SURVEYING THE GENETIC RISK LANDSCAPE OF AMYOTROPHIC LATERAL SCLEROSIS IN THE ERA OF NEXT-GENERATION SEQUENCING

by

Jonathan M. Downie

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Human Genetics

The University of Utah

August 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Jonathan M. Downie**

has been approved by the following supervisory committee members:

Lynn B. Jorde	, Chair	**05/03/2017** Date Approved
Nicola J. Camp	, Member	**05/03/2017** Date Approved
Summer Gibson	, Member	**05/03/2017** Date Approved
Charles L. Murtaugh	, Member	**05/03/2017** Date Approved
Karl V. Voelkerding	, Member	**05/03/2017** Date Approved
Robert B. Weiss	, Member	**05/03/2017** Date Approved

and by **Lynn B. Jorde** , Chair/Dean of

the Department/College/School of **Human Genetics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease, is an adult-onset fatal disease in which the upper and lower motor neurons of the body progressively degenerate. Efforts to understand the pathophysiology of ALS over the past two decades have shown that mutations in genes involved in a wide variety of cellular processes can cause ALS. Patients who develop ALS and have a family history of the disease are termed familial ALS (FALS) and represent 10% of ALS cases. However, similar genetic mutations occur in patients with no family history of ALS, which suggests genetic factors also play a role in sporadic ALS (SALS).

Studies that utilize low-resolution single nucleotide variant (SNV) and microsatellite assays have identified over 30 ALS-associated genes. However, only 68% of FALS and 11% of SALS cases have an identifiable genetic cause. The identification of the genetic factors responsible for these unexplained ALS cases has been challenging because of the technological limitations of SNV and microsatellite assays. The increasing availability of next-generation sequencing (NGS) allows for the potential identification of such elusive disease-causing genetic variants.

The aim of this dissertation is to better understand ALS genetic risk factors using NGS technology and computational methods. The first chapter will review ALS and the importance of genetic factors in its pathogenesis. The analyses presented in Chapter 2 try to determine whether NGS approaches can identify known and potentially novel ALS

genetic risk loci in individual FALS patients. Next, efforts to better understand the importance of known ALS risk loci in SALS pathogenesis will be covered in Chapter 3. Chapter 4 will focus on attempts to find novel ALS risk genes in a cohort of SALS patients. Chapter 5 will focus on the results of functional studies aimed at validating *TP73* as an ALS candidate risk gene. Lastly, Chapter 6 will be focused on determining whether SALS can be caused by deleterious genetic variation shared between distantly related patients. The results of these studies will help to push the understanding of ALS pathogenesis forward towards the ultimate goal of a cure.

This dissertation is first and foremost dedicated to the patients who made these studies possible. Your suffering has not been in vain. I also dedicate this work to Dr. Laurent Degos, Dr. Zhen-Yi Wang, Dr. Paul Shami, and the Huntsman Cancer Institute nursing staff. Without all of you, I would not be alive to write this dissertation. Lastly, I dedicate this dissertation to my parents, Michael and Robin. I love you both to the moon and back.

TABLE OF CONTENTS

Chapters

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to first thank my Ph.D. advisor Dr. Lynn Jorde for providing me the wonderful opportunity to work in his lab. It is impossible to state how much I’ve gained as a scientist and person under his mentorship. Lynn has always fostered a culture of curiosity, cooperation, and excellence in his lab. The research I have been able to perform would not be possible without these assets. I would also like to thank Dr. Summer Gibson for allowing me to collaborate with her in these studies. Together we have made an excellent team and have accomplished great things. I would like to acknowledge both past and present Jorde lab members, including Dr. Tatum Simonson, Dr. Chad Huff, Dr. Wilfred Wu, Scott Watkins, Dr. Justin Tackney, Dr. Brett Kennedy, Julie Feusier, and Kristi Russell for the comradery and helpful advice they have provided over the years. I would like to thank the members of the Utah Genome Project who have been an incredible asset in working with the massive amount of data provided by next-generation sequencing. I would like to give special thanks to Dr. Charles Murtaugh for patiently coaching me through the functional studies presented in this dissertation. I would like to thank the rest of the members of my thesis committee, Dr. Nicola Camp, Dr. Robert Weiss, and Dr. Karl Voelkerding for providing me excellent guidance over the years. Lastly, I would like to acknowledge the funding sources for these projects, including the Utah Genome Project, the University of Utah Neuroscience Initiative, the National Institute of General Medical Sciences, NantOmics, and Biogen.

CHAPTER 1

INTRODUCTION

Next-Generation Sequencing Technology

The process in which the exact nucleotide sequence of a molecule of deoxyribonucleic acid (DNA) is determined is called DNA sequencing. DNA sequencing differs from genotyping methods, which includes single nucleotide polymorphism (SNP) and microsatellite assays, in that genotyping only determines the alleles an individual possesses at a set of preselected loci. DNA sequencing allows researchers and clinicians to determine nearly all of the mutations or genetic variants individuals possess without ascertainment bias. Traditional methods of DNA sequencing, such as Sanger sequencing, are highly accurate in determining the nucleotide state at each base in a DNA sequence. However, traditional methods of DNA sequencing are expensive and cannot be efficiently scaled to sequence multiple loci in the number of individuals needed to answer biologically relevant questions (1). Next-generation sequencing (NGS) technology, such as Illumina sequencing, has greatly increased feasibility of whole-genome and exome (the protein-coding portion of the genome) sequencing by increasing the scalability and speed of sequencing for a fraction of the cost of traditional sequencing (1). As a result, researchers can now utilize NGS technology to answer many unsolved biological questions.

Limitations of SNP and Microsatellite Genotype-Phenotype Associations

Much of the genomic research and investigation of genetic variation in humans over the past decade has been performed by utilizing SNP arrays (2), which are designed to assay SNPs that have a high (>1%) minor allele frequency (MAF) in the general population (3). The degree to which a variant negatively affects fitness is inversely proportional to its frequency (4) (Figure 1.1). Therefore, genotype-phenotype associations identified by SNP arrays consist largely of common variants of small effect size (5). This has limited the ability to identify genotype-phenotype associations with appreciable effect sizes as evidenced by the low rate of reproducibility and inability of known variants to fully account for the heritability of particular traits (5-7). Rare genetic polymorphisms account for the majority of human interindividual genetic diversity (8). As a result, the identification of rare variants that have a large effect on fitness will require NGS to capture rare genetic variation.

Clinical Presentation and Epidemiology of Amyotrophic Lateral Sclerosis

Amyotrophic lateral sclerosis (ALS) is the most common adult-onset motor neuron disease with a prevalence of 3.9 cases per 100,000 individuals in the United States (9). ALS is an incurable and fatal adult-onset condition in which the upper motor neurons of the motor cortex and the lower motor neurons of the spinal cord progressively degenerate (10). This leads to a gradual increase in the symptoms of upper motor neuron (muscle weakness, spasticity, abnormal reflexes) and lower motor neuron (muscle fasciculation and paralysis) dysfunction. ALS was likely first described in 1848 (11), but was not formally defined and recognized as it is today until 1869 (12). It is also known as Lou Gehrig's disease after the famous baseball player who was afflicted by it. Progression can be highly variable but typically occurs over 3-5 years on average, culminating in paralysis, respiratory failure, and death (10, 13). The average age of ALS onset is 46 for individuals with a family history of the disease and 56 for those without a family history of ALS (14). The symptoms typically manifest first in the limbs; however, one-third of cases have a bulbar presentation resulting in difficulties with speech and swallowing (15). There is no current treatment for ALS, but riluzole can prolong median survival by 2-3 months (16).

While ALS is typically considered an isolated motor neuron disease, many patients experience cognitive impairment as well. This typically manifests in the form of frontotemporal dementia (FTD)—which is focal atrophy of the frontal and anterior temporal lobes of the brain—and results in impaired executive function, personality change, and impaired language abilities (17). A subset (15%) of individuals that experience adult motor neuron disease (of which ALS accounts for 75% of cases) also experience FTD, suggesting that there is an overlap in pathophysiology between the two disorders (17).

ALS patients that have a first- or second-degree affected family member are termed familial ALS (FALS) and represent 10% of ALS cases (13). The familial nature of FALS highlights the importance of genetic risk factors in the pathogenesis of the disease. A majority (90%) of ALS cases occur sporadically (SALS) with no previous family history (13). An SALS twin study of patients with no family history of ALS in non-twin relatives has estimated that 60% of SALS risk is genetically determined (18).

Furthermore, a number of the ALS genetic risk factors identified in FALS cases have been found in SALS cases (13), which suggests genetic factors are also important in the pathogenesis of SALS. A substantial proportion of SALS cases are thought to be a result of genetic *de novo* mutations (19). However, it is also possible that SALS could be caused by inherited genetic risk factors, but did not manifest as FALS due to incomplete penetrance or early death/misdiagnosis of carrier family members.

Amyotrophic Lateral Sclerosis Molecular Pathology

Linkage studies and genome-wide association studies (GWAS) have identified genetic variants in over 30 genes to be associated with ALS. Nearly all ALS-causing mutations act in a genetically dominant fashion (20). *SOD1* was the first gene identified, via linkage analysis, to be strongly associated with autosomal dominant inheritance of ALS (21). It is responsible for 12% of FALS and 1% of SALS cases (22). The protein product of *SOD1* (superoxide dismutase 1) is involved in free radical scavenging in cells (20). Mutations in *SOD1* cause the protein to misfold and are targeted for degradation. However, the misfolded protein is able to escape degradation by forming protease-resistant aggregates (23)—which leads to toxic effects on the cellular protein degradation system (24), activates the unfolded protein stress response, initiates axonal retraction, and causes eventual neuronal death (20, 25). These pieces of evidence, in combination with the discovery of mutations in other genes involved in protein degradation—such as *UBQLN2* (26), *SQSTM1* (27), and *VCP* (28)—suggested ALS occurred as a result of failure of the proteasome (20).

However, the discovery of mutations in genes involved in RNA processing and the effects of toxic RNA products has changed the view that ALS results purely from proteostasis dysfunction. For instance, mutations of both *TARDBP* (29) and *FUS* (30) have been discovered to be associated with ALS pathogenesis. *TARDBP* and *FUS* both encode for proteins involved in RNA processing. It is believed that mutant copies of these proteins result in cytoplasmic protein/RNA aggregates and toxic RNA species, leading to cellular dysfunction and death (20). The association of hexanucleotide (G_4C_2) repeat expansions in the first intron of *C9orf72* with ALS further solidified the notion that ALS can also result from ribonucleopathies (31, 32). Normal copies of *C9orf72* contain fewer than 30 G_4C_2 repeats, while mutant copies carry tens to thousands of these repeats (31-33). *C9orf72* accounts for a substantial amount of FALS cases (>40%) and 7% of SALS cases (13). It was recently discovered that mutant copies of *C9orf72* cause disease by impairing its transcription, leading to abortive transcripts with toxic properties that sequester other proteins that can bind to them (34). Interestingly, pathological *C9orf72* hexanucleotide repeat expansions are also thought to account for 25% of patients with isolated FTD and may possibly explain the overlap between ALS and FTD (35). However, the exact mechanism by which this occurs is poorly understood.

In light of both proteopathies and ribonucleopathies being responsible for ALS pathogenesis, it is thought that aggregation of these defective species causes cellular stress and subsequent motor neuron death (20). The association of genes that encode for proteins involved in cytoskeleton arrangement, axonal transport, and neuronal excitation with ALS has further complicated such a model (20). The identification of other genetic causes of ALS will help reconcile how these different causes of ALS converge on a clinically similar phenotype. It will also help to determine if there is a central molecular pathway involved in ALS pathogenesis that can be targeted for therapy.

The Genetic Landscape of Amyotrophic Lateral Sclerosis

Despite the many efforts to search for ALS causing genetic variants, the complete understanding of how genetic factors give rise to ALS is incomplete. For instance, a significant percentage of FALS (32%) and of SALS (72%-89%) cases have no identifiable genetic cause (13, 36) (Figure 1.2). Most of the studies aimed at identifying ALS risk loci largely depended on low-resolution SNP and microsatellite arrays, which cannot directly assay low-frequency and high-effect size variants. As a result, the discovery of additional disease-causing variants might have been missed in previous investigations. Recent studies that have employed NGS approaches have been successful in identifying a number of novel ALS risk loci (19, 37). The success of these approaches suggests that further ALS genetic studies that utilize NGS technology can identify novel ALS risk loci.

The role genetic factors have in ALS pathogenesis is also incompletely understood because it is unclear what proportion of cases are caused by known genetic risk loci. More specifically, there are inconsistent results in the number of SALS cases that have an identifiable genetic cause. The first attempt at estimating the amount of risk known genetic factors contribute towards SALS found that such factors only caused 2.8% of cases (38). This was determined by calculating the percentage of SALS patients who had a coding mutation in at least one of five ALS-associated genes (38). A more recent study found that ALS risk loci are responsible for causing 27.8% of SALS cases (36). This estimate was calculated by finding what proportion of SALS patients had a rare (minor allele frequency <1%) coding mutation in a panel of 17 ALS-associated genes (36). The majority of these types of studies follow a similar protocol where variant presence or rarity is used to determine variant pathogenicity. However, variant rarity is not a sufficient criterion of variant pathogenicity as the majority of rare, nonsynonymous variants are not likely to be pathogenic (39). More accurate estimates of the proportion of SALS patients with an identifiable genetic cause should be achievable by incorporating direct estimates of variant pathogenicity instead of variant rarity alone.

The considerable gaps in our understanding of ALS pathogenesis underscore the importance of applying NGS methods to discover risk loci that have not been detected by previous methods. The subsequent chapters of this dissertation will focus on expanding the knowledge of how genetic factors play a role in ALS pathogenesis by utilizing NGS technology and computational methods. The results of a limited sample size FALS NGS study aimed at identifying ALS risk loci will be presented in Chapter 2. Chapter 3 will focus on obtaining a better understanding of what proportion of SALS is caused by known genetic risk loci by using direct predictions of variant pathogenicity. The findings presented in Chapter 4 were generated from efforts made to identify novel ALS risk loci in an SALS patient cohort. Chapter 5 will outline the results of functional experiments aimed at determining whether *TP73*, a candidate ALS risk gene found in Chapter 4, is involved in ALS pathogenesis. Lastly, the focus of Chapter 6 is on determining whether shared deleterious variants between distantly related patients can give rise to ALS. The findings of these studies will help to better understand the pathogenic mechanisms of ALS and lead the way to potential therapeutics.

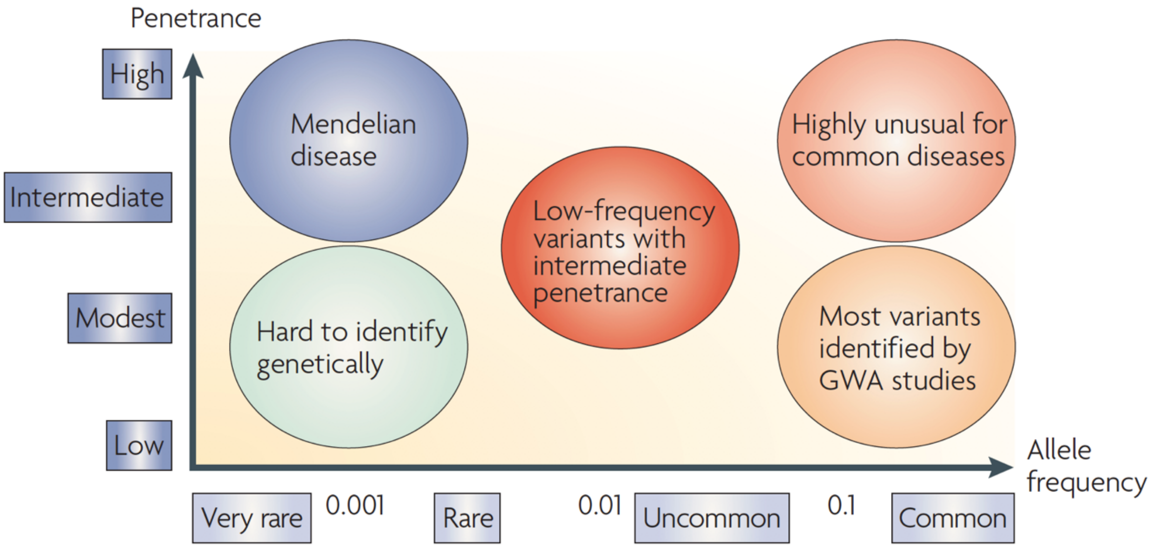

Penetrance
High
Intermediate
Modest
Low
Mendelian disease
Hard to identify genetically
Low-frequency variants with intermediate penetrance
Highly unusual for common diseases
Most variants identified by GWA studies
Allele frequency
Very rare
0.001
Rare
0.01
Uncommon
0.1
Common

Figure 1.1 The allele frequency of a variant is typically inversely proportional to the effect size or penetrance it has on a phenotype. Genotype-phenotype associations using genotyping arrays largely find common, low effect size variants. In contrast, NGS approaches have the ability to detect rare variants with large effect sizes. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, McCarthy et al. 2008; 9(5):356-369, copyright 2008.

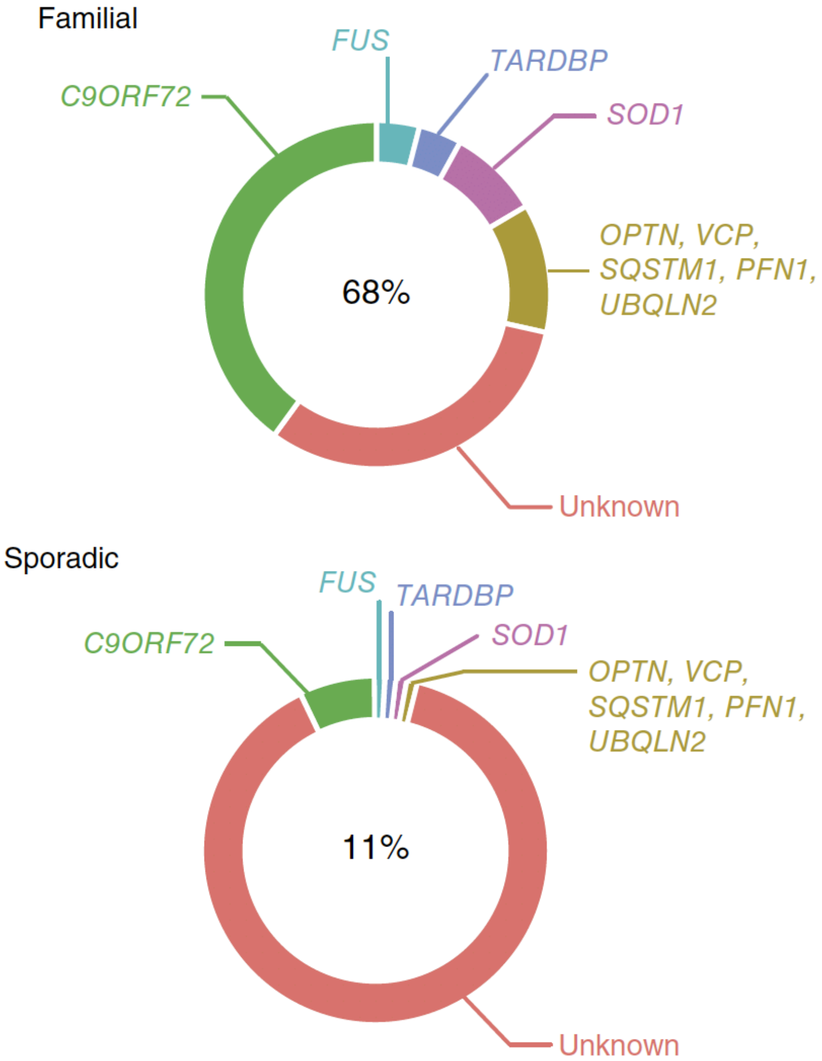


Figure 1.2 The percentage of familial and sporadic ALS cases caused by ALS-associated genes. The size of each bar is proportional to the percentage of cases the gene causes. The number inside each circle is the percentage of ALS cases with an identifiable genetic cause. Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: Nature Neuroscience, Renton et al. 2014; 17(1):17-23, copyright 2014.

References

1. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* 17(6):333-351.

2. LaFramboise T (2009) Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Res* 37(13):4181-4193.

3. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866):1100-1104.

4. McCarthy MI, et al. (2008) Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat Rev Genet* 9(5):356-369.

5. Manolio TA, et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461(7265):747-753.

6. Nebert DW, Zhang G, Vesell ES (2008) From human genetics and genomics to pharmacogenetics and pharmacogenomics: Past lessons, future directions. *Drug Metab Rev* 40(2):187-224.

7. Ward LD, Kellis M (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* 30(11):1095-1106.

8. Tennessen JA, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64-69.

9. Mehta P, et al. (2014) Prevalence of amyotrophic lateral sclerosis-United States, 2010-2011. *MMWR Surveill Summ* 63(suppl 7):1-14.

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

11. Aran F (1848) Research on an as yet undescribed disease of the muscular system (progressive muscular atrophy). *Arch Gén Méd* 24:15-35.

12. Charcot J-M, Joffroy A (1869) *Deux cas d'atrophie musculaire progressive: Avec lésions de la substance grise et des faisceaux antéro-latéraux de la moelle épinière* (V. Masson, Paris, France).

13. Renton AE, Chio A, Traynor BJ (2014) State of play in amyotrophic lateral sclerosis genetics. *Nat Neurosci* 17(1):17-23.

14. Kinsley L, Siddique T (1993) Amyotrophic Lateral Sclerosis Overview. *GeneReviews(R)*, eds Pagon RA, et al. (University of Washington, Seattle, WA).

15. Chio A, et al. (2009) Epidemiology of ALS in Italy: A 10-year prospective

population-based study. *Neurology* 72(8):725-731.

16. Miller RG, Mitchell JD, Moore DH (2012) Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). *Cochrane Database Syst Rev* 3:CD001447.

17. Lillo P, Hodges JR (2009) Frontotemporal dementia and motor neurone disease: Overlapping clinic-pathological disorders. *J Clin Neurosci* 16(9):1131-1135.

18. Al-Chalabi A*,* et al. (2010) An estimate of amyotrophic lateral sclerosis heritability using twin data. *J Neurol Neurosurg Psychiatry* 81(12):1324-1326.

19. Chesi A*,* et al. (2013) Exome sequencing to identify de novo mutations in sporadic ALS trios. *Nat Neurosci* 16(7):851-855.

20. Robberecht W, Philips T (2013) The changing scene of amyotrophic lateral sclerosis. *Nat Rev Neurosci* 14(4):248-264.

21. Rosen DR*,* et al. (1993) Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 362(6415):59-62.

22. Chio A*,* et al. (2008) Prevalence of SOD1 mutations in the Italian ALS population. *Neurology* 70(7):533-537.

23. Ciechanover A, Kwon YT (2015) Degradation of misfolded proteins in neurodegenerative diseases: Therapeutic targets and strategies. *Exp Mol Med* 47:e147.

24. Bendotti C*,* et al. (2012) Dysfunction of constitutive and inducible ubiquitin-proteasome system in amyotrophic lateral sclerosis: Implication for protein aggregation and immune response. *Prog Neurobiol* 97(2):101-126.

25. Saxena S, Cabuy E, Caroni P (2009) A role for motoneuron subtype-selective ER stress in disease manifestations of FALS mice. *Nat Neurosci* 12(5):627-636.

26. Deng HX*,* et al. (2011) Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. *Nature* 477(7363):211-215.

27. Fecto F*,* et al. (2011) SQSTM1 mutations in familial and sporadic amyotrophic lateral sclerosis. *Arch Neurol* 68(11):1440-1446.

28. Johnson JO*,* et al. (2010) Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron* 68(5):857-864.

29. Sreedharan J*,* et al. (2008) TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science* 319(5870):1668-1672.

30. Kwiatkowski TJ, Jr.*,* et al. (2009) Mutations in the FUS/TLS gene on

chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* 323(5918):1205-1208.

