INVESTIGATING AMYOTROPHIC LATERAL SCLEROSIS CANDIDATE GENES

By

Sandrine Kim Kiow Chan Moi Fat

A thesis submitted to Macquarie University
for the degree of Masters of Research
Department of Biomedical Science
Faculty of Medicine and Health Sciences
January 2019

Typeset in $\LaTeX\,2_{\varepsilon}$.

Declaration

I wish to acknowledge the following assistance in the research outlined in this project:

Whole-exome sequencing data was generated at Macrogen, Korea. Family FALS147 whole-exome sequencing analysis and *in silico* scoring of candidate genes were performed by Jennifer Fifita. The resulting whole-exome candidate genes were also validated by Jennifer Fifita prior to this project.

Whole-genome sequencing data was generated at the Kinghorn Centre for Clinical Genomics, Sydney. Raw data was processed by collaborators at CSIRO and annotated by Emily McCann.

The gene_search.sh, gene_variant_search.sh and shared_variant.sh Unix scripts were previously written by Emily McCann. The SALS_subset.sh script in R studio was developed by Ingrid Tarr.

Some dot blotting rounds, and western blotting was carried out by Alison Hogan and Natalie Grima for the CRISPR project.

DNA sequencing was carried out by Macrogen Sequencing, Korea.

All other research described in this project is my own original work and has not been submitted in whole or part for a degree at any university.

S.Chan

Acknowledgements

I would first like to express my deepest gratitude to my supervisors, Prof. Ian Blair, Dr Jennifer Fifita, and Dr Shu Yang, for all their support during this research year. Thank you, Ian, for welcoming me into your team, and always providing me with expert guidance and advice despite your busy schedule. Jenn, I cant thank you enough for everything you have done for me. You have been an incredible mentor, with great patience and knowledge, and I certainly could not have made it through without you. Shu, thank you for sharing your love and expertise of cell biology with me, your help was vital for the successful completion of this project.

To the rest of the Blair group, Kelly Williams, Alison Hogan, Emily McCann, Natalie Grima, Ingrid Tarr, Elisa Cachia, Sarah Furlong, Owen Watson, and Sharlynn Wu, I really enjoyed working and getting to know you all. Emily, you never got tired of my silly questions and always managed to keep me calm. Natalie, you have been such a fun and helpful lab mate. Alison, your support during the CRISPR project made the whole process less stressful. And of course, to my MRes friends, Owen and Sharlynn, it was great having you around to share stress, jokes, and snacks together.

Also, I need to acknowledge my friends and relatives, who have provided me with huge moral support. Dr LayPeng Tan, you are such an influential role model, I admire your passion, courage, hard work, and strong devotion to your students. To my cousins Martine and Alison, thanks for all the joys, laughs, movies, and food sessions together! A special mention to my boyfriend Ronnie, who has always supported me.

Last but not least, I am very grateful to my loving parents, who have always encouraged me to aim for higher than the stars. You have always believed in me and made so many sacrifices to provide me with the life that I currently have. To my sister Carine, you have always been here for me, as a best friend and confidant. Words are not enough to describe how thankful I am to have such a wonderful family.

Abstract

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease of motor neurons. Approximately 10% of cases are classified as familial and 90% sporadic, although a proportion of sporadic cases have an unrecognised family history. This thesis aimed to identify novel ALS genes from familial and sporadic cases. Whole-exome and whole-genome sequencing analysis identified 22 candidate gene variants in an ALS family, which were investigated using an *in silico* analysis pipeline including protein predictions, conservation, genic tolerance, tissue expression and gene function. Three top priority variants, in *CYB5R3*, *DCAF7*, and *SAV1*, were assessed using *in vitro* toxicity and localisation assays, with one candidate, CYB5R3, displaying a differential cytoplasmic protein expression pattern. Whole-genome data from 635 sporadic ALS patients were interrogated for ALS candidate genes (n = 21), and ten novel variants were identified, including a novel *CYB5R3* variant, which rated strongly from *in silico* tools. Lastly, this project sought to develop a CRISPR-Cas9 protocol to generate ALS cell models for future use in the *in vitro* analysis pipeline. Overall, a combination of genetic, *in silico* and *in vitro* pipelines were used to identify candidate genes in an ALS kindred and apparently sporadic patients. Novel ALS genes will further our understanding of disease biology and contribute to development of diagnostics and treatments.

Conflict of Interest Statement

We declare that there are no conflicts of interest and this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

S.Chan

Sandrine Kim Kiow Chan Moi Fat

Professor Ian Blair

Dr Shu Yang

J Fifita

Dr Jennifer Fifita

Contents

List of Figures

List of Tables

1

Introduction

1.1 Amyotrophic lateral sclerosis

Motor neuron disease (MND) is an umbrella term for a group of conditions that are variably caused by the progressive degeneration of motor neurons. Motor neurons are specialised nerve cells that control muscle movement. Upper motor neurons (UMNs) originate in the motor cortex of the brain, descend to the brain stem and spinal cord and control the lower motor neurons (LMNs). The LMNs originate in the brain stem and spinal cord and directly innervate the muscles. There are four main subtypes of MND; namely amyotrophic lateral sclerosis (ALS), primary lateral sclerosis (PLS), progressive bulbar palsy (PBP) and progressive muscular atrophy (PMA) (Figure 1.1). PLS affects the UMNs, while PMA affects LMNs in the spinal cord. Degeneration of UMNs leads to muscle spasticity, weakness and paralysis, while degeneration of LMNs leads to muscle atrophy, weakness, and flaccidity (**Kiernan et al.**, 2011). In PBP, the LMNs of the brainstem are the most affected, causing difficulty in chewing and swallowing. ALS is the most common type of MND and it affects both the UMNs and LMNs (**Rowland and Shneider**, 2001). It has the most rapid disease course and shortest survival of typically 3-5 years. ALS has great phenotypic heterogeneity where age of onset, site of onset, and progression of disease can vary widely.

ALS also shares clinical and pathological features with frontotemporal dementia (FTD) as well as other degenerative diseases. FTD is a neurodegenerative disease characterised by frontotemporal cerebral atrophy (**Englund et al.**, 1994), which leads to changes in personality, behaviour and language. Up to 50% of ALS patients show deficits in frontal executive skills (**Lomen-Hoerth et al.**, 2003), and about 20% meet the criteria for co-morbid ALS-FTD (**Taylor et al.**, 2016). ALS and FTD also show commonalities with white matter disease, where patients experience issues with balance, walking, and performing two or more activities simultaneously. For example, ALS and ALS-FTD patients can exhibit white matter changes such as tissue atrophy, also seen in white-matter disease (**Lillo et al.**, 2012). The high degree of overlap between ALS, FTD and perhaps other neurodegenerative diseases such as white-matter disease, suggest that numerous conditions may form part of a continuum of neurodegenerative disease (**Fecto and Siddique**, 2011).

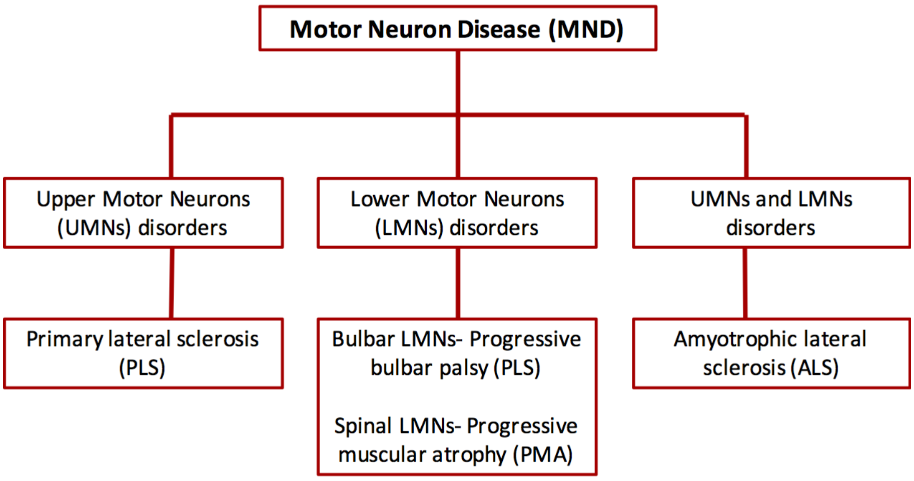

Motor Neuron Disease (MND)
Upper Motor Neurons (UMNs) disorders
Lower Motor Neurons (LMNs) disorders
UMNs and LMNs disorders
Primary lateral sclerosis (PLS)
Bulbar LMNs- Progressive bulbar palsy (PLS)
Spinal LMNs- Progressive muscular atrophy (PMA)
Amyotrophic lateral sclerosis (ALS)

Figure 1.1: **The MND subtypes.** There are four main MND subtypes. PLS affects mostly the UMNs, PBP predominantly the bulbar LMNs, PMA the spinal LMNs and ALS affects both the UMNs and LMNs.

The prevalence of ALS in Australia is about 8.7 per 100,000 people(recorded in 2015 by MND Australia, `www.mndaust.asn.au`), although prevalence is highly variable between different ethnic populations (**Cronin et al.**, 2007). Approximately 10% of ALS cases have a family history and are considered as familial (FALS), while the remaining 90% of cases are classified as sporadic (SALS). Most FALS cases display an autosomal dominant pattern of inheritance, but rare autosomal recessive and X-linked inheritance has also been identified (**Al-Chalabi et al.**, 2012; **Taylor et al.**, 2016).

To date, the only proven causes of ALS are gene mutations.

Several environmental factors have also been implicated as risk factors contributing to ALS (**Bozzoni et al.**, 2016). Proposed risk factors include exposure to heavy metals, electromagnetic fields, physical activity, pesticides and B-N-methylamino-L-alanine (BMAA). There has been mixed evidence regarding the role of these environmental factors on disease. Of these, exposure to pesticides and to BMAA have the strongest evidence from *in vivo* studies, but both await compelling epidemiological evidence (**Bozzoni et al.**, 2016; **Cox et al.**, 2018).

Although the underlying pathological processes of ALS are poorly understood, several pathogenic mechanisms have been proposed. These include RNA dysfunction, protein homeostasis, oxidative stress, impaired axonal transport, neuroinflammation, apoptosis, and mitochondrial dysfunction (**Ling et al.**, 2013; **Maurel et al.**, 2018). The main pathological feature of ALS is aggregation of ubiquitinated misfolded proteins in the motor neurons that are undergoing neurodegeneration (**Leigh et al.**, 1989; **Neumann et al.**, 2006). In 2006, it was found that most protein inclusions contain the TAR-DNA binding protein, TDP-43 (**Neumann et al.**, 2006). TDP-43 pathology has also been implicated in other neurodegenerative diseases, such as Parkinson's disease, Alzheimer's disease, FTD, and Huntington's disease (**Neumann et al.**, 2006; **Schwab et al.**, 2008; **Wilson et al.**, 2011). Moreover, experimental studies have observed that aggregates of mutant ALS proteins are structurally different to wild type proteins, with mutant proteins more abundant in detergent-insoluble rather than detergent-soluble fractions of the cell (**Basso et al.**, 2006, 2009).

Research in ALS has been progressing rapidly, but there is still a lack of early diagnostic tools and effective treatments for ALS. Screening of known gene mutations and observation of progressive symptoms are currently used. Clinicians use the El Escorial criteria, which examines the UMN and LMN symptoms to determine levels of diagnostic certainty in the patient (**Brooks et al.**, 2000). Most patients are only diagnosed after UMN and LMN symptoms have been fully established. Therefore, there is an urgent need to implement new methods for early diagnosis. For example, the identification of blood/specimen biomarkers, and a greater genetic mutation repertoire (**Talbot**, 2009). To-date, there are only two drugs that have been licensed by the USA-FDA for ALS treatment, Riluzole and Radicava (Edaravone), but both have very limited efficacy (**Cruz**, 2018; **Miller et al.**, 2012). A greater understanding of disease pathogenesis and identification of the full genetic aetiology of ALS can aid the development of diagnostic tests, targeted therapeutic interventions and gene therapy in the future (**Marangi and Traynor**, 2015), to delay disease progress and extend the length and quality of life of ALS patients.

1.2 Genetics and pathology of ALS

Genetic mutations remain the only proven cause of ALS and mutations in more than 22 genes have been identified, accounting for about 60% of Australian FALS and 10% of Australian SALS **McCann et al.** (2017). Despite rapid progress in research, 40% of FALS cases and 90% of SALS cases still have an unknown genetic cause (**Renton et al.**, 2014), highlighting the need for gene discovery research.

In 1993, mutations in the first ALS gene, *SOD1*, were identified in ALS families in a previously identified disease-linked genetic loci (**Rosen et al.**, 1993; **Siddique et al.**, 1991). The encoding protein, superoxide dismutase (SOD1) is involved in ER stress and protein homeostasis. The next major disease gene discovery in ALS was 15 years later, when ALS mutations were identified in the gene encoding TDP-43 (*TARDBP*), also through linkage analysis and direct sequencing of FALS and SALS patients. The discovery of *TARDBP* mutations closely followed the identification of TDP-43 pathology in ALS (**Sreedharan et al.**, 2008). The next year, ALS mutations in *FUS* were discovered using genome-wide linkage analysis (**Vance et al.**, 2009). FUS and TDP-43 are both DNA and RNA binding proteins involved in RNA metabolism. In 2011, hexanucleotide repeat expansions in *C9orf72*, encoding for another protein involved in RNA processing, were identified as the most common known cause of ALS, accounting for about 40% of FALS cases (**DeJesus-Hernandez et al.**, 2011; **McCann et al.**, 2017). This repeat expansion was implicated in ALS following genome-wide linkage studies carried out on large ALS-FTD families. Therefore, family-based linkage studies, genome-wide association studies (GWAS) and more recently next-generation sequencing (NGS) studies have all been used to find new ALS gene mutations, as well as new variants in ALS susceptibility genes.

In recent years, the improvements in NGS technologies have dramatically changed human genetics research in medicine, including ALS. NGS refers to whole-genome sequencing (WGS), whole-exome sequencing (WES) and targeted gene sequencing. NGS can either be used alone, or in combination with traditional genetic linkage analysis. Many FALS genes, including *CHCHD10*, *MATR3*, *PFN1*, *CCNF*, *TBK1* and *UBQLN2*

(**Bannwarth et al.**, 2014; **Deng et al.**, 2011; **Freischmidt et al.**, 2015; **Johnson et al.**, 2014; **Williams et al.**, 2016; **Wu et al.**, 2012) have recently identified using NGS technologies (Table 1.1). As such, in combination with linkage analysis, the genetic cause of most large ALS families has been solved, with typically small families remaining. These smaller families have insufficient power for significant linkage and segregation analysis, and therefore NGS analysis typically results in the identification of numerous candidate gene variants. To address this issue, this project used a three-stage novel gene discovery pipeline to prioritise and find the disease-causing gene variant in a small ALS family (Section 1.3).

After the identification of novel ALS genes in FALS, other ALS patients with unknown genetic mutation, including SALS, can be screened. Using this method, mutations in many familial ALS genes such as *FUS*, *TARDBP*, and *C9orf72* have been found in sporadic ALS cases (**Sreedharan et al.**, 2008; **Vance et al.**, 2009; **Williams et al.**, 2016). Thus, gene mutations that cause FALS can also contribute to SALS mutations discovery. So far, more than 30 genes have been implicated in ALS, as listed in the Amyotrophic Lateral Sclerosis Online Genetics Database (`http://alsod.iop.kcl.ac.uk/`). These disease genes have given researchers great insight into the molecular origins and pathogenic mechanisms of the disease. Known ALS genes can broadly categorised into two groups; genes involved in protein homeostasis, and genes involved in RNA homeostasis. The following section will thus provide a brief history of gene mutations found in FALS, their functions and how they helped our understanding of ALS pathogenesis (Table 1.1).

ER stress

The endoplasmic reticulum (ER) is an important organelle that plays a role in protein folding and lipid synthesis. When proteins are misfolded and accumulate within the ER, the unfolded protein response (UPR) signalling system is activated (**Kaufman**, 2002). If the UPR cannot resolve the defects, it will trigger apoptosis (**Hitomi et al.**, 2004; **Kaufman**, 2002). ER stress and UPR dysfunction have been implicated in motor neuron degeneration in ALS following analysis of mutant *SOD1* mouse models and spinal cord tissues (**Atkin et al.**, 2008; **Kieran et al.**, 2007; **Sasaki**, 2010).

SOD1

SOD1 was identified in 1993 following genetic linkage studies in ALS families (**Siddique et al.**, 1991), which identified a locus at 21q22.1-22.1. Mutation screening of

SOD1 revealed 11 *SOD1* missense mutations in 13 ALS families (**Rosen et al.**, 1993). To date, there have been over 180 *SOD1* mutations reported in ALS, spanning all five exons of the gene (**Boylan**, 2015). Although numbers vary among different ethnicities, mutations in *SOD1* account for approximately 20% of FALS cases, corresponding to about 1-2% of all ALS cases (**Boillée et al.**, 2006; **Boylan**, 2015). Inheritance of most *SOD1* mutations is autosomal dominant. Patients with *SOD1* mutations show ubiquitinated protein aggregates in affected neurons for which SOD1 is the main protein component. Interestingly, they are usually TDP-43 negative. ALS linked *SOD1* mutations have been shown to cause oxidative stress in the ER, proteasome and mitochondria, which may affect protein folding and eventually cause cell death by initiating the UPR (**Atkin et al.**, 2008; **Turner and Atkin**, 2006).

RNA metabolism

RNA metabolism includes RNA transcription, processing, splicing and transport. Dysfunction of RNA metabolism was first proposed as a pathogenic mechanism in ALS after the discovery of TDP-43 as a major component of ubiquitinated inclusions in ALS patients. Since then, other ALS genes implicated in RNA metabolism including, *FUS* (**Vance et al.**, 2009), *C9orf72* (**DeJesus-Hernandez et al.**, 2011), *ANG* (**Greenway et al.**, 2006), *HNRNPA1* (**Kim et al.**, 2013) and *MATR3* (**Johnson et al.**, 2014) have also been identified.

TARDBP

The identification of TAR DNA-binding protein (TDP-43) as the main component of ubiquitinated protein aggregates in the affected neurons of the majority of ALS patients (**Neumann et al.**, 2006) greatly improved our understanding of ALS pathology. Soon after, genetic studies using linkage and direct sequencing found autosomal dominant missense mutations in the gene encoding TDP-43, *TARDBP*, from an Australian ALS family and two sporadic cases from the UK and Australia (**Sreedharan et al.**, 2008). Linkage to the *TARDBP* locus (1p36) was confirmed, and over 50 *TARDBP* mutations have since been reported in ALS (**Boylan**, 2015). This accounts for approximately 4% of FALS and 1% of SALS (**Boylan**, 2015; **Renton et al.**, 2014). TDP-43 is a ubiquitously expressed heterogeneous nuclear ribonucleoprotein (hnRNP), involved in RNA processing (**Wang et al.**, 2004). TDP-43 is important for transcriptional regulation, splicing regulation, micro-RNA processing and regulation of RNA localisation and translation (**Lagier-Tourenne et al.**, 2010).

FUS

In 2009, more evidence for the involvement of abnormal RNA metabolism in ALS was provided when two independent research groups used genome-wide linkage analysis to identify a total of 14 different mutations in *FUS* from 25 ALS families (**Kwiatkowski et al.**, 2009; **Vance et al.**, 2009). *FUS* mutations account for approximately 4% of FALS and 1% of SALS patients (**Boylan**, 2015; **Renton et al.**, 2014). Inheritance of *FUS*-linked ALS is mostly autosomal dominant. *FUS* encodes fused-in-sarcoma, a protein that is ubiquitously expressed and is functionally similar to TDP-43. ALS cases with *FUS* mutations have protein aggregates in the motor neurons that are positive for FUS, but negative for TDP-43. Similar to TDP-43, FUS is also a heterogeneous nuclear ribonucleoprotein (hnRNP) that plays a role in RNA metabolism including, regulation of transcription, splicing, processing of RNA and controls export to the cytoplasm (**Lagier-Tourenne et al.**, 2010).

