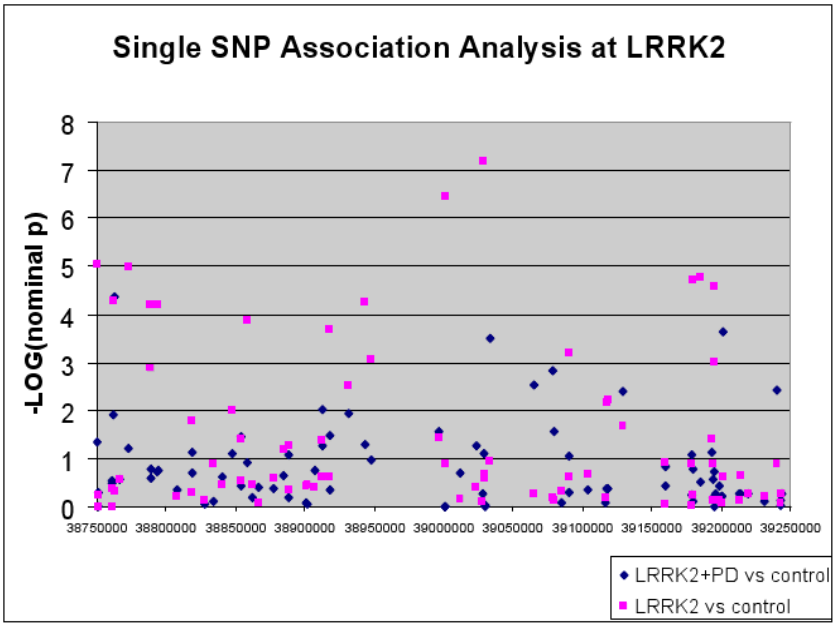


Figure 2.4: Manhattan Plot of the $-\log_{10}$(p-values) of the replication IPD GWAS. SNPs in the *LRRK2* region are in green.

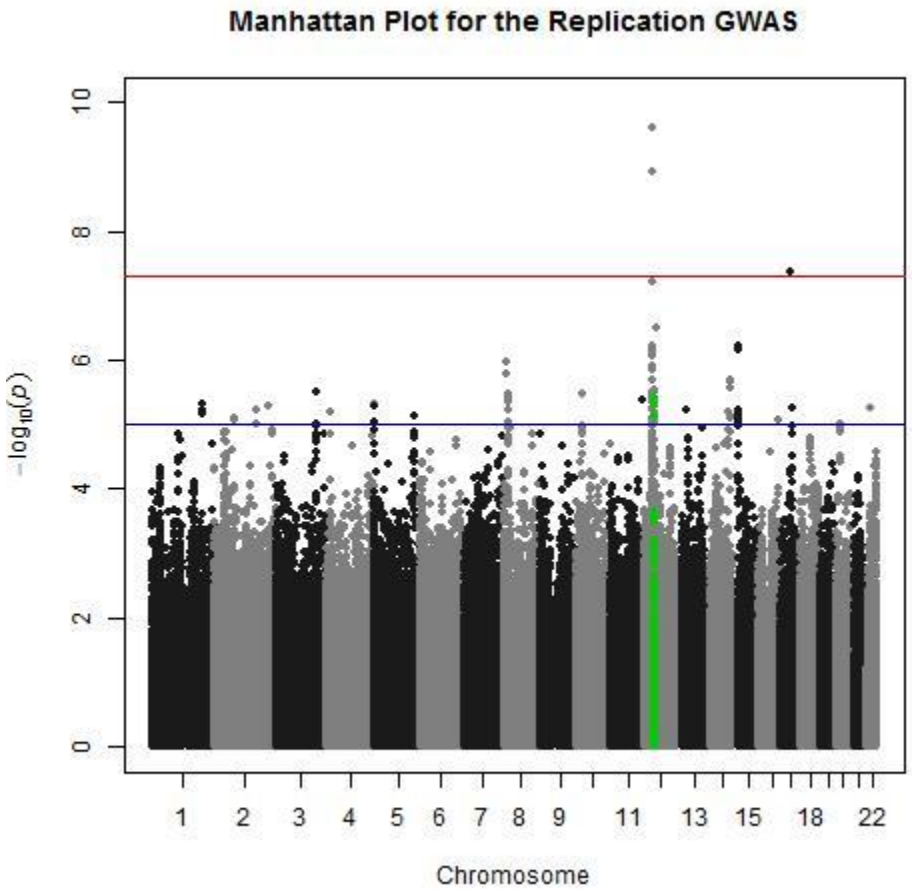


Figure 2.5: Plot of the Association Signal Across the *LRRK2* region

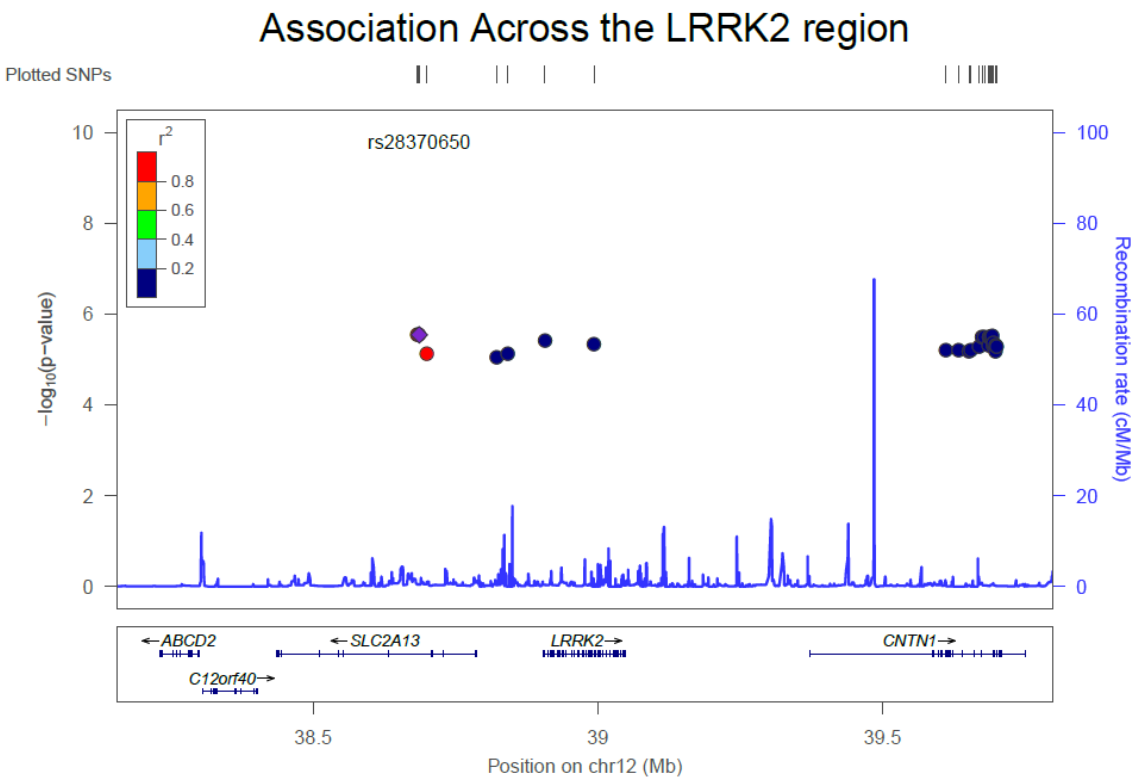


Table 2.1: Significant SNPS ($p<1.0x10^{-5}$) in the initial IPD GWAS

SNP	Chromosome	Position	Alleles	P-value	Locus
rs11167267	20	33572687	G/T	9.21E-07	*LOC729591*
rs4443273	4	14417782	C/T	1.50E-06	
rs794148	4	110178444	A/C	5.78E-06	*COL25A1*
rs181311	14	47408911	G/T	8.14E-06	

Table 2.2: Significant SNPs ($p<1.0x10^{-5}$) in the replication IPD GWAS

SNP	Chromosome	Position	Alleles	P-value	Locus
rs7534906	1	197798537	A/G	7.02E-06	*CAC1S*
rs12409138	1	197799441	C/T	4.67E-06	*CAC1S*
rs17454947	1	197801418	A/C	6.02E-06	*CAC1S*
rs10520244	2	79672427	C/T	8.31E-06	*CTN2*
rs10520245	2	79672517	G/T	8.13E-06	*CTN2*
rs16852806	2	167632117	C/T	9.55E-06	*XIRP2*
rs11889300	2	167632527	A/G	9.52E-06	*XIRP2*
rs11896680	2	167633994	C/T	5.78E-06	*XIRP2*
rs7578258	2	216607237	A/G	5.04E-06	
rs1724725	3	159385098	C/T	3.12E-06	*RSRC1*
rs827152	3	159392089	C/T	9.44E-06	*RSRC1*
rs13128422	4	17155671	C/T	6.52E-06	
rs13128864	4	17155877	C/G	6.60E-06	
rs7666115	4	189593305	C/T	4.68E-06	
rs323653	5	2693466	C/T	8.84E-06	
rs10035114	5	2702027	A/T	5.01E-06	
rs6863801	5	2702263	A/G	5.00E-06	
rs17294992	5	157459405	C/T	7.59E-06	
rs7462014	8	10944461	A/C	1.63E-06	*XKR6*
rs7821584	8	10952970	G/T	1.07E-06	*XKR6*
rs13261970	8	19173262	C/G	9.18E-06	
rs1193308	8	19177828	C/T	5.81E-06	

rs1193296	8	19187657	A/G	4.50E-06	
rs1514701	8	19206223	A/G	3.44E-06	
rs1193304	8	19208810	A/G	3.74E-06	
rs10741067	10	25160382	A/G	3.36E-06	
rs11220605	11	126146741	C/T	4.00E-06	*KIRREL3*
rs17563276	12	31564296	C/T	5.38E-06	*MGC24039*
rs10844094	12	32065781	G/T	6.11E-07	
rs7316376	12	32070167	A/G	1.20E-06	
rs12230201	12	32074419	C/T	8.96E-06	
rs10466825	12	32075914	A/G	5.83E-06	
rs12426455	12	32077228	C/T	2.06E-06	
rs10844098	12	32078551	A/G	7.19E-07	
rs12309730	12	32079488	A/C	7.08E-07	
rs7316771	12	32117395	A/G	1.19E-09	
rs1995303	12	32539025	C/T	6.24E-08	
rs17609576	12	32565912	G/T	3.89E-06	*FGD4*
rs10506095	12	32691394	A/G	2.49E-10	
rs11052837	12	33723272	G/T	1.42E-06	
rs10844685	12	33738527	A/G	1.32E-06	
rs10844691	12	33747193	A/G	1.33E-06	
rs10128817	12	33750401	A/G	1.33E-06	
rs12303803	12	33757777	A/G	1.33E-06	
rs10844695	12	33760364	C/T	1.33E-06	
rs10844696	12	33763068	C/G	1.33E-06	
rs11052866	12	33763963	C/T	1.33E-06	

rs12303090	12	33790590	C/T	8.81E-07	
rs12304892	12	33800536	C/G	3.09E-06	
rs11052894	12	33802361	A/G	3.11E-06	
rs10844701	12	33803535	C/T	3.11E-06	
rs1917788	12	33805295	C/T	3.56E-06	
rs1917787	12	33805365	C/T	3.59E-06	
rs10844704	12	33812449	A/C	3.92E-06	
rs8181672	12	33816074	C/T	7.28E-07	
rs7358615	12	33823930	A/C	4.08E-06	
rs11052909	12	33828400	A/T	4.36E-06	
rs12372058	12	36483324		8.32E-06	
rs11181339	12	36608538	G/T	9.44E-06	
rs4520663	12	36693099		7.69E-06	
rs11181857	12	36699522	C/G	9.45E-06	
rs11181864	12	36701491		9.45E-06	
rs11174697	12	38682716	C/T	2.88E-06	*SLC2A13*
rs28370649	12	38685416	A/G	2.87E-06	*SLC2A13*
rs28370650	12	38686215	A/T	2.86E-06	*SLC2A13*
rs28370664	12	38699572	A/C	7.69E-06	*SLC2A13*
rs17442721	12	38822040	C/G	9.07E-06	
rs17483919	12	38841841	C/T	7.54E-06	
rs17465751	12	38907148	C/T	3.88E-06	*LRRK2*
rs17443882	12	38993127	C/T	4.73E-06	*LRRK2*
rs1866996	12	39611343	G/T	6.37E-06	*CNTN1*
rs11179174	12	39633654	A/T	6.31E-06	*CNTN1*

rs1442190	12	39651907	A/G	6.43E-06	*CNTN1*
rs11179263	12	39655042	A/G	6.18E-06	*CNTN1*
rs278914	12	39670040	G/T	5.15E-06	*CNTN1*
rs12296475	12	39674979	A/G	3.13E-06	*CNTN1*
rs9652040	12	39679144	C/G	3.14E-06	*CNTN1*
rs278903	12	39686710	C/T	4.88E-06	*CNTN1*
rs278902	12	39687471	A/G	3.14E-06	*CNTN1*
rs278901	12	39688545	A/G	3.15E-06	*CNTN1*
rs280355	12	39690372	C/T	3.14E-06	*CNTN1*
rs157126	12	39692596	A/G	2.93E-06	*CNTN1*
rs397967	12	39694401	A/G	4.79E-06	*CNTN1*
rs278900	12	39695078	A/T	4.27E-06	*CNTN1*
rs280357	12	39698403	C/T	6.50E-06	*CNTN1*
rs280361	12	39699334	C/T	5.01E-06	*CNTN1*
rs280362	12	39700624	A/C	5.23E-06	*CNTN1*
rs11169141	12	48432726	A/G	3.17E-07	*TEGT*
rs9548650	13	38640007	A/C	5.81E-06	
rs1775671	14	86373673	A/G	6.45E-06	
rs8016073	14	94095240	C/T	7.48E-06	
rs910353	14	94098498	A/G	7.66E-06	*SERPI4*
rs2284655	14	94101183	A/G	8.07E-06	*SERPI4*
rs4905223	14	94106776	A/G	2.73E-06	
rs1998243	14	94107549	C/G	2.04E-06	
rs761536	14	94109674	C/T	2.02E-06	
rs1955652	14	94111231	A/G	2.02E-06	

rs8022491	14	94114330	A/G	2.07E-06	
rs1983657	14	94117046	A/T	2.12E-06	*SERPI5*
rs4392030	15	24573424	C/G	6.16E-06	
rs7182514	15	24575314	A/G	6.25E-06	
rs7165224	15	24575429	C/T	6.28E-06	
rs9324132	15	24579670	C/G	8.01E-06	
rs7174912	15	24579934	C/G	9.11E-06	
rs7170353	15	24582371	C/T	6.74E-07	
rs7497827	15	24584497	G/T	7.77E-06	
rs11161340	15	24584610	A/G	9.88E-06	
rs7495536	15	24584887	G/T	9.87E-06	
rs7183073	15	24585118	G/T	7.05E-06	
rs4906904	15	24585178	A/G	7.00E-06	
rs4906905	15	24585338	A/T	6.99E-06	
rs4906906	15	24585616	A/C	5.84E-07	
rs8028779	15	24585824	C/T	6.95E-06	
rs4632102	15	24586127	C/T	6.81E-06	
rs6576616	15	24586375	A/G	9.41E-06	
rs6576617	15	24586468	A/G	9.39E-06	
rs7495131	15	24587031	A/G	9.26E-06	
rs7170111	15	24587839	A/G	9.29E-06	
rs6576619	15	24597498	A/T	5.86E-06	
rs4889373	16	80305103	A/G	8.74E-06	
rs3887424	17	37279744	C/T	4.29E-08	*ACLY*
rs9897679	17	49236622	C/T	5.69E-06	

rs6047628	20	21722338	C/G	9.53E-06	
rs13057533	22	28935285	C/T	5.42E-06	*LOC729980*

Chapter 3

A Pilot GWAS of Idiopathic Focal Dystonia identifies an Association Signal at *RNF213*

3.1 Introduction

The genetics of dystonia, a group of neurological disorders in which the muscles contract involuntarily, are still largely unexplained. The disease is characterized by considerable clinical heterogeneity. There are both early onset and late onset forms of the disease. Furthermore, the disease can be classified as being generalized or focal; focal dystonias can affect muscles near the extremities, eyes, and vocal cords, among others. The extent to which this disease is heritable is unclear. There is currently no cure for dystonia, and treatment is only available to remediate it symptoms; thus, establishing an understanding of the genetic causes of dystonia is crucial for identifying the pathways involved in this disease and potential therapeutic targets.