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

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

33. Rutherford NJ, et al. (2012) Length of normal alleles of C9ORF72 GGGGCC repeat do not influence disease phenotype. *Neurobiol Aging* 33(12):2950 e2955-2957.

34. Haeusler AR, et al. (2014) C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* 507(7491):195-200.

35. Majounie E, et al. (2012) Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: A cross-sectional study. *Lancet Neurol* 11(4):323-330.

36. Cady J, et al. (2014) Amyotrophic lateral sclerosis onset is influenced by the burden of rare variants in known amyotrophic lateral sclerosis genes. *Ann Neurol*.

37. Cirulli ET, et al. (2015) Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* 347(6229):1436-1441.

38. Kwon MJ, et al. (2012) Screening of the SOD1, FUS, TARDBP, ANG, and OPTN mutations in Korean patients with familial and sporadic ALS. *Neurobiol Aging* 33(5):1017 e1017-1023.

39. Li MX, et al. (2013) Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet* 9(1):e1003143.

CHAPTER 2

THE IDENTIFICATION OF AMYOTROPHIC LATERAL SCLEROSIS GENETIC RISK FACTORS IN SMALL SAMPLE SIZE NEXT-GENERATION SEQUENCING STUDIES

Introduction

The use of NGS technology and genomic information in the healthcare setting is likely to radically change how physicians make clinical decisions in a variety of different contexts. For example, the drug vemurafenib can be used to treat melanoma tumors with positive genomic tests for the *BRAF*:p.V600E mutation, which results in improved patient survival rates (1). However, the use of genome sequencing results by clinicians can be extremely challenging due to sheer amount of data yielded by NGS. Whole-exome sequencing results from a single individual can return thousands of genetic variants to be interpreted (2). As a result, methods that prioritize variants based on their probable functional consequences are required to reasonably interpret NGS data. Methods, such as *VAAST* (3), are able to prioritize variants by determining whether any genes in the genome are more burdened by deleterious variation in patients versus healthy control individuals. However, these methods are underpowered to find significant gene associations to properly prioritize variants in studies consisting of one or a few patients. *PHEVOR* is a method that analyzes patient phenotypic information to prioritize variants that likely give rise to the disease of interest (4). When used in conjunction with the predictions of variant pathogenicity that come from *VAAST*, genetic risk variants can be identified in small sample size sequencing studies (4). When applied to NGS data from individual ALS patients and small ALS kindreds, such an approach should be able to identify both known and novel ALS genetic risk loci.

A majority of ALS risk loci have been identified by low-resolution linkage analysis and GWAS based on common SNP arrays. However, these risk factors only account for 68% of FALS cases (5), which suggests there are unidentified risk loci. NGS approaches allow for the detection of rare variants that are likely to have a large impact on phenotypic traits and diseases. DNA has been collected from a number of FALS patients seen at the University of Utah. The focus of this chapter is aimed at analyzing these DNA samples to determine whether NGS approaches are useful in identifying known and novel ALS risk factors in limited sample size studies. The approaches and results of these efforts can serve as a model for future researchers and clinicians to interpret genomic data from similar cohorts. Furthermore, any novel candidate risk loci identified from these studies serve as intriguing targets for subsequent functional studies to determine their role in ALS pathogenesis.

Materials and Methods

Dr. Summer Gibson (Department of Neurology; University of Utah) has collected DNA samples from FALS probands and their family members seen at the University of Utah motor neuron disease clinic. Individuals were selected for genetic study based on whether there was an already known genetic cause for their disease. Six unrelated FALS samples were selected for analysis. Additionally, an unaffected mother and affected son pair were selected for analysis. This pair is considered an FALS pedigree due to their family history and the mother having two affected sons. Two female siblings with ALS were also selected for analysis. Another five individuals affected by primary lateral sclerosis (PLS), which is a subtype of ALS where only the upper motor neurons are affected, were also selected. In total, 15 individuals (9 FALS, 1 unaffected family member, and 5 PLS samples) from 13 different families were selected for sequencing.

These samples were whole-exome sequenced (6) by using the Agilent SureSelectXT Human All Exon V5+UTR exome capture kit and the Illumina HiSeq 125 base-pair paired-end sequencing platform (7). These samples were sequenced to a depth of 60-80X coverage. Raw sequenced reads that were obtained from the sequencer in FASTQ format were aligned to the Genome Reference Consortium human genome 37 (GRCh37) using the Burrows-Wheeler Aligner MEM algorithm (8, 9). These aligned reads were then processed with the SAMtools software (10) to generate aligned, coordinate sorted BAM files. Optical and PCR read duplicates were marked and removed from further analyses using Picard Tools (http://broadinstitute.github.io/picard/) to eliminate any potential biases resulting from duplicate reads. Single nucleotide variant (SNV) genotypes were called using the Genome Analysis Toolkit (GATK) v3.0+ variant pipeline (11-13) (Figure 2.1). Genome and variant quality were assessed via FastQC software, GATK's variant quality score recalibration (VQSR) metrics, and principal components analysis (PCA).

Obtained genotype calls were processed through the *VAAST* pipeline (14) to prioritize genes in each individual (or family, where applicable) that possess potentially pathogenic variants. *VAAST* is a tool that combines variant frequency information and amino acid substitution scores to determine whether any genes in the genome are significantly more affected, or burdened, by deleterious variation in cases versus controls (3, 15). The *VAAST* pipeline first annotates variants from each patient to determine which variants have a potential functional impact (silent, missense, nonsense, splice-site variants, etc.) on a gene. Variants were then selected for further analysis based on which were possessed by affected and unaffected individuals, where applicable. Lastly, *VAAST* was performed on each individual or intersected family dataset to determine which genes are negatively impacted by genetic variation. A background file containing variant population frequencies derived from the 1000 Genomes Project (16), the NHLBI Exome Sequencing Project, and the Complete Genomics diversity panel was used as a control dataset.

The *VAAST* ranked list of genes negatively impacted by deleterious variation for each patient or family was then analyzed by *PHEVOR* (4). This was done to select genes that will likely give rise to ALS when impacted by harmful variation. *PHEVOR* does this by first collecting genes previously shown to be associated with a phenotype as provided by the Human Phenotype Ontology (HPO) (17). *PHEVOR* then traverses multiple gene ontologies—such as the Gene Ontology, Mammalian Phenotype Ontology, and the Disease Ontology—using genes from the HPO gene list to find ontology nodes, and the genes contained in them, likely to be associated with the phenotype in question. This leads to the potential identification of genes previously associated with the phenotype in question and novel disease-causing gene candidates. These results are combined with variant prioritization results (such as from *VAAST*) to find and rank genes according to the degree that they are likely damaged and associated with the phenotype in question.

PHEVOR—using the HPO terms "Abnormality of the motor neurons," "Atrophy/Degeneration involving motor neurons," "Frontotemporal dementia," and "Amyotrophic lateral sclerosis"—was applied to the *VAAST* results of each studied individual/family to generate an initial candidate gene list. A literature search was then performed on the top 20 gene candidates to identify a potential cause of disease for the individual/family in question.

Results and Discussion

SOD1 was ranked as the top gene in the combined *VAAST* and *PHEVOR* analysis for FALS sample S27 (Table 2.1), who was selected as a validation control because they possessed a pathogenic *SOD1*:p.His44Arg variant (18). Not surprisingly, the *C9orf72* repeat expansion was not detected in sample S26 (the only *C9orf72* sample selected for sequencing) due to the inability of Illumina short-read sequencing to detect such repeats.

The top-ranking gene for patient S1 was *FIG4* (Table 2.1)*,* which has been previously shown to be causative for ALS (19). However, the particular variant (*FIG4*:p.Thr34Met) this patient possesses has not been previously described before within the context of ALS pathogenesis.

The eighth ranked gene for patient S4 was *CPEB2* (Table 2.1). *CPEB2* is thought to bind and regulate the translation of specific mRNAs (20), which is a common characteristic of ALS risk genes (21). The protein encoded by *CPEB2* has two RNA-recognition motifs and a prion-like domain that predisposes this protein to aggregation (22). These characteristics are also very common to ALS risk genes (22), which makes *CPEB2* a very intriguing ALS disease gene candidate.

Patient S5 possessed a *TP53INP2*:p.Trp71Cys variant that is only seen in one other individual in the Exome Aggregation Consortium (ExAC), giving it a global allele frequency of $9.083*10^{-6}$ (Table 2.1). *TP53INP2* is thought to be critical to autophagy processes in mammalian cells by acting as a scaffold protein at the autophagosome membrane (23), which is one of the cellular processes disrupted in ALS (21). Further, *TP53INP2* transgenic models have shown that muscle-specific expression of this gene has a role in muscle wasting (24), which is a key feature of ALS.

Patient S10 showed *HTRA2* as a possible disease causing gene candidate (Table 2.1). Intraneuronal inclusions of *HTRA2,* which is a serine protease that promotes apoptosis, have been reported within the context of ALS (25).

A variant in *SETX*, which is a DNA/RNA helicase known to cause ALS (26), was listed as possible disease gene candidate for patient S11 (Table 2.1).

The top-ranking gene for patient S12 was the gene *CSF1R* (colony stimulating factor 1 receptor) (Table 2.1). The protein product of this gene is the receptor for colony stimulating factor 1 and is critical in many processes including microglial proliferation and differentiation in the brain (27). Missense mutations of *CSF1R* have been previously shown to be causative for a disorder known as autosomal dominant diffuse leukoencephalopathy with spheroids (27). This disease shares many of the same clinical features as ALS including frontotemporal dementia, muscle weakness, and fasciculations (28). Interestingly, this patient showed signs that could be suggestive of leukoencephalopathy on MRI. This result highlights how diseases with similar signs and symptoms as ALS can potentially confound ALS genetic studies.

The analysis uncovered that FALS sample S14 possessed an *SOD1*:p.Ile114Thr (18) variant and an *ANG*:p.Lys41Ile (29) variant (Table 2.1), which are both known to be pathogenic for ALS. Furthermore, this individual has a *PSEN1*:p.Val94Leu variant, which occurs at the same amino acid position as a known early-onset Alzheimer risk variant (30). However, the amino acid change itself was different. Future work will have to be performed to determine if these multiple pathogenic mutations work together to cause poorer clinical outcomes.

ACTRT2, which is an actin associated protein thought to be involved in cytoskeleton organization (31), was the 10th ranked gene in sample S17 (Table 2.1). A region that includes *ACTRT2* has been previously associated with ALS (32).

A gene named *EHMT1* was the 14th ranked gene for Patient S25 (Table 2.1). *EHMT1* is a histone methyltransferase that is part of the E2F6 complex, which acts to repress transcription of specific gene targets (33). *EHMT1* was previously described as an ALS candidate gene in an experiment involving exome-sequencing of an ALS mother-father-proband trio pedigree (34). Table 2.1 summarizes the gene candidates explained above.

While a number of candidate risk genes were identified from this analysis, there were some shortcomings and areas that require improvement. The analysis was only limited to SNPs because no normal healthy control samples were available to be jointly called with the FALS samples. Insertion and deletion genotype calls for a sample can vary between variant calling runs, which can lead to false positives in analyses like *VAAST*. Joint variant calling with publicly available exome or whole genome sequencing datasets could be incorporated into the analysis to allow for the use of indel information. There was also a lack of genotype information from related individuals of the probands in question, which would allow for variant filtering and reduction of false positive genes.

Further sequencing of these individuals would allow for more accurate results.

Despite these shortcomings, a known or novel candidate risk gene was identified in 10 of the 13 (76.9%) FALS and PLS individuals/families from the combined *VAAST* and *PHEVOR* analysis. These results suggest that NGS technology, when combined with proper variant prioritization methods, can be very useful in identifying disease risk loci in small patient cohorts. The results also show NGS and variant prioritization methods can help clinicians sift through large genomic datasets to identify potentially actionable targets. Functional tests that rapidly determine if a candidate risk variant affects normal gene function will likely be needed for NGS testing to be useful in the clinical setting. Future work will be required to functionally validate if and how these novel candidate risk genes affect ALS pathogenesis.

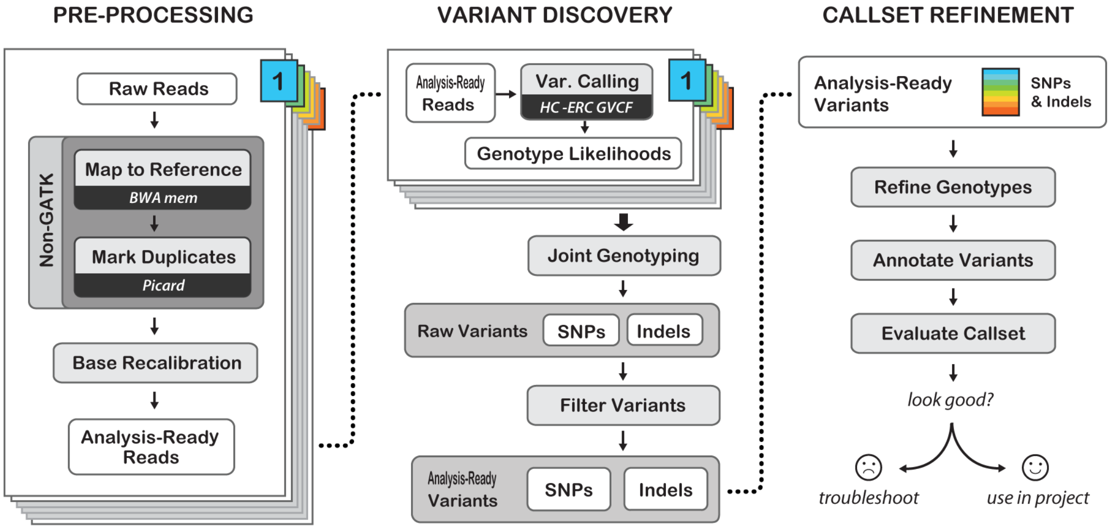


Figure 2.1 A diagram of the Genome Analysis Toolkit (GATK) pipeline version 3.0+. BWA and Picard Tools were used to perform the read mapping and duplicate marking steps, respectively. Used with permission from the Broad Institute, http://www.broadinstitute.org/gatk/.

Table 2.1 The most interesting disease gene candidates for each sample or family after the *VAAST*/*PHEVOR* analysis and literature search. The nucleotide changes responsible for each candidate gene are listed in the variant column. Variant coordinates correspond to the GRCh37 reference genome.

Sample	Gene	Variant	dbSNP ID	Variant	*PHEVOR* rank/*VAAST* p-value
S1	*FIG4*	6:110036315 C>T	rs375691683	Thr34Met	1/0.00895
S4	*CPEB2*	4:15004505 G>A	None	Gly70Ser	8/0.000899
S5	*TP53INP2*	20:33297128 G>C	rs200318321	Trp71Cys	6/0.000899
S10	*HTRA2*	2:74757348 T>C	rs150047108	Leu72Pro	16/0.0119
S11	*SETX*	9:135218103 A>C	rs145438764	Leu158Val	15/0.038
S12	*CSF1R*	5:149447846 C>A	None	Val520Phe	1/0.000899
S14	*SOD1*	21:33039672 T>C	rs121912441	Ile114Thr	1/0.000899
S14	*PSEN1*	14:73637697 G>T	rs63750831	Val94Leu	2/0.00279
S14	*ANG*	14:21161845 A>T	rs121909536	Lys41Ile	10/0.00334
S17	*ACTRT2*	1:2939276 G>T	rs369911854	Trp342Cys	10/0.000899
S25	*EHMT1*	9:140622895 G>A	rs144871446	Arg215Gln	14/0.00613
S27	*SOD1*	21:33036161 A>G	rs121912435	His44Arg	1/0.000899

References

1. Chapman PB, et al. (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* 364(26):2507-2516.

2. Lek M, et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285-291.

3. Hu H, et al. (2013) VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol* 37(6):622-634.

4. Singleton MV, et al. (2014) Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am J Hum Genet* 94(4):599-610.

5. Renton AE, Chio A, Traynor BJ (2014) State of play in amyotrophic lateral sclerosis genetics. *Nat Neurosci* 17(1):17-23.

6. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* 17(6):333-351.

7. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218):53-59.

8. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints.* 1303:3997.

9. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.

10. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.

11. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491-498.

12. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297-1303.

13. Van der Auwera GA, et al. (2013) From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43(11.10):1-33.

14. Kennedy B, et al. (2014) Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data. *Curr Protoc Hum Genet* 81:6 14 11-25.

15. Yandell M, et al. (2011) A probabilistic disease-gene finder for personal genomes. *Genome Res* 21(9):1529-1542.

16. 1000 Genomes Project Consortium, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56-65.

17. Kohler S, et al. (2014) The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Res* 42(Database issue):D966-974.

18. Rosen DR, et al. (1993) Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 362(6415):59-62.

19. Chow CY, et al. (2009) Deleterious variants of FIG4, a phosphoinositide phosphatase, in patients with ALS. *Am J Hum Genet* 84(1):85-88.

20. Kurihara Y, et al. (2003) CPEB2, a novel putative translational regulator in mouse haploid germ cells. *Biol Reprod* 69(1):261-268.

21. Robberecht W, Philips T (2013) The changing scene of amyotrophic lateral sclerosis. *Nat Rev Neurosci* 14(4):248-264.

22. King OD, Gitler AD, Shorter J (2012) The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. *Brain Res* 1462:61-80.

23. Nowak J, et al. (2009) The TP53INP2 protein is required for autophagy in mammalian cells. *Mol Biol Cell* 20(3):870-881.

24. Sala D, et al. (2014) Autophagy-regulating TP53INP2 mediates muscle wasting and is repressed in diabetes. *J Clin Invest* 124(5):1914-1927.

25. Kawamoto Y, et al. (2010) HtrA2/Omi-immunoreactive intraneuronal inclusions in the anterior horn of patients with sporadic and Cu/Zn superoxide dismutase (SOD1) mutant amyotrophic lateral sclerosis. *Neuropathol Appl Neurobiol* 36(4):331-344.

26. Chen YZ, et al. (2004) DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *Am J Hum Genet* 74(6):1128-1135.

27. Rademakers R, et al. (2012) Mutations in the colony stimulating factor 1 receptor (CSF1R) gene cause hereditary diffuse leukoencephalopathy with spheroids. *Nat Genet* 44(2):200-205.

28. Sundal C, Wszolek Z (1993) Adult-Onset Leukoencephalopathy with Axonal Spheroids and Pigmented Glia. *GeneReviews(R)*, eds Pagon RA, et al. (University of Washington, Seattle, WA).

29. Greenway MJ, et al. (2006) ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nat Genet* 38(4):411-413.

30. Jacquier M, et al. (2000) Presenilin mutations in colombian familial and sporadic AD sample. *Neurobiol Aging* 21:176.

31. Heid H, et al. (2002) Novel actin-related proteins Arp-T1 and Arp-T2 as components of the cytoskeletal calyx of the mammalian sperm head. *Exp Cell Res* 279(2):177-187.

32. Mok K, et al. (2013) Homozygosity analysis in amyotrophic lateral sclerosis. *Eur J Hum Genet* 21(12):1429-1435.

33. Ogawa H, Ishiguro K, Gaubatz S, Livingston DM, Nakatani Y (2002) A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells. *Science* 296(5570):1132-1136.

34. Chesi A, et al. (2013) Exome sequencing to identify de novo mutations in sporadic ALS trios. *Nat Neurosci* 16(7):851-855.

CHAPTER 3

THE EVOLVING GENETIC RISK FOR SPORADIC AMYOTROPHIC LATERAL SCLEROSIS

The following chapter is a manuscript co-authored by Summer B. Gibson, Spyridoula Tsetsou, Julie E. Feusier, Karla P. Figueroa, Mark B. Bromberg, Lynn B. Jorde, Stefan M. Pulst.

Summer B. Gibson and I contributed equally to this work and are listed as co-first authors. I was responsible for writing the manuscript and performing the statistical analyses within it. This manuscript was accepted for publication to *Neurology* on March 17th, 2017 and was published on July 18th, 2017. The research article can be found in Gibson and Downie et al. (2017) The evolving genetic risk for sporadic ALS *Neurology* 18;89(3):226-233 and is available at http://www.neurology.org/content/89/3/226. *Neurology* has given me permission to include the text of the manuscript and has confirmed that no formal license is required from their publisher.

Abstract

Objective

To estimate the genetic risk conferred by known amyotrophic lateral sclerosis (ALS)–associated genes to the pathogenesis of sporadic ALS (SALS) using variant allele frequencies combined with predicted variant pathogenicity.

Methods

Whole exome sequencing and repeat expansion PCR of *C9orf72* and *ATXN2* were performed on 87 patients of European ancestry with SALS seen at the University of Utah. DNA variants that change the protein coding sequence of 31 ALS-associated genes were annotated to determine which were rare and deleterious as predicted by MetaSVM. The percentage of patients with SALS with a rare and deleterious variant or repeat expansion in an ALS-associated gene was calculated. An odds ratio analysis was performed comparing the burden of ALS-associated genes in patients with SALS vs 324 normal controls.

Results

Nineteen rare nonsynonymous variants in an ALS-associated gene, 2 of which were found in 2 different individuals, were identified in 21 patients with SALS. Further, 5 deleterious *C9orf72* and 2 *ATXN2* repeat expansions were identified. A total of 17.2% of patients with SALS had a rare and deleterious variant or repeat expansion in an ALS-associated gene. The genetic burden of ALS-associated genes in patients with SALS as predicted by MetaSVM was significantly higher than in normal controls.

Conclusions

Previous analyses have identified SALS-predisposing variants only in terms of their rarity in normal control populations. By incorporating variant pathogenicity as well as variant frequency, we demonstrated that the genetic risk contributed by these genes for SALS is substantially lower than previous estimates.

Introduction

Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease of the upper and lower motor neurons, which eventually leads to death within an average of 3–5 years[1] after symptom onset. ALS is classified as familial (FALS) when a clear family history of ALS exists and sporadic (SALS) when it does not. No clinical features reliably distinguish FALS from SALS. Genetic research on ALS has largely been focused on FALS, which represents 10% of ALS cases.[1] Most FALS is inherited in autosomal dominant fashion. However, this transmission pattern can be complicated by the early death of unrecognized affected family members due to non-ALS causes, misdiagnoses in older affected individuals, small family sizes, incomplete penetrance of genetic risk factors, and the development of disorders associated with ALS, such as frontotemporal dementia (FTD). Thus, sporadic and familial forms of ALS can be difficult to distinguish, and much remains unknown about the roles of genetic factors in FALS and especially in SALS. The discovery of the pathogenic $(G_4C_2)_n$ hexanucleotide repeat expansion of *C9orf72* in a large percentage of FALS and SALS patients,[2-4] as well as the identification of other ALS genes in patients with SALS,[5, 6] has highlighted the importance of genetic risk factors in SALS pathogenesis. The significance of genetics in SALS is further supported by ALS genome-wide association studies, which estimate the heritability of

ALS to be at least 21.0%.[7]

With the growing affordability and avail-ability of next-generation sequencing technologies, along with the advent of specific treatments for certain genetic forms of ALS,[8] it is increasingly important to understand the genetic factors in causing SALS. Currently, there is considerable variation in estimates of the percentage of SALS cases caused by genetic variants, ranging from 11%[5] to 28%[9] in populations of European ancestry. This variation is due largely to differences in estimation methods. In one large study[9], the percentage was derived by calculating the portion of SALS cases with a rare (minor allele frequency [MAF] <0.01), protein-altering variant in a set of known ALS genes. Using variant rarity as the main criterion for pathogenicity may have inflated the risk estimate as the majority of rare nonsynonymous variants are not thought to be pathogenic.[10]

In this study, we sought to better estimate the percentage of SALS cases that have an identifiable genetic factor likely responsible for disease pathogenesis. To address this, a joint approach utilizing both allele frequency and variant pathogenicity prediction was used to determine the percentage of SALS cases that possess a potentially deleterious genetic variant in an ALS-associated gene.

Methods

Standard protocol approvals, registrations, and patient consents

The sample collection and study design we performed was approved by the University of Utah Institutional Review Board. Written informed consent for disease-specific genetic studies was obtained from each patient who participated in this study.