C9orf72

In 2006, genome-wide linkage studies carried out on large ALS-FTD families implicated a locus on the short arm of chromosome 9 (**Morita et al.**, 2006). In 2011, the hexanucleotide (GGGGCC) repeat expansion in *C9orf72* was reported at this locus in ALS patients through haplotype analysis (**DeJesus-Hernandez et al.**, 2011) and targeted sequencing (**Renton et al.**, 2011). This pathogenic hexanucleotide repeat expansion is the most common genetic cause of ALS, and has an incidence of about 40% in FALS patients and 3-10% in SALS patients, depending on ethnic group (**McCann et al.**, 2017; **Renton et al.**, 2014). Although the exact function of the C9orf72 protein remains unclear, bioinformatics studies have shown it to have a strong structural homology to the DENN domain (differentially expressed in normal and neoplastic cells) (**Levine et al.**, 2013). C9orf72 protein likely plays a role in DNA and RNA processing pathways, such as the production of proteins from RNA, as well as RNA transport within the cell (**Renton et al.**, 2011), supporting the role of RNA metabolism dysfunction in ALS. More recently, *C9orf72* repeat expansions in ALS have also been linked to increased toxicity and DNA damage (**Farg et al.**, 2017).

UPS and autophagy

There are two major protein degradation systems in cells, the ubiquitin-proteosome system (UPS) and the autophagy system (**Ding et al.**, 2007). Both systems are essential for protein homeostasis by controlling protein quantity and quality. In the UPS system, proteins are first tagged with ubiquitin, then recognised by UPS chaperone proteins that mediate transport to the proteasome for destruction (**Ding et al.**, 2007;

Nandi et al., 2006). Autophagy is usually initiated during cellular stress, such as in the presence of protein aggregates. The autophagy pathway involves degradation and recycling of proteins and organelles by lysosomes that contain proteolytic enzymes (**Banerjee et al.**, 2010). The UPS and autophagy pathways have been widely implicated in neurodegenerative diseases, including ALS (**Banerjee et al.**, 2010). Known ALS genes that are involved in the UPS and/or autophagy include *UBQLN2* (**Deng et al.**, 2011), *CCNF* (**Williams et al.**, 2016), *VCP* (**Johnson et al.**, 2010), *OPTN* (**Maruyama et al.**, 2010) and *TBK1* (**Freischmidt et al.**, 2015).

UBQLN2

A mutation in *UBQLN2*, encoding the ubiquilin 2 protein, was first reported using dense linkage mapping in an ALS-FTD family (**Deng et al.**, 2011). *UBQLN2* mutations are a rare cause of X-linked dominant ALS and ALS-FTD, with reduced penetrance in females, and account for less than 1% of FALS cases. The ubiquilin 2 protein plays a major role in protein homeostasis through the UPS and autophagy. It mediates protein degradation by physically associating with ubiquitin ligases and proteasomes (**D'Angiolella et al.**, 2010). As a result, mutations in this gene prevent the degradation of ubiquitinated proteins and may partly explain the presence of protein aggregates in the motor neurons of ALS patients. Moreover, ubiquilin 2 pathology has also been found in ALS patients who do not carry mutations in *UBQLN2* gene (**Williams et al.**, 2012). Ubiquilin 2 is responsible for the regulation of proteasome autophagy and recycling may be an essential component of the pathways involved in the degeneration of motor neurons (**Renton et al.**, 2014).

CCNF

Recently, our laboratory used NGS and linkage analysis to identify mutations in *CCNF*, encoding for cyclin F, in ALS and FTD patients (**Williams et al.**, 2016). A pathogenic *CCNF* mutation was found in a large Australian ALS-FTD kindred, and additional mutations were identified from international FALS and FTD cohorts from collaborators (**Williams et al.**, 2016). *CCNF* mutations lead to aberrant misfolded proteins and aggregation of proteins including TDP-43 in the neuronal cells. Cyclin F is a member of the cyclin family, vital for regulation of cell cycle transitions. Cyclin F has an F-box motif that allows binding to Skp1 adaptor protein, which is part of the Skp1-Cul1-F-box (SCF) protein complex (**D'Angiolella et al.**, 2010). SCF is an E3-ubiquitin ligase complex that controls ubiquitination of proteins destined for destruction by the UPS (**D'Angiolella et al.**, 2010). Mutations in *CCNF* therefore further implicate UPS dysfunction and abnormal protein homeostasis in ALS pathogenesis.

In conclusion, abnormal RNA metabolism and protein homeostasis are recognised processes in ALS pathogenesis. Both pathways are interlinked and dysfunction in one pathway can affect the other (**Ling et al.**, 2013). *TARDBP*, *FUS*, *HNRNPA1* and *C9orf72* encode for RNA-binding proteins responsible for RNA processing and metabolism. *UBQLN2*, *VCP*, *OPTN* and *TBK1* are all involved in protein homeostasis. Dysfunction in RNA metabolism and protein homeostasis can lead to proteotoxic stress, loss of neuroprotection, protein misfolding and lack of autoregulation (**Ling et al.**, 2013). This then leads to deterioration and death of motor neurons, as typically seen in ALS patients. Other pathways that have been variably implicated in ALS include impaired axonal transport, neuroinflammation, apoptosis, mitochondrial dysfunction, and most recently, DNA damage (**Farg et al.**, 2017; **Ling et al.**, 2013; **Maurel et al.**, 2018).

1.3 Multi-disciplinary pipeline for ALS gene discovery

As described previously, the remaining ALS families with unknown genetic mutations are typically small, and NGS analysis of these families results in the identification of many candidate variants. Evidence for the pathogenic nature of these candidate variants must therefore be assessed rigorously using additional analyses such as a combination of *in silico*, *in vitro* and *in vivo* tools to prevent false-positive findings (**MacArthur et al.**, 2014). Therefore, our laboratory has established a genetic, *in silico* and *in vitro* gene discovery pipeline (Figure 1.2) to address this issue. This pipeline has been successfully applied to MQ1, a small Australian ALS family, leading to the identification of five novel candidate genes, of which a single candidate in *VPS29* ranked highly for both *in silico* and *in vitro* analyses, and is likely to be the causative gene mutation in the family.

The family-based gene discovery workflow is summarised in Figure 1.2. Following NGS and shared variant analysis, filtering steps were carried out to refine the variant list. The remaining candidate variants then underwent an *in silico* predictive analysis pipeline to assess the potential pathogenicity of each variant. *In silico* results were used to score and rank all candidate variants to identify the strongest candidates (most likely to be pathogenic) for *in vitro* functional analysis (Figure 1.2). The *in vitro* pipeline established by our lab aimed to identify specific pathological features of ALS, such as protein mislocalisation, insolubility, aggregation and co-localisation with TDP-43. In the future, genes that show strong pathogenic features using the *in vitro* pipeline can also undergo further *in vivo* mouse and zebrafish studies.

1.4 Generating ALS models using CRISPR-Cas9

To study disease-causing gene mutations, animal and cell models that accurately reflect those genetic changes are needed. Since motor neurons cannot be collected and cultured directly from ALS patients, accurate disease models are essential. This can be achieved using genome-editing tools, such as the clustered regularly interspaced short palindromic repeats (CRISPR) and Cas9 system. The Cas9 protein recognises CRISPR sequences in prokaryote DNA and plays a role in an adaptive immune response to bacteriophage invasion by cleavage of bacteriophage DNA (**Pourcel et al.**, 2005). CRISPR-Cas9 was first reported as a genome editing tool in 2013, (**Cong et al.**, 2013) and has successfully been used to develop animal and cell models of neurodegenerative diseases, including ALS (**Cho et al.**, 2013; **Hwang et al.**, 2013; **Wang et al.**, 2017). A point mutation in *FUS* was corrected in motor neurons differentiated from patient-derived induced pluripotent stem cells (iPSCs) using CRISPR-Cas9 technology (**Guo et al.**, 2017). This showed that correcting the *FUS* mutation restored axonal transport defects seen in the uncorrected patient-derived motor neurons (**Guo et al.**, 2017). CRISPR-Cas9 also has the benefit of being cheaper, more efficient, and easier to design than other genome-editing tools, and has the potential to be used to analyse novel ALS candidate gene variants. Moreover, a CRISPR model contains gene mutations expressed under endogenous conditions and avoids limitations often associated with overexpression models. For example, the greater abundance of the protein may be toxic or cells may gain functions they previously lacked, such as downregulation of DNA repair mechanisms, abnormal cell division, and compromised metabolism and growth **Krämer et al.** (2010).

1.5 Project aims

Gene mutations are the only proven cause of ALS, but about 40% of FALS cases still have an unknown gene mutation. Gene discovery in familial ALS is critical, as these genes are also often found to contribute to sporadic ALS, and have helped us understand ALS disease mechanisms and pathogenesis. Moreover, novel gene discovery can lead to improvements to diagnostic testing, including pre-implantation genetic diagnosis before pregnancy, which can benefit families with a history of ALS. The development of NGS has greatly improved our ability to discover new ALS gene mutations. However, bioinformatics analysis can result in numerous candidate gene variants, and additional *in silico* and *in vitro* pipelines are therefore necessary to rank these genes and assess potential pathogenicity in disease. Functional pipelines established by our lab will be used for this purpose (Figure 1.2). Additionally, novel ALS genes identified in other cohorts, such as those from recent publications or from collaborators, are strong candidate genes that can be screened through the Australian SALS cohort. CRISPR-Cas9 provides the potential to generate new cell lines containing candidate gene mutations expressed under endogenous conditions. These could act as a platform for further functional studies and validation of candidate genes, and form a part of our current gene discovery pipeline.

The goal of this project was to use two main approaches to identify gene mutations that cause ALS in familial and sporadic cases. The first approach seeked to find candidate gene mutations in an ALS family (FALS147), and to assess the potential pathogenic role of these candidate mutations using *in vitro* assays. WES data from family FALS147 has previously been analysed prior to this project. Recently available WGS data was analysed for candidate mutations as part of this project. The second approach seeked to screen strong candidate genes in the SALS cohort to identify known or novel mutations. Additionally, a CRISPR-Cas9 protocol was assessed and optimised to generate knockout cell models of *VPS29*, that could be part of future *in vitro* studies.

More specifically, the aims are listed as follows and are described in Figure 1.3.

1. To screen WGS data from Australian SALS patients for mutations in 18 new ALS genes that have recently been identified by collaborators
2. To perform genetic analysis on family FALS147 by identifying candidate gene mutations in WGS data.
3. To investigate the potential pathogenicity of family FALS147 candidate gene mutations using a functional analysis pipeline. This includes both *in silico* and *in vitro* strategies.
4. To assess a CRISPR-Cas9 protocol for generation of cell models of novel candidate variants.

Table 1.1: Summary of the major ALS genes.

Gene symbol	Name	Discovery method	Year	Protein function	Reference
SOD1	Superoxide dismutase 1	Linkage	1993	Oxidative stress	(**Rosen et al.**, 1993; **Siddique et al.**, 1991)
ANG	Angiogenin	Candidate gene	2006	DNA/RNA processing	(**Greenway et al.**, 2006)
TARDBP	TAR DNA binding protein	Linkage, candidate gene	2008	DNA/RNA processing	(**Kabashi et al.**, 2008; **Sreedharan et al.**, 2008)
FUS	Fused in sarcoma	Candidate gene	2009	DNA/RNA processing	(**Kwiatkowski et al.**, 2009; **Vance et al.**, 2009)
VCP	Valosin containing protein	Candidate gene	2010	Protein homeostasis	(**Johnson et al.**, 2010)
OPTN	Optineurin	Homozygosity mapping	2010	Protein homeostasis	(**Maruyama et al.**, 2010)
C9orf72	Chromosome 9 open reading frame 72	Linkage, GWAS	2011	DNA/RNA processing	(**DeJesus-Hernandez et al.**, 2011; **Morita et al.**, 2006)
UBQLN2	Ubiquilin 2	Family-based NGS	2011	Protein homeostasis	(**Deng et al.**, 2011)
PFN1	Profilin 1	Family-based NGS	2012	Cytoskeleton and cellular transport	(**Wu et al.**, 2012)
HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1	Linkage, candidate gene	2013	DNA/RNA processing	(**Kim et al.**, 2013)
MATR3	Matrin 3	Family-based NGS	2014	DNA/RNA processing	(**Johnson et al.**, 2014)
CHCHD10	Coiled-coil-helix-coiled-coil-helix domain containing 10	Family-based NGS	2014	Mitochondrial function	(**Bannwarth et al.**, 2014)
TBK1	TANK-binding kinase 1	NGS burden analysis	2015	Protein homeostasis	(**Cirulli et al.**, 2015; **Freischmidt et al.**, 2015)
CCNF	Cyclin F	Linkage, family-based NGS	2016	Cell cycle	(**Williams et al.**, 2016)

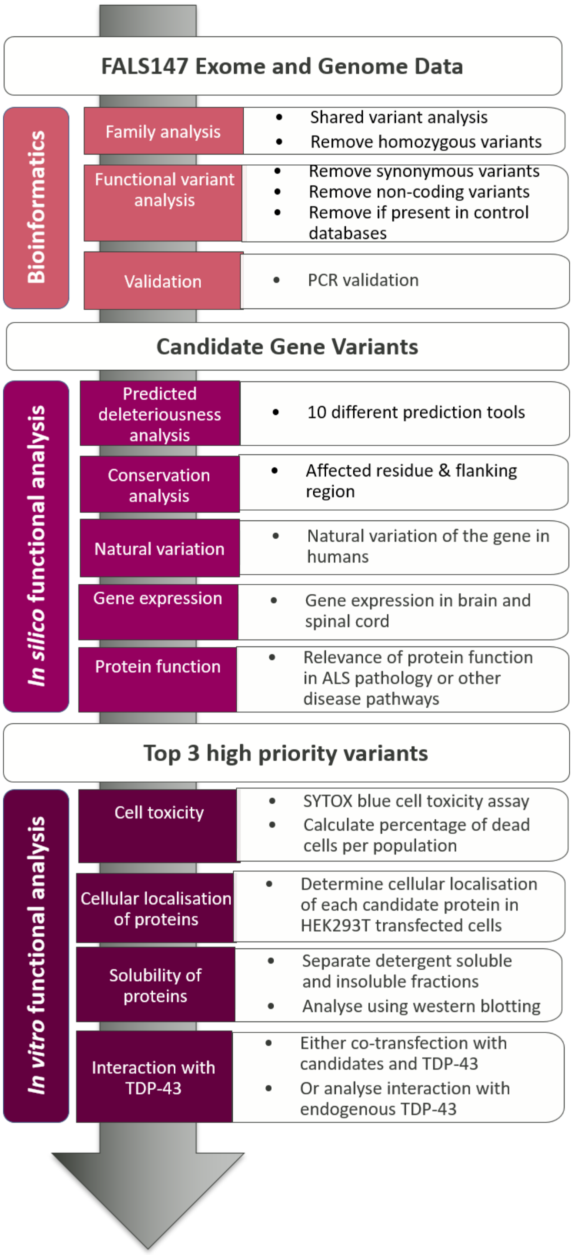

FALS147 Exome and Genome Data
Bioinformatics
Family analysis
Shared variant analysis
Remove homozygous variants
Functional variant analysis
Remove synonymous variants
Remove non-coding variants
Remove if present in control databases
Validation
PCR validation
Candidate Gene Variants
In silico functional analysis
Predicted deleteriousness analysis
10 different prediction tools
Conservation analysis
Affected residue & flanking region
Natural variation
Natural variation of the gene in humans
Gene expression
Gene expression in brain and spinal cord
Protein function
Relevance of protein function in ALS pathology or other disease pathways
Top 3 high priority variants
In vitro functional analysis
Cell toxicity
SYTOX blue cell toxicity assay
Calculate percentage of dead cells per population
Cellular localisation of proteins
Determine cellular localisation of each candidate protein in HEK293T transfected cells
Solubility of proteins
Separate detergent soluble and insoluble fractions
Analyse using western blotting
Interaction with TDP-43
Either co-transfection with candidates and TDP-43
Or analyse interaction with endogenous TDP-43

FIGURE 1.2: **Novel gene discovery pipeline applied to family FALS147.** The main steps of the pipeline included analysis and standard filtering of NGS data, *in silico* ranking, and *in vitro* functional studies of candidate genes.

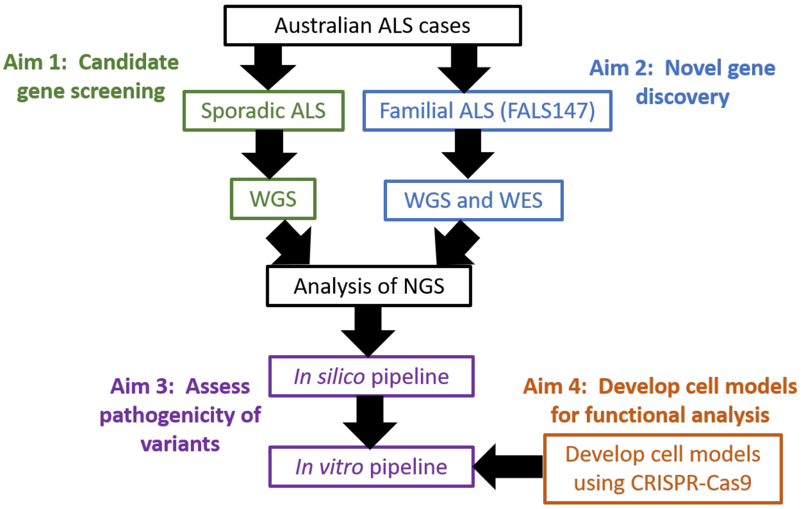

Australian ALS cases
Aim 1: Candidate gene screening
Sporadic ALS
Familial ALS (FALS147)
Aim 2: Novel gene discovery
WGS
WGS and WES
Analysis of NGS
In silico pipeline
Aim 3: Assess pathogenicity of variants
Aim 4: Develop cell models for functional analysis
In vitro pipeline
Develop cell models using CRISPR-Cas9

FIGURE 1.3: **Aims of this project.** Aim 1 used a candidate gene screening approach to screen for recently reported ALS genes in our sporadic ALS cohort. Aim 2 used a novel gene discovery to find new candidate gene mutations from WES and WGS data that contribute to ALS disease. Aim 3 assessed potential pathogenicity of variants using an *in silico* and *in vitro* pipeline. Aim 4 used CRISPR-Cas9 to generate new cell models for novel candidate variants, that could form part of our existing gene discovery pipeline.

2

Subjects and Materials

2.1 Subjects

ALS patients were recruited through the Macquarie Neurology Clinic, the Molecular Medicine Laboratory, Concord Hospital, and the MND DNA Bank of Australia. Patients and family members who visited the Macquarie Neurology Clinic were recruited by staff members of the Macquarie University Neurodegenerative Disease Biobank. All patients, were diagnosed with probable or definite ALS according to El Escorial criteria (**Brooks et al.**, 2000), and were of mostly European descent (Non-Finnish). Family FALS147 was recruited from the Molecular Medicine Laboratory (described in Section 4.1.1). Samples obtained from the Macquarie University Neurodegenerative Disease Biobank were extracted using the QIASymphony automated liquid handling robot and the DSP Midi extraction kid (Qiagen, USA), while samples obtained from the Molecular Medicine Laboratory and MND DNA bank were extracted manually using standard protocols.

Ethics and consent

All patients and control individuals provided informed written consent to be part of genetics research as approved by the Macquarie University Human Research Ethics Committee (HERC) (Approval number 5201600387), and the Sydney South West Area

Health Service (Approval number CH62/6/2011-123-G Nicholson HREC/11/CRGH/179, Title: Research study into identifying new gene mutations for motor neuron disease).