To date, family-based linkage studies have identified a number of loci as being implicated in this disease, especially in primary dystonia. The majority of early-onset, generalized dystonias are believed to share a common genetic etiology. The first of these linked genes, *TOR1A*, which is also known as *DYT1*, was mapped to chromosome 9q34; this gene encodes a protein known as torsinA (1). A three base-pair deletion, known as the GAG deletion, within this gene was shown to be responsible for the majority of primary, early-onset torsion dystonias and was identified as having an autosomal dominant mode of inheritance (2). However, this gene was also found to be linked to focal dystonias (3,4). Furthermore, the *TOR1A* deletion was estimated to have between 30% and 40% penetrance (5-8).

In idiopathic focal dystonias, there is emerging evidence for a genetic etiology (9); however, no genes have been implicated in this disease. There is no clear estimate of the heritability of focal dystonias in families, and the disease can also occur sporadically. Because the likelihood of an affected individual having a family member with the disease is estimated to be approximately 20% (10), it is possible that any causative variants are characterized by low penetrance. This suspected low penetrance, coupled with the uncertain heritability of the condition, makes it difficult to investigate the genetics of this disease through family studies.

To better understand the genetics of focal dystonias, we instead sought to identify the role of common genetic variants. Thus, we aimed to perform the first GWAS of dystonia.

The study was performed in an Ashkenazi Jewish population from New York. The prevalence of primary dystonia is estimated to be 152 per 1,000,000 in the European population, and the prevalence of blepharospasm, a focal dystonia that causes uncontrolled contractions of the eyelids, is estimated to be 36 per 1,000,000 (11). However, because dystonia is believed to be 5-10 times more prevalent in the Ashkenazi than in the overall European population (8, 12), sample recruitment was easier.

Another advantage of examining the genetics of dystonia within this population is that the shared genetic background of the Ashkenazi increased the possibility of finding a shared genetic etiology that originated through a founder's effect. The number of haplotypes present in the Ashkenazi population is lower than what would be expected due to inbreeding and genetic drift. Furthermore, the Ashkenazi genetic background is comprised of both European and African haplotypes (13), which increases the likelihood that the results of the study will be generalizable. It has already been shown that the GAG deletion in *TOR1A* is linked to dystonia in both Ashkenazi and non-Jewish cases, so it is possible that any variants associated with focal

dystonia will be relevant in several populations. Thus, by electing to perform the GWAS within the Ashkenazi population, we increased our power to detect an association.

The study identified two SNPs on chromosome 17q25.3 with genome-wide significance: rs12601730 ($p=2.40x10^{-9}$, OR = 2.169) and rs12603583 ($p=2.61x10^{-9}$, OR=1.942). These markers were in strong LD with each other ($r^2=0.87$) Imputation to infer genotypes identified two more markers with genome-wide significant association signals that were in strong linkage disequilibrium with the two genotyped markers rs4889968 ($p=2.61 \times 10^{-9}$) and rs9902013 ($p=3.84 \times 10^{-9}$). All of these markers were located within an intronic region of *RNF213*, which encodes a protein of unknown function with a RING finger domain and a Walker domain. These findings provide evidence for a genetic etiology for focal dystonias and formulate insights for future research to describe the cause of this disease.

3.2 Methods

Sample recruitment

The samples were recruited in New York, New York and evaluated by Dr. Bressman, a movement disorder specialist, according to current clinical criteria. Only individuals who had four Ashkenazi grandparents were included in the study. The majority of cases were diagnosed with primary, idiopathic blepharospasm, and the rest were diagnosed with other focal dystonias. The controls were either healthy individuals or individuals who had been diagnosed with either idiopathic Parkinson's disease, Crohn's disease or schizophrenia.

Genotyping

All samples were genotyped on the Affymetrix Genome-Wide Human SNP Array 6.0 platform according to the manufacturer's protocol, and the genotype calls were made using the Birdseed V2 algorithm included in the Affymetrix Genotyping Console Software package.

Quality control of SNP Genotyping

Standard quality control measures were used to filter the genotyping calls with the PLINK software (14). Only SNPs with a minor allele frequencies > 0.05 and genotyping rates > 90% that did not violate Hardy-Weinberg equilibrium were included in the association analysis. After removing these SNPs, 634,338 markers remained.

To prevent type I error that resulted from aggressive genotyping calling by the Birdseed algorithm, we plotted the fluorescent probe intensities for each allele for each genotype for SNPs with p-values that were less than 1.0×10^{-5}. These scatter plots allowed us to determine if the genotype calls had been made inappropriately.

Analysis of Population Stratification

Population stratification is known to confound the results of association studies. To assess population stratification, the genomic control inflation factor,λ, was calculated using the EIGENSTRAT software suite (15) by selecting a subset of SNPs that were in linkage dysequilibrium with each other. The median of the observed Chi-square distribution of the association values for these SNPs was divided by the expected median of this distribution to obtain a value of 1.04, which was indicative of negligible population stratification.

To further rule out population stratification, principal components analysis was performed in EIGENSTRAT by comparing the study population with the CEU and YRI HapMap populations. The two principal components were plotted, and our population clustered together, reflecting the homogeneity of the study population.

Power Calculations

The QUANTO software was used to calculate the power of the study, assuming an additive model and a binary outcome. The number of cases needed to achieve 80% power to detect an association for a case:control ratio of 1:14 was computed for minor allele frequencies from 0.05 to 0.50 at increments of 0.025.

Association Analysis

The association analysis was performed with the mach2dat tool, associated with the MACH (Markov Chain Haplotyping) software package (16, 17). An association was considered to have genome-wide significance if the p-value was less than 2.0×10^{-8}; this threshold was determined using Bonferroni's correction ($p \sim 0.05/2{,}500{,}000$ markers).

Imputation

Additional genotypes were imputed using MACH. A set of reference haplotypes was created using the HapMap2 CEU population after first ensuring that the alleles from our study matched the HapMap alleles for SNPs that were found in both data sets (i.e., that they were genotyped on the same strand). The genotypes from the HapMap data were phased to produce a

set of reference haplotypes with the following parameters: --rounds 20, --states 300, --phase, --interim 5, --sample 5, --compact.

Then, this set of reference haplotypes was used to impute additional markers in our study data with the following options in MACH: --rounds 5, --states 200. For genotype inference, only SNPs genotyped on the Affymetrix platform with genotyping rates > 90% and minor allele frequencies > 5% that did not violate Hardy-Weinberg equilibrium ($p < 0.000001$) were included in the input files.

Plots

The Manhattan plot and Q-Q plot were produced using "qqman" in the R software environment. The plot of the association signal and linkage disequilibrium for *RNF213* was produced in Locus Zoom (18) using data from the 1000 Genomes project European population (hg19).

3.3 Results

The sample consisted of 105 cases that had been diagnosed with focal dystonias (mostly blepharospasm) and 1495 controls from the Ashkenazi population. We included individuals with focal dystonias other than blepharospasm because there is evidence that all focal dystonias share a genetic etiology (19). The controls were comprised of both healthy individuals and individuals who were diagnosed with unrelated diseases (Crohn's disease, idiopathic Parkinson's disease, and schizophrenia). Because of the small sample size, the study was expected to have sufficient power to identify associations with odds ratios that were comparatively higher than those found by most recent association studies. Power calculations indicated that the study had 80% power to

detect associations with odds ratios between 1.75 and 2.10, depending on the minor allele frequency.

After genotyping the samples on the Affymetrix Genome-Wide Human SNP Array 6.0 platform, both the subjects and markers were checked for missingness of genotypes. Three subjects, all controls, were excluded for having a genotyping rate less than 90%; the average genotyping rate was 96.1%. Markers were excluded using PLINK on the basis of low genotyping rates (genotyping rate <90%, 55,310 SNPs), low minor allele frequencies (MAF<0.05, 227,747 SNPS) and violating Hardy-Weinberg equilibrium (195,09 SNPS). After these steps, 634,338 autosomal SNPs were analyzed. To rule out the possibility of improper genotype calling by the Birdseed V2 algorithm, the genotype clustering plots of the Affymetrix probe intensities were visually inspected for all SNPs with $p < 1.0 \times 10^{-5}$. This step was necessary because the Birdseed algorithm will frequently assign genotypes to samples when the fluorescent probe intensity is low; thus, we needed to check that this aggressive genotype calling had not occurred. Any SNPs for which it appeared that the majority of genotypes had been called incorrectly were removed from further analysis.

The case-control analysis revealed 2 SNPs on chromosome 17q25.3, rs12601730 ($p=2.40 \times 10^{-9}$, OR = 2.169) and rs12603583 ($p=2.61 \times 10^{-9}$, OR=1.942), that had association signals with genome-wide significance ($p < 2.0 \times 10^{-8}$), as indicated by the Manhattan plot in Figure 3.1. These two markers were in strong linkage disequilibrium with each other ($r^2=0.87$). Both SNPs were located in an intronic region in *RNF213*. This gene encodes a RING finger protein of unknown function.

The Q-Q plot (Figure 3.2) of the association p-values indicates that the observed distribution of p-values only deviated slightly from the expected distribution of p-values. This

deviation largely occurred in the right tail end of the distribution, suggesting that there was negligible Type I error in the association analysis. The genomic control inflation factor, λ, was calculated as 1.04, which suggests that the results were not influenced by population substructure. Population substructure was further ruled out by performing principal components analysis in EIGENSTRAT with the European and African populations from HapMap II. The plot of the first two principal components was indicative of strong homogeneity within the study population (data not shown).

After performing the initial analysis, additional markers were imputed using MACH (MArkov Chain Haplotyping) for a total of 2,500,000 genotyped SNPs. To ensure that markers were not inappropriately excluded we checked for compatibility between the alleles in the HapMap data and the Affymetrix data. Two more markers were found to have associations with genome-wide significance: rs4889968 ($p=2.61 \times 10^{-9}$) and rs9902013 ($p=3.84 \times 10^{-9}$). These SNPs were in the same LD block with the other two significant SNPs on chromosome 17q25.3 (Figure 3.3). Thus, we identified a strong association between *RNF213* and dystonia. Below we posit the role that mutations within this gene could play.

3.4 Discussion

We performed the first genome-wide association study of idiopathic dystonia and found an association signal with genome-wide significance at chromosome 17q25.3. This association signal was located within an intronic region of *RNF213,* which encodes a RING (Really Interesting New Gene) finger protein. While these variants were found within an intronic region, it is possible that they are in LD with rare variants that could disrupt the protein-coding region of this gene. RING finger domains are stable structures between 40 and 60 amino acids that are able

to bind 2 zinc cations. More importantly, RING finger proteins can participate in ubiquitination by acting as E3 ubiquitin ligases. The disruption of ubiquitination, which could prevent proteins from being targeted for degradation, has been shown to play a role in a number of neurological disorders, such as Parkinson's disease and Alzheimer's disease. Recent research has suggested that mutations in *TOR1A*, the gene implicated in primary, early-onset torsion dystonia, disrupt ubiquitin-mediated protein degradation and lead to the formation of protein aggregates within the brain. Thus, mutations in *RNF213* could act similarly by disrupting ubiquitination and preventing appropriate protein degradation.

The RNF213 protein also contains a Walker motif, which could confer ATP-binding activity. Interestingly, *TOR1A* codes for a protein that also contains an ATP-binding domain. The causative GAG-deletion is believed to act by disrupting the ATP-binding activity of the protein.

RNF213 was recently implicated in the first genome-wide association study of Moyamoya disease (MMD) in a Japanese population (20). In patients with MMD, blood vessels in the brain are blocked, causing the formation of auxiliary blood vessels that are weak and prone to hemorrhage. In this particular GWAS, the researchers found an association signal with genome-wide significance for a 7-SNP haplotype that was not in LD with the SNPs implicated in this study. In forwarding the functional characterization of the protein, these researchers performed a preliminary analysis of mRNA and protein expression in several tissues through RT-PCR and Western blotting, respectively. *RNF213* mRNA was highly expressed in the skeletal muscle and the cerebellum, but not in the whole brain. This expression in the cerebellum is of interest because the latest models of dystonia suggest involvement of both the cerebellum and basal ganglia.

At first glance, it would appear that dystonia and MMD are largely unrelated. However, models for both diseases implicate the basal ganglia. Interestingly, there have been a few cases of MMD patients developing secondary movement disorders, including dystonia (21). MMD patients presenting with secondary dystonia were shown to have lesions in the basal ganglia. Thus, mutations in *RNF213* may play a role in both conditions.

More research is needed to characterize the expression of *RNF213* and to verify its role in dystonia. A better understanding of this gene will lead to a better understanding of the pathways and regions of the brain involved in this disease.

Contributions

Within this work, I performed the Birdseed genotyping, quality control of genotyping, development of the genotype correction tool, association analysis, imputation, population stratification analysis, general data analysis, and drafting of the manuscript. Dr. Kirk Wilhelmsen provided oversight and mentorship; he also developed software tools used for the quality control of genotyping calls and imputation. Dr. Scott Chasse assisted with the Affymetrix genotyping. Dr. Susan Bressman and Dr. Laurie Ozelius provided the samples and some of the Affymetrix genotyping files.

3.5 REFERENCES

1. Ozelius LJ, Hewett J, Kramer P, et al. Fine localization of the Torsion Dystonia Gene (DYT1) on Human Chromosome 9q34: Yac Map and Linkage disequilibrium. Genome Research. 1997; 7:483-494.

2. Ozelius LJ, Hewett JW, Page CE, et al. The early-onset torsion dystonia gene (DYT1) encodes an ATP-binding protein. Nat Genet. 1997; 17: 40–48.

3. Calakos N, Patel VD, Gottron M, et al. Functional evidence implicating a novel TOR1A mutation in idiopathic, late-onset focal dystonia. J Med Genet. 2010; 47(9):646-50.

4. Sharma N, Franco RA Jr, Kuster JK, et al. Genetic evidence for an association of the TOR1A locus with segmental/focal dystonia. Mov Disord. 2010; 25(13):2183-7.

5. Bressman SB, de Leon D, Brin MF, Risch N, Burke RE, Greene PE. Idiopathic dystonia among Ashkenazi Jews: evidence for autosomal dominant inheritance. Ann Neurol. 1989; 26:612-620.

6. Bressman SB. Dystonia genotypes, phenotypes, and classification. Adv Neurol. 2004; 94:101–107.

7. Pauls D, Korczyn A. Complex segregation analysis of dystonia pedigrees suggests autosomal dominant inheritance. Neurology. 1990; 40:1107-1110.

8. Zilber N, Korczyn AD, Kahana E, Fried K, Alter M. Inheritance of idiopathic torsion dystonia among Jews. J Med Genet. 1984; 21:13-20.

9. Defazio G, Martino D, Aniello MS, et al. A family study on primary blepharospasm. J Neurol Neurosurg Psychiatry. 2006; 77(2):252-4.

10. Leube B,Kessler, KR,Goecke, T,Auburger, G,Benecke, R. Frequency of familial inheritance among 488 index patients with idiopathic focal dystonia and clinical variability in a large family. Mov Disord. 1997; 12*:* 1000–1006.

11. Warner T, Camfield L, Marsden CD et al. A prevalence study of primary dystonia in eight European countries. J Neurology. 2000**;** 247 (10): 787 – 792.