Participants

Patients with ALS diagnosed at the University of Utah from 2011 to 2013 were invited to participate in genetic studies. All participants were seen by neuromuscular specialists and diagnosed with probable or definite ALS according to revised El Escorial criteria.[11] These patients were followed longitudinally in our motor neuron disease clinic. Patients with SALS were identified as having no self-reported family history of ALS, probable ALS, or FTD. In total, 96 patients with SALS were enrolled in this study. DNA was obtained from whole blood of each participant using the Gentra Puregene Blood Kit (Qiagen, Venlo, Netherlands).

Identification of deleterious *ATXN2* and *C9orf72* repeat expansions

ATXN2 CAG repeat size was determined by fluorescent PCR amplification. Repeat lengths between 29 and 33 were considered to be of intermediate length and deleterious.[12] The detection of *C9orf72* GGGGCC repeat expansions was performed by using previously established repeat primed-PCR and amplicon length analysis criteria.[13]

Whole exome sequencing

Patient DNA was exome enriched by the SeqCap EZ Exome Enrichment Kit v3.0 (Roche [Basel, Switzerland] NimbleGen) and sequenced by the Illumina (San Diego, CA) HiSeq platform to generate 101-bp, paired-end reads that covered target regions to an average depth ranging from 41X to 224X per sample. Reads were then aligned to the GRCh37 reference genome using BWA-MEM v0.7.12. Picard Tools v1.130 was used to perform indexing, coordinate sorting, and duplicate read marking of all aligned genomic reads. Variant calling and quality filtering were performed using Genome Analysis Toolkit's (v3.3-0) HaplotypeCaller and variant quality score recalibration (VQSR) methods.[14] In order to properly power VQSR filtering, 99 CEU (Utah residents [CEPH] with northern and western European ancestry) and 92 GBR (British in England and Scotland) individuals from the 1000 Genomes Project[15] with exome sequencing data were included in the genotyping steps.

Quality control

Utah's population is outbred and genetically resembles other populations of northern European ancestry.[15-17] As a result, we focused our analysis on patients with SALS who were of European ancestry in order to limit population stratification effects. Patient ancestry derived from genetic data is more reliable than self-reported ancestry, which has been used in previous ALS studies.[9, 18] An Admixture[19] analysis (K=3) was performed to determine the genetic ancestry of each patient with SALS. Any participants with less than 90% European ancestry were removed from further analysis. Next, principal components analysis (PCA) was performed using smartpca[20] to remove poor-quality samples. Any sample with an eigenvector value more than 6 SDs from the mean for the first 10 principal components was discarded. Finally, the sex of each patient with SALS was inferred by PLINK 1.9[21] and compared to the reported sex to identify sample identification errors.

Variant annotation

SnpEff[22] (v4.1) was used to identify protein-coding and splice-site altering genetic variants. These variants were then annotated with information from the database for nonsynonymous single nucleotide polymorphism functional predictions (dbNSFP; sites.google.com/site/jpopgen/dbNSFP) v2.9.[23] dbNSFP contains 11 different in silico functional prediction methods that determine which single nucleotide variants (SNVs) are likely to alter protein function. MetaSVM was chosen as the primary method to determine variant pathogenicity as it has been shown to have a better predictive ability than other methods.[24] Insertion, deletion, and splice-site acceptor/donor variants were classified as deleterious. Variants were also annotated with European-specific MAF estimates from the Exome Aggregation Consortium (ExAC)[25] by dbNSFP. A manual search of the Amyotrophic Lateral Sclerosis Online Database (ALSoD),[26] the Single Nucleotide Polymorphism database (dbSNP),[27] and the Human Gene Mutation Database (HGMD)[28] was performed to identify known ALS pathogenic variants.

Genetic risk analysis

To determine the proportion of patients with SALS who have a potentially disease-causing variant in a SALS gene, all annotated rare (European MAF <0.001), protein-coding, and splice-site altering variants in 31 ALS-associated genes (*ANG*, *CHCHD10*, *CHMP2B*, *DAO*, *DCTN1*, *ELP3*, *ERBB4*, *EWSR1*, *FIG4*, *FUS*, *GLE1*, *GRN*, *HNRNPA1*, *HNRNPA2B1*, *MATR3*, *NEFH*, *NEK1*, *OPTN*, *PFN1*, *SETX*, *SOD1*, *SPAST*, *SQSTM1*, *SS18L1*, *TAF15*, *TARDBP*, *TBK1*, *TUBA4A*, *UBQLN2*, *VAPB*, *VCP*) were assessed. A MAF of 0.001 corresponds roughly to the European allele frequency of *SOD1*:p.Asp91Ala[29], which is the most common known pathogenic variant we could identify in ALSoD. The proportion of patients with SALS who possessed a rare and deleterious variant, as determined by MetaSVM, in at least 1 of the 31 ALS-associated genes or had a deleterious repeat expansion in *C9orf72* or *ATXN2* was then calculated. The proportion of patients with SALS who possessed a rare variant, deleterious or not, or a repeat expansion in an ALS-associated gene was also calculated as a reference. All variants were assumed to act in a dominant fashion, like most ALS-causing variants.[30]

Odds ratio analysis

Genetic burden analysis determines if there is a difference in the amount of pathogenic variation, or burden, in a set of genes between cases and controls. To determine whether the combination of variant frequency and MetaSVM predictions identified variant pathogenicity better than variant frequency alone, we estimated the excess burden of ALS-associated genes in SALS cases vs healthy controls. To do so, whole exome sequence data from 714 individuals from 181 families of the Simons Simplex Collection (SSC)[31] were analyzed. The SSC dataset contains exome sequence data from children with autism, an unaffected sibling, and their unaffected parents. These samples underwent joint variant calling with the SALS exomes using the same pipeline as described above. Admixture and PCA were performed as previously described on 362 unaffected parents to select for high-quality controls of European ancestry. Variant calls were limited to exome capture regions with at least 53 coverage on average in both the SALS and SSC cohorts. The proportion of SSC controls with a rare and deleterious variant in at least 1 of the same 31 ALS-associated genes was calculated. An odds ratio (OR) analysis was then performed to determine whether the burden of ALS-associated genes is higher in patients with SALS than in normal controls. An OR analysis comparing the genetic burden when only variant frequency was utilized was also performed. The significance of this OR was determined by a one-tailed Fisher exact test.

Results

Patient cohort characteristics

The Admixture (Figure 3.1A) and PCA (Figure 3.1B) results showed that 9 of the 96 patients with SALS possessed significant non-European ancestry or were genetic outliers. No sex mismatches were detected in the data. The 87 patients with SALS of European ancestry were selected for analysis, and characteristics of these patients are detailed in Table 3.1. We selected 324 SSC parents as high-quality European controls because Admixture showed that 38 of the parents were likely non-European.

Known ALS-associated genetic variants

We identified pathogenic *C9orf72* hexanucleotide repeat expansions in 5 of 87 patients with SALS (5.7%). Two patients with SALS (2.3%) possessed ALS-associated trinucleotide repeat expansions in *ATXN2* (31 and 32 repeats in length, respectively). We compared the rare (European MAF <0.001) coding variants in 31 ALS-associated genes uncovered in the SALS cohort to known ALS risk variants contained in ALSoD, dbSNP, and HGMD. This comparison revealed only one known ALS-associated rare SNV (*SOD1*:p.Asp91Ala[29]), which was found in 2 heterozygous patients (Table 3.2).

Potentially novel ALS variants

After examining the 31 ALS-associated genes in our patient cohort, we identified 18 rare coding variants (European MAF <0.001) not previously described in ALS (Table 3.2). Of these variants, 10 were not found in dbSNP (v141). Furthermore, 6 were novel, as they were not found in ExAC, the 1000 Genomes Project dataset,[15] or the National Heart, Lung, and Blood Institute Exome Sequencing Project dataset. One novel single nucleotide frameshift insertion was found in *SQSTM1* (chr5:179263453A>AT). A novel splice-site acceptor variant in *NEK1* (chr4:170428944C>T) was found in 2 patients.

Genetic risk analysis

Among the 31 ALS-associated genes, 19 rare variants were found in 21 patients. The *FIG4*:p.Leu643* variant was the only variant not Sanger-validated due to a lack of high-quality DNA. When combined with the 5 *C9orf72* and 2 *ATXN2* deleterious repeat expansions, 28 rare variants or repeat expansions were found in 25 patients across all ALS-associated genes. Three patients had 2 rare variants in an ALS gene. One patient had a *GLE1*:p.Met134Val (chr9:131277886A>G) variant in addition to a pathological *C9orf72* repeat expansion. Another patient had a *C9orf72* repeat expansion in addition to a rare *SETX*:p.Thr2507Ala (chr9:135140228T>C) missense variant. Finally, one patient possessed *SPAST*:p.Pro42His (chr2:32289025C>A) and *ERBB4*:p.Thr643Ile (chr2:212522497G>A) missense variants.

These 28 rare variants were used to calculate the proportion of patients with SALS who have a rare mutation in at least one ALS-associated gene. A total of 25 patients with SALS (28.7%) had at least one rare variant or pathogenic repeat expansion in an ALS gene when variant deleteriousness was not considered. However, only 4 of the 17 SNVs annotated with MetaSVM were considered deleterious (Table 3.2). As a result, the proportion of patients with SALS with a rare and deleterious SNV or repeat expansion in an ALS-associated gene was 17.2% (15/87 patients; Figure 3.2). Variant predictions from 10 other methods were also used (Table 3S.1 at Neurology.org), which yielded proportions ranging between 14.9% and 21.8%.

OR analysis

The genetic burden of ALS-associated genes in patients with SALS was compared with the burden among 324 SSC controls. Using only rare variant frequency as a criterion for assessing burden, patients with SALS had a modest increase in burden compared to controls (OR 1.90, $p < 0.025$; Table 3.3). However, when variant pathogenicity was added by incorporating MetaSVM results and variant frequency, SALS cases showed a much higher burden in ALS-associated genes compared to controls (OR 4.98, $p < 9 \times 10^{-5}$; Table 3.3). Other variant prediction methods in dbNSFP yielded similar findings, but the OR analysis using MetaSVM predictions resulted in the highest *p* value (Table 3S.2).

Discussion

We report findings from a genomic analysis of 87 patients with SALS of European origin. In total, 28 rare variants were found in 33 ALS genes in our patient cohort. Only one non-repeat variant that has been previously described in ALS pathogenesis was observed (*SOD1*:p.Asp91Ala). This variant is known to cause autosomal recessive ALS and was predicted to be deleterious by MetaSVM. *SOD1*:p.Asp91Ala has also been suggested to act in a dominant fashion; however, few instances of this have been reported.[32] In addition, we identified 18 rare variants in ALS-associated genes that have not been described previously in patients with ALS. Of these 18 variants, 5 either caused a protein loss of function or were predicted to be deleterious by MetaSVM. One is a frameshift variant in the ubiquitin-associated domain of *SQSTM1*. This change likely ablates *SQSTM1*'s ability to bind ubiquitinated substrates, which is often seen in *SQSTM1* variants that cause ALS.[33] *NEK1*:c.1750-1G>A is a novel loss of function SNV that ablates the splice acceptor site of intron 19, which is located approximately in the middle of the gene. *NEK1*:p.Gly646Arg is another damaging variant that was discovered in *NEK1*; however, it does not occur in a defined protein domain.

FUS:p.Gly465Glu is predicted to be damaging by MetaSVM and affects an amino acid one position upstream from previously reported SALS variant (*FUS*:p.Met464Ile).[34] *ERBB4:*p.Gly735Val is an SNV predicted to be deleterious and occurs in the tyrosine kinase domain of erbB-4. The tyrosine kinase function of erbB-4 is required for protein autophosphorylation and triggering downstream signaling cascades upon activation. A variant in the tyrosine kinase domain of erbB-4, which was identified from an FALS family, has been shown to reduce protein autophosphorylation and likely causes ALS.[35] Additional studies will determine the functional importance of these variants on cellular and molecular mechanisms.

We have demonstrated that using variant pathogenicity predictions is more reliable than variant frequency alone to determine the proportion of patients with SALS whose disease is likely caused by a variant in an ALS-associated gene. The relative effect of ALS-associated genes is stronger when variant pathogenicity is considered instead of only variant rarity. This follows from the fact that an appreciable proportion of rare nonsynonymous variants are not predicted to be functionally damaging.[10] Thus, only a subset of rare variants in ALS-associated genes are pathogenic.

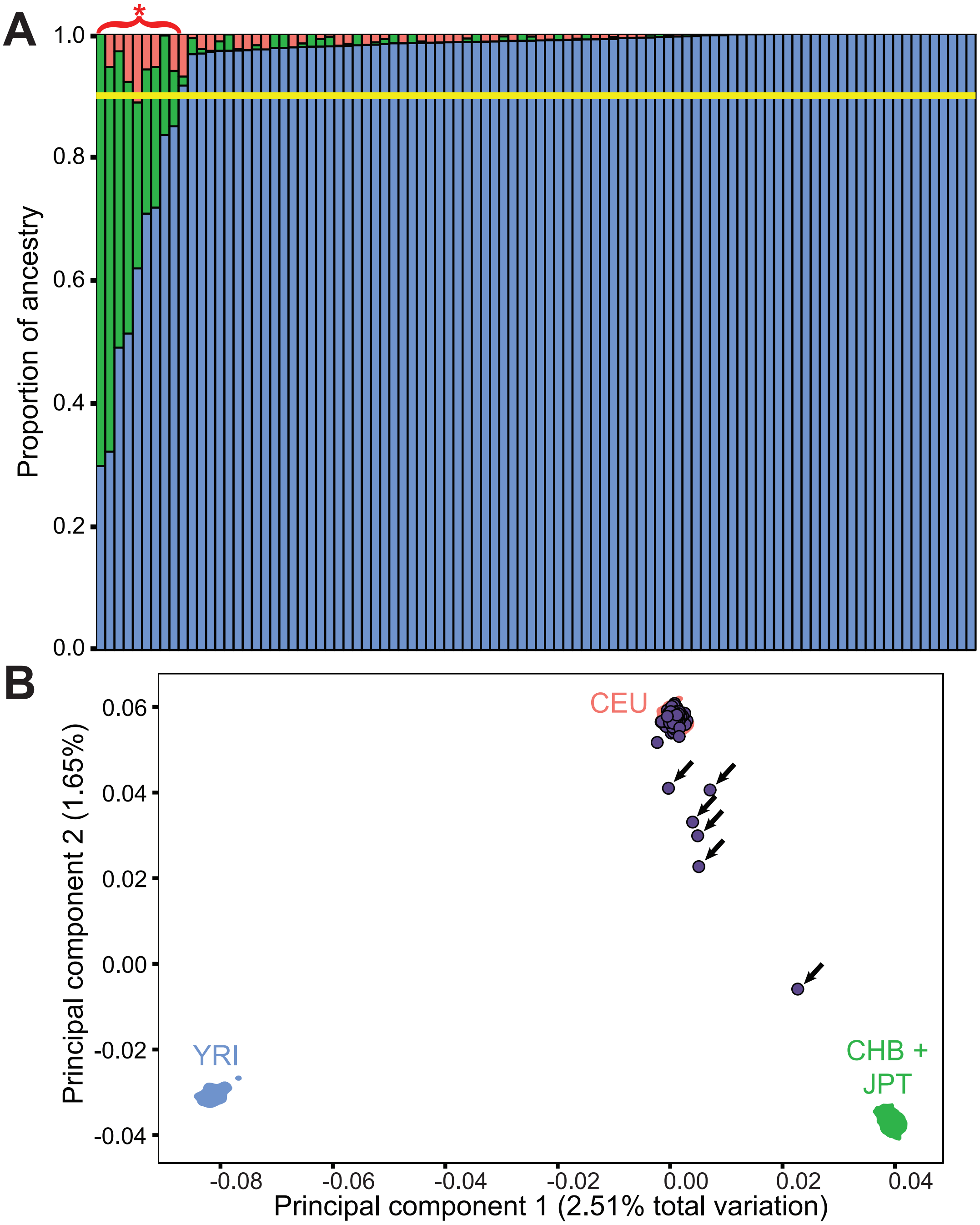
Our approach to estimating the genetic contribution of a large panel of known ALS-associated genes by directly predicting variant pathogenicity differs from earlier approaches. The first attempts at determining the proportion of genetically caused SALS cases did so by calculating the proportion of patients who had a protein-coding variant in a panel of 5–7 ALS-associated genes.[18, 36-38] These analyses yielded estimates ranging from 2.8% to 11%, which are lower than our estimate of 17.2%. A more recent study, in which variant rarity (MAF <1%) was used as the sole criterion for pathogenicity in a panel of 17 ALS-associated genes, concluded that genetic factors may cause 27.8% of SALS cases,[9] a figure similar to our estimate when only variant rarity is considered (Table 3S.1). However, these variants (MAF <1%) are not significantly more common in our patients with SALS than in unaffected controls (OR 1.25, $p > 0.25$), suggesting that many of them are not pathogenic. The same 17 ALS-associated genes are significantly more burdened in patients with SALS than controls when variant rarity (MAF <1%) and pathogenicity (estimated by MetaSVM) are combined (OR 2.61, $p < 0.02$). These OR differences support our conclusion that variant frequency alone is not a sufficient predictor of SALS risk.

Another analysis of 33 ALS-associated genes defined only novel and extremely rare variants (MAF ≈ 0.0002) as pathogenic and found that 14.5% of SALS cases could be attributed to genetic causes.[39] In our sample, the genetic burden of ALS-associated genes in patients with SALS is less when pathogenicity is defined in the same way than when MetaSVM is integrated (OR 2.24, $p < 0.01$ vs OR 5.52, $p < 2 \times 10^{-4}$). These results demonstrate that direct predictions of variant pathogenicity are important for defining genetic risk in SALS and other genetic diseases.

Our results also highlight that genetic factors play an important role in the disease, the clinical relevance of which will become even more important as genetic specific treatments become available. Further, exome or targeted sequencing of patients with SALS and their family members is likely warranted to provide adequate genetic counseling. In addition, our results suggest the distinction between SALS and FALS may be problematic as heritable risk factors are found in a significant proportion of patients with SALS. Future genetic investigations of patients with SALS are needed to broaden the scope of SALS-associated loci. Studies with larger patient cohorts that incorporate measures of variant pathogenicity will also be needed to further pinpoint the proportion of SALS cases with an identifiable probable genetic cause of disease, especially as more ALS-associated genetic loci are discovered.

Our study has several limitations. First, the size of the SALS cohort was limited, especially given the genetically heterogeneous nature of ALS. Second, because we focused on individuals of European ancestry, our findings may not be completely applicable to ALS found in other populations. Third, 13 of the 324 (4.0%) healthy control samples used in this study had at least one rare and deleterious variant in ALS-associated genes as predicted by MetaSVM (Table 3.3; Table 3S.3). The mean age of these individuals was 41.76 (SD 5.92) years, which is much lower than the average age at onset of SALS at 56 years of age.[40] It is possible that some of the control individuals with these variants could develop ALS later in life.

Figure 3.1 Admixture and principal components analysis (PCA) plots show the ancestry and sample quality of the sporadic amyotrophic lateral sclerosis (SALS) cohort. (A) An Admixture plot where each bar represents a patient with SALS (in total 96 patients). The height of each colored bar represents the amount of ancestry each individual derives from. Blue = European (CEU), green = East Asian (CHB + JPT), and red = African (YRI). Individuals with less than 90% European ancestry (yellow bar) were removed from further analysis. The 9 patients with SALS with less than 90% European ancestry are indicated with a red asterisk. (B) PCA plot of all 96 individuals with 1,000 genomes data (CEU = Utah residents [CEPH] with northern and western European ancestry; CHB = Han Chinese in Beijing, China; JPT = Japanese in Tokyo, Japan; YRI = Yoruba in Ibadan, Nigeria). Shaded areas represent the area over which the kernel density of each respective 1000 genomes population spans. SALS samples are listed as purple circles. Arrows indicate non-European individuals who were removed from further analysis.


A
*
1.0
0.8
0.6
0.4
0.2
0.0
Proportion of ancestry
B
CEU
YRI
CHB +
JPT
0.06
0.04
0.02
0.00
-0.02
-0.04
Principal component 2 (1.65%)
-0.08
-0.06
-0.04
-0.02
0.00
0.02
0.04
Principal component 1 (2.51% total variation)

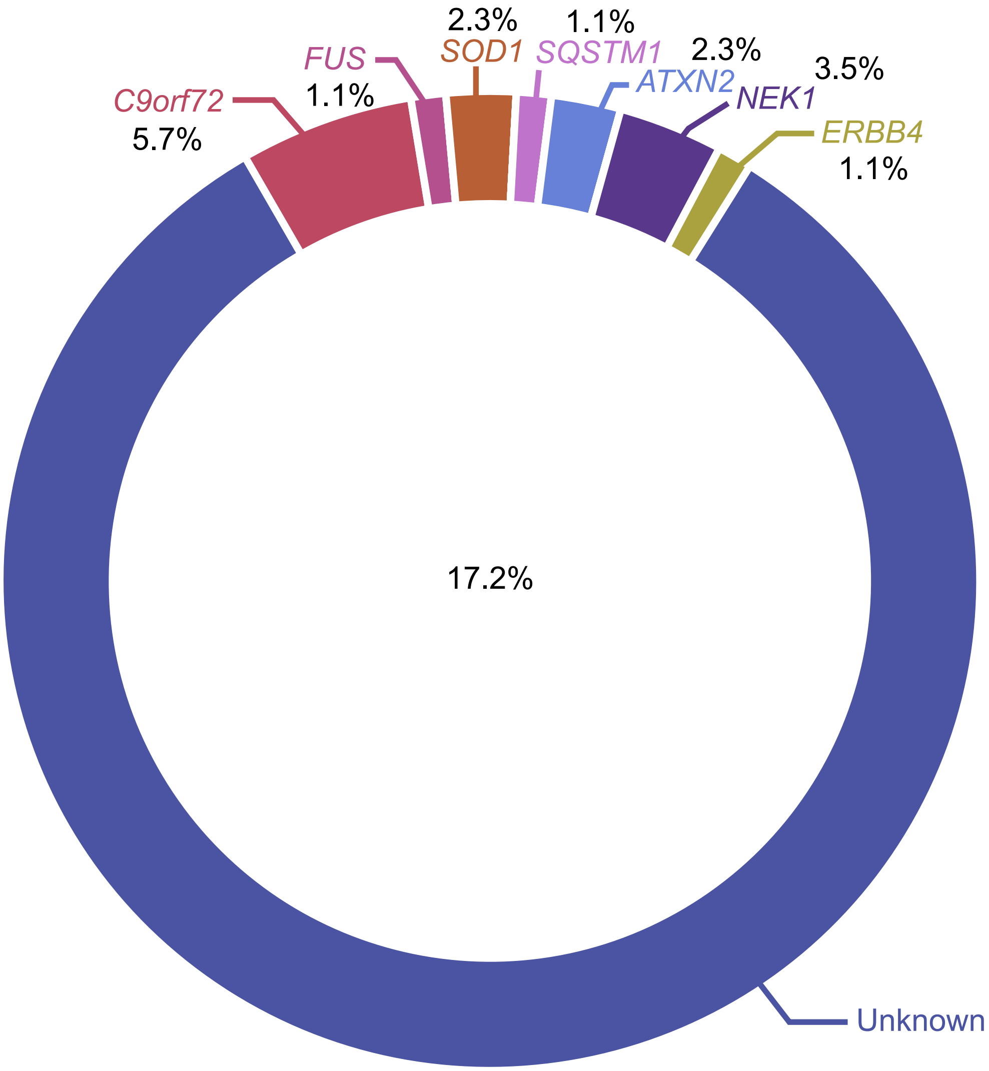


Figure 3.2 Percentage of sporadic amyotrophic lateral sclerosis (SALS) cases with an identifiable genetic variant likely responsible for disease. The percentage next to each gene indicates what percentage of SALS cases have a rare and pathogenic variant in that gene. A majority (82.8%) of SALS cases have no identifiable genetic variants potentially responsible for their disease.