2.2 Datasets

850VCF dataset

Whole-genome sequencing (WGS) was completed on 850 ALS patients including 26 ALS families (23 SOD1 mutation positive and three mutation unknown), 628 SALS, 108 FTD, eight twin sets, and five sample duplicates. WGS data was generated using 30X TruSeq PCR-free v2.5 and HiSeq2000 NGS platform (Kinghorn Centre for Clinical Genomics, Sydney). The WGS data for these individuals was in the variant call format (VCF, named 850VCF), which contains variant specific information such as chromosomal location, reference and alternate alleles, functional categorisation, as well as data quality measures. Genotypes were denoted as 0/0 (wild type homozygous), 0/1 (heterozygous) and 1/1 (variant homozygous). A script written by Ingrid Tarr was used to subset the SALS patients from the complete 850VCF dataset for gene screening approach (Appendix A.2.1).

FALS147 dataset

Whole-exome sequencing (WES) data was also available from family FALS147, and was used for novel gene discovery in this project. WES data was generated using 100X TruSeq exome capture chemistry and HiSeq2000 NGS platform (Macrogen, Korea). This data was also in the form of a VCF file.

2.3 Materials

Main reagents, buffers and growth media

Table 2.1 provides a description of the main reagents, buffers and growth media used for cloning and tissue culture.

Cell culture media

HEK293T cell culture media
DMEM supplemented with 10% FBS

HEK293T cell culture media for CRISPR-Cas9 transfections
DMEM medium supplemented with 10% FBS and 100g/ml Penicillin-Streptomycin

SH-SY5Y cell culture media
DMEM/F12 medium, supplemented with 10%FBS and 100g/ml Penicillin-Streptomycin Solution

2.4 Constructs

pCMV6-Entry construct (OriGene, USA)

The pCMV6-Entry construct is a 4.9kb mammalian expression construct carrying sequences encoding a Myc-DDK tag at the C-terminal (Figure 2.1). The myc tag is a polypeptide protein tag derived from the c-myc gene product, with 54bp in size and 1202Da in molecular mass. A bacterial promoter (Neo^r/Kan^r) confers kanamycin resistance in *Escherichia coli (E. coli)* and expression of the gene insert is controlled by the mammalian CMV promoter located upstream of the Myc-DDK tag sequence.

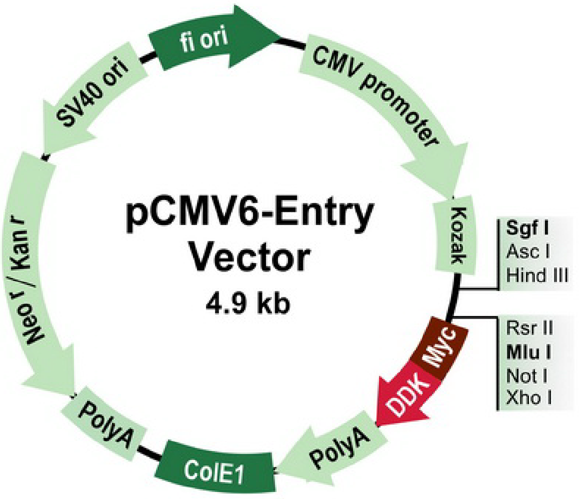


FIGURE 2.1: **Vector map of pCMV6-Entry construct** (https://www.origene.com/).

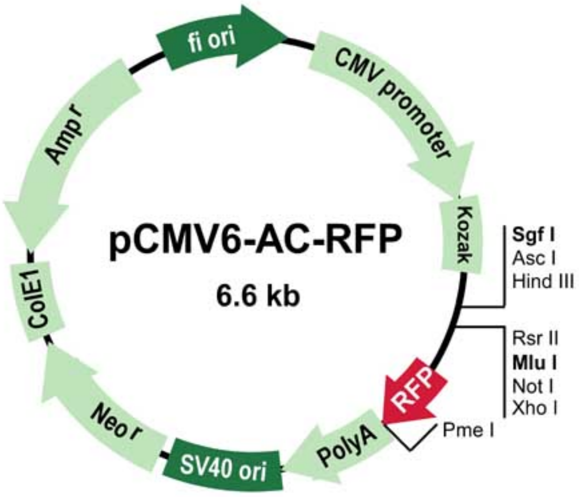


FIGURE 2.2: **Vector map of pCMV6-AC-RFP construct (https://www.origene.com/).**

pCMV6-AC-RFP construct (OriGene, USA)

The pCMV6-AC-RFP construct is a 6.6kb mammalian expression construct with a TurboRFP tag at its C-terminal (Figure 2.2). The RFP (red fluorescent protein) tag is 696bp in length, with a molecular mass of 27kDa and is a red (orange) fluorescent protein with excitation/ emission max = 553/574 nm. The multiple cloning site of the pCMV6-AC-RFP construct contains compatible restriction enzymes to pCMV6-Entry vector (Figure 2.1), allowing easy sub-cloning between the two constructs.

TABLE 2.1: **Materials and reagents used during this project**

Item	Description	Company
Cloning reagents		
T4 DNA ligase	T4 DNA ligase in 10mM Tris-HCI (pH 7.4 at 25°C), 50mM KCl, 1mM DTT, 0.1mM EDTA and 50% glycerol	Promega, USA
2X T4 DNA Ligation buffer	60mM Tris-HCI (pH 7.8); 20mM mgCl2; 20mM DTT; 2mM ATP; 10% polyethylene glycerol	Promega, USA
SOC medium spell out SOC	2% tryptone; 0.5% yeast extract; 10mM NaCl; 2.5mM KCl; 10mM MgCl2 ; 10mM MgSO4; 20mM glucose	Bioline, Australia
Luria Broth (LB) medium	10g/L Bacto-tryptone; 5g Bacto-yeast extract; 5g/L NaCL	Bacto Laboratories
LB agar	15g/L Davis Agar dissolved in LB	Astral Scientific Pty Ltd
Ampicillin	100mg/mL in water, filtered	Sigma Aldrich, USA
Kanamycin	100mg/mL in water, filtered	Sigma Aldrich, USA
PCR reagents		
MyTaq HS Red Mix	MyTaq HS DNA Polymerase and a novel buffer	Bioline, Australia
Exonuclease I (ExoI)	67 mM Glycine-KOH, 6.7 mM MgCl2, 10 mM B-ME	New England Biolabs, USA
Thermosensitive alkaline phosphatase (Tsap)	Tsap in 10mM Tris (pH 7.5 @ 25°C), 100mM NaCl, 50% glycerol.	Promega, USA
SYBR Safe	Cyanine dye	Life Technologies, USA
PCR primers	Oligonucleotides designed for each target region, diluted 100mM in water	Sigma Aldrich, USA
Growth media and buffers for tissue culture		
1X Gibco PBS (Phosphate Buffered Saline)	1.5mM KH2PO4; 155.2mM Cl; 2.7mM Na2HPO4-7H2O (pH 7.2)	Life Technologies, USA
Trypsin-EDTA	0.25% (2.5g porcine trypsin, 0.2g EDTA)	Sigma-Aldrich, USA
Dulbeccos Modified Eagles Medium (DMEM)	4500 mg/L glucose, sodium pyruvate and sodium bicarbonate	Sigma-Aldrich, USA
DMEM - Nutrient Mixture F-12 Ham (DMEM/F12)	15 mM HEPES, sodium bicarbonate, L-glutamine	Sigma-Aldrich, USA
Penicillin-Streptomycin Solution	10,000 units penicillin and 10 mg streptomycin/mL	Life Technologies, USA
Lipofectamine	Lipofectamine and PLUS reagent	Invitrogen, USA
Opti-MEM	HEPES, sodium bicarbonate, with hypoxanthine, thymidine, sodium pyruvate, L-glutamine, and growth factors	Life Technologies, USA

3

Methods

3.1 NGS analysis and bioinformatics

Overview

Novel gene discovery and candidate gene mutation screening approaches were used in this project to identify novel ALS candidate genes (Figure 3.1). Novel gene discovery was carried out to find novel candidate disease genes from family FALS147, while candidate gene screening was used to screen SALS patients for candidate mutations in recently published ALS genes, as well as genes implicated by collaborators or from analysis of family FALS147. Both approaches involved custom filtering steps using R studio, filtering from control databases, Sanger sequencing validation, and *in silico* functional analysis.

3.1.1 Novel disease gene discovery in FALS

Validation of exomic variants in whole-genome data

Analysis of whole-exome data from family FALS147 was carried out prior to this project, resulting in identification of 21 candidate gene variants. The 850VCF dataset containing genome data for family FALS147 was interrogated for previously identified

WES candidate gene variants to ensure that they were present in the genome data. UNIX scripts used are described in Appendix section A.1.

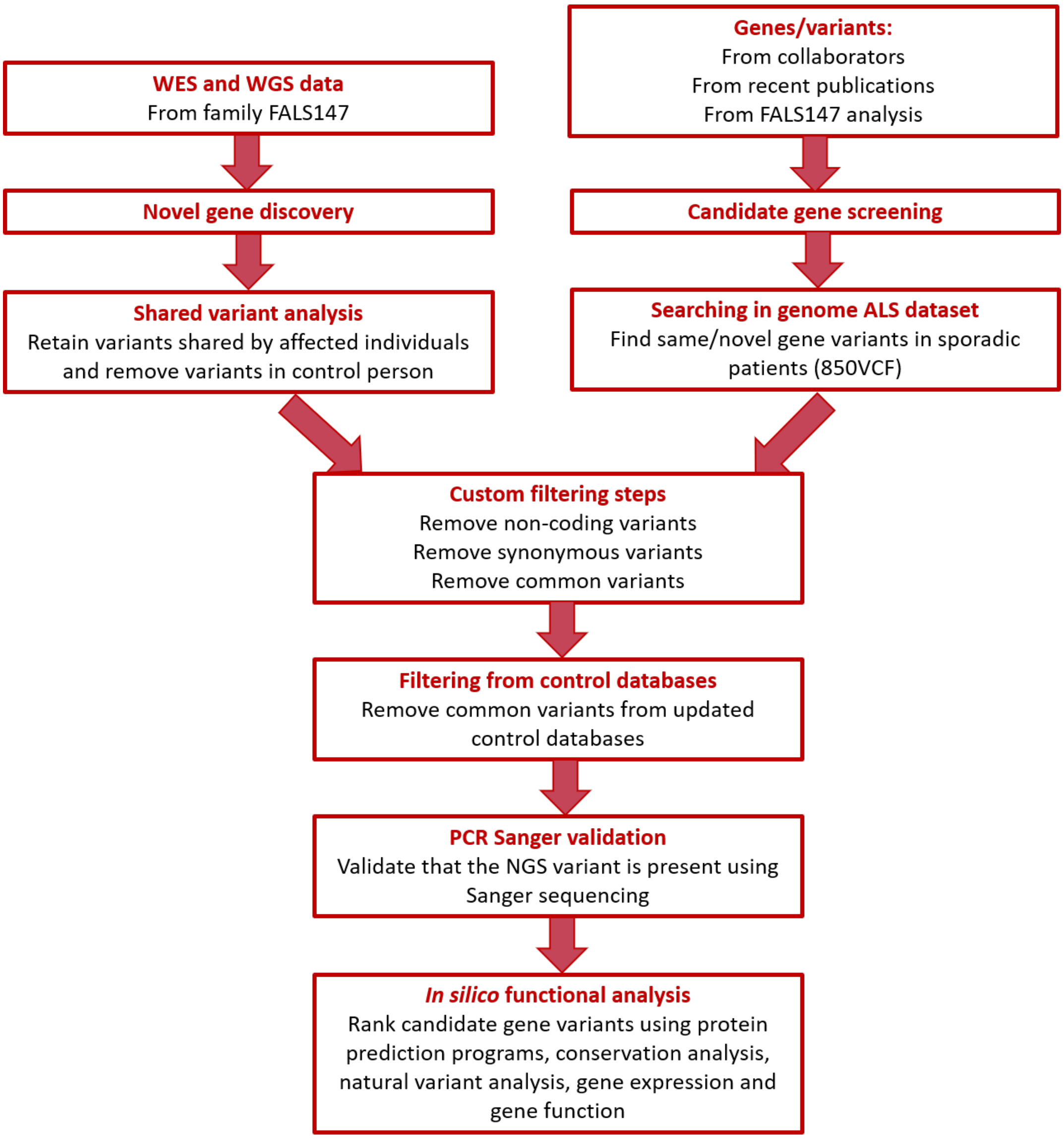

WES and WGS data
From family FALS147
Genes/variants:
From collaborators
From recent publications
From FALS147 analysis
Novel gene discovery
Candidate gene screening
Shared variant analysis
Retain variants shared by affected individuals and remove variants in control person
Searching in genome ALS dataset
Find same/novel gene variants in sporadic patients (850VCF)
Custom filtering steps
Remove non-coding variants
Remove synonymous variants
Remove common variants
Filtering from control databases
Remove common variants from updated control databases
PCR Sanger validation
Validate that the NGS variant is present using Sanger sequencing
In silico functional analysis
Rank candidate gene variants using protein prediction programs, conservation analysis, natural variant analysis, gene expression and gene function

Figure 3.1: **Novel gene discovery and candidate gene screening pipeline.** The pipeline shows the different steps involved in novel gene discovery (family FALS147) and candidate gene screening, including custom filtering steps, Sanger sequencing validation and *in silico* functional analysis.

Shared variant analysis

Custom bioinformatics scripts were used for shared variant analysis on family FALS147 WES and WGS data. Shared variant analysis involved identification of variants that were present in both affected individuals and not present in the control individual. The bioinformatics script used the UNIX **awk** command to find variants with a genotype call of 0/1 (variant present) in both affected siblings, and a call of 0/0 (variant not present) in the control individual (Appendix A.1.5, Table 3.1). As this family showed an autosomal dominant pattern of ALS inheritance, homozygous variants were removed from analysis. Shared variant analysis was followed by custom filtering to exclude non-coding variants and variants with allele frequencies >0.0001 from the public databases as described below.

TABLE 3.1: **Commands used in Unix and R**

Command	**Language**	**Function**
cd	Unix	Navigate to directory/file
X=file.name	Unix	Denote file
awk	Unix	Search
cat	Unix	Concatenate
>	Unix	Save as new file name
|	Unix	Pipe
setwd	R	Navigate to directory/file
X ← file.name	R	Denote file
grep	R	Search
filter	R	Filter data by specific charater string (word)
group by	R	Group similar data together
%>%	R	Pipe

3.1.2 Candidate gene screening

A total of 21 candidate ALS genes identified by our laboratory, collaborators, or recent publications, were screened in the SALS patient subset of the 850VCF (Section 2.2, Table 3.2). UNIX scripts using the **awk** command that were previously developed in our laboratory were adapted to identify all variants present in the full candidate gene sequences (Appendix A.1.1 and A.1.2) or specific candidate gene variants (Appendix A.1.3 and A.1.4) in the SALS dataset.

3.1.3 Custom variant filtering

A script was developed to further filter variants identified from both novel gene discovery and candidate gene screening in R studio (Appendix A.2.2). Firstly, the **filter**

Table 3.2: **Candidate genes screened in the SALS cohort**

Candidate genes	**Reason**
AARS2	White-matter disease gene
ABCD1	White-matter disease gene
ARPP21	ALS gene from collaborators
ARSA	White-matter disease gene
AUH	White-matter disease gene
CLN6	White-matter disease gene
COQ2	White-matter disease gene
CSF1R	White-matter disease gene
CYP27A1	White-matter disease gene
EIF2B5	White-matter disease gene
GBE1	White-matter disease gene
HTRA1	White-matter disease gene
KIF5A	ALS gene from collaborators
LAMA2	White-matter disease gene
LMNB1	White-matter disease gene
PLP1	White-matter disease gene
TREX1	White-matter disease gene
USP7	ALS gene from collaborators
CYB5R3	FALS147 candidate gene
DCAF7	FALS147 candidate gene
SAV1	FALS147 candidate gene

command was used for removal of variants that did not meet Filter=PASS criteria. Next, the **grep** command used to output exonic variants, and remove synonymous variants and variants with allele frequencies >0.0001 from public databases such as dbSNP (`www.ncbi.nlm.nih.gov/projects/SNP/`), 1000Genomes (`www.1000genomes.org/`) and the Exome Aggregation Consortium Database (ExAC, `http://exac.broadinstitute.org/`). Manual filtering was then carried out to remove variants that were found in web-based updated public control databases, including the Medical Genome Reference Bank (`https://sgc.garvan.org.au/initiatives/mgrb`), gnomAD (`http://gnomad.broadinstitute.org/`), and ProjectMinE (`http://databrowser.projectmine.com/`).

3.1.4 PCR validation of variants

All candidate gene variants obtained from bioinformatics analysis were validated in the relevant individuals. Variant validation included design and optimisation of primers, polymerase chain reaction (PCR), gel electrophoresis, PCR clean-up, Sanger sequencing and chromatogram analysis, as detailed below.

Primers design for candidate gene screening

Primers were designed to amplify the target variant base, and around 200bp of flanking sequences, using either Primer3Plus (`http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi`) or ExonPrimer (`https://ihg.helmholtz-muenchen.de/ihg/ExonPrimer.html`). M13F (5′ - TGTAAAACGACGGCCAGT - 3′) and M13R-pUC (5′ -CAGGAAACAG CTATGAC - 3′) universal primer sequences were also added to the primers for optimal sequencing. Primer sets can be seen in Table 3.3.

Polymerase Chain Reaction

Polymerase Chain Reaction (PCR) was used for variant validation, with reactions cycled on a Mastercycler Pro S (Eppendorf). Primers were first optimised as described in Table 3.4. A temperature gradient of 53.4°C, 55.7°C, 58.3°C, 61°C, 63.7°C, 66.1°C, 68°C, and 69.4°C was used on the thermocycling program to find optimal annealing temperature. Where optimal conditions could not be determined by this method, 10X PCR enhancer (Life Technologies) was added to the reaction. The optimal annealing temperature of each primer sets were determined from the intensity of the DNA bands on an agarose gel (below). Next, primer optimisation was carried out using PCR reactions shown in Table 3.4.

Agarose gel electrophoresis

Agarose gel electrophoresis was used to visualise amplified products. Gels were made using 1.5% w/v agarose (Bioline) in 1X Tris-Borate-EDTA (TBE, Life Technologies), with 1X SYBRSafe DNA gel stain (Life Technologies) to allow for visualisation of DNA. Gel lanes were loaded with 3μl of each reaction and ran at 100V for 30-45 minutes. EasyLadder I (Bioline) was used to determine product size and concentration. Gels were visualised using a Safe Imager 2.0 Blue Light Transilluminator (Life Technologies) and imaged using Gel Doc EZ Imager and Image Lab software (Bio-rad).

PCR cleanup

PCR products were cleaned before sequencing to remove excess dNTPs and primers. 0.2μl of both exonuclease I and thermosensitive alkaline phosphatase were added to each PCR product, and incubated at 37°C for 30mins before enzyme deactivation at 65°C for 15mins.

Sequencing analysis

Sanger sequencing of PCR products was carried out using the M13F or M13R-pUC universal primer and conducted using Big-Dye terminator sequencing on an ABI 3730XL sequencer (Macrogen). Candidate genes that yielded messy or inconclusive results using M13F primers were re-sequenced using M13R-pUC primer. Table 3.3 shows the sequencing primer used for each gene primer set. Results were analysed using Sequencher v5.1 software (Gene Codes). Chromatogram results were analysed by comparing them to reference gene sequences obtained from the UCSC Genome browser (`https://genome.ucsc.edu/`). Heterozygous variants were confirmed by observing double sequencing peaks and a reduced peak height compared to wild type alleles, indicating two different bases at that locus.