12. Eldridge R. The torsion dystonia: literature review: genetic and clinical studies. Neurology. 1970; 20:1-78.

13. Hammer MF, Redd AJ, Wood ET, et al. Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. Proc Natl Acad Sci USA 97. 2000;(12):6769-74.

14. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 2007; 81(3): 559–575.

15. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904-909.

16. Li Y, Willer CJ, Ding J, Scheet P and Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34:816-834.

17. Li Y, Willer CJ, Sanna S and Abecasis GR. Genotype Imputation. Annu Rev Genomics Hum Genet. 2009; 10:387-406.

18. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18): 2336.2337.

19. Defazio G, Berardelli A, Hallett M. Do primary adult-onset focal dystonias share aetiological factors? Brain. 2007; 130:1183–1193.

20. Kamada F, Aoki Y, Narisawa A, et al. A genome-wide association study identifies RNF213 as the first Moyamoya disease gene. Journal of Human Genetics. 20011; 56 (1):34–40.

21. Baik JS, Lee MS. Movement disorders associated with moyamoya disease: A report of 4 new cases and a review of literatures. Movement Disorders. 2010;25(10):1482-86.

Figure 3.1: Manhattan plot of the $-\log_{10}$(p-values) of the dystonia GWAS.

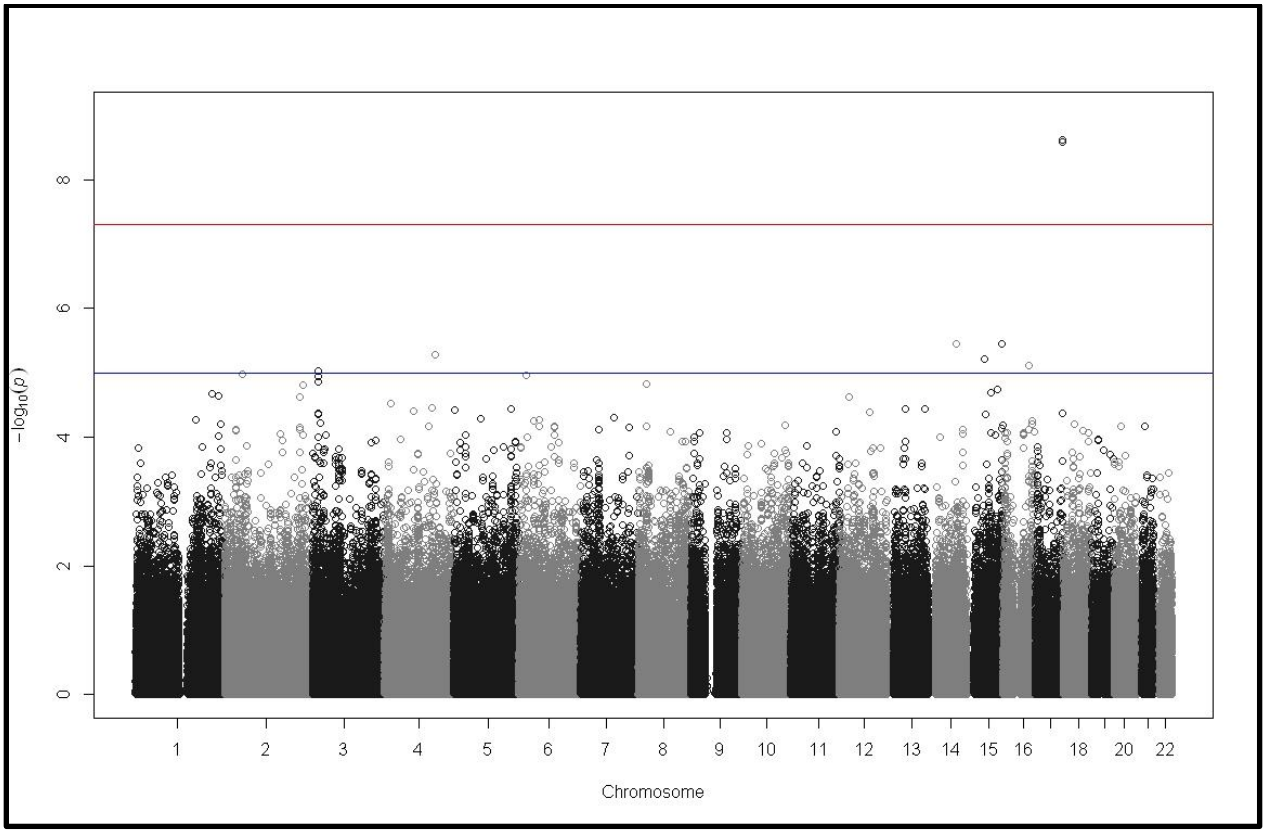


Figure 3.2: Quantile-quantile plot of the observed vs. the expected distribution of p-values for the dystonia GWAS.

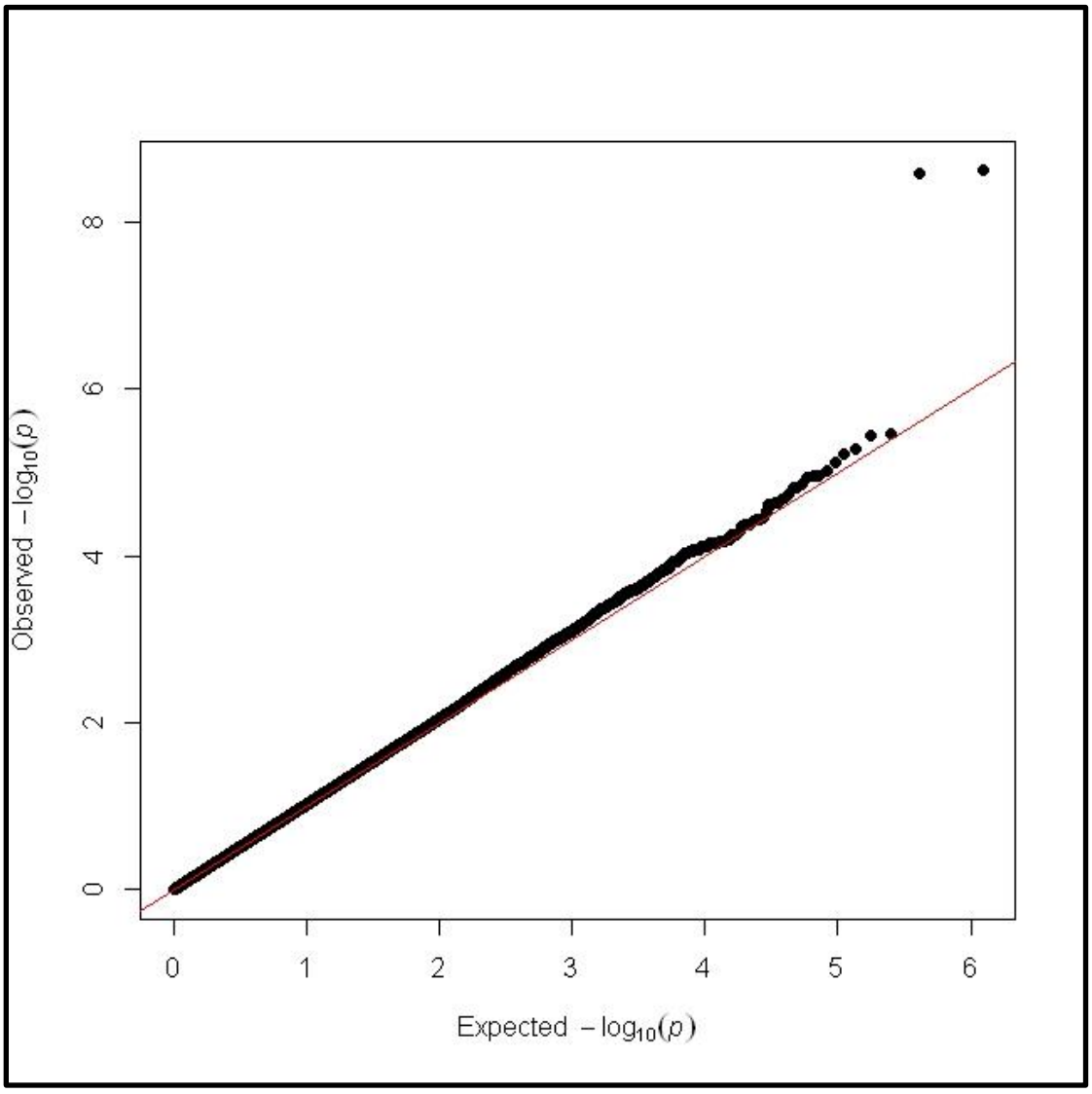
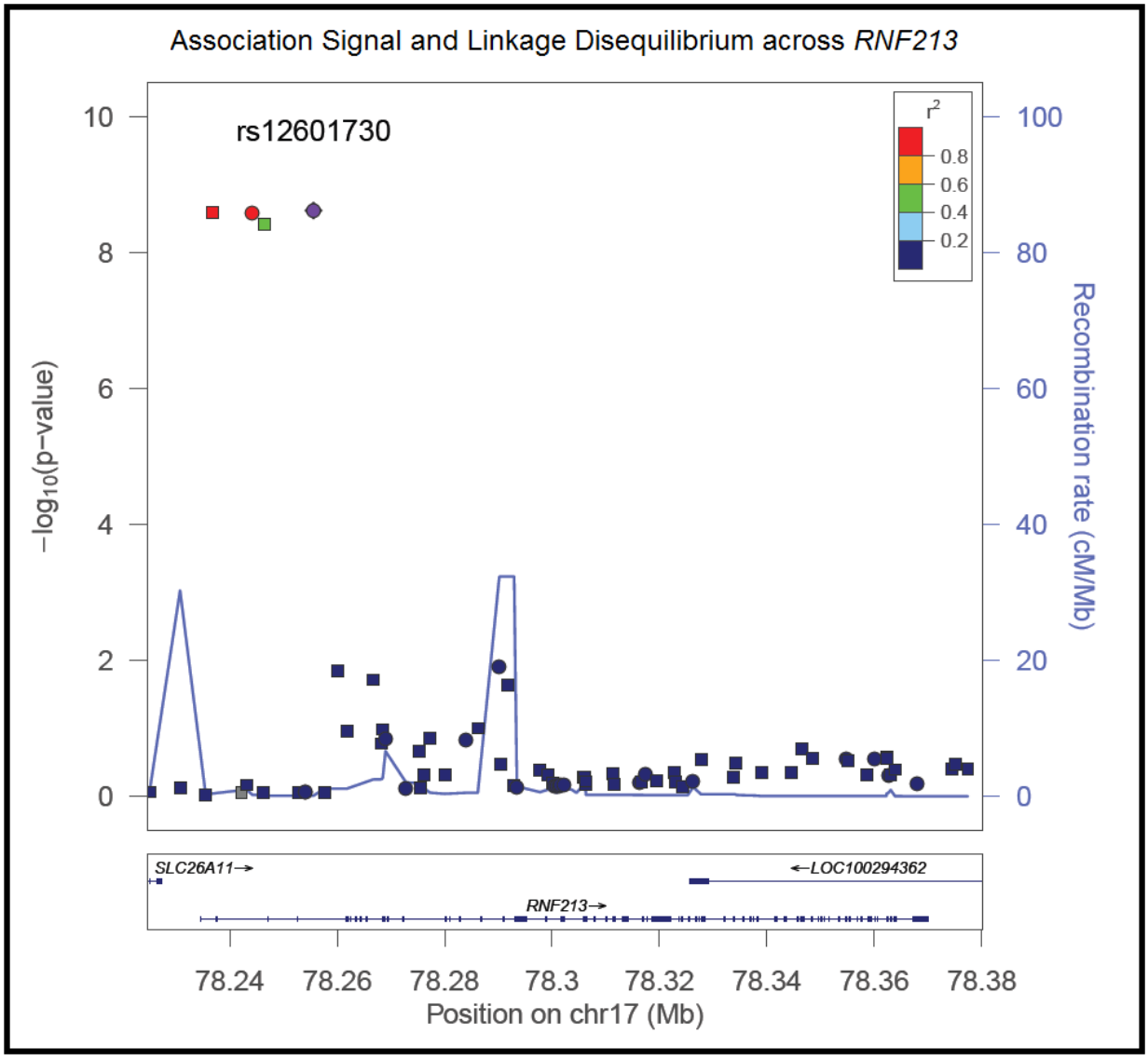


Figure 3.3: Plot of the association signal across *RNF213*. The imputed SNPs are marked as a square, while the genotyped SNPs are marked with a circle. The SNP with the most significant association signal (rs12601730) is displayed in purple.



Chapter 4

Combined Exome and Whole Genome Sequence Analysis of FTD-ALS family San Francisco-A

4.1 Introduction

Recent research strongly suggests that amyotrophic lateral sclerosis (ALS), a neurodegenerative disease characterized by progressive motor neuron loss (1), and frontotemporal dementia (FTD), the second leading cause of pre-senile dementia after Alzheimer's disease (2), are different manifestations of the same genetic cause. Approximately 15% of FTD patients will also have amyotrophy (3), and many patients that present with amyotrophy have loss of frontal lobe cortical neurons in the distribution typical of FTD (4).

The ALS-FTD disease spectrum is genetically heterogeneous. A common cause of familial and sporadic cases of FTD-ALS that segregates with chromosome 9p21.2 is a hexanucleotide repeat (GGGGCC) expansion in the noncoding region of the uncharacterized gene *C9orf72* (5-7). The transcribed product of this expanded repeat has been found in the nuclei of affected individuals; the transcript is translated into an aggregated dipeptide repeat protein (8). The loci responsible for a substantial portion of familial FTD cases are linked to 17q21 and can be attributed to mutations of the microtubule-associated protein tau gene (*MAPT*) (9) or progranulin gene (*GRN*) (10). A few patients with *MAPT* mutations present with amyotrophy, but most present with FTD (11). FTD-ALS, however, is typically not characterized by tau inclusions; the pathology is instead classified as ubiquitin-positive (12). There are other mutations known to cause autosomal dominant FTD and/or ALS, such as those in *TARDBP* (13),

ATXN2 (14), *VCP* (15), *CHMP2B* (16), *FUS* (17), *DCTN1* (18), *SOD1* (19), *ANG* (20), and *VAPB* (21).

Our study addresses the genetics of FTD-ALS in a family, henceforth called San-Francisco A (SF-A), with an atypical pathology. Previously, an autopsy of one family member revealed both tau and alpha-synuclein inclusions (12). While FTD is regularly characterized by tau inclusions, the ubiquitin-positive inclusions in ALS patients can be comprised by SOD1, TDP-43, or FUS. At the time of this report, there was no tissue available for family SF-A for re-evaluation of the presence of these inclusion proteins.

The unique pathology observed in family SF-A suggests that the genetic etiology of FTD-ALS is also unique in this family. Microsatellite data had yielded a linkage signal on chromosome 17q (12). However, the *MAPT* locus was not within the support interval. Furthermore, resequencing of the *MAPT* exons in 3 affected individuals in SF-A did not discover any novel or known sequence changes, suggesting that the tau inclusions may not have been caused by mutations in the tau gene. Therefore, additional studies were needed to identify the mutation responsible for the disease in family SF-A.

To this end, the current study performed whole exome sequencing (WES) of ten family members and whole genome sequencing (WGS) of one affected family member with the hope of discerning the genetic etiology of this form of FTD-ALS. By first sequencing the exons of the entire genome using massively parallel sequencing (MPS) technology, this study sought to identify the causative mutation responsible for the disease in SF-A. While it is possible to sequence the entire genome, 85% of the mutations that explain diseases with simple modes of inheritance are in coding sequences, which comprise approximately 1% of the genome (22-23).

WGS of one subject was pursued in addition to WES to identify potentially interesting variants outside of the coding regions.