Table 3.1 Detailed summary of the sporadic amyotrophic lateral sclerosis cohort before and after selecting for European patients. Abbreviation: ALSFRS-R = Amyotrophic Lateral Sclerosis Functional Rating Scale–revised.
[a]Survival data were available for 78 participants overall and for 72 analyzed.
[b]Rate of progression data were available for 76 participants overall and for 69 analyzed.

Variables	Overall (n=96)	Analyzed (n=87)
Male sex, % (n)	61.5 (59)	62 (54)
Bulbar onset, % (n)	28.1 (27)	27.6 (24)
Age at onset, y, mean ± SD	58.7 ± 12.1	59.3 ± 11.6
Age at onset, y, range	19–83	19–83
Survival, y, mean ± SD[a]	2.9 ± 1.7	3 ± 1.8
Survival, y, range[a]	0.5–12	0.5–12
Rate of progression, ALSFRS-R/y, mean ± SD[b]	−13.1 ± 9.7	−12.4 ± 9.5
Rate of progression, ALSFRS-R/yr, range[b]	−1 to −60	−1 to −60

Table 3.2 The 19 rare nonsynonymous variants found in the 31 amyotrophic lateral sclerosis–associated genes. Abbreviations: dbSNP = Single Nucleotide Polymorphism database; ExAC = Exome Aggregation Consortium; MAF = minor allele frequency; SALS = sporadic amyotrophic lateral sclerosis.

[a] Variants that were considered to be deleterious by MetaSVM. Indel and splice-site variants were automatically considered deleterious.

Chromosome:Position (GRCh37)	dbSNP141 ID	Amino acid change	Gene	MetaSVM prediction	ExAC European MAF	No. of patients with SALS
2:32289025 C>A		p.Pro42His	*SPAST*	Tolerated	0.0	1
2:74593484 T>A		p.Ser883Cys	*DCTN1*	Tolerated	4.50E-05	1
2:212251806 T>C	rs143251275	p.Thr1085Ala	*ERBB4*	Tolerated	4.50E-05	1
2:212483999 C>A[a]		p.Gly735Val[a]	*ERBB4*[a]	Damaging[a]	0.0[a]	1[a]
2:212522497 G>A		p.Thr643Ile	*ERBB4*	Tolerated	0.0001049	1
4:170400673 C>G[a]		p.Gly646Arg[a]	*NEK1*[a]	Damaging[a]	4.54E-05[a]	1[a]
4:170428944 C>T[a]		c.1750-1G>A (Splice variant)[a]	*NEK1*[a]	NA[a]	0.0[a]	2[a]
5:138629745 G>T		p.Ala26Ser	*MATR3*	Tolerated	0.0	1
5:138652744 G>A	rs201075828	p.Ala378Thr	*MATR3*	Tolerated	0.0001978	1
5:179263453 A>AT[a]		p.Glu396fs[a]	*SQSTM1*[a]	NA[a]	0.0[a]	1[a]
6:110106211 T>A		p.Leu643*	*FIG4*	Tolerated	0.0	1
9:131277886 A>G		p.Met134Val	*GLE1*	Tolerated	4.65E-05	1
9:135140228 T>C	rs142303658	p.Thr2507Ala	*SETX*	Tolerated	7.49E-05	1
9:135144866 G>A	rs375949756	p.Pro2433Leu	*SETX*	Tolerated	1.57E-05	1
9:135203159 G>C	rs148604312	p.Gln1276Glu	*SETX*	Tolerated	0.0004495	1
9:135203725 G>A	rs139559547	p.Ser1087Phe	*SETX*	Tolerated	1.50E-05	1
16:31195253 T>A	rs372638663	p.Ser89Thr	*FUS*	Tolerated	3.00E-05	1
16:31202284 G>A[a]	rs141684472[a]	p.Gly465Glu[a]	*FUS*[a]	Damaging[a]	0.0001352[a]	1[a]
21:33039603 A>C[a]	rs80265967[a]	p.Asp91Ala[a]	*SOD1*[a]	Damaging[a]	0.00087[a]	2[a]

Table 3.3 Odds ratio (OR) analyses comparing the genetic burden of amyotrophic lateral sclerosis (ALS)–associated genes of patients with sporadic ALS (SALS) vs controls. Abbreviations: CI = confidence interval; MAF = minor allele frequency; SSC = Simons Simplex Collection.
The incorporation of MetaSVM predictions of deleteriousness shows a much higher genetic burden of ALS-associated genes in patients with SALS than by considering variant rarity alone.

Variant prediction model	No. of patients with SALS with a rare and deleterious mutation in an ALS-associated gene/number without	No. of SSC individuals with a rare and deleterious mutation in an ALS-associated gene/number without	Odds Ratio (95% CI)	*p* Value
Rare (MAF <0.001)	22/65	49/275	1.90 (1.07–3.36)	0.022
Rare + MetaSVM	15/72	13/311	4.98 (2.27–10.94)	8.90×10^{-5}

Table 3S.1 Proportion of SALS cases with a rare and pathogenic variant for each method in dbNSFP.

Variant prediction model	Proportion of SALS patients with a rare *and* pathogenic variant or a pathogenic repeat expansion in an ALS-associated gene
Rare (MAF < 0.001)	28.7%
MutationTaster	21.8%
MutationAssessor	14.9%
SIFT	20.7%
Polyphen2 HVAR	17.2%
Polyphen2 HDIV	19.5%
FATHMM	19.5%
MetaSVM	17.2%
MetaLR	19.5%
LRT	17.2%
PROVEAN	14.9%
CADD	19.5%

Table 3S.2 Genetic burden OR results for each variant prediction model in dbNSFP.

Variant prediction model	Number of SALS patients with a rare and pathogenic mutation in an ALS-associated gene / number without	Number of SSC individuals with a rare and pathogenic mutation in an ALS-associated gene / number without	Odds Ratio (95% CI)	p-value
Rare (MAF < 0.001)	22/65	49/275	1.90 (1.07-3.36)	0.022
Rare + Mutation Taster	18/69	33/291	2.30 (1.22-4.33)	0.009
Rare + Mutation Assessor	13/74	14/310	3.89 (1.75-8.63)	0.001
Rare + SIFT	16/71	31/293	2.13 (1.11-4.11)	0.021
Rare + Polyphen2 HVAR	14/73	31/293	1.81 (0.92-3.58)	0.066
Rare + Polyphen2 HDIV	16/71	33/291	1.99 (1.04-3.81)	0.032
Rare + FATHMM	16/71	22/302	3.09 (1.55-6.19)	0.002
Rare + MetaSVM	15/72	13/311	4.98 (2.27-10.94)	8.90 x 10^{-5}
Rare + MetaLR	16/71	17/307	4.07 (1.96-8.44)	2.40 x 10^{-4}
Rare + LRT	14/73	23/301	2.51 (1.23-5.12)	0.011
Rare + PROVEAN	13/74	19/305	2.82 (1.33-5.97)	0.007
Rare + CADD	16/71	27/297	2.48 (1.27-4.85)	0.008

Table 3S.3 The rare coding variants in ALS-associated genes found in the SSC control cohort.

Chromosome:Position (GRCh37)	dbSNP141 ID	Amino acid Change	Gene	MetaSVM prediction	ExAC European MAF	Number of control individuals
2:74594827 T>C		p.Tyr727Cys	*DCTN1*	Damaging		1
2:74598791 G>C		p.Ala173Gly	*DCTN1*	Tolerated		1
2:212248504 G>A		p.His1255Tyr	*ERBB4*	Tolerated		1
2:212252671 C>T	rs372352845	p.Gly1061Glu	*ERBB4*	Tolerated	6.00E-05	1
2:212530199 G>T	rs200792124	p.Pro574Thr	*ERBB4*	Tolerated	4.52E-05	1
2:212537978 A>T	rs141594820	p.Phe543Ile	*ERBB4*	Tolerated	0.0002407	1
2:212566740 T>C	rs368860175	p.Ile481Val	*ERBB4*	Tolerated	1.50E-05	1
2:213403205 G>A	rs201202926	p.Ala17Val	*ERBB4*	Tolerated	0.0001054	1
4:170458958 A>C		c.1665+2T>G (splice variant)	*NEK1*	NA		1
4:170476956 C>G		p.Gly493Arg	*NEK1*	Tolerated		1
4:170483338 T>C		p.Thr344Ala	*NEK1*	Tolerated	0.0001917	1
5:179250875 C>T		p.Arg107Trp	*SQSTM1*	Tolerated	3.10E-05	1
5:179250908 C>T	rs200152247	p.Pro118Ser	*SQSTM1*	Tolerated	0.0002582	1
6:110037748 C>T		p.Ala89Val	*FIG4*	Tolerated	4.50E-05	1
6:110110877 T>G		p.Leu726Trp	*FIG4*	Tolerated		1
6:110146434 T>C		p.Met897Thr	*FIG4*	Tolerated	1.50E-05	1
7:26237018 C>A		p.Ala73Ser	*HNRNPA2B1*	Tolerated		1
8:27967909 G>A	rs144486746	p.Arg139His	*ELP3*	Tolerated	8.99E-05	1
9:35065261 G>T		p.Pro188His	*VCP*	Damaging		1
9:131285907 C>T	rs146025848	p.Arg227Cys	*GLE1*	Tolerated	0.0008836	1
9:131285938 G>A	rs139953543	p.Arg237Gln	*GLE1*	Tolerated	3.03E-05	1

Table 3S.3 Continued

Chromosome:Position (GRCh37)	dbSNP141 ID	Amino acid Change	Gene	MetaSVM prediction	ExAC European MAF	Number of control individuals
9:131287520 G>A	rs147943229	p.Arg316Gln	*GLE1*	Tolerated	0.0008859	2
9:135140000 A>T	rs368269464	p.Phe2554Ile	*SETX*	Tolerated	0.0001049	1
9:135171367 G>C	rs142917412	p.Gln2000Glu	*SETX*	Tolerated	0.0002548	1
9:135202358 A>T		p.Ser1543Thr	*SETX*	Tolerated		1
9:135202552 G>T	rs143661911	p.Ala1478Glu	*SETX*	Tolerated	0.0005244	2
9:135202889 A>G	rs140147684	p.Ser1366Pro	*SETX*	Tolerated	0.0003896	1
9:135205564 C>T		p.Cys474Tyr	*SETX*	Damaging	1.53E-05	1
9:135205882 G>A		p.Ser368Phe	*SETX*	Tolerated		1
9:135221659 T>C		p.His126Arg	*SETX*	Damaging	7.50E-05	1
10:13151192 C>A		p.Pro24Thr	*OPTN*	Tolerated		1
10:13158282 G>A		p.Gly190Arg	*OPTN*	Tolerated		1
10:13168019 G>T		p.Glu408*	*OPTN*	not scored and excluded		1
12:64879788 G>A		p.Arg444Gln	*TBK1*	Tolerated	4.52E-05	1
12:109283278 C>T	rs201583577	p.Arg115Trp	*DAO*	Damaging	0.0001199	1
12:109288127 G>A	rs200850756	p.Arg199Gln	*DAO*	Damaging	0.0002237	1
16:31193926 C>T		p.Ser44Phe	*FUS*	Damaging		1
16:31193959 ATTC>A		p.Ser57del	*FUS*	NA	0.0002098	1
16:31196366 G>C		p.Gln210His	*FUS*	Damaging		1
16:31196412 G>A		p.Gly226Ser	*FUS*	Damaging	9.81E-05	1
16:31199667 G>A		p.Arg274His	*FUS*	Tolerated		1
16:31201423 C>T		p.Arg377Trp	*FUS*	Tolerated	1.51E-05	1

Table 3S.3 Continued

Chromosome:Position (GRCh37)	dbSNP141 ID	Amino acid Change	Gene	MetaSVM prediction	ExAC European MAF	Number of control individuals
17:42427038 G>A	rs200019356	p.Val90Met	*GRN*	Tolerated	0.0003468	1
17:42429444 G>T	rs63750920	p.Gly414Val	*GRN*	Tolerated	1.51E-05	1
17:42429835 G>A	rs142926942	p.Val514Met	*GRN*	Tolerated	7.51E-05	1
20:60736600 C>T	rs144059766	p.Pro114Ser	*SS18L1*	Tolerated	0.0002373	1
20:60747782 G>A	rs36106901	p.Ala321Thr	*SS18L1*	Tolerated	0.000453	1
20:60749659 G>C		p.Ala375Pro	*SS18L1*	Tolerated		1
21:33032141 A>G		p.Asn20Ser	*SOD1*	Damaging	0.000152	1
22:29881797 A>C	rs148653339	p.Asn390Thr	*NEFH*	Damaging	0.0002549	1 (homozygous)

References

1. Rowland LP, Shneider NA. Amyotrophic lateral sclerosis. N Engl J Med 2001;344:1688-1700.

2. 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;72:245-256.

3. 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. Lancet Neurol 2012;11:323-330.

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

5. Renton AE, Chio A, Traynor BJ. State of play in amyotrophic lateral sclerosis genetics. Nat Neurosci 2014;17:17-23.

6. Cirulli ET, Lasseigne BN, Petrovski S, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science 2015;347:1436-1441.

7. Keller MF, Ferrucci L, Singleton AB, et al. Genome-wide analysis of the heritability of amyotrophic lateral sclerosis. JAMA Neurol 2014;71:1123-1134.

8. van Zundert B, Brown RH, Jr. Silencing strategies for therapy of SOD1-mediated ALS. Neurosci Lett 2017; 636:32–39.

9. Cady J, Allred P, Bali T, et al. Amyotrophic lateral sclerosis onset is influenced by the burden of rare variants in known amyotrophic lateral sclerosis genes. Ann Neurol 2015;77:100-113.

10. Li MX, Kwan JS, Bao SY, et al. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. PLoS Genet 2013;9:e1003143.

11. Brooks BR, Miller RG, Swash M, Munsat TL, World Federation of Neurology Research Group on Motor Neuron D. El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis. Amyotroph Lateral Scler Other Motor Neuron Disord 2000;1:293-299.

12. Neuenschwander AG, Thai KK, Figueroa KP, Pulst SM. Amyotrophic lateral sclerosis risk for spinocerebellar ataxia type 2 ATXN2 CAG repeat alleles: A meta-analysis. JAMA Neurol 2014;71:1529-1534.

13. Akimoto C, Volk AE, van Blitterswijk M, et al. A blinded international study on the reliability of genetic testing for GGGGCC-repeat expansions in C9orf72

reveals marked differences in results among 14 laboratories. J Med Genet 2014;51:419-424.

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

15. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature 2015;526:68-74.

16. McLellan T, Jorde LB, Skolnick MH. Genetic distances between the Utah Mormons and related populations. Am J Hum Genet 1984;36:836-857.

17. Jorde LB. Inbreeding in the Utah Mormons: An evaluation of estimates based on pedigrees, isonymy, and migration matrices. Ann Hum Genet 1989;53:339-355.

18. Lattante S, Conte A, Zollino M, et al. Contribution of major amyotrophic lateral sclerosis genes to the etiology of sporadic disease. Neurology 2012;79:66-72.

19. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655-1664.

20. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;2:e190.

21. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 2015;4:7.

22. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6:80-92.

23. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 2013;34:E2393-2402.

24. Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 2015;24:2125-2137.

25. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285-291.

26. Abel O, Shatunov A, Jones AR, Andersen PM, Powell JF, Al-Chalabi A. Development of a Smartphone App for a Genetics Website: The Amyotrophic Lateral Sclerosis Online Genetics Database (ALSoD). JMIR Mhealth Uhealth 2013;1:e18.

27. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: The NCBI database of genetic variation. Nucleic Acids Res 2001;29:308-311.

28. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 2014;133:1-9.

29. Andersen PM, Nilsson P, Ala-Hurula V, et al. Amyotrophic lateral sclerosis associated with homozygosity for an Asp90Ala mutation in CuZn-superoxide dismutase. Nat Genet 1995;10:61-66.

30. Robberecht W, Philips T. The changing scene of amyotrophic lateral sclerosis. Nat Rev Neurosci 2013;14:248-264.

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

32. Al-Chalabi A, Andersen PM, Chioza B, et al. Recessive amyotrophic lateral sclerosis families with the D90A SOD1 mutation share a common founder: Evidence for a linked protective factor. Hum Mol Genet 1998;7:2045-2050.

33. Rea SL, Majcher V, Searle MS, Layfield R. SQSTM1 mutations--bridging Paget disease of bone and ALS/FTLD. Exp Cell Res 2014;325:27-37.

34. Nagayama S, Minato-Hashiba N, Nakata M, et al. Novel FUS mutation in patients with sporadic amyotrophic lateral sclerosis and corticobasal degeneration. J Clin Neurosci 2012;19:1738-1739.

35. Takahashi Y, Fukuda Y, Yoshimura J, et al. ERBB4 mutations that disrupt the neuregulin-ErbB4 pathway cause amyotrophic lateral sclerosis type 19. Am J Hum Genet 2013;93:900-905.

36. Kwon MJ, Baek W, Ki CS, et al. Screening of the SOD1, FUS, TARDBP, ANG, and OPTN mutations in Korean patients with familial and sporadic ALS. Neurobiol Aging 2012;33:1017 e1017-1023.

37. van Blitterswijk M, van Es MA, Hennekam EA, et al. Evidence for an oligogenic basis of amyotrophic lateral sclerosis. Hum Mol Genet 2012;21:3776-3784.

38. Chio A, Calvo A, Mazzini L, et al. Extensive genetics of ALS: A population-based study in Italy. Neurology 2012;79:1983-1989.

39. Kenna KP, McLaughlin RL, Byrne S, et al. Delineating the genetic heterogeneity of ALS using targeted high-throughput sequencing. J Med Genet 2013;50:776-783.

40. Kinsley L, Siddique T. Amyotrophic lateral sclerosis overview. In: Pagon RA,

Adam MP, Ardinger HH, et al., editors. GeneReviews[Internet]. Seattle: University of Washington; 1993.

CHAPTER 4

THE DISCOVERY OF NOVEL CANDIDATE RISK GENES IN SPORADIC AMYOTROPHIC LATERAL SCLEROSIS

Introduction

The results of Chapter 3 showed that a significant majority (82.8%) of SALS patients lack an identifiable disease causing variant. However, genetic factors have been estimated to account for 60% of SALS risk (1). These pieces of evidence suggest that additional genetic loci that contribute towards the development of ALS remain to be discovered. NGS studies aimed at discovering novel ALS risk loci are more likely to find such variants than previous methods because they can directly assay rare pathogenic alleles. The fact that a number of novel ALS risk loci have been identified by the few NGS studies performed to date (2-5) supports this hypothesis. These studies largely used burden methods to identify genes that are associated with ALS pathogenesis. Unlike GWAS, which tests individual variants for association, burden methods determine whether accumulated mutations in a gene are associated with disease. The aggregation of variants across a gene allows burden methods to achieve better statistical power for rare variant association testing than GWAS (6).

However, a number of the NGS studies which utilized burden methods to find ALS risk genes had methodological flaws that could have impaired their results. For instance, some studies did not control for population stratification between cases and controls (3, 5), which can lead to false positive gene associations due to allele frequency differences between populations (7). Furthermore, these same studies considered all rare variants as pathogenic instead of directly predicting their pathogenicity. Lastly, the MAF thresholds used by these studies is restrictive enough that variants known to be pathogenic for ALS would be incorrectly considered benign. All of these factors could cause false positive and false negative ALS risk gene associations.

The aim of this chapter is to identify novel ALS risk loci by performing *VAAST* (8) and *PHEVOR* (9) on the same sequenced SALS and SSC individuals analyzed in Chapter 3. To do so, multiple steps will be taken to address the methodological shortcomings of previous ALS burden testing studies. First, PCA will be performed to control for population stratification between SALS patients and healthy controls. Second, burden testing will be performed by *VAAST*, which directly estimates variant pathogenicity. Lastly, a variant frequency filtering threshold that is compatible with the maximum observed allele frequency of variants known to be pathogenic for ALS will be used. These measures could lead to the discovery of ALS risk genes missed by previous burden association tests.

Materials and Methods

The exome sequencing results of 87 European SALS patients and 324 SSC control individuals from Chapter 3 were used for this analysis. *ADMIXTURE* (10) and *smartpca* (7, 11) were previously used to determine that these individuals were of European descent to control for population stratification effects. Genomic regions covered by less than five sequencing reads on average in the SALS and SSC cohorts were omitted from further analysis to control for coverage differences between the two cohorts. Variants with an ExAC (12) European MAF greater than 0.001 were removed from further consideration to reduce false positive gene associations. This value corresponds approximately to the frequency of the most common allele known to cause ALS (13). To identify novel ALS risk genes, a *VAAST* (8) analysis was performed to compare the genetic burden of all genes across the genome of SALS patients to SSC control individuals. Insertion and deletion (indel) variants were not used in the *VAAST* analysis as indel frequency based filtering is difficult due to inconsistency in how they are reported between different datasets (14). Multiple test correction is required when using *VAAST* because it tests for an excess of burden in all genes in the genome. Bonferroni correction was used to account for multiple hypothesis testing. As a result, a p-value of 2.57×10^{-6} was required for a gene to be considered significantly burdened (alpha level = 0.05 and 19,492 genes tested).

The ranked list of burdened genes generated by *VAAST* was then processed by *PHEVOR* (9) to identify genes with similar characteristics to known ALS risk genes. The following Human Phenotype Ontology (15) terms were used by the *PHEVOR* analysis: amyotrophic lateral sclerosis (HP:0007354), abnormal motor neuron morphology (HP:0002450), motor neuron atrophy (HP:0007373), and frontotemporal dementia (HP:0002145). The reranked list of burdened genes from *PHEVOR* was then manually reviewed to identify both known and potentially novel ALS risk genes.

Results and Discussion

No genes in the genomes of SALS patients were determined to be significantly more burdened by deleterious genetic variation than controls by *VAAST*. This result is not surprising due to the genetic heterogeneous nature of ALS. A much larger sample cohort would likely be required for a gene to show a significant excess of burden on a genome-wide level. However, two genes had burden levels that approached genome-wide significance. *VAAST* ranked *THOP1* (p = 1.89 × 10^{-5}; 95% Confidence interval (CI) = 1.24 × 10^{-5}–2.61 × 10^{-5}) as the most burdened gene in the SALS patient cohort compared to the SSC control individuals (Figure 4.1). Two missense variants (chr19:2810761 C>T; *THOP1*:p.T589M and chr19:2805121 G>A; *THOP1*:p.V233M) from two different SALS patients were identified in *THOP1*. *THOP1*:p.V233M is a novel allele because it is not found in the ExAC database. *THOP1*:p.T589M is found at extremely low frequency (MAF = 4.72 × 10^{-5}) in ExAC European individuals.

TP73 was the other gene that possessed a nearly significant level of deleterious variation (p = 2.08 × 10^{-5}; 95% CI = 1.39 × 10^{-5}–2.83 × 10^{-5}) in SALS patients (Figure 4.1). Four different missense variants were found among five separate SALS patients (Table 4.1). All of these variants were found at very low frequency in ExAC European individuals (Table 4.1). *DZIP1L* had the next highest amount of genetic burden with a p-value of 2.30 × 10^{-4} (95% CI = 1.65 × 10^{-4}–3.00 × 10^{-4}).

ALS risk genes that have previously described and contained deleterious variation were identified once the *VAAST* burden results were processed by *PHEVOR*. For example, *SOD1*—which was the first gene to be associated with ALS (16)—was the fifth ranked gene by *PHEVOR* (Table 4.1). The *SOD1*:p.D91A (chr21:33039603 A>C) missense variant, which is known to cause ALS in a recessive (13) and dominant (17) manner, was found in two different SALS patients (Table 4.1). *MAPT*, which has been previously described to be associated with ALS (18), was the third ranked gene resulting from the *VAAST*/*PHEVOR* analysis. A *MAPT* nonsynonymous variant (chr17:44101487 G>A; *MAPT*:p.A743T) was found in a single SALS patient. This variant has not been previously associated with ALS. However, it is found at a very low frequency in European individuals in the ExAC database (MAF = 3.01×10^{-5}).