3.1.5 *In silico* functional analysis and variant prioritisation

In silico tools were used to predict the potential pathogenicity and functional consequences of variants identified from NGS and bioinformatics analysis. All novel putative mutations identified from candidate gene screening and family FALS147 novel gene discovery underwent *in silico* functional assessment by analysing the following: protein prediction and conservation, gene natural variation, tissue expression and gene function (Table 3.6). Family FALS147 WES candidate variants were previously analysed. As such, *in silico* scores were used to prioritise family FALS147 variants, giving priority to those predicted to be deleterious, have expression in brain/spinal cord, and relevant protein function.

Protein prediction programs

Eleven protein prediction programs (Mutation Assessor, Mutation Taster, Polyphen-2, Pon-P2, SIFT, FATHMM, PhD-SNP, PANTHER, SNPs&GO, PROVEAN, and CADD, Table 3.5) were used to predict the potential pathogenicity of variants.

Conservation analysis

Conservation of candidate protein sequences and residues was assessed by aligning (ClustalOmega) protein sequences from the orthologues of multiple species (HomoloGene, National Center for Biotechnology Information, NCBI). The number of identical residues (out of total species analysed) were considered, as well as 10 amino acid residues flanking the candidate residue (percentage of identical residues based on all proteins analysed). PhyloP and PhastCons were also used to provide a conservation score for each candidate residue and gene region respectively (**Pollard et al.**, 2010; **Siepel et al.**, 2005), determined by a positive PhyloP score and a score of 1 for PhastCons.

Natural variation analysis

Natural variation for candidate genes was determined using the residual variation intolerance score (RVIS) (`http://genic-intolerance.org/`) and the ExAC database (missense z score) (**Lek et al.**, 2016; **Petrovski et al.**, 2013). A positive RVIS score indicates more common functional variation than expected (tolerance to variation), while a negative score shows less variation than expected (intolerance to variation). RVIS also reports a percentile score for each gene, which reflects the percentile of most variation intolerant genes a given gene falls within. Z score reflects the deviation of observed counts from expected number, so a positive z score indicates low natural variation (intolerance to variation) and negative a z score indicates high natural variation (tolerance to variation).

Scoring of family FALS147 exome variants

Natural variation of family FALS147 exome variants were scored using the percentage of coding bases with non-synonymous and loss-of-function variants present in ExAC for each gene, and the average MAF (minor allele frequency) of all functional variants. Natural variation was ranked as low if the candidate gene had $<11.5\%$ of bases with a functional variant, average MAF was <0.001, and RVIS score was negative.

Scoring of family FALS147 genome and other candidate gene variants

Natural variation of family FALS147 genome variants, and other gene variants identified during this project, were scored using RVIS score and z score. Natural variation was ranked as low if the candidate gene had a negative RVIS score and positive s score.

Gene expression in neural tissues

Gene expression was assessed using the Human Brain Transcriptome (brain) and the GTEx project (spinal cord) databases (**Kang et al.**, 2011; **Lonsdale et al.**, 2013). Positive gene expression in the brain was recorded for genes with expression levels above 6 on the log-2 scale in the primary motor cortex (M1C) at >40 years of age. Gene expression in spinal cord was recorded in TPM (transcript per million), with TPM greater than 0.5 showing positive gene expression.

Gene function

Gene function was investigated using online databases (GeneCards, Uniprot), as well as interrogation of publications on each gene in relation to involvement in neurodegenerative diseases.

3.2 Generation of wild type and mutant expression constructs for *in vitro* analysis

Overview

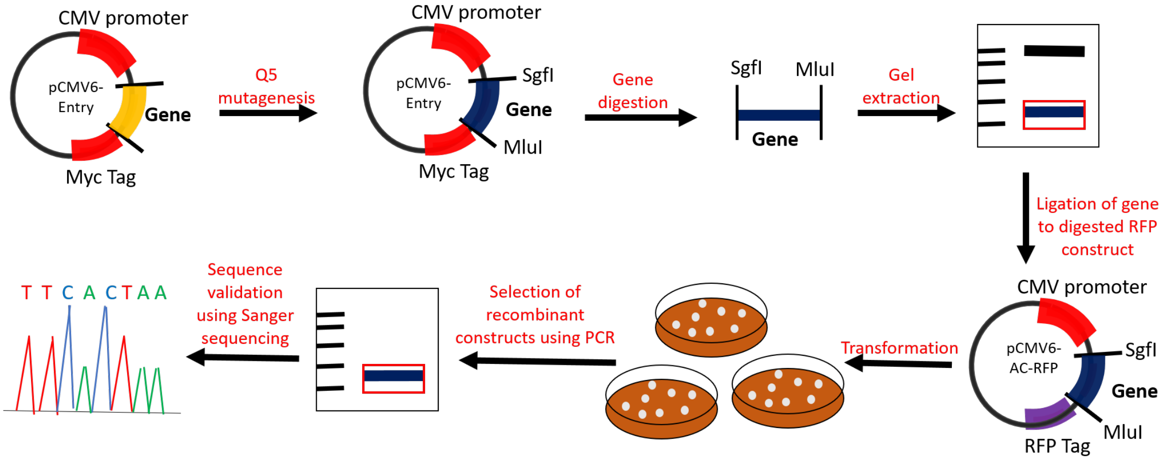


Figure 3.2: **Generation of pCMV6-AC-RFP expression constructs.** This figure summarises the steps taken for the successful generation of pCMV6-AC-RFP constructs expressing *CYB5R3, DCAF7* and *SAV1* wild type and mutant candidate genes. Q5 mutagenesis was first performed to introduce candidate gene variants in pCMV6-Entry construct. This was then followed by digestion of the genes from the pCMV6-entry construct using *SgfI* and *MluI* restriction enzymes, gel extraction of the gene product, ligation to pCMV6-AC-RFP construct, transformation, check-colony PCR, and Sanger sequencing.

Two sets of constructs were generated for *in vitro* expression of family FALS147 candidate genes (*CYB5R3*, *DCAF7* and *SAV1*). Wild type constructs (in pCMV6-Entry) were purchased from OriGene (Section 2.4), and candidate gene variants were introduced using the Q5 mutagenesis kit (New England Biolabs, NEB). Each wild type and variant gene was then subcloned into pCMV6-AC-RFP, also purchased from OriGene (Section 2.4). pCMV6-entry constructs were generated for western blot expression analysis and pCMV6-AC-RFP for cellular localisation and expression analysis as it contains a red fluorophore tag (RFP). The cloning protocol is detailed below and summarised in Figure 3.2.

3.2.1 Site-directed mutagenesis

The standard protocol from Q5 site-directed mutagenesis kit was used to introduce desired single nucleotide changes into the cDNA sequences of the respective expression constructs. The PCR reaction and thermocycler conditions used are described in Table 3.7. Mutagenesis primers were designed using the NEB online tool (`http://nebasechanger.neb.com/`). Primer sequences and optimal annealing temperatures are shown in Table 3.8. Kinase-Ligase-DpnI (KLD) reactions were then carried out using 1μl of the PCR product (Table 3.7 and incubated at room temperature for 5 minutes. The kinase phosphorylates the ends of the PCR products to facilitate ligation with the destination construct by the DNA ligase, and *DpnI* acts to cleave the methylated parental DNA template (wild type construct), keeping the newly synthesised un-methylated mutant strands intact.

3.2.2 Digestion

Constructs were digested with *AsisI* and *MluI* (NEB) in 50μl reactions consisting of 15μg of construct, 10X Cutsmart buffer (NEB), and 1μl of each restriction enzyme. Reactions were incubated for a minimum of 2 hours at 37°C, and heat-inactivated at 80°C for 15 minutes.

3.2.3 Gel extraction and purification

The digested pCMV6-entry gene products were then loaded on a 1% agarose gel (100V for 1hr) for electrophoresis. Products were visualized using a Safe Imager 2.0 Blue Light Transilluminator (Life Technologies) and Easy ladder I (Bioline) was used as a marker. Gel bands of correct sizes were extracted from the gel using a scalpel blade and the digested constructs were purified using Isolate II PCR and Gel kit (Bioline), as manufacturers instructions.

3.2.4 Ligation

For ligation of digest and purified gene inserts and empty pCMV6-RFP, a 10μl ligation reaction, including 1μl T4 DNA ligase and 1μl Ligase 10X buffer was set up for each using a 3:1 insert:construct ratio using the following formula:

$\frac{kb\ of\ insert}{kb\ of\ construct} \times$ ng of construct

Ligation reactions were incubated at room temperature for 3 hours.

3.2.5 Transformation

Ligation reactions were transformed into alpha-select Gold competent *Escherichia coli* cells (E.coli, Bioline) using the heat-shock method. Two microlitres of the ligation reaction were gently mixed with 25μl of thawed competent cells, then cooled on ice for 20 minutes. This was followed by heat shock at 42°C for 45 seconds and immediately put back into ice for another 2 minutes. One-hundred and twenty-five microlitres of SOC medium (Life technologies) was added to the cells and incubated at 37°C with shaking at 200rpm for 1 hour. The transformed cells were subsequently plated (50μl and 100μl) onto LB-agar plates supplemented with either ampicillin (pCMV6-AC-RFP) or kanamycin (pCMV6-entry) to allow for selection of colonies with recombinant constructs. The plates were inverted and incubated at 37°C overnight to allow for growth.

3.2.6 Selection of recombinant constructs

PCR was used to identify positive colonies carrying the target gene. Gene specific primers were designed using Primer3Plus (`http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi`, Table 3.9). Sixteen to thirty-two single colonies from each transformation were picked from the LB agar plates with a pipette tip, resuspended in 5μl MyTaq, 4.2μl water and 0.4μl of each reverse and forward primer for PCR amplification. PCR products were then ran on gel electrophoresis using a 1.5% agarose gel at 100V for 30 minutes and analysed to verify if products of the correct size have been obtained. If bands were not distinct enough, PCR was repeated using the universal T7 forward primer present in the 5$'$ region of construct and the gene specific reverse primer to confirm presence of insert.

3.2.7 Purification and storage of constructs

Following colony PCR validation, positive colonies containing the insert were picked from the LB-agar plate, and cultured in 5mL LB supplemented with 2.5μl kanamycin (100mg/ml) or 5μl ampicillin (100mg/ml) according to construct resistance (Section 2.4). These were incubated overnight at 37°C with 200rpm shaking. The next day, glycerol stocks were generated by mixing 250μl of 100% glycerol (Sigma-Aldrich) with 750μl culture, and stored at 80 °C freezer for long-term storage. The remaining construct DNA was purified using the Plasmid Miniprep Kit (Qiagen, Netherlands) as per standard protocol and eluted into 50μl of water or elution buffer. DNA concentrations were quantitated using a QIAxpert UV/VIS spectrophotometer (Qiagen).

3.2.8 Sequence validation of constructs

All purified constructs were verified by Sanger sequencing, using several primers to cover the whole gene and fluorophore sequence (Table 3.10). As a 3$'$ tag, all pCMV6-AC-RFP constructs were also verified to be in-frame from the end of each gene to the RFP fluorophore tag.

3.3 *In vitro* functional analysis of family FALS147 candidate genes

Overview

In vitro experiments were used to determine potential pathogenicity of family FALS147 candidate gene variants identified from NGS analysis using pCMV6-AC-RFP expression constructs. Experiments included a flow cytometer assay to assess cytotoxicity (SYTOX Blue assay) and confocal microscopy to determine cellular localisation of candidate proteins, their interaction with endogenous TDP-43, and the formation of protein aggregates. Another aim of this project was to create an ALS cell model using CRISPR-Cas9 genome editing. CRISPR-Cas9 was used to generate a gene knockout model of VPS29, a strong candidate gene previously identified in our laboratory.

3.3.1 Cell culture maintenance

Two cell lines were used in this study. The human neuroblastoma line, SH-SY5Y, was used to optimise transfection efficiency. The human kidney fibroblast cell line, HEK293T, was used to transfect candidate genes for the visualization of protein localisation. A flow cytometry based assay was used to assess cytotoxicity. The CRISPR-Cas9 method was also assessed and implemented.

Cells were grown in T75 flasks with appropriate culture media (refer to Section 2.3), until approximately 80% confluent before subculture. To passage cells, medium was removed and cells were washed with PBS. Cells were then detached with 1ml trypsin by incubating for about 2-3 minutes at 37°C and resuspended in 5ml media to neutralize trypsin. For SH-SY5Y cells, culture media and PBS wash were also collected to recover floating population. Cells were collected by centrifugation for 5 minutes at room temperature at 1,105 rcf. The supernatant was removed, and cell pellet was resuspended in 5mL culture medium. SH-SY5Y cells were split at a 1:5 ratio into a new T75 flask containing 10mL of pre-warmed culture media. HEK293T cells were usually split at a 1:20 ratio. CRISPR-transfected cells were regularly monitored and were transferred from 96-well plates to 12-well plates when the confluency reached 80-90%. Briefly, culture media was removed, and the cells were washed with PBS, trypsinised (20μl) and transferred to 12-well plates containing 2ml culture media. All cell lines were incubated at 37°C with 5% CO_2 in a HERACELL 150i CO_2 humidified incubator (Thermo Fisher).

3.3.2 Cell seeding for transfections

Before cell seeding, cells were trypsinised, centrifuged and collected as described above. Cell counts were done using the Scepter 2.0 Cell Counter (Millipore) and appropriate volumes of cells were seeded to obtain different densities based on experimental need. For transfection optimisation in SH-SY5Y, cells were plated at $1x10^5$ in 24-well plates with 1mL of medium, for a next day transfection. For visualization of gene expression and aggregate formation, HEK293T cells were seeded on 12mm glass cover slips in 24-well plates at a density of $5x10^4$ in 1ml medium. For cytotoxicity assay, HEK293T cells were seeded at $4x10^5$ in 6-well plates, with 2mL culture medium in each well. The same cell seeding conditions were used for CRISPR-Cas9 transfections. Cells were cultured for at least 24 hours in a humidified incubator at 37°C with 5% CO_2 before transfections.

3.3.3 Transfections

All constructs were diluted to 500ng/μl in TE buffer. Lipofectamine LTX with PLUS reagent and Opti-MEM media (Life Technologies) were used to transfect cells for *in* *vitro* analysis of candidate gene variants.

To determine optimal transfection efficiency in SH-SY5Y cells, several reagent volumes were used, including 500ng, 750ng or 1,000ng DNA, 0.5μl, 0.75μl or 1μl of Lipofectamine PLUS, and various volumes of Lipofectamine LTX reagent, as summarised in Figure 3.3. Lipofectamine LTX was diluted in Opti-MEM media (Life Technologies) and incubated for 5 minutes before use.

Transfection conditions for HEK293T cells were previously determined by our laboratory and are presented in Table 3.11. Cells were transfected in duplicate (microcopic visualisation) or triplicate (SYTOX blue assay) with pCMV6-AC-RFP wild type or variant candidate gene constructs (*SAV1*, *CYB5R3* and *DCAF7*). The DNA and Lipofectamine PLUS was mixed to 1X Opti-MEM and incubated for 5 minutes at room temperature. The diluted DNA was then combined with the Lipofectamine LTX/OPTI-MEM mix for each reaction and incubated at room temperature for 20 minutes. In the meantime, culture medium on cells was aspirated, and washed once with PBS. Culture media was then added (500μl for 24-well plates, or 2mL for 6-well plates). One hundred microlitres and 500μl of the DNA/Lipofectamine mix were added to each well of the 24-well and 6-well plates respectively in a dropwise fashion, and gently rocked for mixing. The cells were incubated at 37°C with 5% CO_2. After 24 hours, HEK293T media was added to the plates (500μl for 24-well plates and 1mL for 6-well plates). Cells were incubated for a minimum of 48 hours following transfections before ready for downstream experiments.

The CRISPR reagents (Sigma-Aldrich) used for transfection are summarized in Table 3.12 using a previously determined transfection ratio of 4:4:1 (20mM component: 20mM component: 5μlg component). CrRNA (VPS29-0-51-CCA), trRNA and Cas9 protein were gently mixed together and incubated on ice for 30 minutes. Trans-IT CRISPR transfection reagent (Sigma-Aldrich) was diluted 1/40 in Opti-MEM and added to the RNA/Cas9 mixture. This was followed by another 15 minutes incubation at room temperature. Finally, 300μl of Trans-IT/CRISPR complex was added to each well of a 6-well plate. The cells were incubated at 37°C with 5% CO_2 for at least 48 hours.

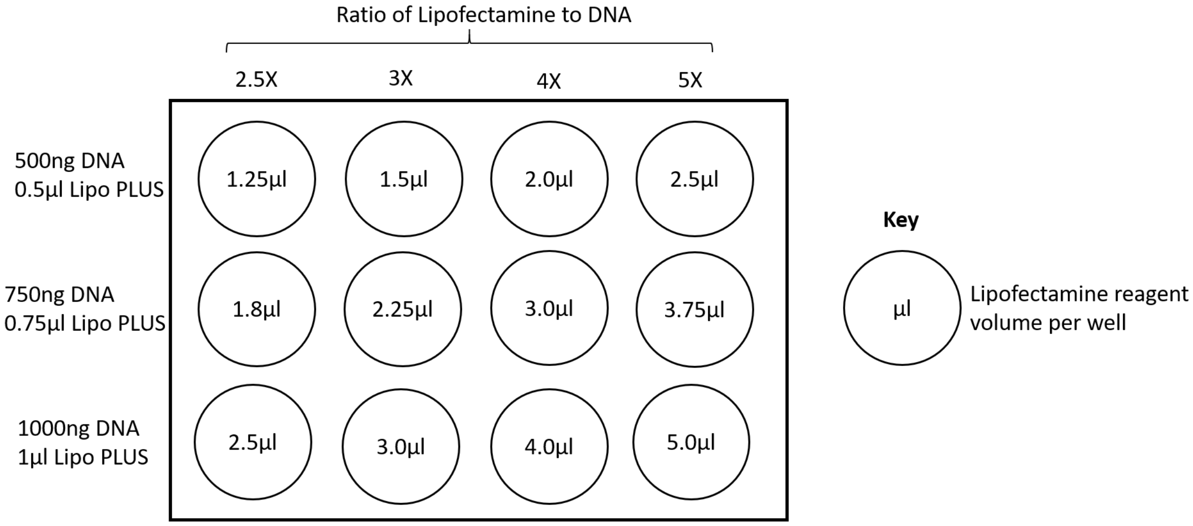


FIGURE 3.3: **Details of SH-SY5Y transfection optimisation**. To determine optimal transfection conditions, different ratios and volumes of Lipofectamine-LTX, Lipofectamine LTX PLUS reagent and construct DNA were used, as shown in this figure.

3.3.4 Endogenous TDP-43 immunofluorescent staining

To visualize protein expression and interaction with endogenous TDP-43 under the confocal microscope, immunofluorescent staining was performed. Transfected cells were fixed and permeabilized with 250μl 4% paraformaldehyde (PFA) and 0.2% Triton X-100 diluted in PBS (PBS), for 15 minutes at room temperature. Cells were then washed 3 X 5 minutes PBS washes. Non-specific antibody binding was blocked with 5% normal goat serum (NGS, Abacus ALS) in PBS (blocking media) for 30 minutes at room temperature. Cells were then incubated with 1/500 TDP-43 primary antibody (rabbit polyAB TDP43, Cosmo Bio, Catalogue no. TIP-PTD-P02) in blocking media overnight at 4°C. This was followed by 3 X 5 minutes PBS washes, and incubation with 1/250 goat anti-rabbit AlexaFluor 488 secondary antibody (Green fluorescence, Invitrogen) in blocking media for 1 hour at room temperature. After another set of 3 X 5 minutes PBS washes, the coverslips were mounted using ProLong Gold Antifade Mountant with the nuclear marker DAPI (Invitrogen) and cured overnight before imaging.

3.3.5 Confocal microscopy

Cells were visualised using a Zeiss LSM 880 Confocal microscope, with the ZEN Pro Imaging Software by Zeiss (Germany). Images were acquired with a 63X oil-immersion objective. Images were merged and transfected cells were counted using the cell counting tool in Fiji ImageJ software (`https://fiji.sc/`).