After submitting the samples for MPS, a family informant interviewed ten years after the last assessment indicated that two individuals, who were previously believed to be affected, that had substantial behavioral problems had not progressed over several years and have now been lost to clinical follow-up.

In this study, WES and WGS failed to detect a rare splice junction, missense or nonsense mutation in the affected family members for any genes on 17q, known genes that cause FTD and/or ALS, and the entire exome. Furthermore, repeat-primed PCR did not detect the presence of the *C9orf72* hexanucleotide repeat expansion. Multipoint linkage analysis with 163,082 polymorphisms detected by WES now excludes 17q; with the updated phenotype data, a linkage signal on chromosome 9q32-34 that segregated with the disease (LOD=1.8) was identified. While this LOD score was below the threshold conventionally used to prove linkage, the rest of the genome was exluded. In the potentially linked region on chromosome 9q, a mutation not detected in thousands of genomes was shared by all the affected individuals in *CRB2*; the protein encoded by this gene has been shown to bind presenilin and suppress the activity of the gamma-secretase complex. While the c.C3459T mutation is synonymous, it may affect gene expression. Furthermore, WGS of one affected individual to identify additional variants of interest in regions beyond the targeted exome identified a 27-bp deletion (g.61699_61725del27) in *LAMC3*, a gene involved in the development of the cerebral cortex. Thus, WGS and WES failed to identify a likely causal mutation, which suggests that the mutation responsible for FTD-ALS in family SF-A could not be detected by the MPS approaches used in this study.

4.2 Methods

Subjects

The pedigree investigated in this study had been previously characterized by Wilhelmsen et al. (12). Recently, a family representative and the spouses of several family members that have died in the interval were interviewed. All but two of the previously identified affected family members progressed and died of FTD/ALS. Two family members (listed as III-1 and III-7 in the pedigree in Figure 4.1) with behavioral disturbances reportedly did not progress over several years and have been lost to follow-up. One of these individuals is now believed to have had longstanding behavioral disturbances that were possibly due to viral encephalitis. DNA was extracted from whole blood samples for 10 family members using a DNA isolation kit (Puregene; Gentra Systems Inc, Minneapolis, MN).

C9orf72 Locus Genotyping

DNA from ND07489 was obtained from Corielle to act as a positive control. Repeat-primed PCR was performed largely as described in Renton et al. and Dejesus-Hernandez et al. (5-6). The PCR fragment lengths were analyzed with an ABI 3730xl genetic analyzer and the Peak Scanner software.

Massively Parallel Sequencing

For the WES, genomic DNA samples from ten subjects (see Figure 4.1) were selectively enriched for exomic DNA using the SureSelect Human All Exon 50 Mb kit (Agilent, Santa Clara, CA). Illumina Hiseq2000 adapter sequences (Illumina, San Diego, CA) were ligated to the DNA fragments. Library preparation was then conducted according to the manufacturer's

instructions. Next, 76-bp paired end reads were generated with the Illumina HiSeq2000 platform and CASAVA v1.8 pipeline.

For the WGS, genomic DNA for one affected subject (listed as II-6 in Figure 4.1) was submitted to massively parallel sequencing on the Illumina HiSeq 2500 platform after library preparation, and 100-bp paired end reads were generated.

The reads were aligned to the reference genome (hg19) using the Burrows-Wheeler aligner (BWA) (24). The resulting SAM (Sequence Alignment/Map) file was sorted using SAMtools and converted to a BAM (binary SAM) file (25). Duplicate reads were removed using Picard (26). The resulting BAM files were processed with the GATK pipeline v.2.0 for multi-sample variant detection (27). To improve the quality of genotype calls, the genotypes for the 10 exomes were called in concert with the genotypes from a set of 100 exomes from unaffected individuals of European descent using the GATK Unified Genotyper (28). The variants were annotated using the SeattleSeq 134 pipeline and the pipeline constructed by the bioinformatics group at the Renaissance Computing Institute (RENCI, www.renci.org) directed by KCW that was constructed for the University of North Carolina at Chapel Hill NHGRI-funded NCGENES project, which is assessing the utility of WES in medical genetics.

Linkage Analysis

To create the marker set (163,082 single nucleotides polymorphisms) used for linkage analysis, the PLINK software was used to identify markers in linkage equilibrium from WES data for 100 individuals of European descent (29). Markers were further excluded for having a minor allele frequency less than 5%.

The Merlin software was used to confirm the expected relationship between family members (30). Next, Merlin's genotype error detection algorithms were used to examine genotype data, blind to phenotype data, to identify and exclude rare markers inconsistent with Mendelian segregation. Multipoint linkage analysis was performed in Merlin, assuming an autosomal dominant pattern of inheritance with complete penetrance. This pattern of inheritance was plausible because individuals from all generations were affected, and there was male-to-male transmission of the disease. The frequency of the potential disease-conferring allele was set to 0.001. For the genetic map files, the genetic distances were derived from the Rutgers Map Project (31).

Haplotype Analysis

Haplotype analysis of chromosome 17q was performed using the fastIBD function in Beagle with unphased data (32).

Causal Variant Determination

After performing the multipoint linkage analysis, variants with a minor allele frequency greater than 5% in dbSNP or 1000 Genomes data were excluded from further analysis. Variants that were present in the four affected individuals, but not in any of the unaffected individuals, were considered candidates. Variants were stratified based on whether they were expected to change the primary amino acid sequence or affect splicing.

RNA Secondary Structure Analysis

The secondary structure of the mRNA for *CRB2* was investigated using the RNA folding form application on the mFold web server at http://mfold.rna.albany.edu/?q=mfold/RNA-Folding-Form (33).

Exonic Splicing Regulatory Element Analysis

The wildtype sequence and the sequence with the point mutation in exon 11 of *CRB2* were submitted to the ESRsearch tool at http://esrsearch.tau.ac.il/ (34).

4.3 Results

The exomes of ten individuals from family SF-A were analyzed with targeted capture and sequencing (Figure 4.1). A previous study of this family had indicated the presence of linkage on chromosome 17q21, which was distal to *MAPT*, and an autopsy of the brain of an affected family member indicated the presence of both tau and alpha-synuclein inclusions. However, resequencing of the *MAPT* exons in 3 affected individuals in SF-A did not discover any novel or known sequence changes, suggesting that the tau inclusions may not have been caused by mutations in the tau gene. Ten years after this initial study, follow-up with a family informant indicated that two individuals who were believed to be affected in this pedigree had not progressed over several years and were lost to follow-up (listed as III-1 and III-7 in the pedigree in Figure 4.1). The informant reported that one of the previously affected individuals with an unspecified behavioral disorder had episodes of what was diagnosed as viral encephalitis during adolescence and again as a young adult. The other patient lost to follow-up had a long history of adult onset behavioral problems. This change in diagnostic status for these two individuals

prompted a review of the previously performed linkage analysis. For the purpose the current study, these two individuals were classified as unaffected.

Thus, a genome-wide linkage analysis was performed with SNP genotypes taken from the WES data assuming an autosomal dominant mode of inheritance. The present multipoint linkage excluded the 17q region, as the LOD scores across the region were negative. However, because of the importance of *MAPT* mutations in neurodegenerative diseases and the discovery of tau inclusions in the brain during an autopsy of a deceased family member in SF-A, variants in or near *MAPT* were investigated. Variants were selected that were present in the genomes of the 4 affected individuals, but not the 4 unaffected individuals. In doing so, a rare, non-synonymous mutation was detected in *LRRC37A2*, which is approximately ~300 kb distal from *MAPT*. A haplotype analysis performed with Beagle, however, indicated that the haplotypes in this region did not segregate with disease.

After excluding the rest of the genome by linkage analysis, a multipoint linkage signal was found in a 20.2 cM region on chromosome 9q32-34 (LOD=1.8) (Figure 4.2). While a LOD score of 1.8 is below the threshold conventionally used to prove linkage, this was the maximum possible LOD score in this family. No LOD scores above 1.0 were detected in any other region in the genome. The region is consistent with linkage and is presently the most likely region to harbor the mutation responsible for disease in SF-A. The discrepancy in the findings from the current study and the previous one can be explained by the change in the diagnostic status of two individuals between the previous report and the current one and by the extreme density of marker data in the current study, which allowed for greater confidence in concluding that the affected individuals shared specific haplotypes.

Interesting variants in the linked region were sought by looking for rare variants or mutations that were shared among the genomes of the affected individuals and missing from those of the unaffected individuals. Rare variants were defined as variants that were either not present in dbSNP, the 1000 Genomes data, or the EVS or that had a minor allele frequency less than 5%. In the linked region, one such variant was found, a single nucleotide variant in exon 11 of *Crumbs Homolog 2* (*CRB2*, c.C3459T*)*.

The mutation in *Crumbs* is not listed in the NHLBI Exome Sequencing Project Exome Variant Server (EVS). *Crumbs* is an intriguing candidate because of its role in Notch signaling and processing the amyloid precursor protein (APP). If the synonymous c.C3459T mutation in *CRB2* is the causal variant in SF-A, this mutation may affect gene expression, e.g., by altering the stability of the mRNA secondary structure. The most stable secondary structure of the primary *CRB2* transcript with c.C3459T is 5.01% less stable in terms of its folding free energy (ΔG=-43.6) than the predicted structure of the wildtype transcript (ΔG=-45.9). Another possibility is that this synonymous mutation could disrupt an exonic splicing enhancer (ESE) or introduce an exonic splicing silencer (ESS). A search for putative ESE and ESS motifs in the region revealed that the variant introduced a putative ESS (TGGTTCTC) [34]. An ESS could contribute to alternative mRNA splicing (35).

Because the LOD score for chromosome 9q was below the conventional threshold needed to prove linkage, the entire exome was interrogated for variants of interest that segregated with the disease, but no additional variants of interest were detected. It is also likely that the causative mutation for FTD-ALS in family SF-A is simply not located in the exomic regions that were sequenced. The hexanucleotide repeat expansion in the intronic region of *C9orf72* is an example of a mutation that causes many familial and sporadic cases of FTD-ALS. The region containing

this locus did not segregate with the disease (Figure 4.2), which suggested that this locus was not implicated in family SF-A. Repeat-primed PCR revealed that this family did not carry the *C9orf72* repeat expansion (Figure 4.3). Repeat-primed PCR was carried out for all 10 family members that were submitted to exome sequencing. The plots of the fluorescence peaks for the PCR fragments indicated the presence of fewer than 30 GGGGCC repeats in this region for all family members, which is considered normal.

To investigate the possibility that the causative variant occurred outside of the exome, the entire genome of one affected subject (II-6) was sequenced. The WGS data confirmed the previously detected *CRB2* mutation and also detected a 27-bp deletion(g.61699_61725del27) in an intronic region of *LAMC3*. These findings are summarized in Table 4.1. The MAF of this variant is not available in dbSNP, 1000 Genomes, or the EVS. The protein encoded by *LAMC3*, laminin subunit gamma-3, is expressed in the brain and plays a role in shaping the cerebral cortex.

In summary, a multipoint linkage signal (LOD=1.8) was detected on chromosome 9q32-34 for FTD-ALS with tau and alpha-synuclein aggregates in family San Francisco-A. In the affected individuals, this region contained a rare variant in *CRB2*, a gene in the gamma-secretase pathway.

4.4 Discussion

In the current study, a novel locus on chromosome 9q32-34 segregated with a familial form of FTD-ALS. While several studies have linked chromosome 9p to the disease, only one other study has observed linkage with chromosome 9q (36). In particular, several studies have cited a hexanucleotide repeat expansion in *C9orf72* (9p21.2) as being implicated in both familial

and sporadic forms of the disease. However, the *C9orf72* region failed to segregate with the disease and was not detected by repeat-primed PCR in this family. These findings underscore the genetic heterogeneity of the FTD-ALS disease spectrum.

This study also illustrates the advantages and limitations of WES in the study of Mendelian diseases. WES data did provide greater resolution for performing genome-wide linkage analysis, but this approach failed to identify a potentially causative splice, nonsense, or missense mutation. Thus, it is likely that the causative mutation was not located within the exomes of the affected family members.

While the rare variant in *Crumbs homolog 2 (CRB2*) discovered in the affected subjects in SF-A was synonymous, this variant could help explain the pathology of FTD-ALS. In humans, CRB2 inhibits the cleavage activity of the gamma-secretase complex; this protease complex is responsible for the cleavage of the transmembrane domains of proteins including APP and various Notch receptors. A study by Mitsuishi et al. found that CRB2 binds presenilin, one of the four main components of the gamma-secretase complex (37), and inhibits the cleavage of APP. APP is cleaved by the gamma-secretase complex into amyloid-beta, which is the primary component of amyloid plaques found in the brains of Alzheimer's patients. Knockdown of endogenous CRB2 led to an increase in gamma-secretase cleavage activity (37). In Drosophila, the Crumbs protein has been shown to attenuate Notch-mediated transcriptional activation by preventing the cleavage of the Notch receptor by the gamma-secretase complex (38).

Thus, this synonymous mutation in *CRB2* could, via the gamma-secretase complex, affect the cleavage of APP and the Notch receptor. While the role of amyloid-beta in Alzheimer's disease has been thoroughly investigated, more recently, Notch has also been implicated in

neurodegenerative pathways. Specifically, Notch-1 signaling has been shown to suppress the formation of new motor neurons in zebrafish (39), and the loss of motor neurons is a hallmark of ALS. Furthermore, TDP-43, which is now believed to be the primary component of ubiquitin inclusions in the brains of patients with FTD-ALS, has been shown to upregulate several Notch target genes in *Drosophila* (40). If the mutation in *CRB2* found in this study modifies protein function and prevents CRB2 from suppressing gamma-secretase cleavage activity, this mutation could lead to a deregulation of Notch signaling and the subsequent loss of motor neurons.

The 27-bp deletion detected by WGS in an intronic region of *LAMC3* is interesting because of the role that this gene plays in the development of the cerebral cortex (41). However, it is unclear whether this variant affects the expression or amino acid sequence of this gene. A functional characterization of the detected mutation could determine whether this variant contributes to the etiology of FTD-ALS in family SF-A.

Thus, a likely causative mutation was not found by either WES or WGS. The ability to genotype rare variants did lead to the identification of potentially interesting variants in *CRB2* and *LAMC3*; however, these variants will need to be characterized to determine whether they contribute to the phenotype in family SF-A. Current MPS strategies are limited in their ability to detect repeat expansions, which have been shown to play a role in neurodegenerative conditions. In the future, it would be prudent to investigate whether a repeat expansion could explain the genetic etiology of FTD-ALS in family SF-A.

Contributions

Within this study, I performed the repeat-primed PCR experiment, the alignment and variant calling of the exome and whole genome sequencing data, the linkage analysis, the

haplotype analysis, the causal variant analysis, the *in silico* functional characterization of the variants, and the drafting of the manuscript. Drs. Kirk Wilhelmsen, Loren Alving, and Bruce Miller identified the family, performed the work described in the previous report, and provided oversight of the current report. Dr. Joshua Sailsbery assisted with the bioinformatics analysis. Dr. Piotr Mieczowski and the High Throughput Sequencing Facility at UNC-Chapel Hill performed the library preparation and generated the Illumina sequence data.

4.5 REFERENCES

1. Rowland LP and Shneider NA. Amyotrophic lateral sclerosis. *N Engl J Med.* 2001; 344(22):1688-1700.

2. Talbot K and Ansorge O. Recent advances in the genetics of amyotrophic lateral sclerosis and frontotemporal dementia: common pathways in neurodegenerative disease. *Human Molecular Genetics.* 2006; vol. 15, no. Review Issue No. 2, R182-R187.

3. Strong MJ, Grace GM, Freedman M, et al. Consensus criteria for the diagnosis of frontotemporal cognitive and behavioural syndromes in amyotrophic lateral sclerosis. *Amyotroph Lateral Scler*. 2009;10(3):131-146.