Two candidate ALS risk genes were also identified by the combined *VAAST* and *PHEVOR* analysis. The top ranked gene resulting from this analysis was *MFN2* (*VAAST* p-value = 1.80×10^{-3}; 95% CI = 1.29×10^{-3}–2.36×10^{-3}). Four missense variants found in *MFN2* from four separate SALS patients were identified (Table 4.1). Two of these variants (chr1:12049301 G>A; *MFN2*:p.A26T and chr1:12062061 T>C; *MFN2*:p.V354A) were novel as they were not found in ExAC. One of the variants in *MFN2* (chr1:12064096 T>G; *MFN2*:p.F403C) was found in a single Latino individual in ExAC, but not in any European individuals. *MFN2* encodes for the Mitofusin-2 protein, which is important in maintaining proper mitochondrial dynamics, such as mitochondrial fusion (19). Loss of function of mitofusin-2 and mitofusin-1, which is a paralogue of mitofusin-2, leads to a complete lack of mitochondrial fusion and greatly reduced cellular respiration (20). Mutations in *MFN2* have been previously shown to cause Charcot-Marie-Tooth Neuropathy Type 2A (21), which is a hereditable axonal peripheral sensorimotor neuropathy characterized by motor and sensory loss mostly in the lower extremities (22). Dysfunction of mitofusin-2 has also been suspected to play a role in ALS pathogenesis because altered mitochondrial dynamics are seen in the disease (23). Interestingly, it has been reported that a patient with a mutation in *MFN2* developed co-occurring Charcot-Marie-Tooth Neuropathy Type 2A and ALS (24). These findings suggest *MFN2* could be involved in the development of ALS. Functional studies will be required to determine whether the *MFN2* variants seen in our SALS patient cohort are deleterious.

The other candidate ALS risk gene identified was *TP73*, which was the second ranked gene by the *VAAST* and *PHEVOR* analysis. In contrast, *PHEVOR* reduced the ranking of *THOP1* to seventh despite being ranked higher than *TP73* by *VAAST*. This suggests that *TP73* is both burdened by deleterious variation and clinically relevant to neurodegenerative disease. Interestingly, mice that possess one *tp73* null allele ($tp73^{+/-}$) show neurodegenerative signs that are similar to those found in ALS (25). These pieces of evidence make *TP73* a very attractive candidate for further study within the context of ALS. The next chapter (Chapter 5) of this dissertation will focus on experiments aimed at determining whether deleterious variants in *TP73* are involved in ALS pathogenesis.

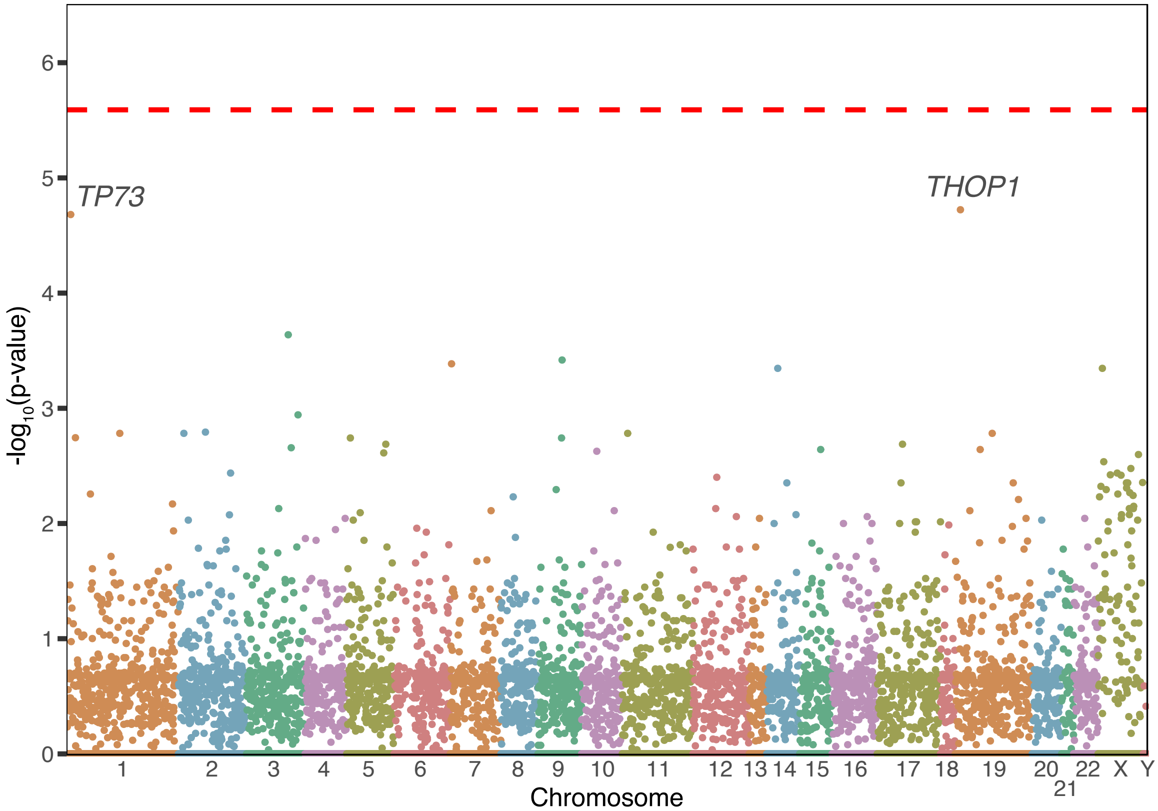


Figure 4.1 A Manhattan plot of the *VAAST* burden test results. Each dot shows the genomic position (x-axis) and *VAAST* p-value (y-axis) of each gene in the genome. The red line indicates the p-value threshold for genome-wide significance ($p = 2.57 \times 10^{-6}$). The only genes which possessed burden levels approaching genome-wide significance were *THOP1* and *TP73*.

Table 4.1 The top five ranked genes from the *PHEVOR* analysis. The specific variants that are contributing genetic burden are listed next to each gene they occur in. All variants are listed according to their genomic position (Chromosome:GRCh37 position) and the nucleotide change they result in. *TP73* was the only gene that was ranked high in both the *VAAST* and *PHEVOR* analyses. ExAC NFE MAF stands for Exome Aggregation Consortium non-Finnish European minor allele frequency. * indicates a variant was not found in non-Finnish Europeans but was found in one Latino individual from ExAC. † indicates a variant which was found in two different SALS patients.

PHEVOR rank	Gene	*VAAST* rank/p-value	Variants	ExAC NFE MAF
1	*MFN2*	14/1.80 × 10^{-3}	1:12049301 G>A	0.0
			1:12062061 T>C	0.0
			1:12064096 T>G	0.0*
			1:12069725 G>A	1.35 × 10^{-4}
2	*TP73*	2/2.08 × 10^{-5}	1:3640007 G>A	1.57 × 10^{-5}
			1:3647534 C>T	2.65 × 10^{-4}
			1:3647609 C>T	1.60 × 10^{-5}
			1:3649488 G>A†	3.78 × 10^{-4}
3	*MAPT*	1001/1.59 × 10^{-1}	17:44101487 G>A	3.01 × 10^{-5}
4	*GRIA3*	26/3.33 × 10^{-3}	X:122387214 T>C	4.17 × 10^{-5}
5	*SOD1*	396/4.71 × 10^{-2}	21:33039603 A>C†	8.70 × 10^{-4}

References

1. Al-Chalabi A, et al. (2010) An estimate of amyotrophic lateral sclerosis heritability using twin data. *J Neurol Neurosurg Psychiatry* 81(12):1324-1326.

2. Chesi A, et al. (2013) Exome sequencing to identify de novo mutations in sporadic ALS trios. *Nat Neurosci* 16(7):851-855.

3. Cirulli ET, et al. (2015) Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* 347(6229):1436-1441.

4. Kenna KP, et al. (2016) NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat Genet* 48(9):1037-1042.

5. Smith BN, et al. (2014) Exome-wide rare variant analysis identifies TUBA4A mutations associated with familial ALS. *Neuron* 84(2):324-331.

6. Lee S, Abecasis GR, Boehnke M, Lin X (2014) Rare-variant association analysis: Study designs and statistical tests. *Am J Hum Genet* 95(1):5-23.

7. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904-909.

8. Hu H, et al. (2013) VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol* 37(6):622-634.

9. Singleton MV, et al. (2014) Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am J Hum Genet* 94(4):599-610.

10. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655-1664.

11. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.

12. Lek M, et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285-291.

13. Andersen PM, et al. (1995) Amyotrophic lateral sclerosis associated with homozygosity for an Asp90Ala mutation in CuZn-superoxide dismutase. *Nat Genet* 10(1):61-66.

14. Tan A, Abecasis GR, Kang HM (2015) Unified representation of genetic variants. *Bioinformatics* 31(13):2202-2204.

15. Kohler S, et al. (2017) The Human Phenotype Ontology in 2017. *Nucleic Acids*

Res 45(D1):D865-D876.

16. Rosen DR, et al. (1993) Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 362(6415):59-62.

17. Al-Chalabi A, et al. (1998) Recessive amyotrophic lateral sclerosis families with the D90A SOD1 mutation share a common founder: Evidence for a linked protective factor. *Hum Mol Genet* 7(13):2045-2050.

18. Poorkaj P, et al. (2001) TAU as a susceptibility gene for amyotropic lateral sclerosis-parkinsonism dementia complex of Guam. *Arch Neurol* 58(11):1871-1878.

19. Santel A, Fuller MT (2001) Control of mitochondrial morphology by a human mitofusin. *J Cell Sci* 114(Pt 5):867-874.

20. Chen H, Chomyn A, Chan DC (2005) Disruption of fusion results in mitochondrial heterogeneity and dysfunction. *J Biol Chem* 280(28):26185-26192.

21. Zuchner S, et al. (2004) Mutations in the mitochondrial GTPase mitofusin 2 cause Charcot-Marie-Tooth neuropathy type 2A. *Nat Genet* 36(5):449-451.

22. Zuchner S (1993) Charcot-Marie-Tooth Neuropathy Type 2A. *GeneReviews(R)*, eds Pagon RA, et al. (University of Washington, Seattle, WA).

23. Shi P, Gal J, Kwinter DM, Liu X, Zhu H (2010) Mitochondrial dysfunction in amyotrophic lateral sclerosis. *Biochim Biophys Acta* 1802(1):45-51.

24. Marchesi C, et al. (2011) Co-occurrence of amyotrophic lateral sclerosis and Charcot-Marie-Tooth disease type 2A in a patient with a novel mutation in the mitofusin-2 gene. *Neuromuscul Disord* 21(2):129-131.

25. Wetzel MK, et al. (2008) p73 regulates neurodegeneration and phospho-tau accumulation during aging and Alzheimer's disease. *Neuron* 59(5):708-721.

CHAPTER 5

TP73, A NOVEL AMYOTROPHIC LATERAL SCLEROSIS CANDIDATE RISK GENE

Introduction

The *TP73* (Tumor Protein P73) gene encodes for the p73 protein (1), which is part of the p53 family of tumor suppressor proteins. This protein family also includes p53 and p63 (2). The p53 family of proteins are transcription factors that modulate the expression of their target genes to facilitate a number of biological and developmental processes, such as cell-cycle arrest, apoptosis, and cellular differentiation (2). Many of the cellular hallmarks of tumorigenesis and cancer result from dysfunction of these processes (3). As a result, mutations that ablate the tumor suppressive function of p53 are commonly seen in a wide variety of cancers (4). The p73 protein is also mutated or deleted in some cancers, such as neuroblastoma (1). However, somatic mutations of p73 cause cancer to a much lesser degree than p53 (2).

The p73 protein possess a NH_2-terminal transactivation (TA) domain, a central DNA-binding domain, and an COOH-terminal oligomerization domain like the other members in the p53 protein family (5). The transactivation domain of p73 induces transcription once p73 has bound to a target gene via its DNA-binding domain (6). The oligomerization domain is responsible for forming active tetramers with other p73 proteins and p53 family members (7).

Transcripts of *TP73* and its family members are expressed in a wide variety of unique isoforms (6). At least three NH_2-terminal and nine COOH-terminal alternative splicing events are known to occur (5). These different *TP73* transcripts vary in the genes they target and how much they can induce expression, which suggests they are involved in different cellular processes (8). For instance, the longest COOH-terminal splice isoform of *TP73*, p73α, possesses a sterile alpha motif (SAM) domain not found in other isoforms (6). SAM domains typically mediate protein-protein interactions and are thought to be involved in the regulation of cellular differentiation. This supports the notion that p73α has a unique role in development compared to other p73 isoforms (9).

Alternative transcripts of *TP73* can also be generated by utilizing an internal promoter in intron 3 that skips the first three exons of the gene. These transcripts result in an NH_2-terminally truncated p73 protein (ΔN-p73) (5). Unlike full length NH_2-terminal p73 proteins (TA-p73), ΔN-p73 isoforms are unable to directly induce the expression of gene targets because they lack a TA domain (2). The ΔN-p73 protein inhibits the tumor suppressive and apoptotic functions of TA-p73 and other p53 family members in a dominant-negative fashion (2). ΔN-p73 does this by binding and sequestering TA-p73, forming less active ΔN-p73/TA-p73 complexes, and outcompeting p53 and TA-p73 for target gene binding sites (2). ΔN-p73 is classified as an oncogene because it inhibits the proapoptotic and tumor suppressive functions of TA-p73 and p53 (2). The oncogenic role of ΔN-p73 has been shown in mouse embryonic fibroblasts (MEFs) that overexpress the gene. ΔN-p73 overexpressing MEFs show significantly increased growth rates (10). Furthermore, MEFs that coexpress ΔN-p73 and oncogenic Ras undergo cellular transformation (10). Conversely, TA-p73 is responsible for mediating E2F-1-induced apoptosis in p53$^{-/-}$ MEFs (11). These pieces of evidence show that TA-p73 and ΔN-p73 have opposing roles which must be properly regulated in order to maintain normal functioning cells.

The complicated role *TP73* (*Trp73* in mice) plays in organism development and maintenance largely comes from animal models. *Trp73*$^{-/-}$ mice were one of first animal models used to test the developmental role of p73 (12). *Trp73*$^{-/-}$ mice were generated by deleting the DNA-binding domain of the protein, which is found in both TA-p73 and ΔN-p73 (12). Unlike *Trp53*$^{-/-}$ mice, which develop spontaneous tumors (13), *Trp73*$^{-/-}$ mice develop a number of developmental abnormalities, including hippocampal dysgenesis, hydrocephalus, chronic infections, abnormal pheromone sensing capabilities, and other neuronal defects (12). *Trp73* isoform specific knockout mice have further defined the specific developmental roles of TA-p73 and ΔN-p73. For instance, TA-p73$^{-/-}$ mice—which were generated by deleting the exons that encode for the TA domain—develop spontaneous tumors, less severe hippocampal dysgenesis, and infertility. However, these mice lacked the central nervous system (CNS) atrophy seen in p73$^{-/-}$ mice (14, 15). In contrast, ΔN-p73$^{-/-}$ mice—which lack the ΔN-p73 specific exon 3'—do show signs of cortical atrophy and neurodegeneration like p73$^{-/-}$ mice (15, 16). Interestingly, aged (18 months) *Trp73*$^{+/-}$ mice show signs that are reminiscent of those found in ALS, such as muscle weakness, motor cortex atrophy, and an abnormal reflex (17). However, the impact loss of functional p73 has on motor neurons was never directly studied in these animals. The primary p73 isoform found in developing neurons is ΔN-p73 and is required for neuronal resistance to apoptotic insults (18, 19). Together, these findings suggest that p73 is a critical neuronal survival and developmental factor that potentially has a role in

ALS pathogenesis.

While the existing mouse data provide useful preliminary evidence that *TP73* is a causal ALS gene, further mouse work would have several drawbacks and limitations. No *Trp73*$^{+/-}$ animals are readily available, which would require the retrieval of *Trp73*$^{+/}$ embryonic cells to establish colonies. In addition, the ALS-like phenotype was only reported in aged (18 months) *Trp73*$^{+/-}$ mice (17). An alternative approach is to model ALS and *TP73* function in zebrafish (*Danio rerio*), an organism well established at the University of Utah for studying human disease genes and whose rapid, economical breeding makes it particularly suitable for functional analysis of potential causal variants. Previous studies have established zebrafish as an ALS model. For instance, morpholino knockdown of the zebrafish *FUS* homolog, *fus*, leads to impaired locomotor ability and reduced spinal motor neuron axon length, which are consistent with an ALS-like phenotype (20). Interestingly, this phenotype could be rescued with coexpression of wild-type (WT) human *FUS*, but not by with coexpression of the most common human ALS-related *FUS* point mutations (20). A similar approach where zebrafish *tp73* is genetically manipulated could serve as useful model to test what role p73 has in motor neuron function and development, especially since *TP73* interacts with *FUS* (21).

Zebrafish *tp73* is highly expressed in the central nervous system like it is in humans and mice. As in the mouse, knockdown of *tp73* in zebrafish (by morpholino technology) results in olfactory and telencephalon defects due to impaired neuronal development and survival (22), although the status of motor neurons and motor system function in *tp73* knockdowns has not been addressed. However, morpholinos can have a number of off-target effects that complicate gene-specific studies (23). Thus, the exact effects zebrafish *tp73* knockout and patient specific mutations of *TP73* have on motor neuron development and morphology can be addressed by CRISPR/Cas9.

The results of Chapter 4 showed that at least five SALS patients seen at the University of Utah possessed a rare and potentially pathogenic variant in *TP73*. The established role *TP73* has in neuronal survival and maturation, the development of ALS-like symptoms in aged *Trp73*$^{+/-}$ mice, and the presence of deleterious *TP73* variants in SALS patients all suggest that *TP73* could be a ALS risk gene. The aim of this dissertation chapter is to determine if *TP73* potentially has a role in the pathogenesis of ALS. To do so, I will determine whether more deleterious *TP73* variants can be identified in other ALS patient cohorts. I also will determine the developmental and morphological effects loss of functional p73 has on motor neurons using a zebrafish (*Danio rerio*) animal model. The results of these investigations will help to elucidate whether p73 has a role in ALS pathogenesis.

Materials and Methods

Screening of ALS patients for deleterious *TP73* variants

If *TP73* is involved in ALS pathogenesis, additional rare and deleterious variants should be discovered by screening additional ALS patients. The *VAAST* and *PHEVOR* analysis used in Chapter 4 only considered SNVs. As a result, it is possible that insertion and deletion variants in *TP73* may exist in the 87 SALS exome-sequenced patient cohort studied in that chapter. These SALS patients were screened for rare (ExAC MAF < 0.001) insertion and deletion variants that change the protein-coding sequencing of *TP73*. Nine individuals from the same SALS cohort were previously identified as non-European and were not assessed for *TP73* variants of any kind. These nine SALS patients were screened for all rare variants that change the normal *TP73* amino acid sequence. Any additional *TP73* variants identified among these patients, along with the five *TP73* SNVs found in Chapter 4, were Sanger sequenced to validate their presence.

Another University of Utah ALS cohort, which was whole-genome sequenced at an average coverage of 60X (analyzed in Chapter 6), was screened for rare *TP73* coding variants. This cohort is comprised of 70 ALS patients and eight unaffected relatives. Of these 70 ALS patients, 26 were also found in the 96 patient SALS cohort and were removed from further analysis. In total, 44 additional ALS patients were screened for variants that alter the *TP73* protein-coding sequence. Sanger sequencing was also used to validate any *TP73* coding variants identified in this cohort.

A large ALS cohort consisting of over 2,800 whole-exome sequenced patients (24) was used to search for additional rare SNVs, insertions, deletions, and splice-site altering variants in *TP73* via the ALS Data Browser (ALSdb, http://alsdb.org). Information about the location and number of patients with a p73 variant is available through ALSdb. However, patient specific genotypes and DNA are not accessible. As a result, any coding variants in *TP73* identified from this cohort could not be validated by Sanger sequencing.

The effects of p73 loss of function on neuronal development and morphology

The CRISPR/Cas9 system was used to determine whether knockout of *tp73* in zebrafish leads to an ALS-like phenotype. Zebrafish *tp73* null alleles were created by developing guide RNA (gRNA) targeting sequences to *tp73* exon 4 (5'-TGTATTGGAAGGGATGGCCGggg-3'; target site = uppercase and protospacer adjacent motif = lowercase. Exon 4 of *tp73* encodes for part of the DNA binding domain of zebrafish p73. While it is unclear if ΔN-p73 exists in zebrafish (25), all p73 isoforms possess the DNA-binding domain The *tp73* CRISPR RNA and Cas9 protein were diluted to 450 ng/μl using DNase-free water and injected into one cell stage zebrafish embryos to generate *tp73* mutant animals. *Tg(hb9:Gal4-UAS:GFP)* and *mnx1*:GFP transient transgenic embryos were used for these injections. The promoter sequences of the *hb9* and *mxn1* genes were used to drive the expression of green fluorescent protein (GFP) specifically in motor neurons for visualization. The proportion of injected zebrafish that possessed a mutated copy of *tp73* was determined by high resolution melting (HRM) analysis.

Injected *tp73* mutant zebrafish were then assessed for motor neuron dysfunction. To do so, confocal microscopy images of GFP fluorescence in motor neurons were taken at 72 hr postfertilization (hpf). The primary axon length and number of motor neurons in *tp73* mutant fish were compared to uninjected WT zebrafish to determine the impact loss of *tp73* has on motor neurons. ImageJ and NeuronJ were utilized to quantify the number and primary axon length of motor neurons in *tp73* and WT zebrafish.

Results

Discovery of additional rare *TP73* coding sequence variants

One additional rare *TP73* variant was identified in the 87 European SALS cohort upon searching for insertion and deletion variants. This variant (chr1:3646605 CCATGAACAAGGTGCACGGGGG>C; *TP73*:p.PMNKVHGG413-420P) is a rare (ExAC European MAF = 1.17 × 10^{-4}) 21 base pair and seven amino acid in-frame deletion in exon 11 (14 total exons) of *TP73* (Figure 5.1). No additional *TP73* variants were uncovered when the nine non-European patients from the SALS cohort were screened. A single rare (ExAC European MAF = 4.86×10^{-5}) missense SNV (chr1:3647559 G>A; *TP73*:p.A472T) was found in the 44 whole-genome sequenced University of Utah ALS patient cohort. In total, six unique variants that affect the protein-coding sequence of *TP73* were found in seven ALS patients out of the 140 screened patients (Table 5.1). Sanger sequencing confirmed the presence of all but one of the *TP73* coding sequence variants. More specifically, the presence of a rare *TP73* SNV (chr1:3649488 G>A; *TP73*:p.V586M) was confirmed in only one of the two ALS patients it was first identified in (Figure 5.2).

An additional 17 rare SNVs that alter the *TP73* coding sequence were found in the ALSdb patient cohort (24). Additionally, an in-frame deletion (chr1:3646679 AGTT>A; *TP73*:p.SS438-439T), which occurs in exon 11, was found in the ALSdb cohort. Between the University of Utah ALS and ALSdb cohorts, 24 different rare *TP73* coding sequence variant sites were found. Four of the 24 *TP73* coding sequence variants result in a synonymous substitution in TA-p73α (Ensembl transcript ID: ENST00000378295) and ΔN-p73α (ENST00000378288) (Figure 5.1 and Table 5.1). However, these variants cause nonsynonymous changes in some p73 isoforms, such as ΔN-p73γ (ENST00000378280). All of 22 of *TP73* amino acid sequence altering SNVs were classified as deleterious by MetaSVM (26). A summary of all *TP73* coding sequence variant sites can be found in Table 5.1 and Figure 5.1.