3.3.6 Flow cytometry

BD LSRFortessa X-20 flow cytometer (BD Biosciences) was used to investigate transfection efficiency in SH-SY5Y cells and to measure the percentage of dead cells during SYTOX blue assay analysis. Forward scatter and side scatter lights (FSC and SSC) are used to investigate the population of interest based on size and internal complexity. As shown in Figure 3.4, FSC measures light scattered less than 10° and SSC records light scattered at a 90° angle as a cell passes through the laser beam. FSC gives information about the size of the cell, while SSC is related to the internal cellular complexity, such as size of nucleus, cell membrane and granularity.

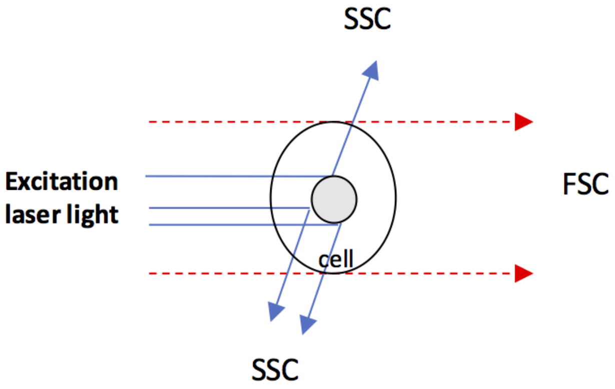


FIGURE 3.4: **Diagram showing forward scatter and side scatter lights for flow cytometery.** FSC measures light scattered less than 10° as a cell passes through the laser beam and gives information about cell size, while SSC records light scattered at a 90° angle as a cell goes through and shows internal cellular complexity, such as nucleus, cell membrane and granularity.

Flow cytometry to investigate transfection efficiency in SH-SY5Y and HEH293T cells

SH-SY5Y cells transfection optimisation

To investigate transfection efficiency in SH-SY5Y cells, cells were first transfected using different conditions (Section 3.3.3). Non-transfected cells were used as controls. Culture media and a PBS wash were collected before cells were detached with 300μl trypsin and incubated for about 2-3 minutes at 37°C. Trypsin was deactivated with 2ml media in each well, and the trypsin/media solution was transferred to the cell media/PBS solution before centrifugation at 1,105 rcf for 5 minutes. The supernatant was discarded and cells were resuspended in 1mL PBS and transferred to flow cytometry tubes. BD LSRFortessa X-20 flow cytometer was then used to measure cell transfection. Gating was carried out using non-transfected cells to exclude debris and unwanted particles (P1), while including transfected cells (P2) (Figure 3.5). To select P2 gating, a value of 0.5% was selected to account for background fluorescence. Analysis was carried out to filter transfected cells positive for RFP fluorescence (excitation/ emission = 553/574 nm) to determine transfection efficiency using different transfection conditions. Ten thousand cells were collected for each group, and number of transfected cells were counted (Q2 population). Data was analysed using BD FACSDiva software V8.0.1 (BD Biosciences).

Wild type and mutant gene transfections in HEK293T cells

Population size of wild type and mutant transfected cells were also compared to ensure that transfection efficiency did not differ between wild type and mutant constructs of each candidate gene. Flow cytometry was used to measure transfection efficiency for the different groups using the same protocol and gating parameters as for SH-SY5Y transfection efficiency analysis, above (Figure 3.5). However, to account for any differences in efficiency during SYTOX blue assay analysis, 10,000 transfected cells were included for each group.

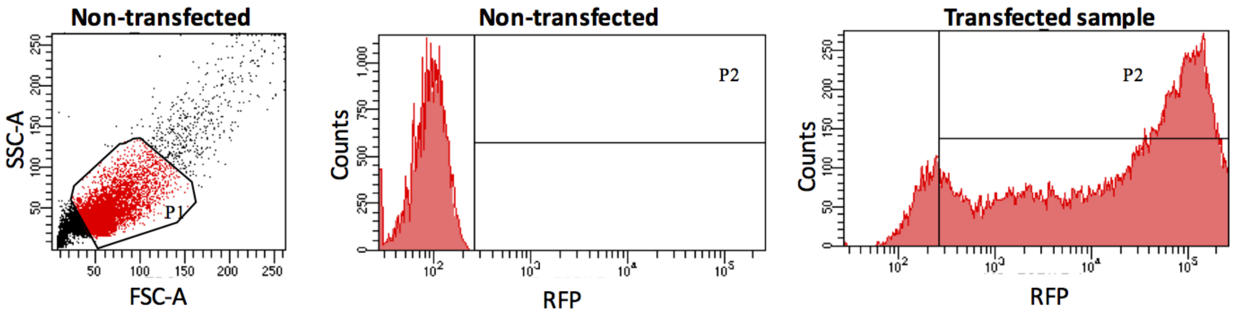


Figure 3.5: **Flow cytometry population gating to assess transfection efficiency**. To assess transfection efficiency, P1 was first determined using SSC and FSC from non-transfected sample to include eliminate debris. Next, P2 was selected to identify transfected cells (RFP fluorescence) using non-transfected cells (0.5% cells). The same P2 was used as a reference to find transfection efficiency for other transfected samples.

Flow cytometry assay to assess cytotoxicity

To measure cell death following candidate gene expression, HEK293T cells were first transfected with empty construct, wild type and mutant (*SAV1*, *CYB5R3* and *DCAF7*) pCMV6-AC-RFP constructs as above. Non-transfected cells and the cells transfected with the empty pCMV6-AC-RFP construct were used as controls. Cells were collected as described previously and transferred to flow cytometry tubes. 1X SYTOX Blue dead cell stain (Invitrogen) was added to each cell suspension and incubated in the dark for a minimum of 10 minutes. SYTOX Blue cell stain is a high-affinity nucleic acid stain that easily enters cells with damaged plasma membranes (dead cells), while leaving healthy cells unstained. SYTOX fluorescence analysis was performed using BD LSRFortessa X-20 flow cytometer. To identify cell populations, debris and unwanted particles were first excluded from the analysis by gating population of cells with similar FSC and SSC measurements (Figure 3.6, P1). Secondly, population gating is carried out to include transfected cells positive for RFP fluorescence (excitation/ emission = 553/ 574 nm) and SYTOX blue positive cells (excitation/emission 444/480 nm) (Figure 3.6, Q2 marked by the arrow). Ten thousand RFP positive cells were collected for each group and the number of transfected dead cells were counted (Q2 population). This percentage of transfected dead cells was then compared between wild type and mutant group.

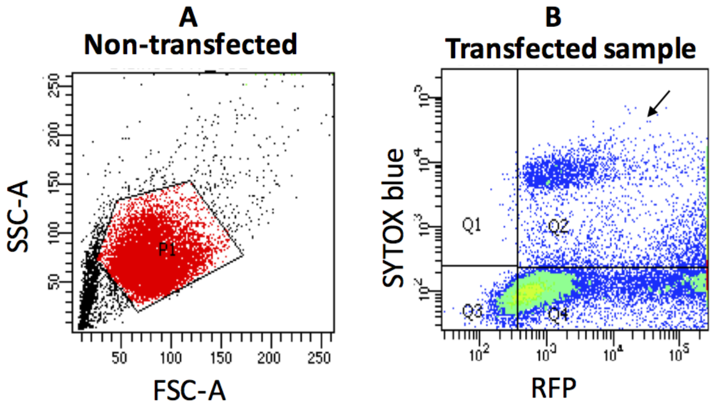


FIGURE 3.6: **Flow cytometry population gating for SYTOX blue toxicity assay.** SSC and FSC data from transfected and SYTOX treated cells were collected and analysed using the FACSDiva software. (A) First, population gating was performed using SSC and FSC to eliminate debris and unwanted particles from the analysis (P1). (B) P1 was subsequently gated to analyse cells that were positive for RFP and SYTOX blue fluorescence (Q2, shown by arrow). P2 was the population of interest and was counted to determine cytotoxicity differences between wild type and mutant group.

3.3.7 Statistical analyses

All statistical analyses required for this project were carried out using Prism v6 software (GraphPad). Statistical significance was calculated using Students *t test* to compare the difference between different groups. All values in the bar charts are shown as mean ± standard error of the mean (S.E.M) (*$p<0.05$; **$p<0.01$, *** $p<0.001$, ****$p<0.0001$).

3.4 Generation of a *VPS29* knockout cell line using CRISPR-Cas9

To generate a *VPS29* knockout cell line, cells were plated, transfected with CRISPR reagents, sorted with fluorescent-activated cell sorter (FACS), and checked for VPS29 protein expression using dot blotting technique. The steps are summarized in Figure 3.8. CRISPR-Cas9 system has three main components, Cas9 protein, CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA), all purchased from Sigma Aldrich (Figure 3.7). CrRNA and trRNA together form the guideRNA, and combine with Cas9 to target and cut DNA specific to the crRNA. There was an Atto488 fluorescent tag (excitation 488nm, green) conjugated to the tracrRNA, to allow for flow sorting and visualization of transfected cells. All three components were transfected into HEK293T cells using Trans-IT CRISPR transfection reagent as described in Section 3.3.3.

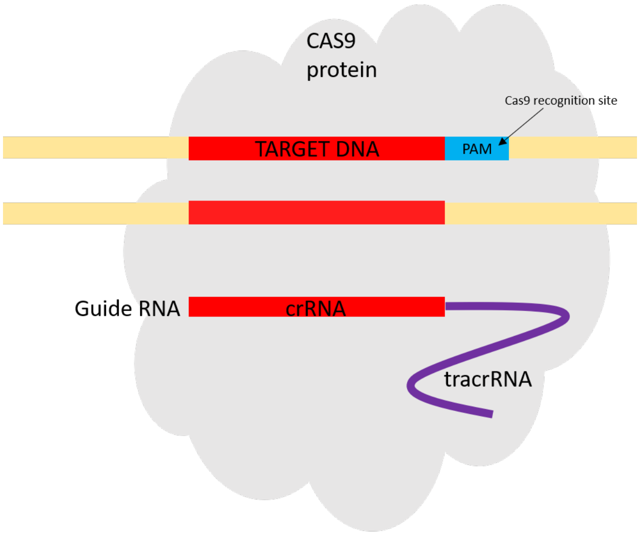


Figure 3.7: **CRISPR-Cas9 components.** The three main components are the Cas9 protein, CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA). crRNA and tracrRNA both form part of the guide RNA, where the crRNA recognises and binds to target RNA, which has a complementary sequence. Cas9 protein targets the PAM sequence and will only cut DNA at this location.

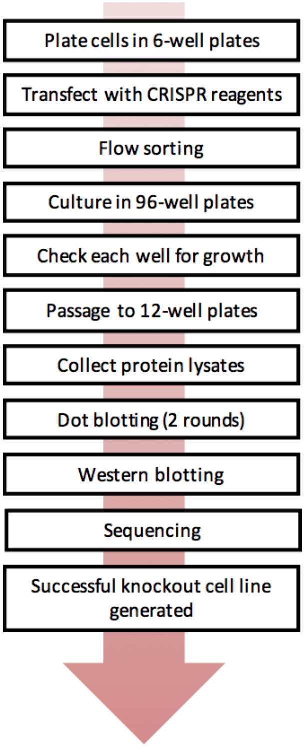


Plate cells in 6-well plates
Transfect with CRISPR reagents
Flow sorting
Culture in 96-well plates
Check each well for growth
Passage to 12-well plates
Collect protein lysates
Dot blotting (2 rounds)
Western blotting
Sequencing
Successful knockout cell line generated

FIGURE 3.8: **Pipeline for the generation of a CRISPR-Cas9 *VPS29* knockout cell line.** Following flow sorting, only colonies that were healthy and originated from a single cell were maintained. Dot blotting and western blotting was used to discard cells that still showed VPS29 expression.

3.4.1 Fluorescence-activated cell sorting

FACSMelody (BD Biosciences) was used to sort CRISPR-Cas9 transfected cells based on the Atto488 fluorescent tag (excitation 488nm) conjugated to the tracrRNA. Prior to sorting, transfected HEK293T cells were washed in PBS, trypsinised, and centrifuged at 1,105 rcf for 5 minutes. The cells were washed with 5 ml PBS twice by resuspension and centrifugation at 1,105 rcf. The pellet was then resuspended in 1mL PBS. Seventy microlitres of PBS was pre-added to each well of a 96-well plates for collection. Single cell sorting was performed using FACSMelody and cells that were positive for Atto488 fluorescence were collected in two 96-well plates containing one transfected cell per well, while cells testing negative were discarded (Figure 3.9).

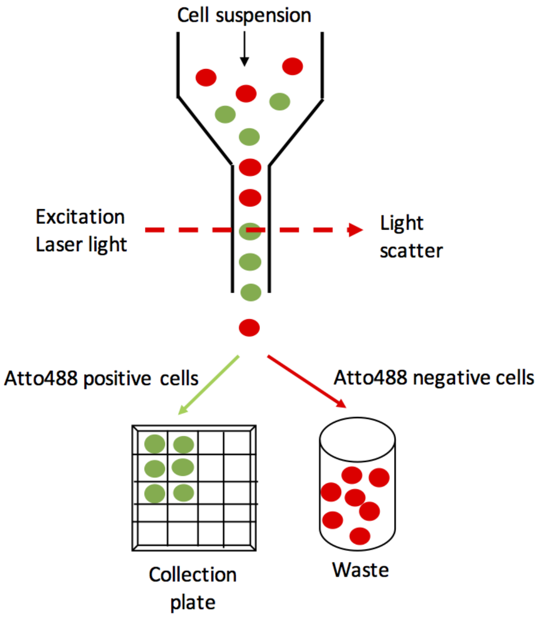


FIGURE 3.9: **Fluorescence-activated cell sorting using the Atto488 fluorescent tag.** Cells passing through the laser beam were monitored, and sorted based on the Atto488 fluorescent tag. Atto488 positive cells were sorted into a 96-well collection plate, while Atto488 negative cells were discarded.

3.4.2 Cell lysis

Atto488 positive cells passaged to 12-well plates (section 3.3.1) were lysed for dot blotting to determine VPS29 protein expression. Media was first removed, and cells were washed with PBS, trypsinised (200μl), and 800μl of culture media was added to each well. New 12-well plates were prepared with 2mL media, and 1/10 (100μl) of the trypsinised cell suspension was added to the new plate to maintain the colony, while the remaining 9/10 (900μl) was placed in 1.5ml Eppendorf tubes for cell lysis. The tubes were spun at 1,000 rpm for 5mins to remove the supernatant, before resuspending the pellet in 1mL PBS. Samples were spun again to remove the PBS and kept on ice to slow protein degradation. One hundred and fifty microlitres of RIPA buffer containing 10X phosphatase inhibitor (Sigma Aldrich) and 25X protein inhibitor (Sigma Aldrich) was added to lyse the cells. Samples were rotated on a rotating wheel in a cold room for 30 minutes before proceeding to dot blotting.

3.4.3 Expression analysis by dot blotting

Two microlitres of each lysed sample was slowly pipetted onto a nitrocellulose membrane. The membrane was then allowed to air dry for about 30 minutes. Non-specific sites were blocked with Odyssey blocking buffer (Li-Cor Biosciences) for about 1 hour at room temperature on a rocking platform. Blocking solution was removed and kept for future use, while the membrane was incubated with 1/500 dilution primary anti-goat polyclonal VPS29 antibody (Abcam, catalogue no. ab51972) with agitation for 1 hour, followed by 3 X 5 minutes washes with tris buffered saline with tween-20 (TBS-T). Secondary antibody incubation (Donkey anti-goat lgG, 1/20,000, Li- Cor Biosciences) was also carried out for 1 hour, followed by another 3 X 5 minutes washes with TBS-T, 1 X 5 minutes with TBS, and 1 X 5 minutes with water. The membrane was lightly dried before imaging with automatic settings provided by Odyssey CLx imaging system (Li- Cor Biosciences) and analysed with Image Studio software. The colonies with visibly reduced expression were maintained, while the rest was discarded.

TABLE 3.3: **Primer sequences, PCR product size, optimal PCR reaction conditions, annealing temperatures and sequencing primers for gene primer sets used for candidate gene screening**

Primer name	Sequence 5′ to 3′(with M13F or M13R-pUC)	PCR product size	Optimised conditions	T^A	Sequencing primer
ABCD1_Ex1 M13_F	TGTAAAACGACGGCCAGTTTCAACTGCTGCCCCAGG	580bp	MyTaq	66.1	M13F
ABCD1_Ex1 M13_R	CAGGAAACAGCTATGACAGTAGAGGCGGTAGGCGTG				
AUH_Ex10 M13_F	TGTAAAACGACGGCCAGTTGGAGTAAGAAGGATGAGTTTGGTTAC	406bp	MyTaq	63.7	M13R pUC
AUH_Ex10 M13_R	CAGGAAACAGCTATGACTGAAGTACACGGATGGGTCC				
C2CD4A_Ex2 M13_F	TGTAAAACGACGGCCAGTGAGCGGAGACTGGCTTCTC	181 bp	MyTaq	68	M13R pUC
C2CD4A_Ex2 M13_R	CAGGAAACAGCTATGACCTCGTCCATCCCTGCTTCT				
CYB5R3_Ex1 M13_F	TGTAAAACGACGGCCAGTGGTTAGCTAGGGAAGGGCTC	257bp	MyTaq	68	M13F
CYB5R3_Ex1 M13_R	CAGGAAACAGCTATGACACCCCTCCTCAAATACCCAC				
CYB5R3_Ex4 M13_F	TGTAAAACGACGGCCAGTAGTGGGTTGACAAGACCCAG	247bp	MyTaq	63.7	M13F
CYB5R3_Ex4 M13_R	CAGGAAACAGCTATGACTCCACATGGGCTGTTGC				
DCAF7_Ex4 M13_F	TGTAAAACGACGGCCAGTAGCTTGCATGAGAACCTTGG	245bp	MyTaq	63.7	M13F
DCAF7_Ex4 M13_R	CAGGAAACAGCTATGACCAAGCTGACCTCTTCCTGTG				
GBE1_Ex8 M13_F	TGAGTTTCATCAACCAACTGAG	513bp	MyTaq	63.7	M13R pUC
GBE1_Ex8 M13_R	CACGAATTATAGCATATTCAGCAG				
KIF5A_Ex18_19 M13_F	TGTAAAACGACGGCCAGTTGGAGTTCTGGTCACAACGTG	424bp	MyTaq	63.7	M13R pUC
KIF5A_Ex18_19 M13_R	CAGGAAACAGCTATGACTACAGCAAGGTGACAGGACC				
LAMA2_Ex2 M13_F	TGTAAAACGACGGCCAGTTTTGGGTTACTTTAATGCTCCG	359bp	MyTaq	63.7	M13F
LAMA2_Ex2 M13_R	CAGGAAACAGCTATGACTTAATGCAACTTGGCCTCAAC				
NUTM2A_Ex2 M13_F	TGTAAAACGACGGCCAGTCTGTCTGTGTTCACGGCTCT	206 bp	MyTaq	68	Messy
NUTM2A_Ex2 M13_R	CAGGAAACAGCTATGACCCCACCTCTGTCCTCATCTG				
NUTM2A_NEW M13_F	TGTAAAACGACGGCCAGCTCTCCGTGGAGGAACAGAG	359bp	MyTaq	63.7	M13F
NUTM2A_NEW M13_R	CAGGAAACAGCTATGACAGCCATCCTGTTCTGTCACC				
TERF2IP_Ex1 M13_F	TGTAAAACGACGGCCAGTGGAGCTGGAGGCCTATCG	176 bp	MyTaq+enhancer	63.7	M13R pUC
TERF2IP_Ex1 M13_R	CAGGAAACAGCTATGACGGGCATTTTCCTTCACGTAG				
USP7_Ex21 M13_F	TGTAAAACGACGGCCAGTTTCTGGCCATATTAAACGTTAGAC	256bp	MyTaq	61	M13R pUC
USP7_Ex21 M13_R	CAGGAAACAGCTATGACGCTCGGGATCTAGGTACAAATG				
ZNF580_Ex2 M13_F	TGTAAAACGACGGCCAGTGTGCAGCTGGAGGAGGAG	381bp	MyTaq+enhancer	66.1	M13F
ZNF580_Ex2 M13_R	CAGGAAACAGCTATGACTGTGTGGGTCAGACTTGGTG				