4. Murphy JM, Henry RG, Langmore S, Kramer JH, Miller BL, Lomen-Hoerth C. Continuum of frontal lobe impairement in amyotrophic lateral sclerosis. *JAMA Neurology*. 2007;64(4):530-4.

5. Renton AE, Majounie E, Waite A, et al; ITALSGEN Consortium. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked FTD-ALS. *Neuron*. 2011;72(2):257-268.

6. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*. 2011;72(2):245-256.

7. Majounie E, Renton AE, Mok K, et al. Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *The Lancet Neurology*. 2012; 11 (4) 323–330.

8. Mori K, Weng S-M, Arzberger T, et al. The *C9orf72* GGGGCC Repeat Is Translated into Aggregating Dipeptide-Repeat Proteins in FTLD/ALS *Science*. 15 March 2013; 339: 1335-1338.

9. Hutton M, Lendon CL, Rizzu P, *et al.* Association of missense and 5'-splice-site mutations in *tau* with the inherited dementia FTDP-17. *Nature.* 1998; 393(6686): 702–705.

10. Baker M, Mackenzie IR, Pickering-Brown SM, et al. Mutations in progranulin cause tau-negative frontotemporal dementia linked to chromosome 17. *Nature. 2006;* 442(7105):916-9.

11. Wilhelmsen KC, Lynch T, Pavlou E, Higgins M, Nygaard TG. Localization of disinhibition-dementia-parkinsonism-amyotrophy complex to 17q21-22. *Am J Hum Genet.* 1994; 55(6): 1159-65.

12. Wilhelmsen KC, Forman MS, Rosen HJ, et al. 17q-Linked Frontotemporal Dementia–Amyotrophic Lateral Sclerosis Without Tau Mutations With Tau and α-Synuclein Inclusions. *Arch Neurol*. 2004;61(3):398-406.

13. Neumann M, Sampathu DM, Kwon LK, et al. Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. *Science*. 2006:314:130-133.

14. Elden AC, Kim H-J, Hart MP, et al. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature*. 2010: 466:1069-75.

15. Johnson JO, Mandrioli J, Benatar M, et al. Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron*. 2010; 68(5):857-64.

16. Skibinski G, Parkinson NJ, Brown JM, et al. Mutations in the endosomal ESCRTIII-complex subunit CHMP2B in frontotemporal dementia. *Nat Genet*. 2005 Aug; 37(8):806-8.

17. Kwiatkowski TJ, Bosco DA, LeClerc LA, et al. Mutations in the FUS/TLS Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis. *Science*. 2009; 323(5918): 1205-8.

18. Munch C, Rosenbohm A, Sperfeld AD, et al. Heterozygous R1101K mutation of the DCTN1 gene in a family with ALS and FTD. *Ann. Neurol*. 2005;58:777–780.

19. Abe K, Aoki M, Ikeda M, Watanabe M, Hirai S, Itoyama Y. Clinical characteristics of a familial amyotrophic lateral sclerosis with Cu/ZN superoxide dismutase gene mutations. *J Neurol Sci*. 1996; 136(1-2): 108-16.

20. Greenway MJ, Andersen PM, Russ C, et al. ANG muations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nat Genet*. 2006; 38(4):411-3.

21. Nishimura AL, Mitne-Neto M, Silva HCA, et al. A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. A*m J Hum Genet*. 2004; 75(5):822-831.

22. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461 (7261): 272–276.

23. Choi M, Scholl UI, Ji W, et al.Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci*. 2009;106 (45): 19096–19101.

24. Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*. 2009; 25:1754-60.

25. Li H., Handsaker B., Wysoker A., et al. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. 2009: 25, 2078-9.

26. Picard. 18 May 2009. Available at http://picard.sourceforge.net, Accessed 8 April 2013.

27. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297-303.

28. DePristo M, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*. 2011;43:491-498.

29. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*. 2007 September; 81(3): 559–575.

30. Abecasis GR, Cherny SS, Cookson WO and Cardon LR. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*. 2002;30:97-101.

31. Matise TC, Chen F, Chen W, et al. A second-generation combined linkage physical map of the human genome. *Genome Res*. 2007;17:1783-6.

32. Browning, S. R. and B. L. Browning. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. American *Journal of Human Genetics*. 2007;81:1084-1097.

33. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. 2003;31(13), 3406-15.

34. Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T and Ast G. Comparative analysis identifies exonic splicing regulatory sequences - the complex definition of enhancers and silencers. *Mol Cell*. 23 June 2006; 22:769-781.

35. Zhang XH and Chasin LA. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*. 2004;18(11):1241-50.

36. Hosler, B A, Siddique T, Sapp PC, et al. Linkage of familial amyotrophic lateral sclerosis with frontotemporal dementia to chromosome 9q21-q22. *JAMA*. 200. 284: 1664-1669.

37. Mitsuishi Y, Hasegawa H, Matsuo A, et al. Human CRB2 inhibits gamma-secretase cleavage of amyloid precursor protein by binding to the presenilin complex. *Journal of Biological Chemistry*. 2010;285(20):14920–14931.

38. Herranz H, Stamataki E, Feiguin F, Milán M. Self-refinement of Notch activity through the transmembrane protein Crumbs: modulation of gamma-secretase activity. The EMBO Reports. 2006;7(3):297–302.

39. Dias TB, Yang YJ, Ogai K, Becker T, Becker CG. Notch signaling controls generation of motor neurons in the lesioned spinal cord of adult zebrafish. *Neurosci*. 2012 Feb 29;32(9):3245-52.

40. Zhan L, Hanson KA, Kim SH, Tare A, Tibbetts RS. Identification of Genetic Modifiers of TDP-43 Neurotoxicity in Drosophila. *PLOS One*. 2013; 8(2): e57213.

41. Barak T, Kwan KY, Louvi A, et al. Recessive LAMC3 mutations cause malformations of occipital cortical development. Nature Genetics. 2011 Jun;43(6):590-4.

Figure 4.1: Pedigree affected with FTD-ALS. Subjects labeled with a "+" were analyzed with exome sequencing. The diagnostic status of subjects III:1 and III:7 had changed since the previous study; both of these subjects were previously believed to be affected but have had no progression since the last report. Therefore, they are now considered to be unaffected.

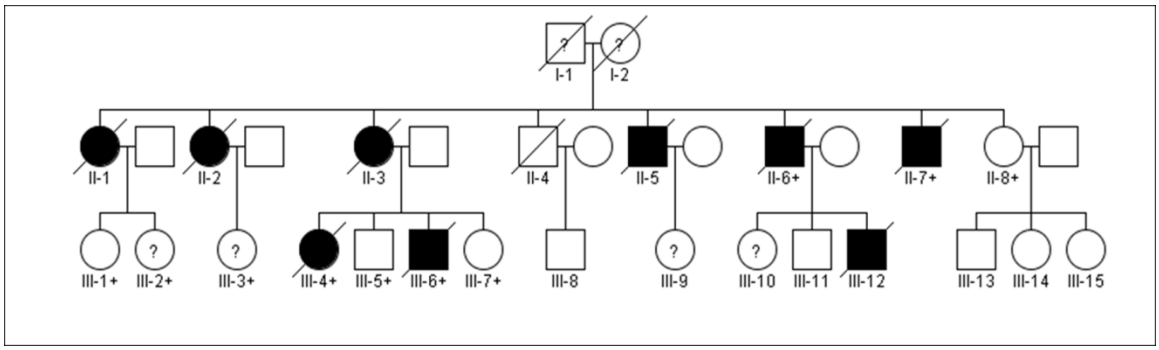


Figure 4.2: Multipoint LOD scores across chromosome 9 for FTD-ALS family San Francisco-A. A LOD score of 1.8, assuming an autosomal dominant mode of inheritance, was seen across a 20 cM region of 9q32-34. While this LOD score is below the threshold conventionally required to prove linkage, the rest of the genome was excluded.

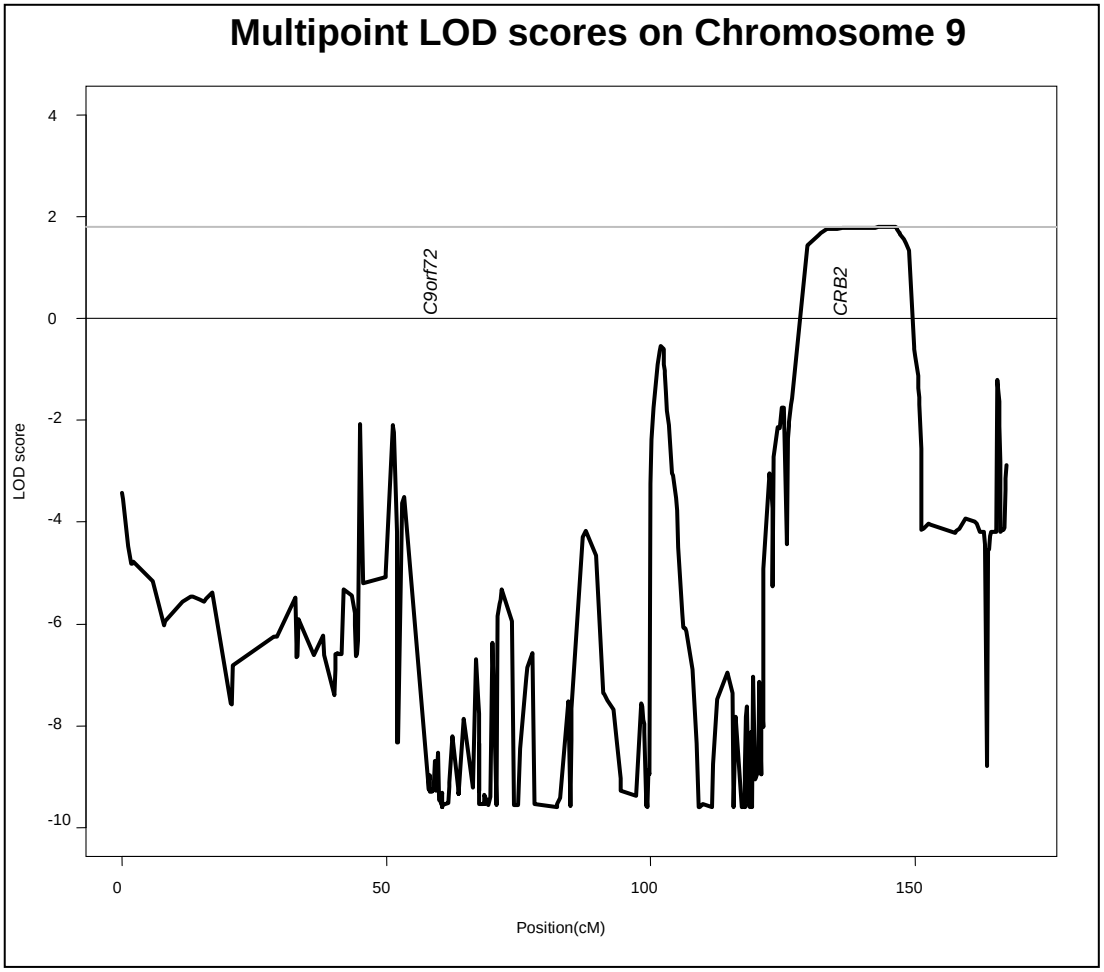
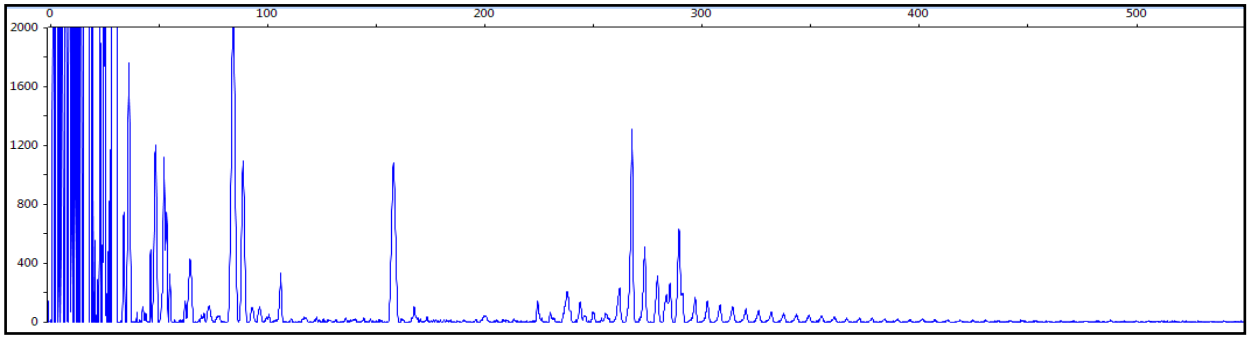
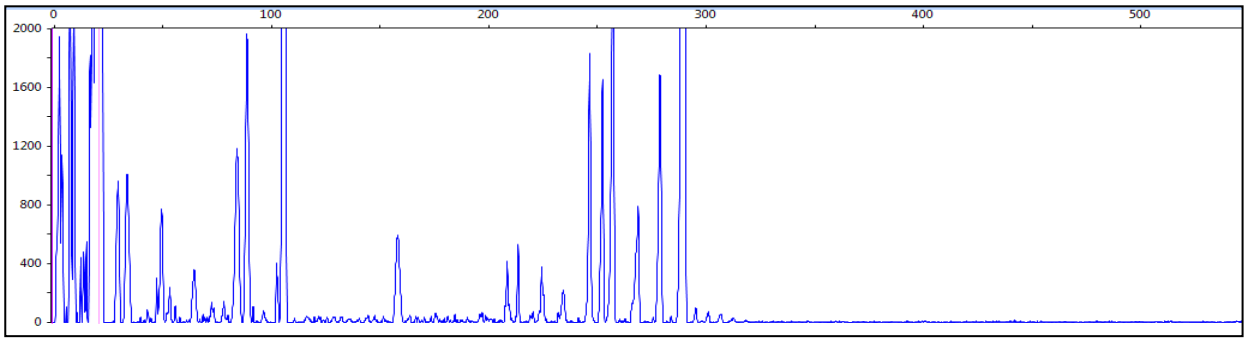


Figure 4.3: Repeat-Primed PCR of the *C9orf72* hexanucleotide repeat. The x-axes correspond to the size of the PCR products in base pairs, and the y-axes correspond to the fluorescence units. The sawtooth pattern of declining peaks extending past 300 bp indicates the presence of the repeat expansion for the positive control. The height of the peaks for the subjects from SF-A indicates that the alleles for these subjects feature fewer than 30 GGGGCC repeats, which is considered normal.

C9orf72 repeat expansion positive control



Subject II-6 (affected)



Subject II-8 (unaffected)

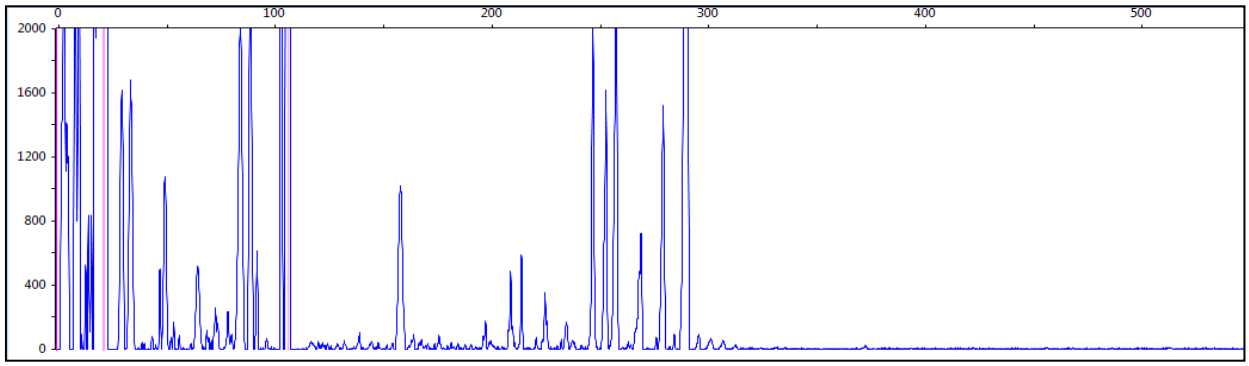


Table 4.1: Variants of Interest on Chromosome 9 in FTD-ALS family San Francisco-A

Locus	Variant	Location	Present?	Evidence
C9orf72	Hexanucleotide (GGGGCC)expansion	Intronic/Promoter Region	No	Repeat-primed PCR
CRB2	c.C3459T	Exonic	Yes	Whole Exome Sequencing
LAMC3	g.61699_61725del27	Intronic	Yes	Whole Genome Sequencing

Chapter 5

Whole Genome and Whole Exome Sequence Analysis of a Family Affected by a Microcoria Myopathy

5.1 Introduction

Muscular dystrophies are characterized by significant genetic and clinical heterogeneity. Nicholl described a British family with members affected by a dominantly inherited proximal myopathy with tubular aggregates, manifesting as chronic progressive muscle weakness and microcoria (pin-point pupils) (1). Congential microcoria, also known as congenital miosis (MCOR, OMIM 156600), is a rare disease characterized by hypoplasia of the dilator pupillae muscle of the iris in both eyes, resulting in pupils that are less than 2 mm in diameter. A family from the United States was identified by Dr. Fan with microcoria and progressive proximal skeletal muscle weakness; the symptoms in this family were consistent with a limb-girdle muscular dystrophy (LGMD).