The effect of *TP73* loss of function on motor neuron development and morphology

HRM analysis found the CRISPR/Cas9 system was working at high efficiency as nine out of 10 injected zebrafish possessed mutated copies of *tp73* on average (Figure 5.3). Sanger sequencing of multiple injected animals confirmed the presence of loss of function frame-shift mutations near the CRISPR/Cas9 cut site in exon 4 (data not shown). Dorsal mounting and confocal microscopy imaging of *Tg(hb9:Gal4-UAS:GFP)* zebrafish that had been injected with *tp73* targeting CRISPR/Cas9 showed a significant reduction (*p*-value < 0.005) in the number spinal motor neurons present compared to uninjected and tyrosinase (*tyr*) CRISPR/Cas9 injected controls (Figures 5.4A and 5.4B). Lateral mounting of *tp73* mutant *mnx1*:GFP zebrafish—which were generated by transient transgenesis—demonstrated a significant reduction (*p*-value < 0.05) in the primary axon length of spinal motor neurons (Figures 5.5A and 5.5B). The length of secondary axon branches of primary axons was also significantly (*p*-value < 0.05) reduced (Figure 5.5C). Interestingly, indirect TUNEL staining of *tp73* mutant zebrafish showed significantly increased apoptosis (*p*-value < 0.05) of motor neurons compared to WT controls (Figures 5.6A and 5.6B).

Discussion

The aim of this chapter was to determine if *TP73* is involved in ALS pathogenesis. To do so, efforts were made to find if deleterious *TP73* occur in the general ALS patient population. In total, 24 rare *TP73* coding variants were found among ~2,900 ALS patients, which is similar to the prevalence of many known ALS disease genes. Such a finding supports the hypothesis that *TP73* is involved in the ALS disease process.

Next, experiments were performed in zebrafish to determine what role p73 has in motor neuron survival and development. The results of these efforts demonstrated that loss of p73 significantly reduced motor neuron survival and development. Such a finding also supports notion that *TP73* is an ALS disease gene.

Since *SOD1* mutations were found to be associated with ALS in 1993 (27), impairment of a number of different cellular processes have been shown to play a role in the disease (28). However, transcription factors that drive cell survival and developmental pathways have not been previously implicated in ALS. Dysfunction of such factors in ALS would be expected as motor neuron death is a central component to the clinical manifestations of the disease. Our results—which show deleterious *TP73* protein-coding variants occur in an appreciable proportion of ALS patients and loss of *tp73* impairs motor neuron survival and development in zebrafish—indicate *TP73* is likely involved in ALS pathogenesis. This finding potentially contributes to the overall efforts aimed at narrowing the ALS missing heritability gap. Our data also provide evidence that neuronal survival factors could be an important piece to the incomplete puzzle of ALS molecular pathology. To definitively show whether p73 is involved in the ALS disease process, future work will be needed to determine whether patient specific variants can rescue the observed zebrafish motor neuron phenotype. Overall, these contributions will assist in understanding the ALS genetic risk landscape and help pave the way forward to an eventual a cure for the disease.

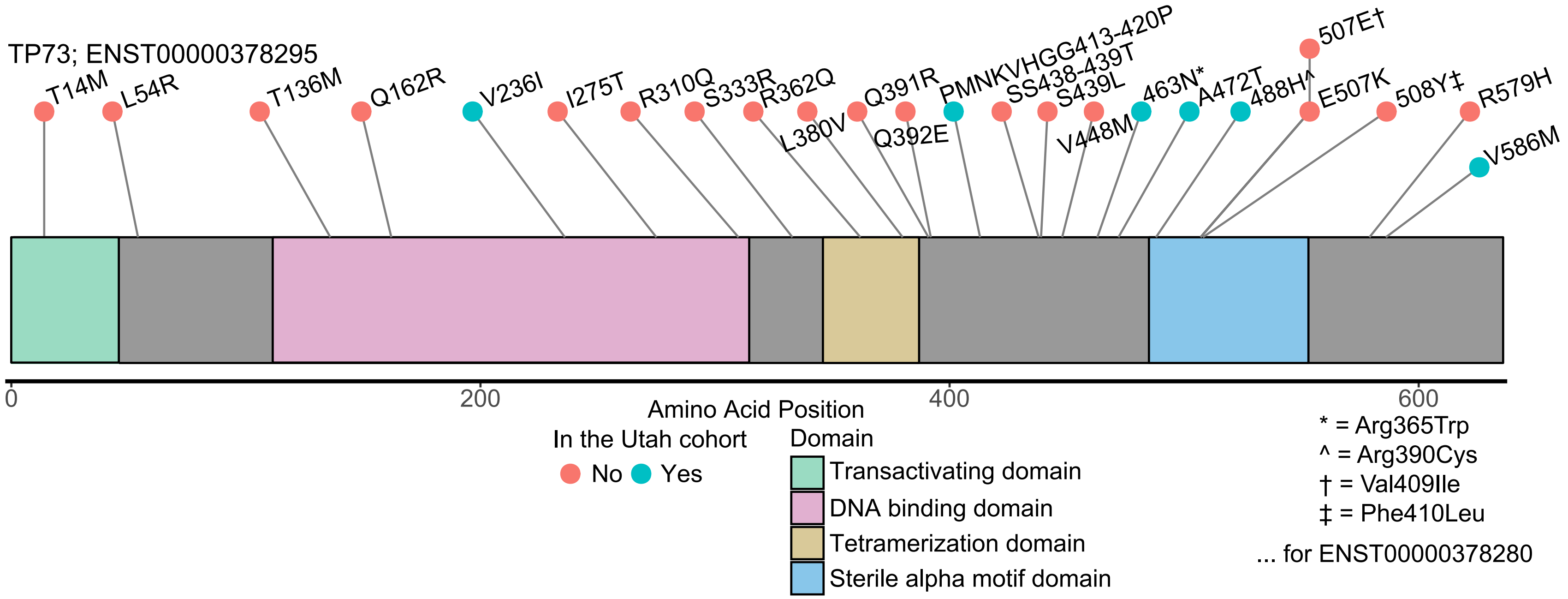


Figure 5.1 A schematic of where the 24 rare (ExAC European MAF < 0.001) amino acid alerting-variants found across all studied ALS cohorts occur in the primary structure of TA-p73α. Eleven of these variants are found within the four functional domains of TA-p73α. Six of the 24 *TP73* variants were found in the University of Utah ALS patient cohorts. Four of the nonsynonymous SNVs are not found in p73α proteins, but do exist in p73γ isoforms due to splicing differences.

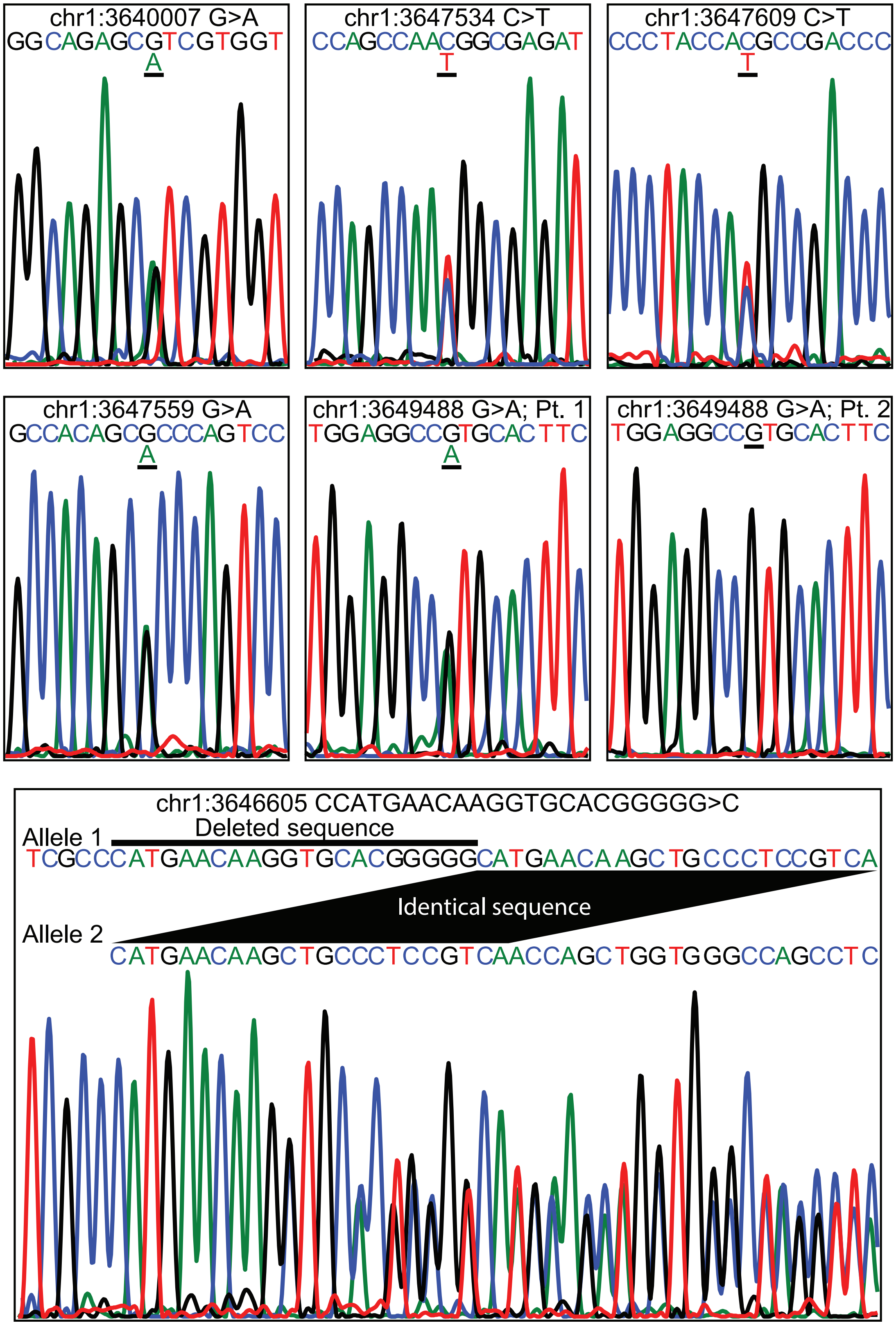


Figure 5.2 Sanger sequencing results of the seven *TP73* variants found in the University of Utah ALS patient cohorts. Chr1:3649488 G>A was verified by Sanger sequencing in one of the two patients who were reported to have it by NGS. Pt. stands for patient.

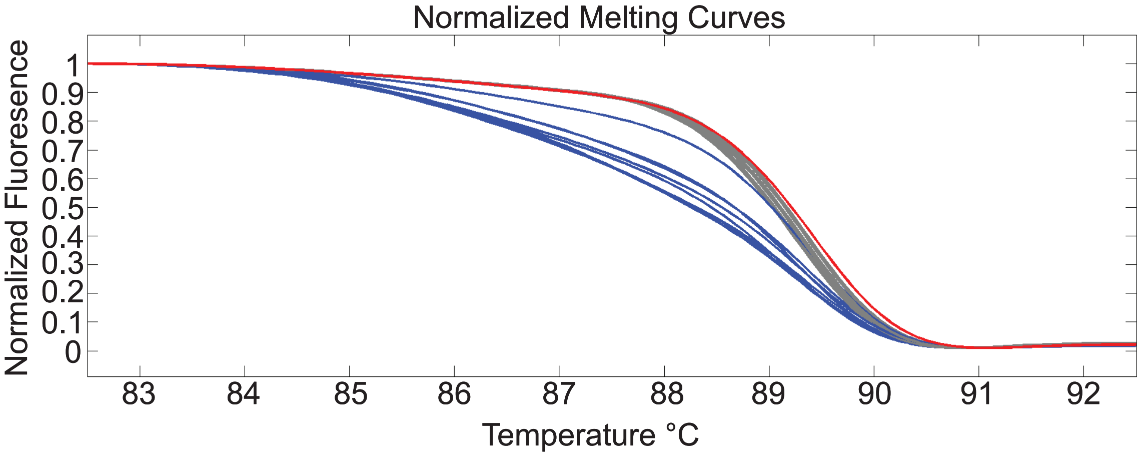


Figure 5.3 A high resolution melting curve of the PCR products covering the *tp73* exon 4 site targeted by CRISPR/Cas9. Grey curves indicate the melting pattern of uninjected wild-type zebrafish. The blue and red curves indicate zebrafish that have been injected with the *tp73* targeting CRISPR/Cas9. The leftward shift of the blue curves indicate zebrafish that have mutant copies of *tp73*. Only one injected zebrafish did not undergo mutagenesis (red curve). These results indicate CRISPR/Cas9 was able to induce *tp73* mutagenesis at high efficiency.

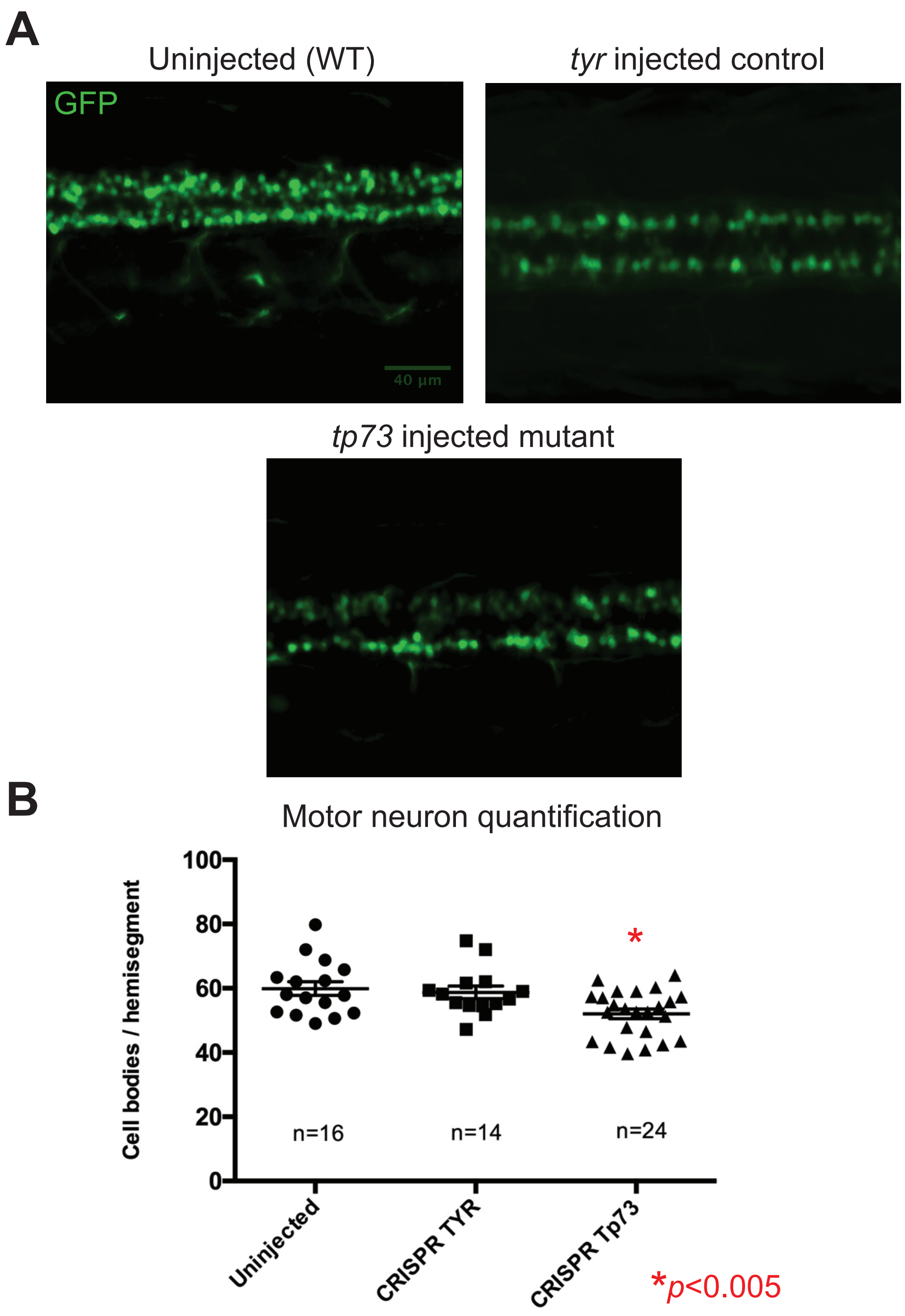


Figure 5.4 Loss of *tp73* function is detrimental to the number of spinal motor neurons present in *Tg(hb9:Gal4-UAS:GFP)* zebrafish at 72 hpf. (A) Dorsal confocal microscopy images (taken at 10x; 5µm/step, 21 steps) of wild-type (WT) uninjected control, tyrosinase (*tyr)* CRISPR/Cas9 injected control, and *tp73* CRISPR/Cas9 mutant zebrafish. A reduced number of GFP-positive motor neurons can be seen in *tp73* zebrafish. (B) The number of GFP-positive motor neurons in *tp73* zebrafish is significant lower (*p*-value < 0.005) than both WT uninjected and injected tyrosinase control zebrafish. The number of zebrafish tested is listed under each respective bar. The error bars indicate the standard error of the mean.

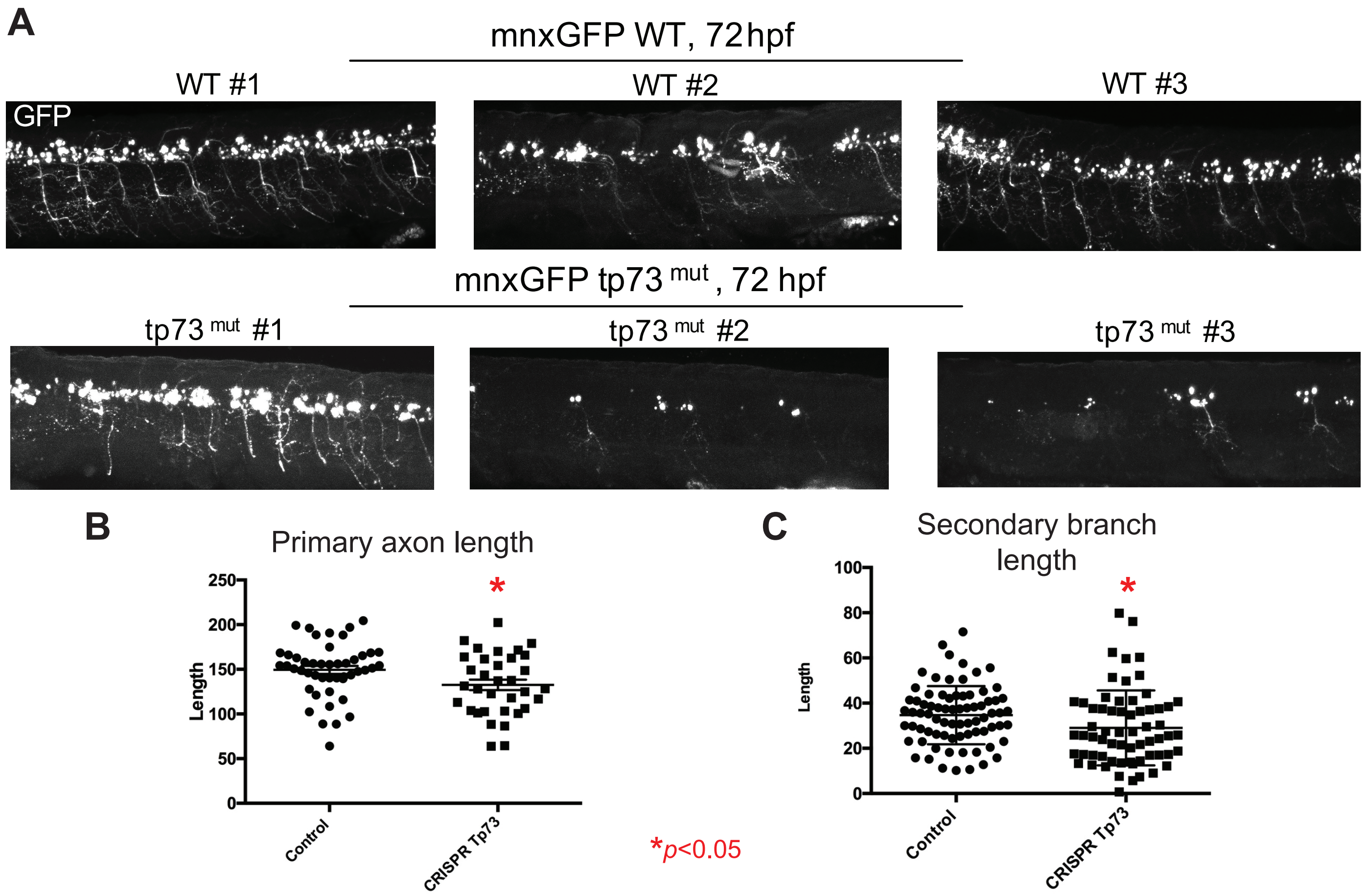


Figure 5.5 Loss of *tp73* function negatively impacts axon development of spinal motor neurons in transient transgenic *mnx1*:GFP zebrafish. (A) Lateral confocal microscopy GFP images of three different WT and *tp73* CRISPR/Cas9 mutant zebrafish. The spinal motor neuron axons of *tp73* mutant zebrafish appear to be shorter and disordered in their arrangement. (B and C) The length of primary axons (B) and secondary axon branches (C) of spinal motor neurons in *tp73* mutants is significantly lower (*p*-value < 0.05) compared to uninjected WT zebrafish.

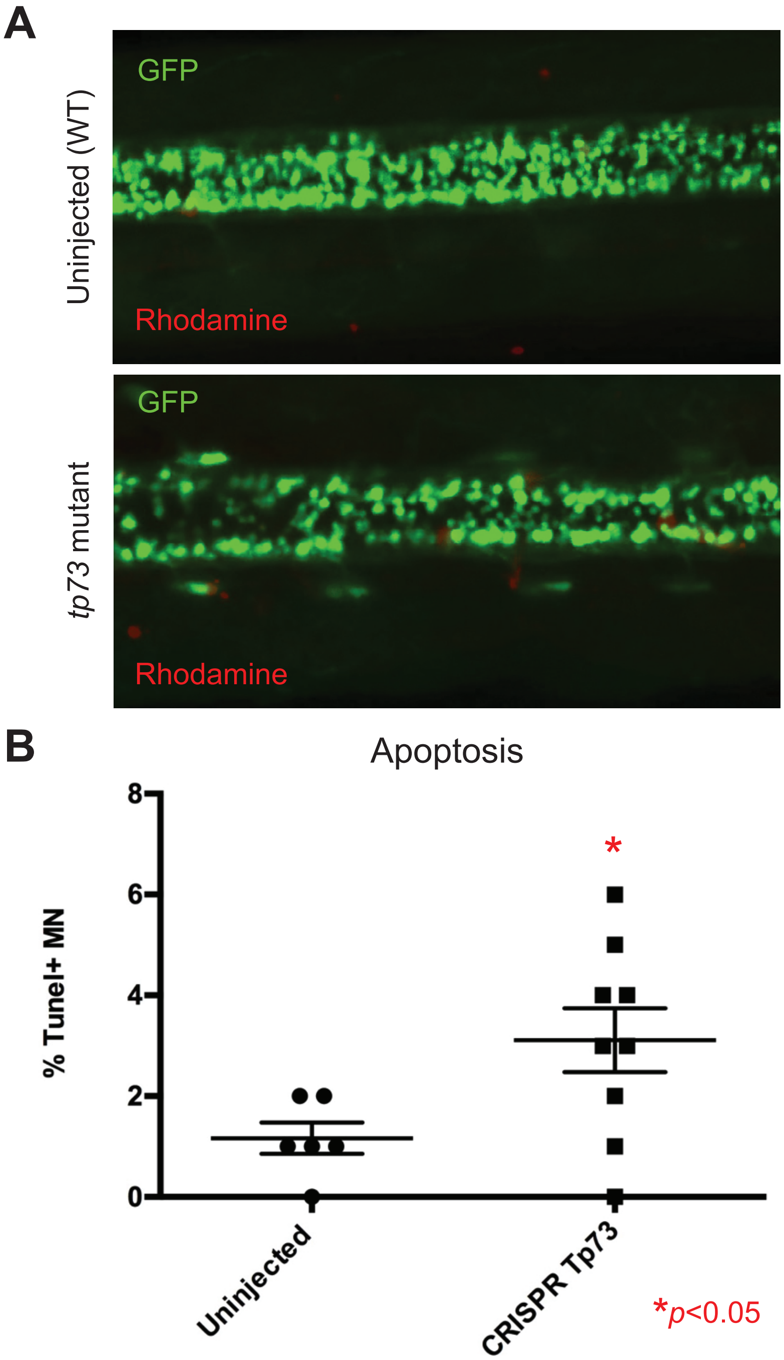


Figure 5.6 Loss of *tp73* function results in increased motor neuron apoptosis in *Tg(hb9:Gal4-UAS:GFP)* zebrafish at 72 hpf. (A) Dorsal confocal microscopy images of wild-type (WT) uninjected control and *tp73* CRISPR/Cas9 mutant zebrafish. Red rhodamine fluorescence is used to measure apoptosis. A yellow overlap of red and green GFP fluorescence indicates motor neurons undergoing apoptosis. (B) The number of apoptotic spinal motor neurons is significantly increased (*p*-value < 0.05) in *tp73* mutants compared to uninjected controls.