TABLE 3.4: **PCR conditions for primer optimisation and variant validation**

Reagent	**PCR optimisation**	**PCR validation**
Water	3.7	7.4
MyTaq	5	10
Primer Forward	0.4	0.8
Primer Reverse	0.4	0.8
DNA	0.5	1
Total volumes	10 (μl)	20 (μl)

Temperature (°C)	**Time**	**Cycles**
95	3 min	1 X
95	15 sec	35 X
50-70	15 sec	35 X
72	15 sec	35 X
72	5 min	1 X

TABLE 3.5: **Details of the *In silico* pipeline**

Analysis	**Tool**	**Description**	**Scores**	**Reference**
Protein prediction	Mutation Assessor	sequence conservation	high, medium, low ,neutral	(**Reva et al.**, 2011)
Protein prediction	Mutation Taster	sequence conservation	disease causing, polymorphism	(**Schwarz et al.**, 2010)
Protein prediction	Polyphen-2	homology information plus structural and functional annotation of proteins	probably damaging, possibly damaging, benign	(**Adzhubei et al.**, 2010)
Protein prediction	Pon-P2	amino acid features, conservation, functinoal annotations	pathogenic, neutral, unknown tolerance	(**Niroula et al.**, 2015)
Protein prediction	SIFT	sequence conservation	damaging, tolerated	(**Ng and Henikoff**, 2003)
Protein prediction	FATHMM	function prediction	deleterious, tolerated	(**Shihab et al.**, 2013)
Protein prediction	PhD-SNP	homology information plus structural and functional annotation of proteins	disease-related, neutral polymorphism	(**Capriotti et al.**, 2006)
Protein prediction	PANTHER	sequence conservation	disease, unclassified, neutral	(**Thomas and Kejariwal**, 2004)
Protein prediction	SNPs&GO	homology information plus structural and functional annotation of proteins	disease, neutral	(**Calabrese et al.**, 2009)
Protein prediction	PROVEAN	function prediction	deleterious, neutral	(**Choi and Chan**, 2015)
Protein prediction	CADD	score that intergrates multiple annotations into one metric	ranking scores (>20 is high)	(**Kircher et al.**, 2014)
Conservation	PhyloP	p-values based on an alignment and model of neutral evolution	positive (conserved), negative (accelerated)	(**Pollard et al.**, 2010)
Conservation	PhastCons	probability that a nucleotide is from a conserved element	0-1, 1(conserved)	(**Siepel et al.**, 2005)
Conservation	NCBI homologene	contains protein sequence alignments for different species	1A, 1B, 2, 3, 4	(**Pruitt et al.**, 2005)
Natural variation	z score	calculated using the deviation of observed counts from expected number	positive (low variation), negative (high variation)	(**Lek et al.**, 2016)
Natural variation	RVIS score	rank genes in terms of deviation of functional genetic variation relative to expectation	positive (high variation), negative (low variation)	(**Petrovski et al.**, 2013)
Gene expression	HBT	contains transcriptome data from human brain	log-2 signal intensity score >6 (expressed)	(**Kang et al.**, 2011)
Gene expression	GTEx project	contains details about human gene expression and regulation	reads per trascripts per million (TPM) values >0.5 (expressed)	(**Lonsdale et al.**, 2013)

Table 3.6: **Details of the *in silico* scoring system**

***In silico* analysis Score**	**Description**	
Predicted deleteriousness	Fraction	shows the number of prediction tools that determined the candidate to be deleterious
Conserved residue	1A	candidate residue is conserved in all species and n=9 or more
	1B	candidate residue is conserved in all species and n <9
	2	candidate residue is not conserved in 2 or less species
	3	candidate residue is not conserved in 3 or more species
	4	candidate residue is not conserved in 3 or more species, and the substituted residue is present in a species
Conserved flanking region	Percentage	percentage of conservation across species of 10 residues flanking the candidate residue
PhyloP and PhastCons	2	conserved PhyloP score only
	1	conserved PhastCons score only
	0	neither PhyloP or PhastCons have a conserved score
Natural variation for family FALS147 exome candidate genes	Low	candidate gene has <11.5% of bases with a functional variant, the average MAF is <0.001, and the gene has a negative RVIS score
	Medium	candidate gene has one or two of the above conditions
	High	candidate gene has >11.5% of bases with a functional variant, the average MAF is >0.001, and the gene has a positive RVIS score
Natural variation for other candidate gene screening	Low	candidate has RVIS negative and z score is positive
	Medium	candidate has either RVIS negative OR z score positive
	High	candidate has RVIS positive and z score negative
Expression in brain and spinal cord	No	not expressed in HBT (<6) and GTEx Project (<0.5)
	Low	low expression in either HBT (6-8) and GTEx Project (0.5-10)
	Yes	medium to high expression in both HBT (8+) and GTEx Project (11+)

Table 3.7: **Q5 site-directed mutagenesis PCR conditions**

Reagent	**Volume(μl)**
Q5 2X master mix	12.5
10mM F primer	1.25
10mM R primer	1.25
10ng plasmid template	1
dH_2O	9
Total volume	25μl

Temperature	**Time**	**Cycles**
98°C	30 sec	1×
98°C	10 sec	25×
T^A	30 sec	25×
72°C	3 mins	25×
72 °C	2 mins	1×

KLD reaction	**Volume(μl)**
PCR product	1
2 X KLD reaction buffer	5
10X KLD enzyme mix	1
Nuclease free water	3
Total volume	10μl

Table 3.8: **Q5 site-directed mutagenesis primer information**

Primer name	**Sequence 5′ to 3′**	**T^A**
Q5_CYB5R3_p.H95Y_F	CAAGGACACCtATCCCAAGTTTC	60°C
Q5_CYB5R3_p.H95Y_R	AAGTAAACCTTGATGACCAG	
Q5_DCAF7_p.V252I_F	GATTCTAGATaTCCGGGTTCC	60°C
Q5_DCAF7_p.V252I_R	ACCACCTCCATTCCATCC	
Q5_SAV1_p.D219E_F	ATTATATAGAaCATAACACAAATACAAC	55°C
Q5_SAV1_p.D219E_R	ATTTTCTCCCTCTCATTG	

Table 3.9: **Primers used for selection of recombinant constructs**

Oligonucleotide name	Sequence 5′ to 3′	Product size
SAV1_F_primer	ACGAGCCCCTGTGAAATATG	186
SAV1_R_primer	CACTGCTGTCTCTGCTTTCG	
CYB5R3_F_primer	CTACCTCTCGGCTCGAATTG	179
CYB5R3_R_primer	TGTCTCCAATCTGCATGCTC	
T7_F_primer	AATACGACTCACTATAG	348
CYB5R3_R_primer	TGTCTCCAATCTGCATGCTC	
DCAF7_F_primer	CACCTTTGACCACCCATACC	180
DCAF7_R_primer	AGGGGAGCACAGAAATCAGA	
T7_F_primer	AATACGACTCACTATAG	571
DCAF7_R_primer	AGGGGAGCACAGAAATCAGA	

Table 3.10: Primers used to validate constructs using Sanger sequencing

Oligo name	**Sequence 5′ to 3′**	**Gene**
T7 Primer	AATACGACTCACTATAG	CYB5R3
CYB5R3_cDNA_568bp_part_F	AGAAGGACATCCTGCTGCGA	
Internal_pCMV6_entry_R	TCTGTTCAGGAAACAGCTATGA	
T7 Primer	AATACGACTCACTATAG	SAV1
SAV1_cDNA_1127bp_F	TGCCCAACAACATGGAAAAAA	
SAV1_cDNA_470bp_part_F	ATATTATGAATACAACCATGAT	
DCAF7_cDNA_571bp_part_F	ACATGTTTGCCTCTGTGGGTGCT	DCAF7
DCAF7_cDNA_1000bp_F	TACAACAACTGCCTGGAGATA	
DCAF7_cDNA_793bp R	TTGACACATGCTCGATGG	

Table 3.11: **Transfections reagents and volumes for 24 and 12-well plates**

DNA mix	**24-well plate (μl)**	**12-well plate (μl)**
Construct DNA (500ng/μl)	2	10
Lipofectamine Plus reagent	1	5
Opti-MEM	47	235
Total volumes	50	250

Lipofectamine mix	**24-well plate (μl)**	**12-well plate (μl)**
Lipofectamine LTX	3	15
Opti-MEM	47	235
Total volumes	250	250

Table 3.12: **Reagents for CRISPR-Cas9 transfection**

Reagents	**Volume per well (6-well plate)**
crRNA - 20mM	7.2
tracRNA - 20mM	7.2
Cas9 protein - 0.5ug	1.8
Opti-MEM	275
Trans-IT-CRISPR	6.9
Total volume	298.1

4

Results

4.1 NGS analysis and bioinformatics

4.1.1 Family FALS147 information

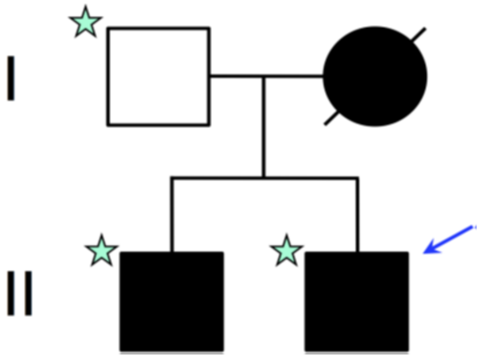


FIGURE 4.1: **Pedigree for family FALS147.** Family FALS147 consists of two generations with autosomal dominant inheritance of ALS. DNA was available from individuals marked with stars, all of whom underwent WES and WGS. Females are represented by circles, and males by squares. Black filled symbols indicate individuals affected by ALS, a diagonal strikethrough shows a deceased individual. The arrow represents the family proband.

Family FALS147 is a two generation ALS kindred, which was previously screened and shown to be negative for all known ALS gene mutations (Figure 4.1). Family FALS147 comprises two affected children, one affected parent, and one married-in parent, consistent with autosomal dominant inheritance of ALS. Exome and genome sequencing data was available from the two affected children and the one married-in control.

4.1.2 Validation of exomic variants in whole-genome data

Twenty-one candidate gene variants resulting from family FALS147 WES analysis (identified prior to this project) were screened in the whole-genome data (Appendix Table A.1). With one exception (*ISM2*), all candidate gene variants were present in the whole-genome data from the two affected FALS and absent in the married in control parent.

4.1.3 Analysis of family FALS147 whole-genome sequencing data

Whole-genome sequencing data from family FALS147 was subjected to shared variant analysis (Section 3.1.1) to identify variants that were present in both affected individuals and absent in the married in control parent. In total, 683613 shared variants were identified. Of these, 3214 variants were identified within genes or within 1kb upstream and downstream from coding sequences. These included coding variants (0.49%) and non-coding variants in introns and splice sites (Figure 4.2). Filtering of coding variants, shown in Figure 4.3, led to the identification of 15 novel heterozygous non-synonymous candidate variants in 11 genes (Appendix Table 4.1). Analysis across the WGS dataset identified extremely high allele-counts (variants present in >90% of screened samples that are likely to be false-positives) for 11 variants, which were subsequently removed. In contrast, four candidate gene variants (located in *NUTM2A*, *C2CD4A*, *TERF2IP*, and *ZNF580*) were found only in the two affected individuals from family FALS147. These four novel candidate variants were screened through updated control databases and underwent Sanger sequencing for validation (Sections 4.1.5 and 4.1.6).

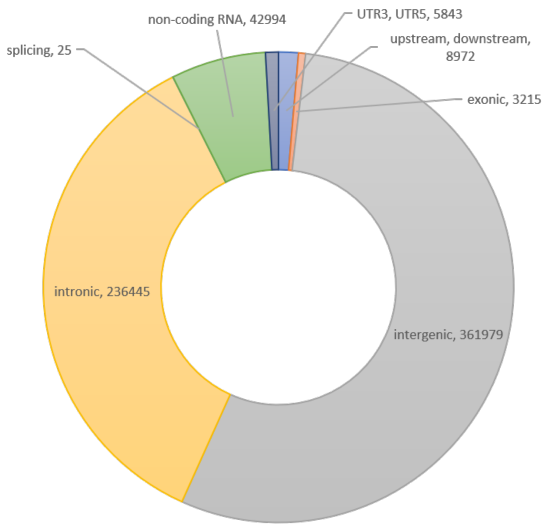
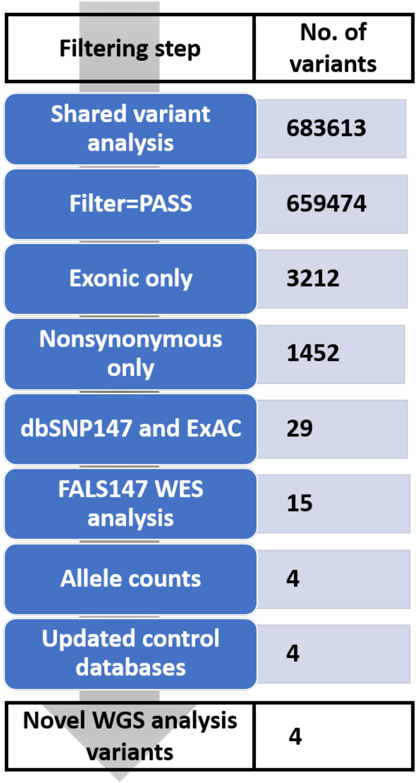


FIGURE 4.2: **Types of shared variants identified in family FALS147.** Most of the shared variants were intergenic (55%), followed by intronic (36%) and non-coding RNA (7%). There are only 3215 exonic variants (0.49%).

TABLE 4.1: **Family FALS147 candidate gene variants identified from WGS analysis**

Gene	Location	Accession no	Variant (DNA)	Variant (Protein)	Obtained from
MST1L	1:17085995	NM_001271733	c.C902G	p.A301G	FALS147 whole-genome analysis
NBPF19	1:145366193	NM_001351365	c.C8548A	p.Q2850K	FALS147 whole-genome analysis
OR2T33	1:248436972	NM_001004695	c.T145G	p.W49G	FALS147 whole-genome analysis
OR2T12	1:248458736	NM_001004692	c.T145G	p.W49G	FALS147 whole-genome analysis
FAM8A1	6:17601035	NM_016255	c.A395G	p.H132R	FALS147 whole-genome analysis
FAM8A1	6:17601040	NM_016255	c.G400A	p.G134S	FALS147 whole-genome analysis
FAM8A1	6:17601041	NM_016255	c.G401T	p.G134V	FALS147 whole-genome analysis
FAM8A1	6:17601044	NM_016255	c.T404C	p.L135P	FALS147 whole-genome analysis
FAM8A1	6:17601058	NM_016255	c.G418A	p.A140T	FALS147 whole-genome analysis
NUTM2A	10:88988120	NM_001099338	c.C483A	p.H161Q	FALS147 whole-genome analysis
CDHR5	11:618998	NM_001171968	c.C1543G	p.P515A	FALS147 whole-genome analysis
OR6S1	14:21109726	NM_001001968	c.C125G	p.T42R	FALS147 whole-genome analysis
C2CD4A	15:62359942	NM_207322	c.G130A	p.D44N	FALS147 whole-genome analysis
TERF2IP	16:75682129	NM_018975	c.G349A	p.A117T	FALS147 whole-genome analysis
ZNF580	19:56154346	NM_001163423	c.C472T	p.R158C	FALS147 whole-genome analysis



Filtering step	No. of variants
Shared variant analysis	683613
Filter=PASS	659474
Exonic only	3212
Nonsynonymous only	1452
dbSNP147 and ExAC	29
FALS147 WES analysis	15
Allele counts	4
Updated control databases	4
Novel WGS analysis variants	4

Figure 4.3: **Filtering of family FALS147 WGS data.** Novel candidate variants identified after custom bioinformatics filtering analysis of WGS data from family FALS147. Shared variant analysis identified 683613 variants. Custom filtering resulted in the identification of four novel candidate gene variants that were not previously identified from analysis of WES data.

4.1.4 Candidate gene screening in SALS whole-genome sequencing data

A small but significant proportion of apparently SALS cases are known to carry pathogenic mutations in genes that were originally linked to FALS. As such, the top 3 candidate genes identified from WES analysis of family FALS147 (*CYB5R3*, *DCAF7* and *SAV1*), as well as from recent publications or collaborators (18 genes, Table 3.2) were screened in the SALS WGS dataset to identify additional novel gene variants that may contribute to disease. Analysis of the top three family FALS147 candidate genes (*SAV1*, *CYB5R3*, and *DCAF*), led to the discovery of four additional gene variants. Moreover, 20 new candidate variants were identified after candidate gene screening and custom filtering of the 18 genes obtained from publications or collaborators. These novel variants are shown in Appendix Table A.2.

4.1.5 Updated control database filtering

All variants identified from the novel gene discovery pipeline in family FALS147 as well as candidate gene screening were further filtered using NGS data from publicly available control datasets, including GnomAD, MGRB and ProjectMinE. Three of the 21 family FALS147 exome candidate variants (*CCDC50*, *IL17RE*, and *LRRC70*) were removed after control screening. Four (of 24) gene variants identified from gene screening from sporadic cohorts were also removed. Summaries of candidate gene variants filtered out and maintained following control variant filtering are shown in Tables 4.2 and 4.3.

TABLE 4.2: **Details of candidate genes filtered out by updated control databases**

Gene	**GnomAD**	**MGRB**	**ProjectMinE**
CCDC50 (c.A811G)	Yes (3 Non-Finnish)	No	No
IL17RE (c.C64A)	Yes (5 Non-Finnish)	No	No
LRRC70 (c.A1490T)	Yes (2 Non-Finnish)	No	No
AUH (c.C973G)	Yes (1 African and 4 Non-Finnish)	Yes	No
CYP27A1 (c.G1309T)	Yes (2 Non-Finnish)	Yes	No
LAMA2 (c.C5937G)	Yes (3 Non-Finnish)	Yes	No
LAMA2 (c.C7513T)	Yes (4 Non-Finnish and 2 others)	Yes	No

TABLE 4.3: **Novel variants from candidate gene screening**

Gene	**Patient sample**	**Acession no**	**Codon change**	**AA change**
DCAF7	ch17:61657233	NM_005828	c.G457C	p.G153R
CYB5R3	ch22:43026934	NM_007326	c.C218A	p.P73H
CYB5R3	ch22:43040403	NM_001171660	c.G113A	p.S38N
CYB5R3	ch22:43040433	NM_001171660	c.A83G	p.Q28R
ABCD1	chrX:152990929	NM_000033	c.G208C	p.V70L
AARS2	chr6:44272510	NM_020745	c.G1624T	p.D542Y
AUH	chr9:93976697	NM_001698	c.C953T	p.T318I
AUH	chr9:94123951	NM_001698	c.A221C	p.E74A
GBE1	chr3:81643080	NM_000158	c.C1087T	p. L363F
GBE1	chr3:81754764	NM_000158	c.G144T	p.R48S”
GBE1	chr3:81810577	NM_000158	c.T92A	p.L31H”
HTRA1	chr10:124248938	NM_002775	c.G573C	p.K191N”
LAMA2	chr6:129371083	NM_000426	c.A133C	p.N45H
LAMA2	chr6:129511430	NM_000426	c.T1548A	p.D516E
LAMA2	chr6:129571323	NM_000426	c.A1849C	p.T617P
LAMA2	chr6:129591890	NM_000426	c.C2444T	p.S815F
LAMA2	chr6:129641769	NM_000426	c.G4145T	p.C1382F
LAMA2	chr6:129766925	NM_000426	c.A6388T	p.I2130L
KIF5A	chr12:57969899	NM_004984	c.A2053G	p.K685E
USP7	chr16:8994445	NM_003470	c.G2251C	V751L

4.1.6 Validation of ALS candidate variants by direct DNA sequencing

Validation of family FALS147 candidate variants

Candidate variants identified from NGS data required validation by direct DNA sequencing. Primers designed to amplify candidate gene variants were successfully optimised (Table 3.3) and used for PCR amplification and Sanger sequencing. All family FALS147 variants were validated. Exomic variants were previously validated by other group members. Among the four novel genomic variants obtained from WGS analysis, *ZNF580*, *C2CD4A*, and *TERF2IP* variants were confirmed to be present in both affected siblings while absent in the control parent, while *NUTM2A* had ambiguous results. An example of successful validation is shown in Figure 4.4 for *C2CD4A*. Table 4.4 shows the final list of all PCR validated candidate gene variants remaining for family FALS147, following filtering from control databases.