Previous studies have linked chromosome 13q31-33 to microcoria (2,3); however, microcoria is also believed to be characterized by genetic heterogeneity(4-5). Thus, this study set out to identify the locus responsible for this familial case of microcoria myopathy through a conventional genome-wide linkage study. The linkage study was followed by whole exome sequencing (WES) and whole genome sequencing (WGS) to identify putative causative mutation(s). Furthermore, WES was performed for 7 exomes from unrelated patients, all from

different families, presenting with either a tubular aggregate myopathy or familial pinpoint pupils.

Genome-wide linkage studies have identified over 1,200 chromosomal regions containing rare variants that contribute to the etiology of various disorders, most of which are Mendelian. Although the advent of genome-wide single nucleotide polymorphism (SNP) genotyping has increased the resolution achieved by linkage studies, it is uncommon for the linkage support interval not to contain hundreds of genes. While there has been a modicum of success in understanding the underlying genetics of more complex phenotypes than those engendered by single mutation Mendelian diseases, new approaches are needed to determine the genes responsible for the majority of human diseases (6).

The realization of the goals of the Human Genome Project and the advent of next-generation, massively parallel sequencing technologies may be the key to improving our understanding of the genetics of disease. Massively parallel sequencing (MPS) allows for the identification of causal mutations in human diseases, greatly enhancing the resolution of traditional linkage and association studies. WES has enabled the successful identification of many mutations that cause single-gene, Mendelian disorders (7-10).

This study integrated genome-wide linkage analysis and MPS in order to determine the mutation responsible for a familial microcoria myopathy. Linkage suggested that the causative mutation was located on Chromosome 5q35. WES revealed that the affected individuals in the pedigree all shared a missense mutation in *C5orf60* and that 5 of 7 unrelated subjects presenting with either a myopathy with tubular aggregates or pinpoint pupils also had missense mutations at this locus. In the 7 unrelated subjects, WES also identified additional mutations in genes previously implicated in myopathies including *MACF1, ANO-5, PLEC, TTN, SBCB,* and *FLNC*.

5.2 Methods

Subjects and Samples

A family affected by a myopathy characterized by progressive muscle wasting and microcoria was identified and assessed by Dr. Fan (Figure 5.1).

Genomic DNA (gDNA) was extracted from peripheral blood samples from 12 family members using a Puregene Blood Core kit (Qiagen, Valencia, California) according to the manufacturer's instructions, and the concentration of DNA was quantified with a spectrophotometer (NanoDrop, Wilmington, Delaware).

Seven DNA samples from subjects, all from different families, from the United Kingdom, identified by Dr. Houlden, who were affected by tubular aggregate myopathies or microcoria were used in the exome sequencing phase of the study.

This study was conducted in accordance with the guidelines of the Institutional Review Board of the University of North Carolina at Chapel Hill (IRB Study #07-1001). All subjects provided informed consent.

SNP Genotyping

All gDNA samples from the affected family were genotyped on the Affymetrix Genome-Wide Human SNP Array 6.0 platform according to the manufacturer's protocol, and the genotype calls were made using the Birdseed V2 algorithm included in the Affymetrix Genotyping Console software package (Affymetrix, Santa Clara, California).

Standard quality control measures were used to filter the single nucleotide polymorphism genotyping calls with the PLINK software(11). Only potentially informative SNPs (i.e., those for

which all subjects did not have the same genotype) with genotyping rates > 90% were included. After filtering, there were 442,393 SNPs left for analysis.

Genome-wide Linkage Analysis

Singlepoint linkage analysis was performed in Merlin(12), assuming a fully penetrant, autosomal dominant model of disease. This model of inheritance was plausible because there were affected subjects in all generations. The genotypes were checked for Mendelian inconsistencies using the Merlin error checking function (--error). Because linkage disequilibrium (LD) between SNPs can inflate LOD scores, LD was modeled in Merlin before performing the analysis. To exclude false positive linkage signals that were the result of incorrect genotyping calls, the genotyping cluster plots for SNPs located in the chromosomal regions with the highest LOD scores were examined (described above) Once incorrectly called markers had been excluded, multipoint linkage analysis was performed in the regions of interest with Merlin with error checking and linkage disequilibrium modeling.

Haplotype Analysis

The haplotypes for each subject were phased in Merlin, and the haplotype that was shared among the affected subjects in the region with the highest LOD score was determined.

Whole Genome Sequencing

Five micrograms of gDNA from 1 affected subject was used to construct a shotgun library for MPS. An Illumina Genome Analyzer II (Illumina, San Diego, CA) was used to

generate 76 bp paired-end reads with the standard primer, according to the manufacturer's recommendations.

Whole Exome Sequencing

Twelve exomes were sequenced through targeted capture and MPS. The targeted capture was performed with the Nextera Exome Capture Kit (Illumina, San Diego, CA). Briefly, 10 ng of genomic DNA were simultaneously fragmented and tagged with adaptors. After cleaning up the resulting "tagmented" DNA with bead size selection, the sequencing primers were ligated to the DNA fragments. The sample was then PCR-amplified and denatured to allow for the hybridization of biotinylated probes to the targeted regions of the exome. The targeted DNA library was then captured with streptavidin beads and eluted. The hybridization and elution steps were repeated. An Illumina HiSeq 2500 on the rapid run setting was used to generate 100-bp paired end reads.

Read Alignment, Variant Calling, and Variant Annotation

The sequence reads were separately mapped against the reference genome, UCSC assembly h19 (NCBI build 37.1), using the Burrows-Wheeler Aligner (BWA) software (13). The paired-end reads were then mapped to each other with BWA. The resulting SAM (Sequence Alignment/Map) file was then sorted, converted into a BAM (binary SAM) file, and indexed using the SAMtools software package (14). Reads resulting from duplicate PCR molecules were removed with the MarkDuplicates function in the Picard software package (15). The resulting BAM file was indexed and submitted to the Genome Analysis Toolkit (16) Variant Detection pipeline. After realigning around indels, the variants were called in the Unified Genotyper and

listed in a variant call format (VCF) file for further analysis.(17) The VCF file was annotated using the SeattleSeq annotation pipeline.

Because common variants were not believed to explain the disease, variants were excluded for having a minor allele frequency greater than 0.5%. The minor allele frequencies were derived from the Exome Sequencing Project (ESP) Exome Variant Server data and/or from the 1000 Genomes Project data. The analysis was limited to variants that were shared by all 5 affected family members. Priority was given to variants that were determined to be potentially damaging; this included variants that were rated as being possibly or probably damaging by Polyphen as well as splice site variants and nonsense variants.

5.3 Results

Clinical History

A family affected by a mild to moderate limb-girdle muscular dystrophy and microcoria was identified; the two conditions co-existed in all affected family members. All affected subjects also presented with calf muscle hypertrophy. There were phenotypical variations in this family; two male subjects (II-1 and II-3) had moderate to severe LGMD and became wheelchair bound in their early 50's and late 40's respectively. I-1 had a milder LGMD phenotype; she was still ambulatory in her 70's. All these affected members had moderately elevated creatine kinase (CK) levels, in the range of 1100 – 2400. Both I-1 and II-1 had muscle biopsies that showed increased fiber size variation, increased central nuclei, some degenerating and regenerating fibers. No inflammatory infiltrate. No aggregates seen. However, these two muscle biopsies were limited, and most were adipose tissue. Every affected subject had microcoria that was less severe in infancy/early childhood, but that became pinpoint while they were of school age.

Ophthalmological examination revealed normal iris structures without stromal thinning, pupils at a size of 1 mm baseline, minimal pupillary constriction to bright light and minimal dilation in darkness. Only 1 mm dilation of the pupils OU to Mydriacyl and Neosynephrine suggests damage localized to the pupillary sphincter muscles and the pupillary dilator. There were also additional phenotypes that included severe migraines in subject II-4 and epilepsy in III-1.

Genome-wide Linkage Analysis

A genome-wide linkage analysis was performed using single nucleotide polymorphism data to determine where the potential causative mutation(s) for the microcoria myopathy might be localized. The pedigree suggested an autosomal dominant model of disease as there were affected subjects in each generation.

The results of the genome-wide linkage analysis are presented in Figure 5.2. An approximately 24 cM chromosomal region, on Chromosome 5q35, showed linkage to microcoria myopathy (LOD =1.8) While this LOD score is below the conventional threshold for proving linkage, the rest of the genome had been excluded. All other regions of the genome had LOD scores less than 0. Haplotype analysis allowed the confirmation of the affection status of the subjects because the haplotype sharing in the region of interest was exclusive to the affected subjects. The affected subjects shared a haplotype that was approximately 25 cM in length. The approximate expected length of sharing between 2 related subjects is defined by the equation 200 cM/(2n+1), where n= the number of meioses.(18) Thus, the determined length of sharing was reasonably close to the predicted haplotype length of 28.71 cM.

Whole Genome Sequencing

WGS was performed for 1 subject from the affected family to identify potentially causative mutations (see Figure 5.1) An Illumina Genome Analyzer II was used to generate 76 bp paired-end reads that were then mapped to the reference genome with the BWA software. Overall, 1,014,987,470 reads were obtained for the subject (henceforth Subject I-1) The average coverage throughout the genome was 24.52X, and in the linked region on chromosome 5, the average coverage was 36.30X for Subject I-1.

Once the base pairs had been called and aligned, the search for potentially causative variants excluded variants with a minor allele frequency greater than 0.5%. The variant analysis focused on the distal region of Chromosome 5 because the linkage analysis had excluded the rest of the genome. In this particular region, there were 7 rare variants for Subject I-1. The variants were further stratified based on their predicted effect. For the purpose of this study, "potentially damaging" mutations included any missense variants that were predicted to be either possibly or probably damaging by Polyphen as well as splice variants and nonsense variants.

The subject was found to have a potentially damaging missense mutation in *C5orf60* (c.97C>T) that resulted in the conversion of a proline to a serine (p.Pro33Ser) This mutation was classified as "probably damaging" by Polyphen and was not present in the ESP database, suggesting that this variant was not present in the general population. Furthermore, this variant was not present in hundreds of exomes sequenced at UNC-Chapel Hill. The mutation occurred in the first exon of *C5orf60;* within this particular exon, the ESP database indicated that there were only two other rare variants, but both of these variants were predicted to be benign. In the other exons in *C5orf60*, there was only one rare variant that was predicted to be potentially damaging

in the ESP database; the p.L92S variant was predicted to be possibly damaging and occurred in one sequenced exome. Figure 3 displays the variants found in this study, as well as the other rare missense variants at this locus.

Whole Exome Sequencing

To supplement the findings of the WGS, WES was carried out on the 5 affected subjects (see Figure 5.1) and 7 unrelated subjects (listed as Subjects 6-12) from the UK also presenting with either tubular aggregate myopathies or microcoria. The exome sequence data were aligned to the reference genome and called for variants. The mean coverage for the entire exome for all samples was 25.82X; additional coverage data are available in Table 5.1. Variants with a minor allele frequency less than 0.5% and a predicted Polyphen effect of possibly or probably damaging, as well as splice site variants and nonsense variants, were given priority in the analysis.

The 5 related subjects all shared the c.97C>T variant detected in *C5orf60* by WGS. In the region of interest on chromosome 5, no other rare variants were shared that were predicted to be potentially damaging. In addition, the entire exome was interrogated for rare, potentially damaging variants that were shared by all affected family members. One such variant was found in *MYOM1*: p.S618T. However, the affected individuals did not share a haplotype in this region on chromosome 18; thus, this variant is likely to be a sequencing error.

The exome of one of the unrelated subjects, Subject 12, also possessed the c.97C>T mutation in *C5orf60*. Another mutation in *C5orf60*, c.64G>C, was detected in the exomes of Subjects 6, 8, 10, 11, and 12. These mutations were not present in the ESP database or in

hundreds of exomes sequenced at UNC-Chapel Hill. Thus, 5 of the 7 unrelated subjects had at least one potentially damaging mutation in *C5orf60*.

As 2 of the unrelated subjects did not have potentially damaging mutations in *C5orf60*, the exomes of the 7 unrelated subjects were further analyzed for variants that could be implicated in a myopathy or microcoria. Several potentially interesting variants were found. For example, three subjects had rare, potentially damaging variants (c.299A>G, c.3731G<T, c.8941G>T) in *MACF1* (microtubule-actin cross-linking factor 1); mutations in this gene have recently been implicated as a novel cause of myopathies(19). Variants were also detected in genes that have been implicated in muscular dystrophies by linkage studies: *ANO-5*, *FLNC*, *NEB*, *PLEC*, *SGCB* and *TTN*. It should be noted that the related individuals from the study family did not have any mutations in these genes or other genes previously implicated in LGMDs that segregated with the disease. These results are further summarized in Table 5.2.

5.4 Discussion

This study mapped a familial microcoria myopathy locus to chromosome 5q35. This discovery is novel because congenital microcorias had previously been linked to 13q31-33, though genetic heterogeneity has been indicated by other groups as well (4,5) .

The results of the WGS detected a missense mutation (c.97C>T) in the first exon of *C5orf60* that was predicted to be probably damaging by Polyphen. This mutation was not present in the ESP database. Data from the ESP indicated the presence of only one other rare, potentially damaging variant at the *C5orf60* locus. The infrequent occurrence of other rare, potentially damaging mutations at this locus strengthens the possibility that the mutation found in this study is indeed causative. This evidence was further bolstered by the discovery of another potentially

damaging mutation in the first exon of this gene in 5 unrelated subjects presenting with tubular aggregate myopathies or microcoria. Both of the amino acid residues affected by the mutations in this study are located within a predicted transmembrane domain in the translated protein.

Transmembrane proteins play an important role in muscle tissue. For example, the dystrophin glycoprotein complex connects the cytoskeleton to the extracellular matrix. In muscle cells, the sarcoglycan proteins, which each have a single transmembrane domain, are part of this complex. Mutations in the sarcoglycan genes can cause limb-girdle muscular dystrophies. Thus, the presence of a transmembrane domain in the product of *C5orf60* suggests that it might play a similar role in muscle.

C5orf60 is predicted to encode a single-pass membrane protein; the full length product is predicted to be 353 amino acids in length. Currently, there is evidence at the transcript level for this protein; this evidence includes both expressed sequence tags (EST) and poly-adenylation sequencing (PA-seq) data from the NCBI EST and SRA databases, respectively. A BLAST search for the first exon against PA-seq data for human skeletal muscle (Accession number: SRX208129) indicated that the transcript for this gene is expressed in human skeletal muscle. More research is needed to determine which transcripts for this gene are expressed at the protein level and to understand how mutations in this gene might contribute to the microcoria myopathy phenotype.