Table 5.1 A summary of the 24 rare variants found among all studied ALS patients that alter the normal *TP73* protein-coding sequence. The variants are listed according to their GRCh37 genomic position (Chromosome:Position) and the nucleotide change they cause. All amino acid positions are relative to TA-p73α (ENST00000378295). Four of the variants result in an amino acid substitution in only in some p73 isoforms, such as ΔN-p73γ (ENST00000378280). † indicates the amino acid substitution that occurs in ΔN-p73γ.

Variant	dbSNP141 ID	Amino acid change	ExAC NFE MAF	In Utah cohort?
1:3598970 C>T		T14M	0.0	No
1:3599719 T>G	rs377512486	L54R	1.51×10^{-5}	No
1:3624333 C>T		T136M	1.54×10^{-5}	No
1:3638640 A>G		Q162R	0.0	No
1:3640007 G>A		V236I	1.57×10^{-5}	Yes
1:3643770 T>C		I275T	0.0	No
1:3644278 G>A		R310Q	2.68×10^{-4}	No
1:3644706 C>G	rs202005425	S333R	2.77×10^{-4}	No
1:3645901 G>A	rs200330726	R362Q	7.58×10^{-5}	No
1:3645954 T>G		L380V	0.0	No
1:3645988 A>G		Q391R	1.52×10^{-5}	No
1:3645990 C>G		Q392E	0.0	No
1:3646605 CCATGAACAAGGTGCACGGGGG>C		PMNKVHGG413-420P	1.17×10^{-4}	Yes
1:3646679 AGTT>A		SS438-439T	0.0	No
1:3646683 C>T		S439L	0.0	No
1:3646709 G>A		V448M	0.0	No
1:3647534 C>T	rs150268231	463N (R365W†)	2.65×10^{-4}	Yes
1:3647559 G>A	rs369342367	A472T	4.86×10^{-5}	Yes
1:3647609 C>T		488H (R390C†)	1.60×10^{-5}	Yes
1:3648061 G>A		E507K	0.0	No
1:3648063 G>A	rs143515986	507E (V409I†)	1.51×10^{-5}	No
1:3648066 T>C	rs143442213	508Y (F410L†)	1.36×10^{-4}	No
1:3649468 G>A	rs376429700	R579H	2.17×10^{-4}	No
1:3649488 G>A	rs138694448	V586M	3.78×10^{-4}	Yes

References

1. Kaghad M, et al. (1997) Monoallelically expressed gene related to p53 at 1p36, a region frequently deleted in neuroblastoma and other human cancers. *Cell* 90(4):809-819.

2. Melino G, De Laurenzi V, Vousden KH (2002) p73: Friend or foe in tumorigenesis. *Nat Rev Cancer* 2(8):605-615.

3. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: The next generation. *Cell* 144(5):646-674.

4. Collavin L, Lunardi A, Del Sal G (2010) p53-family proteins and their regulators: Hubs and spokes in tumor suppression. *Cell Death Differ* 17(6):901-911.

5. Moll UM, Slade N (2004) p63 and p73: Roles in development and tumor formation. *Mol Cancer Res* 2(7):371-386.

6. Costanzo A, et al. (2014) TP63 and TP73 in cancer, an unresolved "family" puzzle of complexity, redundancy and hierarchy. *FEBS Lett* 588(16):2590-2599.

7. Davison TS, et al. (1999) p73 and p63 are homotetramers capable of weak heterotypic interactions with each other but not with p53. *J Biol Chem* 274(26):18709-18714.

8. Jancalek R (2014) The role of the TP73 gene and its transcripts in neuro-oncology. *Br J Neurosurg* 28(5):598-605.

9. Levrero M, et al. (2000) The p53/p63/p73 family of transcription factors: Overlapping and distinct functions. *J Cell Sci* 113 (Pt 10):1661-1670.

10. Petrenko O, Zaika A, Moll UM (2003) deltaNp73 facilitates cell immortalization and cooperates with oncogenic Ras in cellular transformation in vivo. *Mol Cell Biol* 23(16):5540-5555.

11. Irwin M, et al. (2000) Role for the p53 homologue p73 in E2F-1-induced apoptosis. *Nature* 407(6804):645-648.

12. Yang A, et al. (2000) p73-deficient mice have neurological, pheromonal and inflammatory defects but lack spontaneous tumours. *Nature* 404(6773):99-103.

13. Donehower LA, et al. (1992) Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature* 356(6366):215-221.

14. Tomasini R, et al. (2008) TAp73 knockout shows genomic instability with infertility and tumor suppressor functions. *Genes Dev* 22(19):2677-2691.

15. Pozniak CD, et al. (2002) p73 is required for survival and maintenance of CNS

neurons. *J Neurosci* 22(22):9800-9809.

16. Wilhelm MT, et al. (2010) Isoform-specific p73 knockout mice reveal a novel role for delta Np73 in the DNA damage response pathway. *Genes Dev* 24(6):549-560.

17. Wetzel MK, et al. (2008) p73 regulates neurodegeneration and phospho-tau accumulation during aging and Alzheimer's disease. *Neuron* 59(5):708-721.

18. Walsh GS, Orike N, Kaplan DR, Miller FD (2004) The invulnerability of adult neurons: A critical role for p73. *J Neurosci* 24(43):9638-9647.

19. Pozniak CD, et al. (2000) An anti-apoptotic role for the p53 family member, p73, during developmental neuron death. *Science* 289(5477):304-306.

20. Kabashi E, et al. (2011) FUS and TARDBP but not SOD1 interact in genetic models of amyotrophic lateral sclerosis. *PLoS Genet* 7(8):e1002214.

21. Wang T, Jiang X, Chen G, Xu J (2015) Interaction of amyotrophic lateral sclerosis/frontotemporal lobar degeneration-associated fused-in-sarcoma with proteins involved in metabolic and protein degradation pathways. *Neurobiol Aging* 36(1):527-535.

22. Rentzsch F, Kramer C, Hammerschmidt M (2003) Specific and conserved roles of TAp73 during zebrafish development. *Gene* 323:19-30.

23. Kok FO, et al. (2015) Reverse genetic screening reveals poor correlation between morpholino-induced and mutant phenotypes in zebrafish. *Dev Cell* 32(1):97-108.

24. Cirulli ET, et al. (2015) Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* 347(6229):1436-1441.

25. Satoh S, Arai K, Watanabe S (2004) Identification of a novel splicing form of zebrafish p73 having a strong transcriptional activity. *Biochem Biophys Res Commun* 325(3):835-842.

26. Dong C, et al. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 24(8):2125-2137.

27. Rosen DR, et al. (1993) Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 362(6415):59-62.

28. Taylor JP, Brown RH, Jr., Cleveland DW (2016) Decoding ALS: From genes to mechanism. *Nature* 539(7628):197-206.

CHAPTER 6

HOW “SPORADIC” IS SPORADIC AMYOTROPHIC LATERAL SCLEROSIS?

Introduction

The results of Chapter 3 demonstrated that heritable genetic factors are found in a significant proportion (17.2%) of SALS patients. The presence of genetic risk factors in SALS patients calls into question how appropriate the definitions of FALS and SALS are. Furthermore, it is unclear where these heritable ALS risk factors originate from. One possible mechanism could be through nonsynonymous *de novo* mutations, which are thought to play a role in SALS pathogenesis (1). In an ALS trio study, 38% of SALS patients had at least one identifiable *de novo* variant that altered the protein coding sequence of a gene (2). However, none of the *de novo* mutations identified in these patients occurred in a known ALS risk gene. As a result, it is unclear how important *de novo* mutations are in the development of ALS. Another mechanism by which ALS genetic risk factors could occur in SALS patients is through the inheritance of such variants, which went undetected in carrier family members due to incomplete penetrance, misdiagnosis, early death, or death to non-ALS causes. If such a mechanism occurred to a large extent, it would be expected that distantly related SALS patients would share an ALS genetic risk factor by descent. However, there have been no attempts to identify such events to date, which is likely due to the lack of genealogical data between distantly related SALS patients.

Genetic studies performed at the University of Utah have the unique ability to conduct such analyses because of resources like the Utah Population Database (UPDB). The UPDB contains genetic, medical, and genealogical records from over 9 million individuals from as far back at the late 18th century. The use of the UPDB has led to the discovery of a number of disease genes, such as *APC* as a cause of familial adenomatous polyposis (3). Distantly related patients can be identified through the UPDB. These related patients can then be consented to participate in genetic studies to identify regions of the genome that are identical by descent (IBD). Regions that are IBD between patients could potentially harbor a shared disease risk variant. As a result, the identification of genomic regions that are IBD focuses the search space for potentially pathogenic variants. Analyses that identify regions of the genome that are shared by distantly related patients can also help to understand how genetic risk factors occur in seemingly sporadic cases.

Understanding the mechanisms by which genetic risk factors arise in SALS is critical to providing adequate healthcare and proper genetic counseling to patients and family members. The identification of shared segments of the genome between distantly related patients will help to better understand how these heritable risk factors arise in SALS. Furthermore, these shared genomic segments will allow for the potential discovery of novel ALS risk loci, which may account for the missing heritability seen in SALS (1, 4, 5). In this chapter, I will attempt to identify shared regions of the genome between distantly related ALS patients using the Shared Genomic Segments (SGS) analysis method—which is developed by the Nicola Camp Laboratory—and determine whether they harbor any ALS risk variants. These efforts will help to elucidate whether SALS can be caused by the inheritance of previously unrecognized genetic risk factors or not.

Methods

DNA was collected from 72 ALS patients and eight unaffected relatives by Dr. Summer Gibson (Department of Neurology, University of Utah). The collected DNA samples were Illumina whole-genome sequenced to an average coverage of 60X with 150-bp paired-end reads by NantOmics as part of the Heritage 1K Project. Genomic reads from each sequenced individual were then aligned to the GRCh37 reference genome using the BWA-MEM aligner (6). The aligned genomic reads from all 80 individuals underwent joint variant calling with 95 long-lived individuals (longevity cohort) and 291 European individuals (CEU (Utah Residents (CEPH) with Northern and Western European Ancestry) and GBR (British in England and Scotland)) from the 1000 Genomes Project (7) using the Genome Analysis Toolkit (GATK; v.3.4-46) best practices guidelines (8-10). Of the 95 individuals in the longevity cohort, 76 were determined to be unrelated (results not shown) by Dr. Deborah Neklason (Division of Genetic Epidemiology, University of Utah) and were used as healthy controls. The genotypic sex of each individual was imputed by *PLINK2* (11) and compared to the reported sex to identify any sample identification errors. Furthermore, the genotypes of each sample were compared to each other to identify any unexpected relatedness that may be indicative of a sample labeling error using *KING* (12). The UPDB identified that 36 of the 72 patients with ALS are distantly related (6–14 degrees of separation) to at least one other sampled patient and form 19 distinct pedigrees. Variant calls from these 36 ALS patients and 76 unrelated individuals from the longevity cohort were used for further analysis. The variants from these samples were intersected with biallelic SNPs from 99 CEU, 103 CHB (Han Chinese in Bejing, China), 104 JPT (Japanese in Tokyo, Japan), 108 YRI (Yoruba in Ibadan, Nigeria) samples from the 1000 Genomes Project (7). A *FlashPCA* (v1.2.5) principal components analysis (13) was performed on the intersected variant calls to identify any poor quality samples. An *ADMIXUTRE* (v1.3.0) analysis (14) was also performed to select for individuals with more than 80% European ancestry to avoid population stratification effects.

The distantly related ALS and longevity control cohort samples that passed all of the quality control steps were used for the SGS discovery analysis. To perform the SGS identification pipeline, SNP array genotype data—which was assayed by the Illumina 2.5 Omni array platform—was first gathered from 283 Europeans (CEU, GBR, and FIN (Finnish in Finland)) belonging to the 1000 Genomes Project to be used as additional control information. Biallelic autosomal SNPs found on the Illumina OmniExpress 700K marker genotyping array were extracted from the ALS, longevity, and 1000 Genomes Project cohorts to select for common SNPs. Common SNPs were selected to avoid premature breaks of a shared segment that may be caused by rare variants or sequencing errors. These biallelic autosomal SNPs from the ALS, longevity, and 1000 Genomes Project were then intersected and merged to be analyzed by SGS.

The SGS method aims to find regions of the genome that are significantly shared between distantly related individuals by first identifying segments that are identical-by-state (IBS). IBS is found by determining where in the genome the genotypes between samples are sequentially the same. SGS then calculates whether the length of an IBS region is longer than what is expected for that part of the genome to determine whether the sharing of a segment is statistically significant (15). The expected length of a shared segment is established by simulating Mendelian inheritance with local recombination rates, which are based upon the linkage structure seen in a control cohort and the Rutgers linkage map (16). The thresholds for determining what is a significant shared segment between case individuals is determined by the use of a technique described by Lander and Kruglyak (17).

The normal control recombination structure of the genome was established using linkage disequilibrium statistics from the longevity cohort and the European 1000 Genome Project individuals. An initial pass of 10,000 SGS simulations was performed for each of the distinct pedigrees formed by the ALS patients. SGS simulations are performed on each chromosome for each possible subset of patients that could possibly share a genomic segment. For instance, four sets of 10,000 SGS simulations are performed for each chromosome for a pedigree that consists of patients A, B, and C (subsets: A-B-C, A-B, A-C, B-C). An additional 1 million simulations were performed on any chromosome from a subset that possessed a segment longer than the null distribution 99.98% of the time. Segments of the genome that were significantly shared or suggestive of sharing (false positive rate of 1 segment per genome) between distantly related individuals were identified using the threshold method described above.

Regions of the genome that were significant or suggestive of sharing were then investigated for potential disease causing variants. All variants in the ALS patient cohort were annotated for their functional impact on the genome, ExAC (18) non-Finnish European (NFE) MAF, and European 1000 Genomes Project MAF using Ensembl's

Variant Effect Predictor (VEP; v83) (19). *Vcfanno* (20) was used to further annotate these variants with Genome Aggregation Database NFE MAF information. The Genome Aggregation Database (gnomAD) provides genome-wide allele frequency estimates by analyzing whole-genome sequencing data from 15,496 individuals (http://gnomad.broadinstitute.org/). A search was performed for rare variants, which are more likely to have a large effect size (21), shared by all individuals that possessed a region significant or suggestive of sharing. Alleles that could not be confidently emitted (no-calls) by the GATK HaplotypeCaller were considered to be the variant allele. Rare variants were defined as those with a gnomAD NFE MAF less than 0.001, an ExAC NFE MAF less than 0.001, and a European 1000 Genomes Project MAF less than 0.01. Variants that were multiallelic, occur at a frequency greater than 0.4—which is approximately the prevalence of pathogenic *C9orf72* repeat expansions in FALS patients (1)—in the ALS cohort, were located in a low-complexity region (22), or were marked as low-quality by gnomAD were discarded to reduce the number of false positive candidate variants. The VEP annotations of any rare variant that met these filtering criteria were then used to determine their functional impact.

Results

One sample labeling error was detected when the imputed genotypic sex of the 36 distantly related ALS patients was compared to the reported sex. The imputed genotypic sex of patient 15-0022906 was determined to be male; however, they were reported to be female. Further, the *KING* relationship analysis found that patient 15-0022906 was genetically identical to patient 15-0022867, who was one of the eight unaffected individuals in the total ALS cohort. These results suggest that individual 15-0022867 was sequenced twice at the expense of patient 15-0022906. As a result, patients 15-0022906 and 15-0022912, who was a distant relative with ALS, were removed from further analysis. The principal components analysis showed that none of the distantly related ALS patients or longevity cohort individuals were genetic outliers (Figure 6.1). The *ADMIXTURE* results showed that these same individuals all had greater than 80% European ancestry (Figure 6.2).

The dataset used for the SGS analysis consisted of 559,941 autosomal SNP genotypes from 393 individuals once the three cohorts were intersected and merged. The control recombination structure of the genome was determined from linkage disequilibrium statistics generated from the 76 longevity cohort individuals and 283 European 1000 Genomes samples. After an initial pass of 10,000 SGS simulations for each possible patient subset, 163 out of 770 (21.2%) total chromosomes possessed a segment that was longer than the null distribution 99.98% of the time. After performing 1 million additional simulations on these segments and determining the p-value thresholds, 46 regions from six patient subsets were found to be significant or suggestive for distant sharing. Two regions from two different patient subsets exceeded the significance threshold. However, one of these significant regions was likely a false positive because it was 39.3Mb in length and spanned a centromere. Two of the 45 regions suggestive of SGS were also likely false positives as they were 19.1Mb and 12.0Mb in length and occurred either in a centromere or telomere. These false positives were removed from further analysis, which left 43 significant or suggestive SGS regions for study (Table 6.1).

The smallest genomic region that was significant or suggestive of sharing between distantly related ALS patients was 0.29MB. In contrast, the largest segment was 2.83 Mb in length. The average genomic length of regions that were significant or suggestive of distant sharing was 1.21Mb (standard deviation = 692kb) with a median length of 0.88Mb (Figure 6.3).

A 1.06Mb region at chr18:66544756-67600990 was the only segment found to be significantly shared, which involved three ALS patients (15-0022918, 15-0022895, and 15-0022914). Three genes—*CCDC102B*, *DOK6*, and *CD226*—are found within this region. A missense variant in *CD226* (chr18:67531642 T>C; *CD226*:p. S307G) was the only nonsynonymous variant possessed by all three ALS patients who shared this segment. However, this variant is found at a gnomAD NFE MAF of 0.46. Furthermore, no rare variants that met the necessary filtering criteria and were shared among the three distantly related ALS patients were found in the 1.06Mb region.

Among all of the regions suggestive or significant for sharing between ALS patients, 326 unique genes were found. None of these genes were found in the ALS risk gene list in Chapter 3. No shared rare (MAF < 0.001) or semi-rare (ExAC NFE and gnomAD NFE MAF < 0.01) protein-coding sequence altering variants were found in any of the 326 genes. Two shared rare noncoding variants that met the filtering criteria were found among the 43 shared genomic segments. The first of these variants was chr14: 40968670 C>T and was found in patients 15-0022891, 15-0022900, and 15-0022911. However, inspection of the genomic reads at this position showed only patient 15-0022911 possessed this variant. The other passing noncoding variant was chr7:3983016 C>CCCA and was found in patients 15-0022918, 15-0022895, 15-0022914. This variant is likely a false positive as the genomic reads supporting this variant also mapped to another chromosome. Three semi-rare (MAF < 0.01) noncoding variants—which were shared by patients 15-0022869, 15-0022919, and 15-0022913—were found within a 67kb region of each other. These variants were chr4:41755833 GA>G, chr4:41821729 A>G, and chr4:41822967 C>T. Two long noncoding RNAs (lncRNAs), RP11-227F19.1 and RP11-227F19.2, are encoded in the 67kb region the three shared variants were located in. Furthermore, chr4: 41755833 GA>G is located 5kb upstream from *PHOX2B*, which is a transcription factor involved in the development of specific autonomic neuron populations (23). No other shared semi-rare variants were identified from this analysis.

Discussion

The experiments performed in this chapter sought to determine whether SALS could be caused by inherited genetic factors not recognized in other family members. To accomplish this goal, SGS analyses were conducted to detect regions of the genome that were likely shared by distantly related ALS patients, which may harbor ALS risk variants. In total, 43 regions with suggestive or significant signs of distant sharing were identified between six different patient subsets. However, none of these regions contained any known ALS risk genes, which is the opposite finding expected if these regions played a role in disease pathogenesis. This result suggests that older unrecognized genetic risk factors do not play an extensive role in causing SALS. Instead, it is likely that *de novo* mutations or the inheritance of recently created variants are the mechanism by which genetic risk factors are found in SALS cases. While some efforts have been made to determine the significance of *de novo* mutations in SALS pathogenesis (2), additional studies with larger patient cohorts will be required to determine whether such a mechanism causes disease.

This chapter also sought to find novel ALS risk loci by searching for rare genetic variants in genomic regions shared by distantly related ALS patients. No rare or semi-rare variants that alter the normal protein-coding sequence of a gene were found among all individuals with a shared genomic segment. Furthermore, only three semi-rare noncoding variants—which were in close proximity (67kb) to each other—in one patient subset were found. The two lncRNAs found in this region have no described function. The one protein coding gene near this region, *PHOX2B*, is known to play a role in the development of a set of autonomic neurons. However, it has not been associated with a motor neuron disease before. No significant H3K27Ac marks are found in this 67kb region, which suggests these variants do not play a large role in regulating gene expression. The lack of shared rare variation in any of the 43 regions identified by SGS further supports the notion that the inheritance of older unrecognized genetic risk variants is not a major cause of SALS. However, these shared regions could harbor common, low-effect size ALS risk variants which cause or modulate disease severity when inherited with other ALS risk factors. Large ALS genome-wide association studies based on whole-genome sequencing data will be required to detect such low-effect size variants.

While these results suggest shared genetic variation between distantly related ALS patients is not a major cause of disease, there were technical limitations to this study. First, the limited sample size of some of the studied pedigrees prevented the identification of shared genomic segments. For instance, no regions were found to be significant or suggestive for sharing between patient subsets comprised of two individuals. In contrast, five of the six subsets with three patients had at least one segment suggestive or significant for distant sharing. This result suggests that SGS has limited ability to detect shared genomic regions in small pedigrees. Sequencing additional affected family members from these small pedigrees will increase the power to detect shared genomic segments. The second limitation of this study was structural variants were not considered when searching for variants in shared genetic segments. The GATK HaplotypeCaller is limited in its ability to detect insertion or deletion variants that are tens to thousands of nucleotide bases long. Structural variants callers, such as *LUMPY* (24) and *Wham* (25), have the ability to detect such variants and could be used to search for structural variants in shared genomic segments. Lastly, the small number of extended ALS pedigrees used in this study limits the ability to definitively say what the inheritance pattern of genetic risk factors is for SALS. It is possible shared genomic segments do play a role SALS pathogenesis, but they weren't observed in our limited patient sample. Larger ALS patient and family member cohorts will be achieved as more individuals are seen and sequenced in the motor neuron disease clinic at the University of Utah. These efforts will help to determine how genetic risk factors arise in SALS, which will subsequently help to better understand and eventually treat the disease.