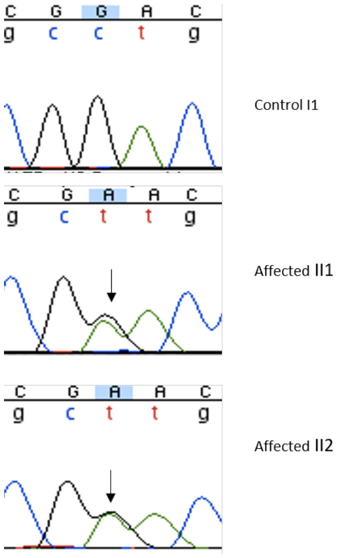


FIGURE 4.4: **Chromatograms showing validation of the family FALS147 *C2CD4A* candidate variant**. The family FALS147 candidate variant in *C2CD4A* (c.G130A , p.D44N) was present in both patient DNA samples, and was absent in the control, and was successfully validated by Sanger sequencing.

TABLE 4.4: **PCR validated family FALS147 candidate gene variants**

Gene	Location	Accession no	Variant (DNA)	Variant (Protein)	Obtained from
PRAMEF12	chr01:12837663	NM_001080830	c.G1373A	p.R458H	FALS147 whole-exome analysis
ZBTB8A	chr01:33059103	NM_001040441	c.A571G	p.K191E	FALS147 whole-exome analysis
TTN	chr02:179546450	NM_133378	c.T29378A	p.I9793N	FALS147 whole-exome analysis
ANKRD31	chr05:74491500	NM_001164443	c.T973C	p.F325L	FALS147 whole-exome analysis
DNAH8	chr06:38879296	NM_001206927	c.G9793T	p.G3265C	FALS147 whole-exome analysis
OR2A25	chr07:143771959	NM_001004488	c.C647G	p.S216C	FALS147 whole-exome analysis
PPP1R9A	chr07:94540079	NM_001166163	c.C654G	p.I218M	FALS147 whole-exome analysis
FGD6	chr12:95604810	NM_018351	c.C250G	p.Q84E	FALS147 whole-exome analysis
NR1H4	chr12:100904749	NM_005123	c.G273T	p.M91I	FALS147 whole-exome analysis
ISM2	chr14:77950679	NM_182509	c.A614G	p.N205S	FALS147 whole-exome analysis
SAV1	chr14:51111611	NM_021818	c.T657A	p.D219E	FALS147 whole-exome analysis
HERC2	chr15:28474716	NM_004667	c.A5010C	p.E1670D	FALS147 whole-exome analysis
DCAF7	chr17:61662588	NM_005828	c.G754A	p.V252I	FALS147 whole-exome analysis
NDEL1	chr17:8347611	NM_001025579	c.G22C	p.D8H	FALS147 whole-exome analysis
ZZEF1	chr17:3961306	NM_015113	c.T5147C	p.V1716A	FALS147 whole-exome analysis
CYB5R3	chr22:43024269	NM_007326	c.C283T	p.H95Y	FALS147 whole-exome analysis
FBLN1	chr22:45928969	NM_006487	c.C571T	p.R191X	FALS147 whole-exome analysis
AFF2	chrX:147743972	NM_001169124	c.T724C	p.S242P	FALS147 whole-exome analysis
NUTM2A	10:88988120	NM_001099338	c.C483A	p.H161Q	FALS147 whole-genome analysis
C2CD4A	15:62359942	NM_207322	c.G130A	p.D44N	FALS147 whole-genome analysis
TERF2IP	16:75682129	NM_018975	c.G349A	p.A117T	FALS147 whole-genome analysis
ZNF580	19:56154346	NM_001163423	c.C472T	p.R158C	FALS147 whole-genome analysis

Validation of variants from other candidate genes

From the remaining 16 variants obtained from gene screening, nine were successfully PCR amplified (Table 4.5). An example of successful primer optimization, PCR and sequencing chromatogram is shown in (Figures 4.5 and 4.6).

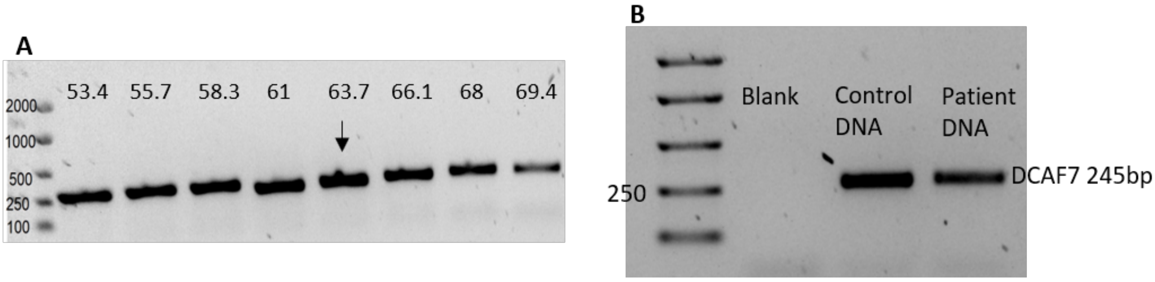


FIGURE 4.5: **Optimisation and validation of the SALS *DCAF7* candidate variant.** Gel electrophoresis image showing (A) optimisation of *DCAF7* primers, and the selected annealing temperature of 63.7 °C (denoted by arrow) for specific amplification of the 245bp PCR product. (B) Successful amplification of the c.G457C variant amplicon in control and patient DNA.

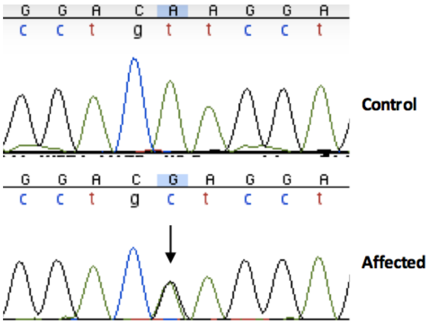


Figure 4.6: **Chromatograms showing validation of SALS *KIF5A* candidate variant**. The *KIF5A* c.A2053G, p.K685E identified in a SALS case was successfully validated as indicated by the double peak present and nucleotide change from A to G in the affected individual (marked by the arrow).

Table 4.5: **PCR validation results from candidate gene screening in SALS and family FALS147 WGS analysis**

Gene	Patient sample	Acession no	Codon change	AA change	Validation results
Candidate gene screening in SALS					
KIF5A	SALS0041	NM_004984	c.A2053G	p.K685E	Yes
USP7	SALS2414	NM_003470	c.G2251C	V751L	Yes
ABCD1	MN201540	NM_000033	c.G208C	p.V70L	Yes
AUH	SALS1097	NM_001698	c.C953T	p.T318I	Yes
GBE1	SALS2162	NM_000158	c.C1087T	p. L363F	Yes
LAMA2	SALS2111	NM_000426	c.A133C	p.N45H	Yes
DCAF7	SALS1085	NM_005828	c.G457C	p.G153R	Uncertain
CYB5R3	SALS0215	NM_007326	c.C218A	p.P73H	Yes
CYB5R3	SALS1427	NM_001171660	c.G113A	p.S38N	Yes
CYB5R3	SALS2270	NM_001171660	c.A83G	p.Q28R	Yes
Family FALS147 WGS candidate genes					
NUTM2A	FALS020183 (control)	NM_001099338	c.C483A	p.H161Q	Uncertain
	FALS020072				Uncertain
	FALS020176				Uncertain
C2CD4A	FALS020183 (control)	NM_207322	c.G130A	p.D44N	No
	FALS020072				Yes
	FALS020176				Yes
TERF2IP	FALS020183 (control)	NM_018975	c.G349A	p.A117T	No
	FALS020072				Yes
	FALS020176				Yes
ZNF580	FALS020183 (control)	NM_001163423	c.C472T	p.R158C	No
	FALS020072				Yes
	FALS020176				Yes

4.1.7 *In silico* functional analysis and variant prioritisation

Following analysis using the bioinformatics filtering pipeline, *in silico* tools were used to predict the potential pathogenicity and functional consequence of candidate gene variants and to rank the variants for subsequent analyses. This comprehensive pipeline included a total of eleven protein prediction programs, as well as analysis of conservation, natural gene variation, gene expression and gene function. *In silico* scoring of family FALS147 exomic variants was carried out prior to this project using the same pipeline. A summary of *in vitro* scores for all candidate genes variants identified during this project is shown in Table 4.8, with additional scores shown in Appendix Tables A.3, A.4, A.6 and A.5.

A novel prioritisation strategy was developed and applied to these results to identify the highest priority candidate variants from family FALS147, leading to the three top ranked candidate gene variants to be *SAV1* c.T657A, *DCAF7* c.G754A and *CYB5R3* c.283T. From the four novel family FALS147 candidate gene variants identified in WGS data (*NUTM2A*, *C2CD4A*, *TERF2IP*, and *ZNF580*), the c.C472T variant in *ZNF580* was the highest ranked candidate, predicted to be deleterious by six protein prediction tools, highly conserved across many species, and *ZNF580* highly expressed both in brain and spinal cord. Moreover, of the 10 variants that were identified using the candidate gene screening approach and ranked *in silico*, four variants (*CYB5R3* p.P73H, *GBE1* p. L363F, *LAMA2* p.N45H , and *AUH*). p.T318I were predicted to be pathogenic by 9 to 11 protein prediction tools (Table 4.8), and were ranked as high priority candidates. Only *CYB5R3* p.P73H ranked highly in most aspects of the *in silico* analysis, making it the strongest candidate gene variant for further analysis. The remaining candidate genes were rated poorly in most aspects of the *in silico* pipeline.

4.2 Generation of wild type and mutant expression constructs for *in vitro* analysis

To assess the pathogenicity of top-priority candidate gene variants identified in family FALS147 (Section 4.1), a small library of expression constructs carrying candidate gene cDNA were generated (Table 4.6) for *in vitro* analysis. To generate the constructs, candidate gene variants were introduced in wild type pCMV6-Entry constructs using site-directed mutagenesis. Wild type and mutant genes were then digested from the pCMV6-Entry construct and ligated to the pCMV6-AC-RFP construct, and validated by Sanger sequencing.

Table 4.6: **Constructs generated in this project**

Construct name	**Construct**	**Tag**	**Candidate gene**	**Mutation**
pCMV6-Entry-CYB5R3-WT	pCMV6-Entry	Myc	CYB5R3	Wild type
pCMV6-AC-RFP-CYB5R3-WT	pCMV6-AC-RFP	RFP	CYB5R3	Wild type
pCMV6-Entry-CYB5R3-Mutant	pCMV6-Entry	Myc	CYB5R3	c.C283T; p.H95Y
pCMV6-AC-RFP-CYB5R3-Mutant	pCMV6-AC-RFP	RFP	CYB5R3	c.C283T; p.H95Y
pCMV6-Entry-DCAF7-WT	pCMV6-Entry	Myc	DCAF7	Wild type
pCMV6-AC-RFP-DCAF7-WT	pCMV6-AC-RFP	RFP	DCAF7	Wild type
pCMV6-Entry-DCAF7-Mutant	pCMV6-Entry	Myc	DCAF7	c.G754A; p.V252I
pCMV6-AC-RFP-DCAF7-Mutant	pCMV6-AC-RFP	RFP	DCAF7	c.G754A; p.V252I
pCMV6-Entry-SAV1-WT	pCMV6-Entry	Myc	SAV1	Wild type
pCMV6-AC-RFP-SAV1-WT	pCMV6-AC-RFP	RFP	SAV1	Wild type
pCMV6-Entry-SAV1-Mutant	pCMV6-Entry	Myc	SAV1	c.T657A; p.D219E
pCMV6-AC-RFP-SAV1-Mutant	pCMV6-AC-RFP	RFP	SAV1	c.T657A; p.D219E

4.2.1 Generation of pCMV6-entry mutant constructs using Q5 mutagenesis

Site-directed mutagenesis was successfully performed to introduce candidate variants into the cDNA sequences of the respective expression constructs, as described in Table 4.6. The presence of each variant in the respective constructs was examined by check-colony PCR (Figure 4.7) followed by Sanger sequencing. For each gene construct, six colonies were sequenced and two colonies each had the desired mutation. These were subsequently sequenced in full (Figure 4.8).

4.2.2 Generation of pCMV6-AC-RFP wild type and mutant constructs

To generate the pCMV6-AC-RFP wild type and mutant constructs required to visualise gene expression, both the empty pCMV6-AC-RFP and the pCMV6-entry constructs were first digested using *SgfI* and *MluI* as shown in Figure 4.9. The digested products that corresponded to the candidate cDNAs (Figure 4.9, bottom band) were then extracted and purified from the gel and ligated to the digested and purified. To ensure ligation and transformation were successful, the cDNA inserts were validated by check-colony PCR using gene specific primers and Sanger sequencing. All primers and expected amplicons are summarised in Table 3.9.As shown in Figure 4.10, all wild type cDNAs and *SAV1*-D219E cDNA were PCR amplified to generate amplicons of the correct size. Mutant *CYB5R3* and *DCAF7* PCRs also generated non-specific bands which, upon sequencing, did not carry the desired gene sequences. Additional

CYB5R3-H95Y and *DCAF7*-V252I mutant colonies were tested using the T7 universal primer and the corresponding gene reverse primer to obtain more specific bands (Figure 4.11). Three additional positive colonies were identified and sequenced, all of which contained correct mutated sequences.

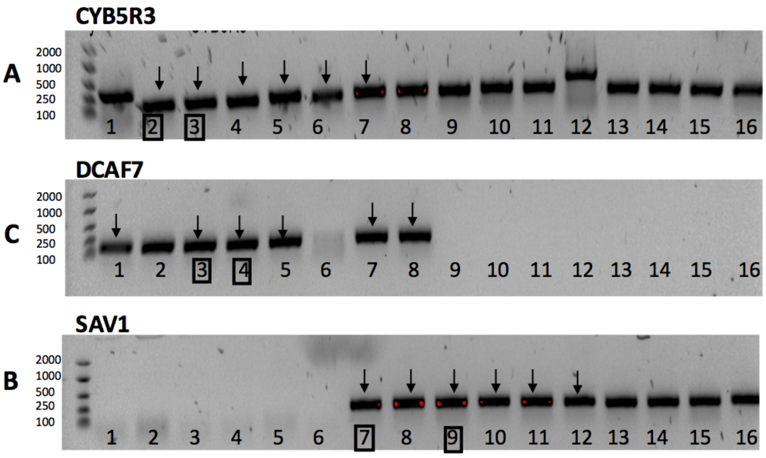


Figure 4.7: **Check-colony PCR to confirm cDNA ligation into pCMV6-entr.** PCR products were amplified using gene specific forward and reverse primers, and electrophoresed. All lanes contain PCR products from a single bacterial colony following transformation. Correct PCR amplicon sizes are 179bp, 180bp, and 186bp for *CYB5R3, DCAF7* and *SAV1* respectively. An empty well represents an empty construct. The arrows represent constructs that were sequenced, while the boxed numbers represent the colonies that were fully sequenced with the desired mutation.

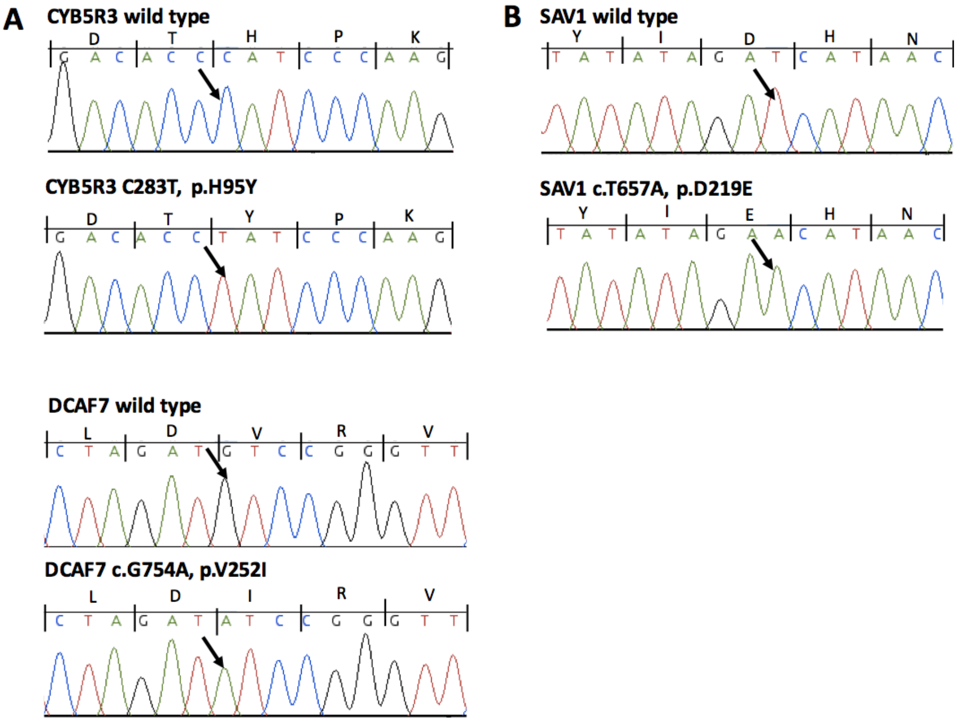


FIGURE 4.8: **Sequencing chromatograms of candidate cDNA in pCMV6-entry.** Sequence chromatograms showing the presence of the desired nucleotide changes in each candidate gene: A; *CYB5R3*, B; *SAV1* and C; *DCAF7*, denoted by the arrows. The top chromatogram represents the reference sequence (or wild type gene), while the lower chromatogram represents the sequence containing the desired variant.

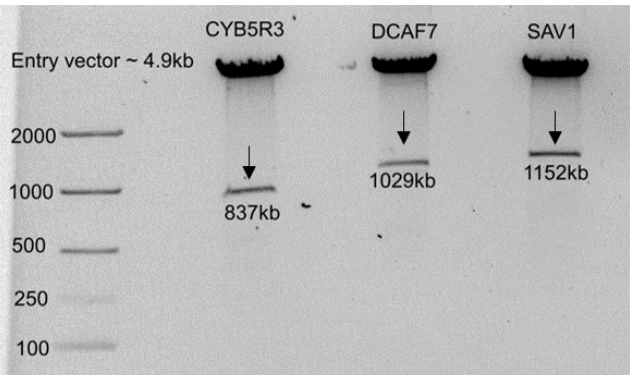


FIGURE 4.9: **Restriction enzyme digests of pCMV6-entry**. Constructs were digested and electrophoresed on agarose gel. Two restriction fragments were generated; the pCMV6-entry construct at 4.9kb size (top bands) and respective candidate cDNA inserts (arrowed). Easy ladder I (Bioline) was used as a marker and arrows indicate the insert that was purified.