Because not all of the subjects had mutations in *C5orf60*, the findings of this study also indicated that this particular myopathy could be characterized by locus heterogeneity. In addition to the mutations found in *C5orf60*, the exome sequencing data for the 7 unrelated subjects indicated the presence of rare, potentially damaging mutations in several different genes, including genes that had previously been implicated in muscular dystrophies and other

myopathies. These findings underscore the genetic heterogeneity of muscular dystrophies. For example, three of the subjects had rare, potentially damaging variants in *Macrophin-1 (MACF1),* a gene which has recently been implicated in myopathies(19). The protein encoded by this gene is similar in structure and function to plectin and dystrophin (20); it acts as a microtubule and actin crosslinking factor. The three detected variants (c.299A>G, c.3731G<T, c.8941G>T) were all classified by Polyphen as being either possibly or probably damaging.

While *in silico* methods might indicate that a discovered mutation is potentially damaging, the effect of a mutation must be assessed at the level of the individual amino acid residue and at the level of the gene as a whole. Several methods have been developed to determine how conserved a residue is and the predicted effect that a change in the residue will have on the protein. At the level of the gene, it has become clear that some genes are more "tolerant" of mutations than others (21). For example, many rare, potentially damaging variants have been detected in olfactory receptor genes, but these variants are typically not pathological. By examining the occurrence of other rare, potentially damaging variants near the mutations detected by exome sequencing, this study sought to qualitatively determine which genes were more "tolerant" than others. The *C5orf60* locus was found to feature only one other possibly damaging variant besides the study mutations, suggesting that this locus is relatively intolerant to mutation. However, to determine whether the mutations found in *C5orf60* are truly the cause of the microcoria myopathy, *in vivo* functional studies will need to be pursued.

MPS technologies will continue to be used to determine the genetic causes of Mendelian diseases that cannot be resolved by linkage analysis alone. The next challenge lies in characterizing the mutations discovered by MPS approaches to determine which mutations contribute to disease.

Contributions

Within this study, I performed the alignment and variant calling of the exome and whole genome sequencing data, the linkage analysis, the haplotype analysis, the causal variant analysis, the *in silico* functional characterization of the variants, and the drafting of the manuscript. Dr. Kirk Wilhelmsen provided oversight and mentorship. Dr. Joshua Sailsbery assisted with the bioinformatics analysis. Dr. Piotr Mieczowski and the High Throughput Sequencing Facility at UNC-Chapel Hill performed the library preparation and generated the Illumina sequence data. Drs. Zheng Fan and James Howard identified the family and provided the clinical data. Dr. Henry Houlden provided the unrelated samples.

5.5 REFERENCES

1. Nicholl D. An English kindred with a dominantly inherited tubular aggregate myopathy with microcoria. Association of British neurologists' spring meeting, church house conference centre, Westminster, London, 14–16 April 2004. J of Neurology, Neurosurgery & Psychiatry. 2004; 75(8): 1213-1228.

2. Rouillac C, Roche O, Marchant D, Bachner L, Kobetz A, Toulemont P, et al. Mapping of a congenital microcoria locus to 13q31-q32. Am J Hum Genet. 1998; 62(5): 1117-1122.

3. Ramprasad V, Sripriya S, Ronnie G, Nancarrow D, Saxena S, Hemamlini A, et al. Genetic homogeneity for inherited congenital microcoria loci in an Asian Indian pedigree. Molecular Vision. 2005;11: 934-40.

4. Bremner FD, Houlden H, Smith SE. Genotypic and phenotypic heterogeneity in familial microcoria. Br J Ophthalmol. 2004;88:469–473.

5. Nortina S, Lowe J and Wills A. Familial myopathy with tubular aggregates associated with abnormal pupils. Neurology. 2004;63(6): 1111-1113.

6. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33 Suppl:228-237.

7. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010; 42(1):30-35.

8. Johnson JO, Mandrioli J, Benatar M, Abramzon Y, Van Deerlin VM, Trojanowski JQ, et al. Exome sequencing reveals VCP mutations as a cause of familial ALS. Neuron. 2010; 68(5): 857-864.

9. Montenegro G, Powell E, Huang J, Speziani F, Edwards YJ, Beecham G, et al. Exome sequencing allows for rapid gene identification in a charcot-marie-tooth family. Annals of Neurology. 2011; 69(3): 464-470.

10. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of kabuki syndrome. Nat Genet. 2010; 42(9): 790-793.

11. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 2007; 81(3): 559–575.

12. Abecasis GR, Cherny SS, Cookson WO and Cardon LR. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002;30(1):97-101.

13. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009;25:1754-60.

14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009; 25:2078-9.

15. Picard. 18 May 2009. Available at http://picard.sourceforge.net, Accessed 8 April 2013.

16. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297-303.

17. DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43(5):491-498.

18. Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet. 2008;40(9):1068-75.

19. Jørgensen LH, Jensen M-BM, Færgeman NJ, Graakjær J, Jacobsen SV, Schrøder HD. A novel myopathic condition caused by mutation of the macf1 gene locus. Abstract from EMC 2012, Rhodos, Greece.

20. Gong T-W L, Besirli CG, Lomax MI. MACF1 gene structure: a hybrid of plectin and dystrophin. Mammalian Genome. 2001; 12: 852-861

21. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet. 2013;9(8): e1003709.

Figure 5.1: Pedigree of a family presenting with microcoria and progressive muscle weakness in a limb-girdle distribution. A blue star to the left of a subject indicates that the subject's DNA sample was submitted to whole genome sequencing, and a red star indicates that the subject's DNA was submitted to whole exome sequencing.

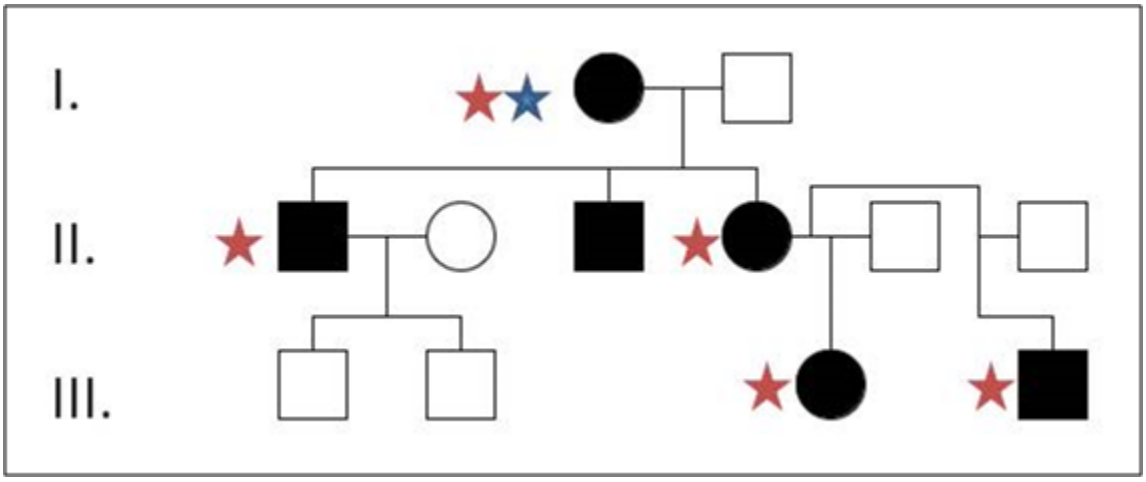


Figure 5.2: Multipoint LOD scores between microcoria myopathy and 515 markers on Chromosome 5. The highest multipoint LOD score (1.8) was seen on 5q35, while the rest of the genome was excluded.

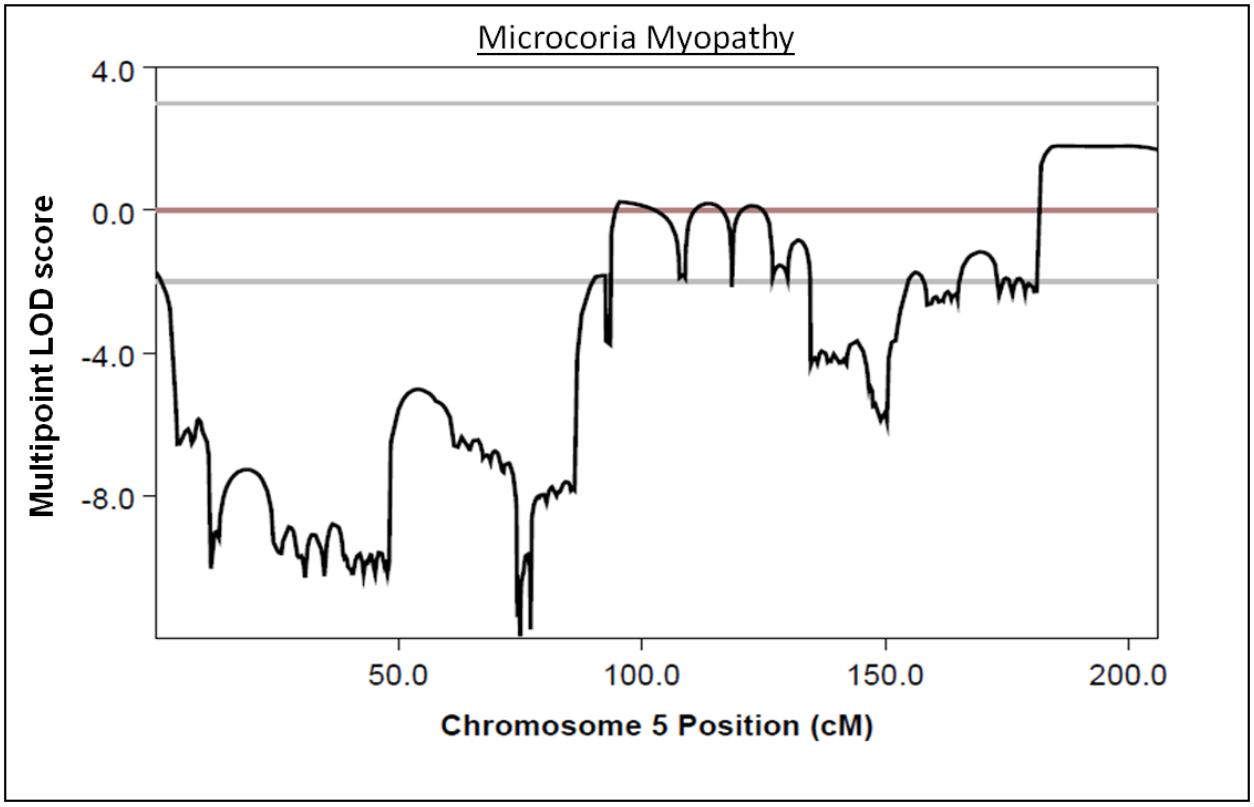


Figure 5.3: Missense mutations detected in *C5orf60* by massively parallel sequencing. Two missense mutations in the first exon of *C5orf60* were detected in a pedigree presenting with a microcoria myopathy as well as 5 unrelated individuals presenting with either a tubular aggregate myopathy or microcoria. These mutations were not in the Exome Sequencing Project Exome Variant Server and were also not detected in hundreds of exomes sequenced at UNC-Chapel Hill. The dashed lines represent the study mutations, and the solid lines represent variants (and their predicted effects) from the Exome Variant Server in the first exon (a) and throughout the entire locus (b) of *C5orf60*. The heights of the variant lines correspond to the negative logs of their minor allele frequencies, except in the case of the study mutations, which were assumed to have a minor allele frequency of 0. The black boxes along the bottom indicate exons, while the horizontal lines indicate introns. The p.P33S mutation was detected in the studied pedigree, and the p.D22H mutation was detected in 5 unrelated subjects presenting with related phenotypes.

Figure 5.3 (continued)

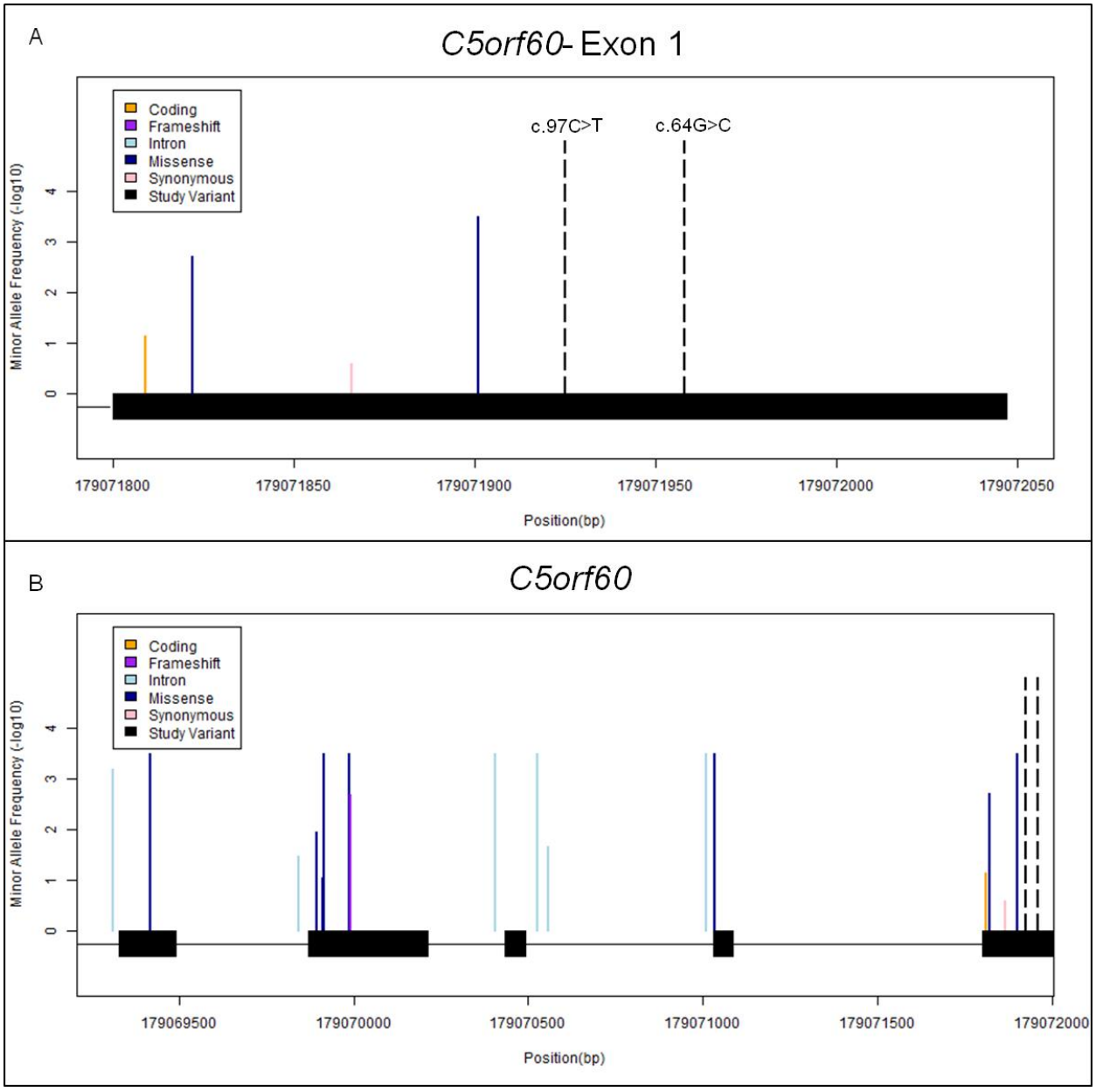


Table 5.1: Mean Coverage of Massively Parallel Sequencing Data

	Exome Sequencing			Whole Genome Sequencing		
Subject	Genome	Chr 5q35	C5orf60	Genome	Chr 5q35	C5orf60
Subject I-1	30.61	32.43	28.83	24.52	36.30	34.83
Subject II-1	24.78	26.90	24.39	---	---	---
Subject II-4	20.39	21.79	18.77	---	---	---
Subject III-3	22.20	23.83	22.08	---	---	---
Subject III-4	27.59	30.60	31.52	---	---	---
Subject 6	28.88	30.65	25.37	---	---	---
Subject 7	30.96	32.66	23.07	---	---	---
Subject 8	25.37	27.80	24.21	---	---	---
Subject 9	21.68	24.38	24.54	---	---	---
Subject 10	28.92	30.80	21.99	---	---	---
Subject 11	30.04	32.43	29.29	---	---	---
Subject 12	18.38	19.09	22.13	---	---	---

Table 5.2: Potentially Damaging Variants Detected by Massively Parallel Sequencing

Locus	Variant	MAF	Family	Sub.6	Sub.7	Sub.8	Sub.9	Sub.10	Sub.11	Sub.12
C5orf60	c.97C>T	0	✓							✓
	c.64G>C	0		✓		✓		✓	✓	✓
ANO-5	c.680G>C	0			✓					
FLNC	c.2068T>C	0.000715		✓						
MACF1	c.299A>G	0.000233							✓	
	c.3731G>T	0.002331				✓				✓
	c.8941G>T	0.002098				✓				
NEB	c.18175A>G	0.000121						✓		
PLEC	c.8201C>T	0.001677					✓			
SGCB	c.800C>T	0				✓				
TTN	c.62066G>A	0.000239								
	c.13879G>T	0.000492			✓					
	c.16603A>T	0.000487			✓					

MAF=minor allele frequency; Sub.=subject

Chapter 6

Conclusions

This work describes approaches for discovering genetic variants that contribute to the etiology of human diseases with complex and simple modes of inheritance through the use of linkage analysis, genome-wide association analysis, and massively parallel sequencing. Linkage analysis and genome-wide association analysis have been successful in identifying thousands of loci that contribute to human traits. The advent of MPS has made it possible to capture nearly all of the variants within the human genome has greatly accelerated the process of identifying causal variants. The studies contained in this work illustrate both the capabilities and limitations of these approaches.