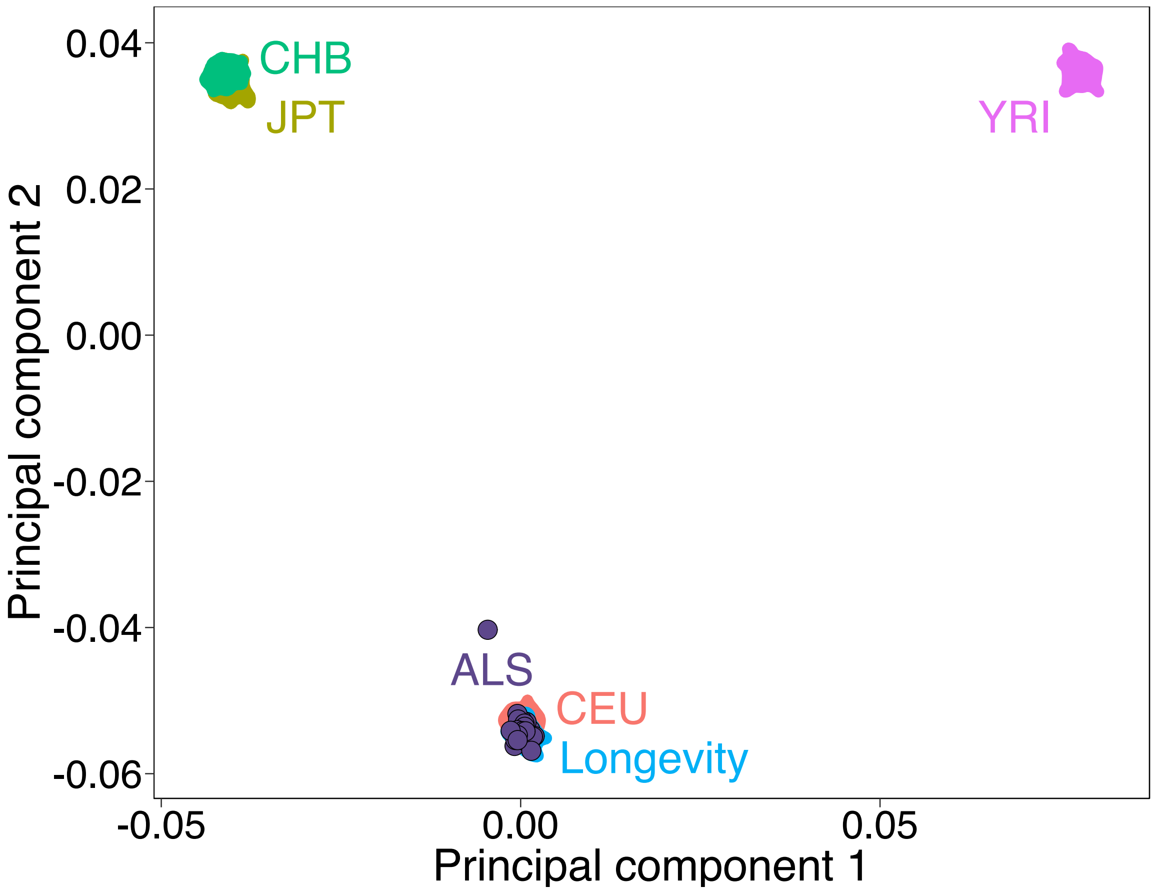


Figure 6.1 A principal components analysis comparing the genetic variance of the ALS, longevity, and selected 1000 Genomes Project samples. None of the sequenced samples were considered to be genetic outliers, which would be a sign of poor sample quality. One ALS sample appeared to have some East Asian admixture.

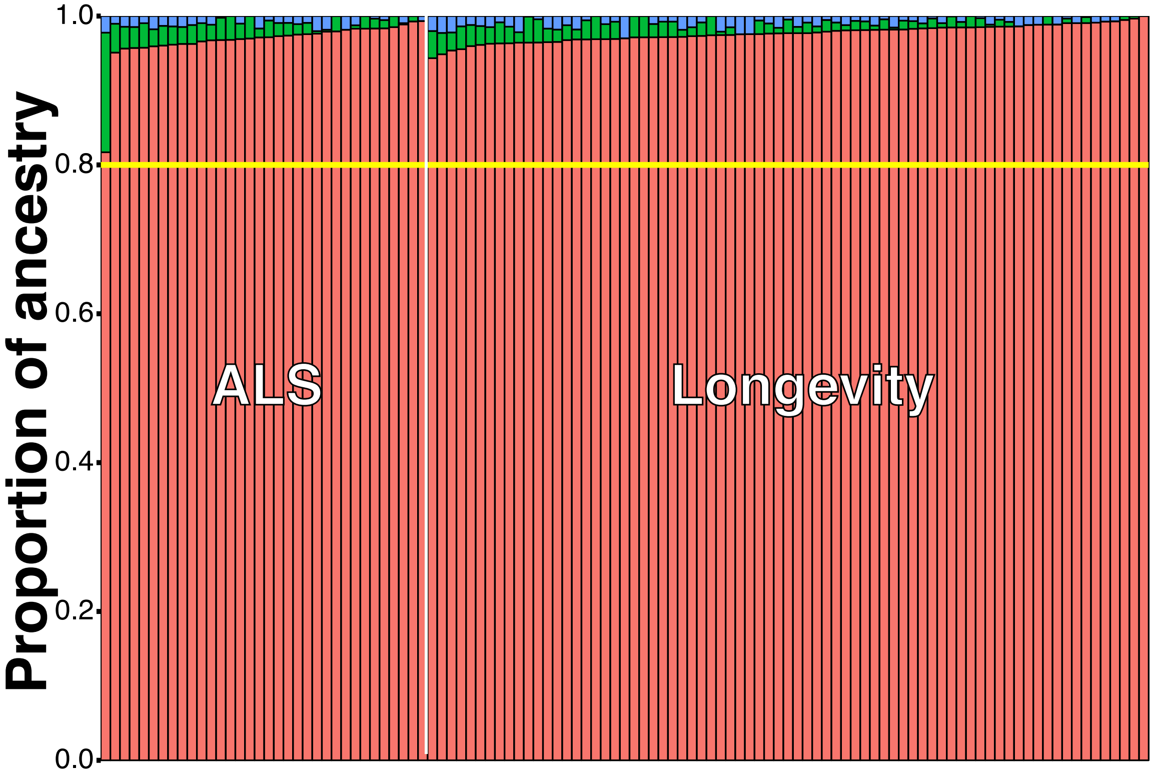


Figure 6.2 An admixture plot where each column represents an individual from the ALS or longevity cohort. The height of each color represents the amount of ancestry each individual has. Red represents European (CEU) ancestry, green represents East Asian (CHB +JPT) ancestry, and blue represents African (YRI) ancestry. The yellow bar indicates the European ancestry proportion cut-off (0.80) to be considered for SGS analysis. All ALS and longevity samples were determined to be largely of European ancestry.

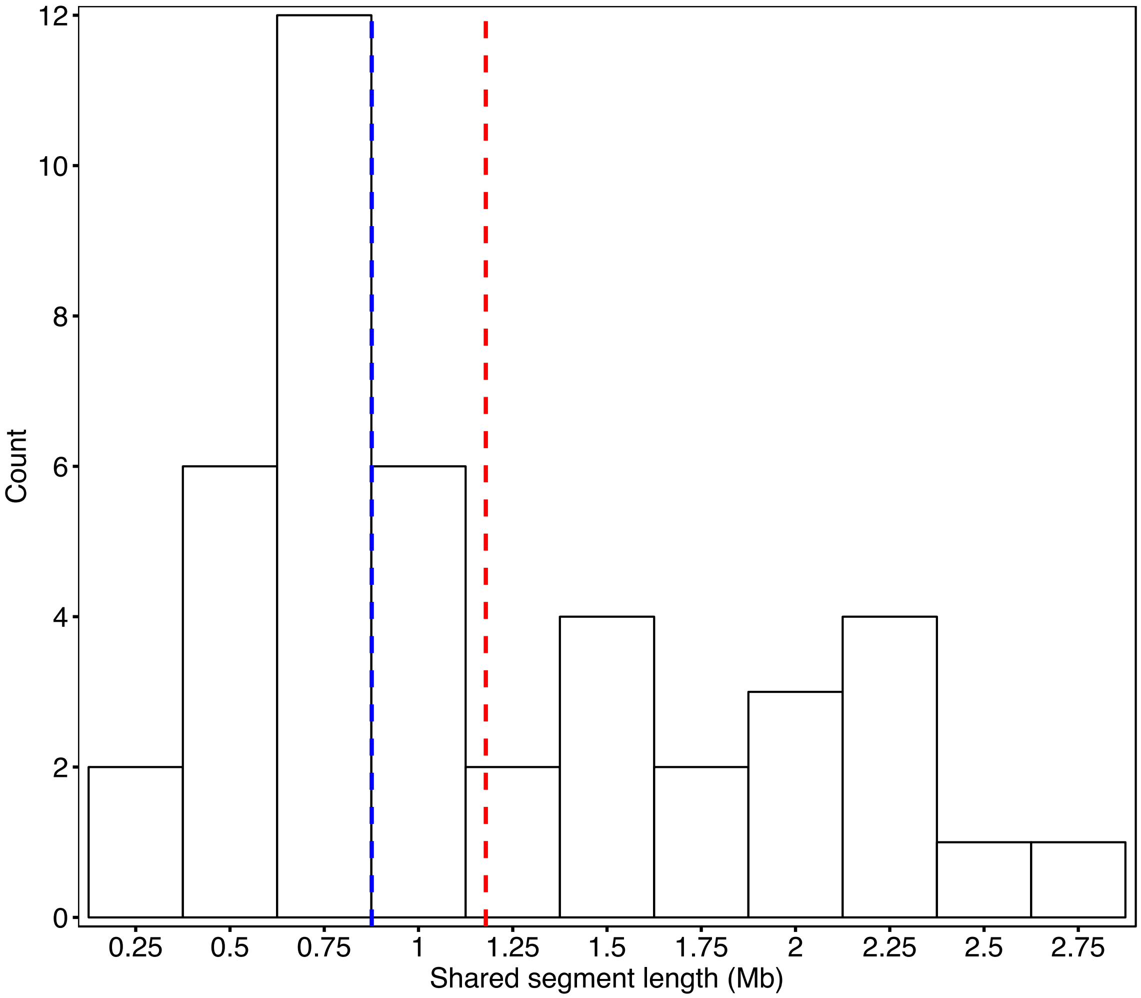


Figure 6.3 A histogram showing the length distribution of all 43 genomic segments with significant or suggestive signs of sharing between distantly related ALS patients. The red dashed line shows the average genome segment length (1.21Mb). The blue dashed line represents the median shared genomic segment length (0.88Mb).

Table 6.1 The 43 genomic regions that were significant or suggestive of sharing between distantly related ALS individuals. The location of each segment is based on the GRCh37 human reference genome (Chromosome:Start-Stop). The ALS patients that possess each SGS region are listed. The p-value indicates the frequency at which simulations find longer shared segments than the reported segment at the same genomic locus. * represents a region with a significant p-value.

Shared segment location	Segment length (base pairs)	p-value	Patients with shared segment
1:45170013-48003120	2833107	2.67×10^{-5}	15-0022918,15-0022895,15-0022914
1:111155596-111929294	773698	7.03×10^{-5}	15-0022928,15-0022894,15-0022930
1:118936322-120197684	1261362	3.27×10^{-5}	15-0022918,15-0022895,15-0022914
1:158425272-159505296	1080024	9.90×10^{-5}	15-0022891,15-0022900,15-0022911
1:173117772-175283118	2165346	9.60×10^{-5}	15-0022869,15-0022919,15-0022913
1:175849783-177458169	1608386	1.07×10^{-4}	15-0022918,15-0022895,15-0022914
2:134657741-136934448	2276707	9.60×10^{-5}	15-0022918,15-0022895,15-0022914
3:13055577-13644175	588598	5.84×10^{-5}	15-0022918,15-0022895,15-0022914
3:106981728-107765391	783663	4.16×10^{-5}	15-0022918,15-0022895,15-0022914
4:40491787-41367088	875301	9.90×10^{-6}	15-0022869,15-0022919,15-0022913
4:41367090-43148952	1781862	7.92×10^{-6}	15-0022869,15-0022919,15-0022913
4:45374306-47382919	2008613	5.84×10^{-5}	15-0022891,15-0022900,15-0022911
4:57741416-58314263	572847	6.73×10^{-5}	15-0022891,15-0022900,15-0022911
4:157235051-158731018	1495967	1.01×10^{-4}	15-0022869,15-0022919,15-0022913
5:2627546-2915300	287754	1.08×10^{-4}	15-0022891,15-0022900,15-0022911

Table 6.1 Continued

Shared segment location	Segment length (base pairs)	p-value	Patients with shared segment
5:23716351-25717661	2001310	6.14×10^{-5}	15-0022928,15-0022894,15-0022930
5:27932381-29516288	1583907	2.77×10^{-5}	15-0022869,15-0022919,15-0022913
5:81960141-82889909	929768	6.44×10^{-5}	15-0022869,15-0022919,15-0022913
5:106149175-107010200	861025	2.08×10^{-5}	15-0022918,15-0022895,15-0022914
5:135296363-136793147	1496784	1.49×10^{-5}	15-0022869,15-0022919,15-0022913
5:150038266-150673386	635120	5.05×10^{-5}	15-0022869,15-0022919,15-0022913
6:158216127-159146870	930743	7.33×10^{-5}	15-0022869,15-0022919,15-0022913
7:3147021-4022952	875931	3.86×10^{-5}	15-0022918,15-0022895,15-0022914
8:134681124-135307547	626423	7.92×10^{-5}	15-0022918,15-0022895,15-0022914
8:134721708-135907055	1185347	1.94×10^{-4}	15-0022916,15-0022928,15-0022894
9:3022593-3838928	816335	4.06×10^{-5}	15-0022891,15-0022900,15-0022911
10:31860346-33689742	1829396	7.92×10^{-6}	15-0022918,15-0022895,15-0022914
10:112186149-112969626	783477	6.04×10^{-5}	15-0022891,15-0022900,15-0022911
10:122480688-123032167	551479	6.53×10^{-5}	15-0022869,15-0022919,15-0022913
12:12042460-12482447	439987	1.25×10^{-4}	15-0022918,15-0022895,15-0022914
12:121794778-124013405	2218627	1.18×10^{-4}	15-0022918,15-0022895,15-0022914
12:129932797-130232896	300099	1.34×10^{-4}	15-0022869,15-0022919,15-0022913
13:28240829-28941059	700230	4.65×10^{-5}	15-0022918,15-0022895,15-0022914

Table 6.1 Continued

Shared segment location	Segment length (base pairs)	p-value	Patients with shared segment
14:40364645-42952897	2588252	8.91×10^{-6}	15-0022869,15-0022919,15-0022913
14:40588542-42719753	2131211	1.88×10^{-5}	15-0022891,15-0022900,15-0022911
14:63322348-65245955	1923607	1.04×10^{-4}	15-0022891,15-0022900,15-0022911
14:98309638-98991119	681481	1.16×10^{-4}	15-0022891,15-0022900,15-0022911
15:33702671-34079431	376760	3.86×10^{-5}	15-0022918,15-0022895,15-0022914
18:43599356-44252273	652917	5.84×10^{-5}	15-0022891,15-0022900,15-0022911
18:66544756-67600990	1056234	4.95×10^{-6}*	15-0022918,15-0022895,15-0022914
18:68828003-69394951	566948	1.29×10^{-4}	15-0022869,15-0022919,15-0022913
18:76687315-77553172	865857	4.26×10^{-5}	15-0022918,15-0022895,15-0022914
19:7084748-7758338	673590	3.56×10^{-5}	15-0022869,15-0022919,15-0022913

References

1. Renton AE, Chio A, Traynor BJ (2014) State of play in amyotrophic lateral sclerosis genetics. *Nat Neurosci* 17(1):17-23.

2. Chesi A, et al. (2013) Exome sequencing to identify de novo mutations in sporadic ALS trios. *Nat Neurosci* 16(7):851-855.

3. Leppert M, et al. (1987) The gene for familial polyposis coli maps to the long arm of chromosome 5. *Science* 238(4832):1411-1413.

4. Keller MF, et al. (2014) Genome-wide analysis of the heritability of amyotrophic lateral sclerosis. *JAMA Neurol* 71(9):1123-1134.

5. Al-Chalabi A, et al. (2010) An estimate of amyotrophic lateral sclerosis heritability using twin data. *J Neurol Neurosurg Psychiatry* 81(12):1324-1326.

6. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

7. 1000 Genomes Project Consortium, et al. (2015) A global reference for human genetic variation. *Nature* 526(7571):68-74.

8. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491-498.

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

10. Van der Auwera GA, et al. (2013) From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43(11.10):1-33.

11. Chang CC, et al. (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* 4:7.

12. Manichaikul A, et al. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867-2873.

13. Abraham G, Inouye M (2014) Fast principal component analysis of large-scale genome-wide data. *PLoS One* 9(4):e93766.

14. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655-1664.

15. Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA (2008) Shared genomic segment analysis. Mapping disease predisposition genes in

extended pedigrees using SNP genotype assays. *Ann Hum Genet* 72(Pt 2):279-287.

16. Matise TC*,* et al. (2007) A second-generation combined linkage physical map of the human genome. *Genome Res* 17(12):1783-1786.

17. Lander E, Kruglyak L (1995) Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nat Genet* 11(3):241-247.

18. Lek M*,* et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285-291.

19. McLaren W*,* et al. (2016) The Ensembl Variant Effect Predictor. *Genome Biol* 17(1):122.

20. Pedersen BS, Layer RM, Quinlan AR (2016) Vcfanno: Fast, flexible annotation of genetic variants. *Genome Biol* 17(1):118.

21. Manolio TA*,* et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461(7265):747-753.

22. Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30(20):2843-2851.

23. Pattyn A, Morin X, Cremer H, Goridis C, Brunet JF (1999) The homeobox gene Phox2b is essential for the development of autonomic neural crest derivatives. *Nature* 399(6734):366-370.

24. Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol* 15(6):R84.

25. Kronenberg ZN*,* et al. (2015) Wham: Identifying Structural Variants of Biological Consequence. *PLoS Comput Biol* 11(12):e1004572.

CHAPTER 7

CONCLUSIONS AND PERSPECTIVES

Just over a decade after ALS was formally defined by Charcot in 1869 (1), it was recognized that ALS had a familial component to its etiology (2, 3). This insight eventually led to the discovery of *SOD1* as the first ALS risk gene in 1993 (4). Since then, a vast number of efforts have been made to determine the genetic risk landscape of ALS. These attempts have largely been in the form of low-resolution linkage studies and common variant genome-wide association studies. While such efforts have been successful in finding a number of ALS risk loci, a large proportion of the heritability seen in ALS is still unaccounted for. This missing heritability is likely due to rare and large-effect size variants, which are not assayed by microsatellite and common SNP genotyping arrays. NGS methods have the ability to detect such rare variants. Until recently, the financial and computational demands needed to perform NGS made widespread adoption of the technology limited. However, the dramatic reduction in cost of NGS since the Human Genome Project (5) —from $2.7B to $1,000 for whole-genome sequencing (6)—has made investigations based of NGS technology feasible. The use of NGS within the context of ALS, and human genetics as a whole, heralds a new era of genetic discovery due the ability to assay rare and structural variation.

The studies performed in this dissertation attempt to better understand the genetic etiology of ALS using NGS approaches. In Chapter 2, efforts were made identify genes and variants that potentially cause FALS in small sequencing studies. A candidate or known risk gene could be identified in a majority of cases in this study, which suggests the use of NGS approaches in the clinic could be useful in providing clinical decision making and genetic counseling. Furthermore, the candidate risk genes identified in Chapter 2 are interesting targets for functional testing to determine if they are involved in ALS pathogenesis.

Chapter 3 of this dissertation attempts to define how large a role known SALS risk loci play in the disease using whole-exome sequencing. Despite numerous attempts to determine what proportion of SALS is caused by known genetic risk factors, no consensus had been reached likely due to methodological flaws. A number of NGS computational approaches were used and developed, including population stratification correction and direct predictions of variant pathogenicity, to correct for these flaws. The results of Chapter 3 showed that these measures were able to derive a more accurate estimation of the amount of risk conferred by known ALS risk loci to the pathogenesis of SALS. Furthermore, this chapter showed that known genetic risk factors do not completely account for the known heritability of SALS.

Chapter 4 was aimed at closing the ALS missing heritability gap by identifying new ALS genetic risk loci. This was done by performing gene burden testing of the whole-exome sequenced cohort analyzed in Chapter 3. While burden testing has been employed in the past to identify ALS risk genes (7), these methods relied on allele frequency as the sole criterion of pathogenicity. Chapter 3 demonstrated that *in silico* predictions of variant pathogenicity capture ALS genetic risk better than variant frequency alone. To improve upon previous ALS risk gene discovery efforts, the *VAAST* method (8)—which incorporates variant frequency and direct predictions of variant pathogenicity—was used in combination with the gene ontology based *PHEVOR* tool (9). This analysis identified two novel ALS candidate risk genes, *MFN2* and *TP73*, in ALS patients that appeared to be burdened by deleterious variation and relevant to the disease phenotype. Both genes would expand the understanding of the disease if they are indeed ALS risk genes. However, functional experimentation is required to determine the role of *TP73* and *MFN2* in ALS.

The efforts made in Chapter 5 build upon that reasoning in order to determine whether *TP73* is involved in ALS pathogenesis. To do so, this chapter first attempted to find whether *TP73* variants were prevalent outside of the initial discovery cohort. Screening of over 2,800 ALS patients found 19 rare and deleterious variants in *TP73* in addition to the five found in the initial discovery cohort. This result suggests that potentially pathogenic variation in *TP73* is not limited to ALS patients seen at the University of Utah. Next, *in vivo* experiments were performed to determine whether *TP73* has a role in the development and function of motor neurons. Loss of p73 function in zebrafish resulted in impaired motor neuron survival and development. Together, these findings strongly link *TP73* to ALS pathogenesis. Future *in vivo* rescue studies will be required to determine if the specific *TP73* variants found in patients cause ALS.

The results from Chapter 3 also demonstrate a large proportion of SALS patients possess an ALS genetic risk factor. However, it is not clear how these genetic risk factors arise in SALS. Understanding how ALS genetic risk factors arise in SALS patients is critical to proper clinical decision making and genetic counseling. Multiple mechanisms have been proposed, including *de novo* mutations. However, it also possible genetic risk factors could be transmitted through multiple generations and inherited by affected individuals, but were previously unrecognized due to incomplete penetrance and mortality due to non-ALS causes. It would be expected distantly related patients would share genomic segments that harbor ALS genetic risk factors if such a mechanism was common. Chapter 6 focuses on determining whether shared genomic segment analysis can identify ALS risk variants. None of the identified regions of the genome shared between distantly related ALS patients appear to contain large effect size variants that could cause ALS. This suggests the multigenerational transmission of unrecognized ALS genetic risk factors is not a common mechanism by which such variants are transmitted to SALS patients. Future studies will be required to determine the importance of other mechanisms by which ALS genetic risk factors are transmitted to SALS patients, such as *de novo* mutations.

Despite many of the insights made by this dissertation, much remains unknown about the genetic etiology of ALS. This is partially due to small sample sizes and limited statistical power to detect the effects of risk variants in many of the performed studies. The overall rarity of ALS makes it difficult to assemble large patient cohorts. A substantial amount of effort will be required to extensively enroll new ALS patients into genetic studies. Furthermore, multicenter collaborations where multiple ALS cohorts are analyzed together by NGS approaches will likely be required to completely understand the genetic etiology of ALS.

The studies performed in this dissertation demonstrate that WGS approaches can make insights into the genetic etiology of ALS that were previously impossible. Furthermore, this work has helped to better understand the importance both known and novel ALS genetic risk loci to disease pathogenesis. The methods and approaches used in this dissertation serve as a model by which future WGS investigations can make important genetic discoveries. The insights into the genetic risk landscape of ALS made by this dissertation have also strengthened the ability to provide proper genetic counseling and clinical care to patients. Lastly, this work will contribute to the final goal of discovering a cure for the harrowing, debilitating, and fatal disease known as ALS.

References

1. Charcot J-M, Joffroy A (1869) *Deux cas d'atrophie musculaire progressive: Avec lésions de la substance grise et des faisceaux antéro-latéraux de la moelle épinière* (V. Masson, Paris, France).

2. Osler W (1880) On heredity in progressive muscular atrophy as illustrated in the Farr family of Vermont. *Arch Med* 4:316-320.

3. Siddique T, Ajroud-Driss S (2011) Familial amyotrophic lateral sclerosis, a historical perspective. *Acta Myol* 30(2):117-120.

4. Rosen DR, et al. (1993) Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 362(6415):59-62.

5. Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860-921.

6. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* 17(6):333-351.

7. Cirulli ET, et al. (2015) Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* 347(6229):1436-1441.

8. Hu H, et al. (2013) VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol* 37(6):622-634.

9. Singleton MV, et al. (2014) Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am J Hum Genet* 94(4):599-610.