Within this work, the GWA studies of idiopathic Parkinson's disease and dystonia demonstrated that the power to detect associations can be increased by reducing genetic and population heterogeneity. The studies of FTD-ALS and microcoria myopathy explored how linkage analysis and MPS approaches can be combined to analyze the genetics of familial diseases. The subsequent section will describe future directions that are specific to the studies described herein and provide context.

6.1 Study-Specific Conclusions and Future Directions

The four studies in this dissertation can be divided into two groups. Chapters 2 and 3 describe GWA studies of complex diseases, while Chapters 4 and 5 describe studies of

Mendelian diseases that incorporate both linkage analysis and MPS. Chapter 1 described how a GWA study of idiopathic Parkinson's disease (IPD) was able to confirm the previously reported association with the *LRRK2* locus. In the preliminary study of 96 cases, 25 G2019S-positive cases, and 96 controls from an Ashkenazi Jewish population, no associations reached genome-wide significance, but some associations were characterized by near genome-wide significance. To complement the findings of this preliminary study, a replication GWA study of 161 cases and 1,404 controls from the same Ashkenazi Jewish population was performed, and additional genotypes were imputed. This replication study was able to detect an association signal that approached genome-wide significance across chromosome 12q12, including the *LRRK2* locus (average p-value=4.85 x 10^{-6}). Furthermore, two SNPs in an intergenic region upstream of *LRRK2* on chromosome 12pq11.21 were found to have association signals with genome-wide significance: rs10506095 (p= 2.49x10^{-10}) and rs7316771(p=1.19x10^{-9}). A previous GWAS of IPD performed in a Japanese population also found significant associations for SNPs that were upstream of *LRRK2*. One possibility is that upstream variants influence the transcriptional regulation of *LRRK2*, which is believed to induce neuronal toxicity or that the variants are in linkage disequilibrium with the causal variants. However, the detection of an association signal upstream of *LRRK2* also opens up the possibility that genes besides *LRRK2* in this region on chromosome 12, such as *SLC2A13*, may be may play a role in susceptibility to IPD.

By reducing population heterogeneity and including a substantial proportion of known *LRRK2* mutation carriers, this study demonstrated how genetic heterogeneity may have prevented previous GWA studies from detecting an association signal in the *LRRK2* region. While GWA studies have moved in the direction of having larger sample sizes, it is crucial to account for population heterogeneity. For example, research has shown that in the Ashkenazi

Jewish population studied in this work, *LRRK2* G2019S mutation carriers share a haplotype from a recent common ancestor. Thus, future studies of IPD in this population should make use of this knowledge by stratifying subjects based on *LRRK2* mutation status. Other studies have exemplified this point by demonstrating that alleles associated with IPD in genes such as *MAPT* and *SNCA* occur at different frequencies in different populations. While GWA studies have identified many common variants in IPD, it will also be crucial to identify rare mutations that contribute to the genetic etiology of this disease. Thus, future studies of IPD will most likely make use of MPS technologies to provide a better understanding of this polygenic disease.

This work also contains one of the first GWA studies of dystonia. While there have been several linkage studies of this group of syndromes, GWA studies posed the challenge of collecting a sufficient number of samples to ensure adequate power. Because of the relatively small sample size, this study was predicted to have enough power to detect a variant with a comparatively larger effect size. As in the GWA study of IPD, the ability to detect an association was also strengthened by the use of a population with less admixture. This study found an association signal with genome-wide significance on chromosome 17q25.3 within an intronic region of *RNF213* at two genotyped markers, rs12601730 ($p = 2.40 \times 10^{-9}$, OR = 2.169) and rs12603583 ($p = 2.61 \times 10^{-9}$, OR=1.942), as well as two imputed markers, rs4889968 ($p = 2.61 \times 10^{-9}$, OR=1.942) and rs9902013 ($p = 3.84 \times 10^{-9}$), which were all in LD. *RNF213* is known to be expressed in the cerebellum, and its expression in the basal ganglia further strengthens its potential involvement in the genetic etiology of dystonia. To confirm this finding, additional samples should be collected for a replication GWA study or for whole genome sequencing to identify potentially causative mutations at this locus and at other loci throughout the genome.

Thus, the two GWA studies in this work illustrated the ability to identify loci of interest for future study by minimizing the genetic heterogeneity of the study population.

This work also explored the application of MPS technologies. Both whole genome and whole exome sequencing were used to analyze the case of frontotemporal dementia with amyotrophic lateral sclerosis (FTD-ALS) in family San Francisco-A. Many recent studies have identified a hexanucleotide repeat expansion in *C9orf72* as being the cause of both familial and sporadic cases of FTD and/or ALS. Linkage analysis and repeat-primed PCR, however, excluded this region in family San Francisco-A. While a previous linkage study of this family using microsatellite data identified a linkage signal on chromosome 17q, the diagnostic status of two individuals had changed since this first report. Thus, the exome sequencing data from FTD-ALS family San Francisco-A were used to perform a multipoint linkage analysis, which excluded 17q and instead potentially implicated chromosome 9q in this family. In this region, all affected family members shared a synonymous mutation in *CRB2*, a gene in the gamma-secretase pathway. No other rare variants were shared by all affected family members in the region of linkage. If the discovered mutation in *CRB2* is in fact causative, it may influence gene expression; functional studies to quantify both RNA and protein expression could test this hypothesis. Additionally, because of the role that repeat expansions have been shown to play in the genetics of neurodegenerative diseases, it may be useful to perform additional whole genome sequencing of subjects from family San Francisco-A on the Pacific Biosciences platform, which produces sequence reads several thousand nucleotides in length. The shorter read lengths (100 bp) produced by the Illumina sequencing technologies used in the current report likely prevented the detection of repeat expansions in the whole genome sequencing data. Thus, functional

characterization of the detected *CRB2* mutation and additional sequencing could help to complete the picture of the genetics of FTD-ALS in family San Francisco-A.

MPS was also used to investigate the genetics of a microcoria myopathy in one family with 5 affected individuals and in 7 unrelated individuals. This study identified a potentially causative variant in *C5orf60* in a three-generation family affected by progressive muscle weakness in a limb-girdle distribution and microcoria. While some previous studies had demonstrated that microcoria was linked to chromosome 13q, other studies had indicated that this condition was characterized by genetic heterogeneity.

To determine whether microcoriawas linked to 13q in the study family, twelve members of the affected family (5 affected and 7 unaffected) were genotyped for 909,622 single nucleotide polymorphisms for a multipoint linkage analysis. The linkage analysis found a multipoint LOD score of 1.8 located within a 25 cM region on Chromosome 5q35. While this LOD score was below the threshold conventionally used as a criterion to declare linkage, the rest of the genome, including chromosome 13q, was excluded. All of the affected subjects shared a haplotype in the region that was not shared by any of the unaffected subjects.

Massively parallel DNA sequencing for the five affected subjects detected a missense mutation on chromosome 5q35 in *C5orf60*: c.97C>T (p.P33S). Whole exome sequencing was also performed for 7 unrelated subjects, all from different families, affected with tubular aggregate myopathies and/or microcoria. The c.97C>T variant in *C5orf60* was found in 1 of the unrelated subjects. Another missense mutation in *C5orf60*, c.64G>C (p.D22H), was present in the exomes from 5 of the unrelated subjects. The exome of 1 unrelated subject had both of these variants. Thus, 5 of the 7 unrelated subjects had at least 1 mutation in *C5orf60*. In the 7 unrelated

subjects, exome sequencing also detected additional mutations or very rare variants in genes previously implicated in myopathies including *MACF1*, *ANO-5*, *PLEC, TTN, SBCB,* and *FLNC*.

These findings are significant because they corroborate the findings of other researchers that microcoria is characterized by genetic heterogeneity and because they demonstrate that mutations in *C5orf60* represent a novel potential cause of microcoria myopathy. Future studies could focus on characterizing the effect that mutations in *C5orf60* have on gene expression and the role that this locus plays in the muscle biology.

6.2 Insights and Context

While GWA studies have implicated thousands of loci in hundreds of phenotypes, several criticisms of the findings of GWA studies have been made. First, while the p-values of GWA signals achieve genome-wide significance, it is difficult to understand what this high level of significance means. It has been speculated that many of these significant association signals may be false positives. There has also been debate over the possibility that these associations are actually the result of rare variants that are in LD with common variants (Dickson et al., 2010). In the study of type 2 diabetes, for example, several genes have been implicated repeatedly by GWA studies, such as *TCF7L2* and *PPARG*. The GWA study of idiopathic Parkinson's disease in this work provided further confirmation of the role of *LRRK2* mutations in the etiology of the disease. The reproducibility of GWA studies points to their robustness and the accuracy of their findings.

Second, part of the dissatisfaction with GWA studies stems from the difficulty in attributing significance to the findings. Even when association signals with genome-wide significance are replicated by multiple studies, it is difficult to know how these associated

variants contribute to the underlying biology of a phenotype. Most of the variants implicated by GWA studies do not cause an obvious change in a protein. Some are even located hundreds of thousands of base pairs from any known gene. While variants implicated in GWA studies that do not cause an amino acid change could simply be "tagging" other variants, these findings may also suggest that the current focus on variants that cause changes in proteins is too narrow. Protein-coding sequences make up less than 2% of the genome, and the findings of GWA studies suggest that the other 98% can play a major role in phenotypic variation. Thus, the findings of GWA studies prompted researchers to investigate how variants can affect gene expression; these expression quantitative trait loci (eQTLs) can either effect the expression of the nearest gene (cis-eQTLs) or a gene located thousands of base pairs away (trans-eQTLs).

Third, despite the ability to genotype approximately one million markers in ever-larger cohorts, GWA studies have not identified common variants with large effect sizes (McClellan and King, 2010). GWA studies have largely not improved medicine's ability to predict an individual's risk for common diseases. However, the discovery of these variants has allowed the implication of previously unexplored biological pathways in common disease and identified pathways that are common to multiple complex diseases. A better understanding of these pathways will likely yield new therapeutic targets. Therefore, even if associations detected by GWA studies did not contribute to the prediction of disease risk, they may provide important insights into the biology and potential treatment of complex diseases.

Despite these criticisms, when considered collectively, GWA studies have greatly advanced the field of human genetics and have comprehensively explored the role of common genetic variation in complex disease. For GWA findings that were replicated by properly designed and sufficiently powered studies, some variant, either rare or common, that is in LD

with the significantly-associated SNP must be present that explains the association. Subsequent resequencing and functional studies can be used to identify these variants and their roles in the genetic etiology of a phenotype.

MPS studies have the ability to identify nearly all the variants in an individual genome and to determine the mutations underlying significant GWA and linkage signals. This great potential, however, comes with its own set of challenges.

Because each sequenced genome results in gigabytes of data, many of the challenges associated with MPS approaches involve data management and bioinformatics. As was discussed in the introduction, an important consideration is the selection of the alignment and variant calling algorithms. In the analysis of MPS data, a paradox has arisen because alignment and variant calling software relies on the similarities between the sequence reads and the reference genome, but researchers are interested in the differences. The problem lies in distinguishing "errors" from actual variants. New alignment algorithms are developed regularly, but more work is needed to understand which of these algorithms provides the best balance between computational speed and accuracy. The inability of the alignment software to unambiguously match a sequence read to the genome can result in the substantial loss of data. As the depth of sequencing and the length of sequence reads increase, this problem will likely be less of an issue. Variant calling software is prone to both type I and type II error. Type I error can be reduced by increasing read depth and performing recalibration steps (Jia et al., 2012), but better variant calling methods are needed to reduce type II error and provide greater sensitivity for rare variants.

Another challenge for MPS studies is the functional characterization of causal variants. The current focus of MPS studies is the exome. Yet the exome only makes up about 1% of the

genome. Emerging data are suggesting that even in the case of Mendelian diseases for which the causal variant is suspected to have a high penetrance, exome sequencing may not be sufficient to identify the causal variant. This observation was exemplified by the FTD-ALS family study in this work; exome sequencing was unable to detect an obvious deleterious mutation. This failure can be explained in one of two ways: either the causal mutation is not located in the exome or the synonymous variant found in *CRB2* in the affected individuals affects gene expression. More effort is needed to identify and to understand the impact of variants besides those that cause a change in the amino acid sequence of a protein.

Furthermore, current approaches to MPS analysis are lacking in the ability to align large copy number variants and repetitive sequences. When a sequence occurs more than once in the reference genome, it is difficult for aligning software to unambiguously match a read to the reference. In the case of CNVs, experience from GWA studies of neuropsychiatric disorders, such as autism and schizophrenia, suggests that novel CNVs can play a very important role in the genetic etiology of a disease. Thus, the scope and analytical power of MPS approaches need to be expanded to account for a wider variety of variants. These ends can be achieved by pursuing whole genome sequencing with longer sequence reads and by improving both *in vivo* and *in silico* methods for characterizing causal variants.

6.3 Final Thoughts

The purpose of this work was to use linkage analysis, association analysis, and MPS approaches to interrogate the genetic etiology of both Mendelian and complex human diseases. The studies presented in this work demonstrate both the potential and drawbacks of these approaches. Linkage and association analysis have collectively improved our understanding of

the biology of thousands of human diseases and paved the path forward for understanding the genetics of human disease. MPS approaches have proven their utility in the identification of causal variants for Mendelian diseases, and their ability to identify causal mutations in complex diseases will be tested in the near future. MPS approaches, in combination with linkage and association analysis, will continue to be leveraged to identify variants that contribute to traits of medical importance.

6.4 REFERENCES

1. Dickson, SP, Wang, K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biology. 2010; 8(1): e1000294.

2. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010;141(2):210-217.

3. Jia P, Li F, Xia J, Chen H, Ji H, Pao W, Zhao Z. Consensus Rules in Variant Detection from Next-Generation Sequencing Data. PloS One. 2012;7(6):e38470.