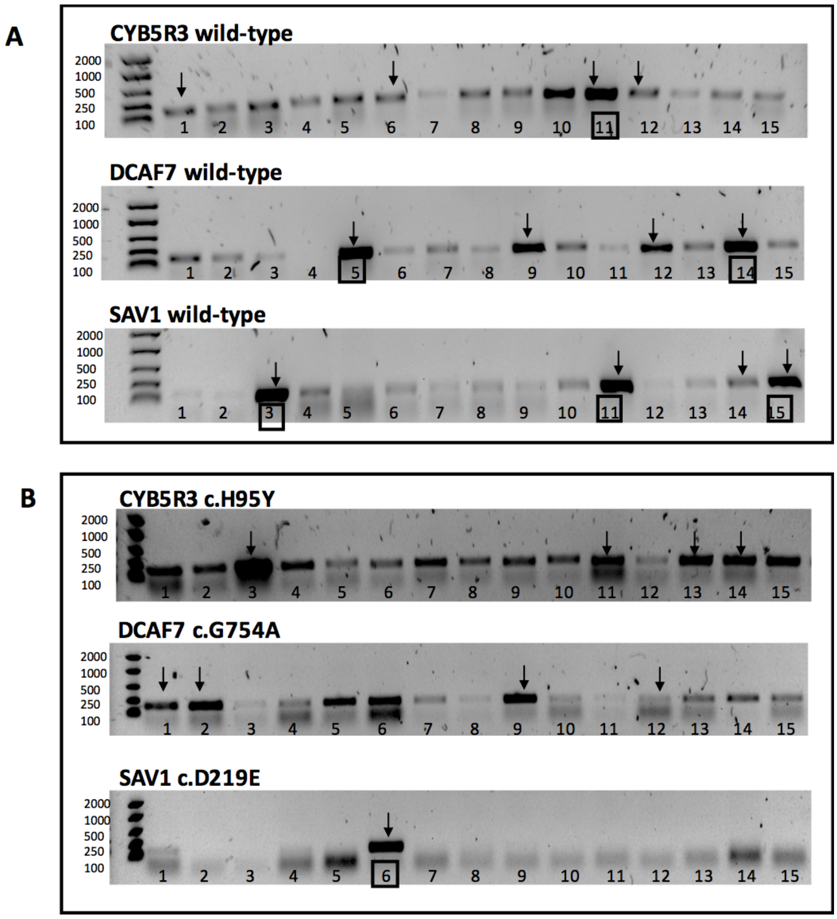


Figure 4.10: **Check-colony PCR to confirm cDNA ligation in pCMV6-AC-RFP.** PCR used gene specific F and R primers to generate amplicons from clones that contained cDNA inserts. Amplicon sizes are 179bp, 180bp, and 186bp for *CYB5R3, DCAF7* and *SAV1* respectively. Each lane represents a PCR using template from a single colony. The arrows represent constructs that were sequenced, while the boxed numbers represent the clones that had the correct desired sequence. Blank wells show colonies with empty constructs.

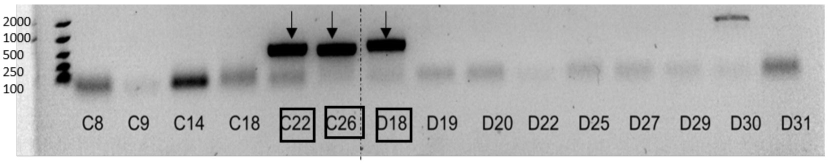


Figure 4.11: **Check-colony PCR using T7 primer to confirm cDNA insert in pCMV6-AC-RFP.** PCR used the universal T7 forward and a gene specific reverse primer to generate amplicons from clones that contained cDNA inserts. Amplicon sizes are 348bp for *CYB5R3* and 571bp for *DCAF7*. Each lane represents a PCR using template from a single colony. The arrows indicate constructs that were sequenced, while the boxed numbers indicate the clones that had the correct desired sequence. Blank wells show colonies with empty constructs.

4.3 *In vitro* functional analysis of family FALS147 candidate genes

4.3.1 Transfection optimisation in SH-SY5Y cells

In order to optimise transfection efficiency in the SH-SY5Y neuronal cell line, different concentrations of transfection reagents were tested together with pCMV6-AC-RFP-empty construct (Section 3.3.3). RFP expression was observed by fluorescence microscopy (excitation/emission = 553/ 574 nm) and transfection efficiency was quantified using flow cytometry at 48 hours post-transfection. Fluorescence microscopy provided a qualitative assessment and indicated that few transfected cells were present regardless of transfection reagent concentration (data not shown). Flow cytometry analysis confirmed this observation with a mean transfection efficiency of 1.6%, as summarised in Table 4.7. Condition 10 (1000ng DNA, 3μl Lipofectamine, and 1μl PLUS reagent) showed the highest efficiency, with 2.7% of cells transfected, whereas condition 1 (500ng DNA, 1.25μl Lipofectamine, and 0.5μl PLUS reagent) showed the lowest efficiency, with only 0.2% of cells transfected. Overall, the transfection efficiency was considered too low for subsequent analysis in SH-SY5Y cells. Instead, with time constraints, the HEK293T cell line (human embryonic kidney) was selected as an alternative due to its reputation for ease of handling, rapid growth and ease of transfection.

TABLE 4.7: **Percentages of transfected SH-SY5Y cells for all transfection conditions**

Sample number	DNA (νg)	Lipofectamine (μl)	PLUS reagent (μl)	% transfected cells
1	500	1.25	0.5	0.2
2	500	1.5	0.5	0.4
3	500	2	0.5	0.8
4	500	2.5	0.5	2.1
5	750	1.8	0.75	1.1
6	750	2.25	0.75	1.4
7	750	3	0.75	2.4
8	750	3.75	0.75	2.1
9	1000	2.5	1	2.1
10	1000	3	1	2.7
11	1000	4	1	2.5
12	1000	5	1	1.8

4.3.2 Transfection efficiency in HEK293T cells

Flow cytometry was also used to determine the transfection efficiency between wild type and mutant genes in HEK293T cells (Figure 4.12). Transfection efficiencies mean (n=3) for the different genes are as follows: CYB5R3-WT (93.6%), CYB5R3-mutant (83.2%), DCAF7-WT (20.4%), DCAF7-mutant (15%), SAV1-WT(88%), and SAV1-mutant (88.7%). No significant difference was found between expression of wild type and mutant for DCAF7 or SAV1. However, transfection efficiency of the CYB5R3 mutant was significantly lower than that of CYB5R3-WT (**$p<0.01$).

4.3.3 Cytotoxicity of candidate proteins

Cytotoxicity was assayed following transfection of each of the three wild type and mutant candidate expression constructs in HEK293T cells. The percentage of dead cells was determined using flow cytometry after staining with SYTOX cell death marker (Figure 4.13). Although non-significant, cells transfected with CYB5R3-H95Y showed a lower proportion of SYTOX blue population compared to CYB5R3-WT (p=0.26). Cells transfected with the DCAF7-V252I also showed a lower proportion of the SYTOX blue cell population compared to DCAF7-WT (p=0.3). The opposite was observed for SAV1, with SAV-D219E showing a higher proportion of dead cells compared to SAV-WT (p=0.39). No observed changes were significant.

4.3.4 Cellular localisation of candidate proteins

To study the effect of each candidate variant on protein cellular localisation, all pCMV6-AC-RFP wild type and mutant constructs were individually transfected into HEK293T cells. Endogenous TDP-43 expression was also investigated. Overexpression of DCAF7 resulted in nuclear localisation of the recombinant proteins, with no observed difference between DCAF7-WT and DCAF7-V252I (Figure 4.14). TDP-43 also localised to the nucleus in both wild type and mutant DCAF7. Overexpression of SAV1 showed cytoplasmic localisation of the recombinant proteins, with no observed difference between SAV-WT and SAV-D219E (Figure 4.15). TDP-43 localisation was mostly nuclear, with occasional cytoplasmic TDP-43 localisation and co-localisation with SAV1 in both wild type and mutant SAV1 expressing cells. Overexpression of CYB5R3-WT showed cytoplasmic localisation of the recombinant protein, whereas overexpressed CYB5R3-H95Y formed large round cytoplasmic inclusions in a subpopulation of cells (Figure 4.16). Quantification of these inclusions was performed on 100 transfected cells. CYB5R3-H95Y expressing cells showed significantly higher proportion of cells with large cytoplasmic inclusions compared to CYB5R3-WT ($p<0.05$, p = 0.015). CYB5R3 did not co-localise with endogenous TDP-43.

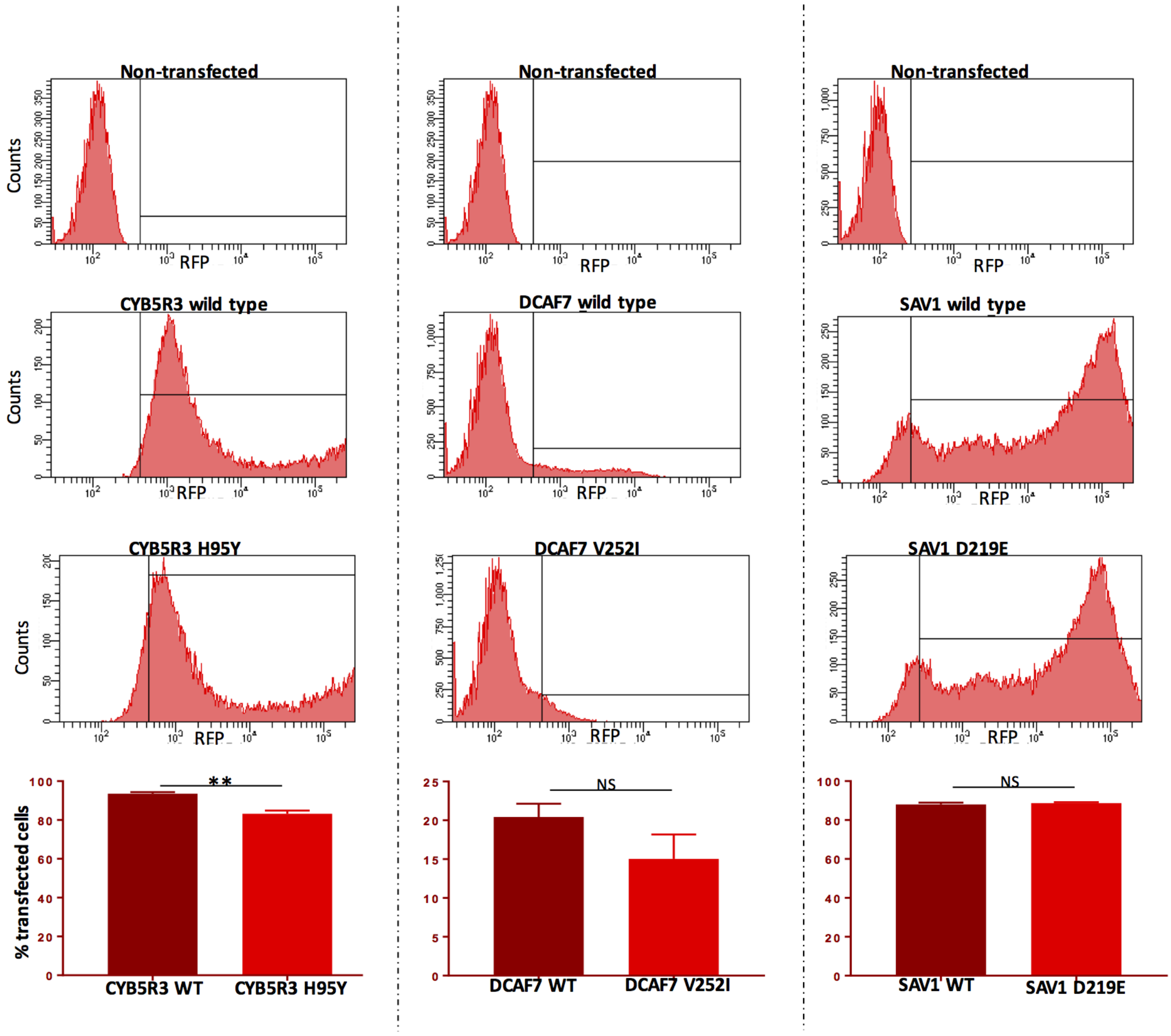


Figure 4.12: **Transfection efficiency of wild type and mutant expression constructs**. Transfection effeciency was determined using RFP flurorescence expression. Non-transfected sample was used as a gating parameter for transfection, and the same gating was used to determine transfection for the other samples. Transfection with CYB5R3-H95Y was significantly lower than with CYB5R3-WT. There was no significant difference between wild type and mutant for DCAF7 and SAV1. The bar chart values represent mean +- S.E.M by pooling 3 samples together. Error bars show S.E.M., (**$p<0.01$).

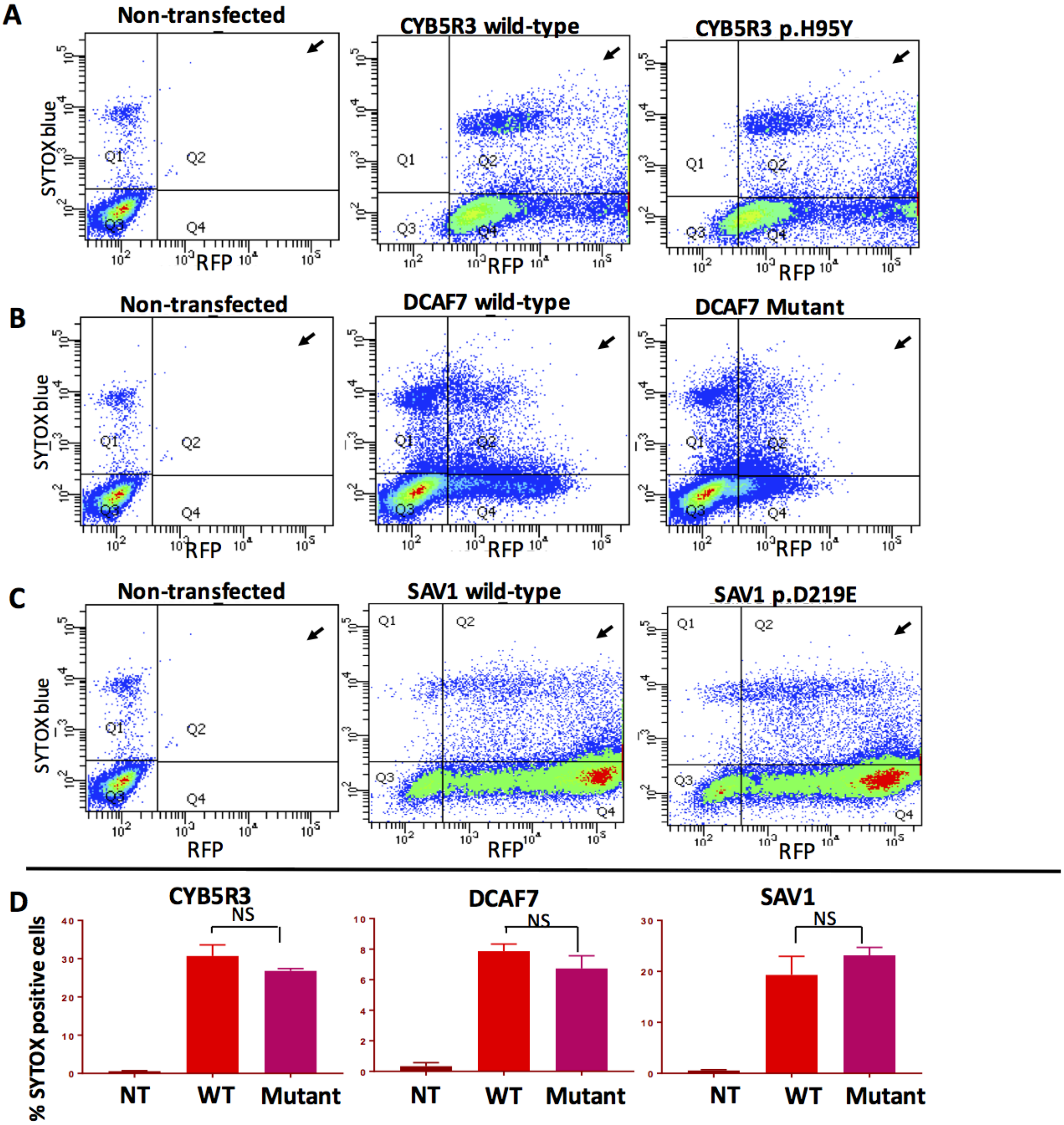


Figure 4.13: **Cytotoxicity of HEK293T cells expressing CYB5R3, DCAF7 and SAV1**. (A, B, C) Representative flow cytometry dot plots of SYTOX blue staining in RFP-positive cell populations (Q2, denoted by the arrow) for wild type and candidate mutant (A) CYB5R3, (B) DCAF7 and (C) SAV1. The population gate was established using non-transfected samples as a control. The x-axis represents RFP expression, y-axis represents SYTOX blue staining. (D) Quantification of SYTOX blue positive populations (Q2). There were no significant differences in toxicity between the wild type and mutant transfected cells. The values represent mean +- S.E.M. Error bars show S.E.M (NS, p >0.05) 10,000 events were measured, n=3).

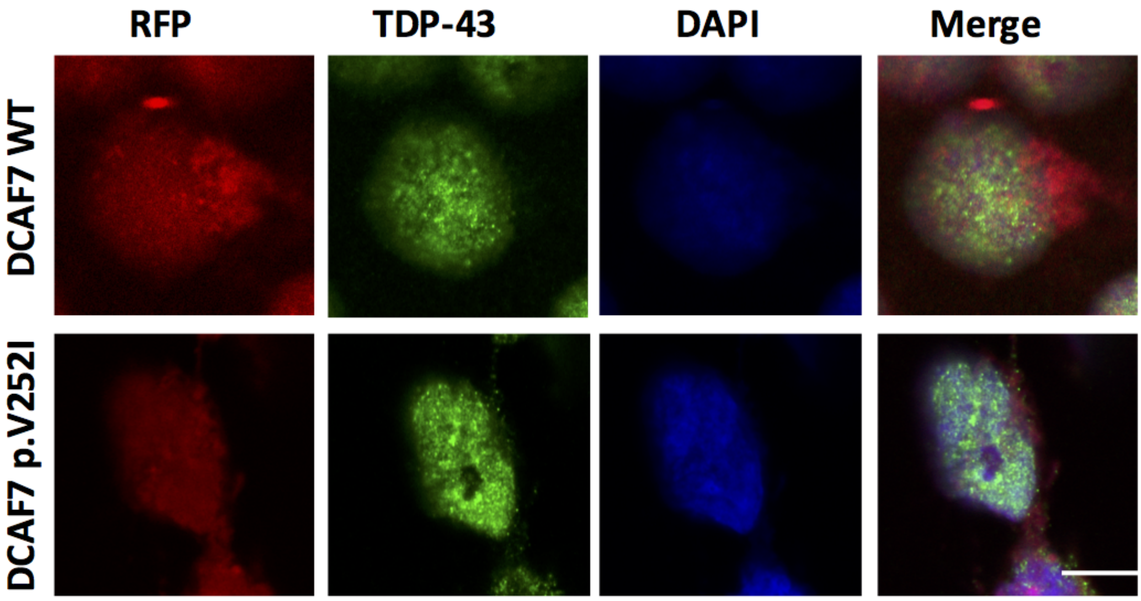


Figure 4.14: (A) **Localisation of DCAF7 in HEK293T cells**. HEK293T cells were transfected with pCMV6-AC-RFP DCAF-WT and DCAF-V252I constructs (red) and stained with nuclear marker DAPI (blue). Cells were also immuno-stained for endogenous TDP-43 (green). DCAF7-WT and DCAF-V252I localised to the nucleus. There was no correlation between TDP-43 and DCAF7 expression, scale bar = 10um

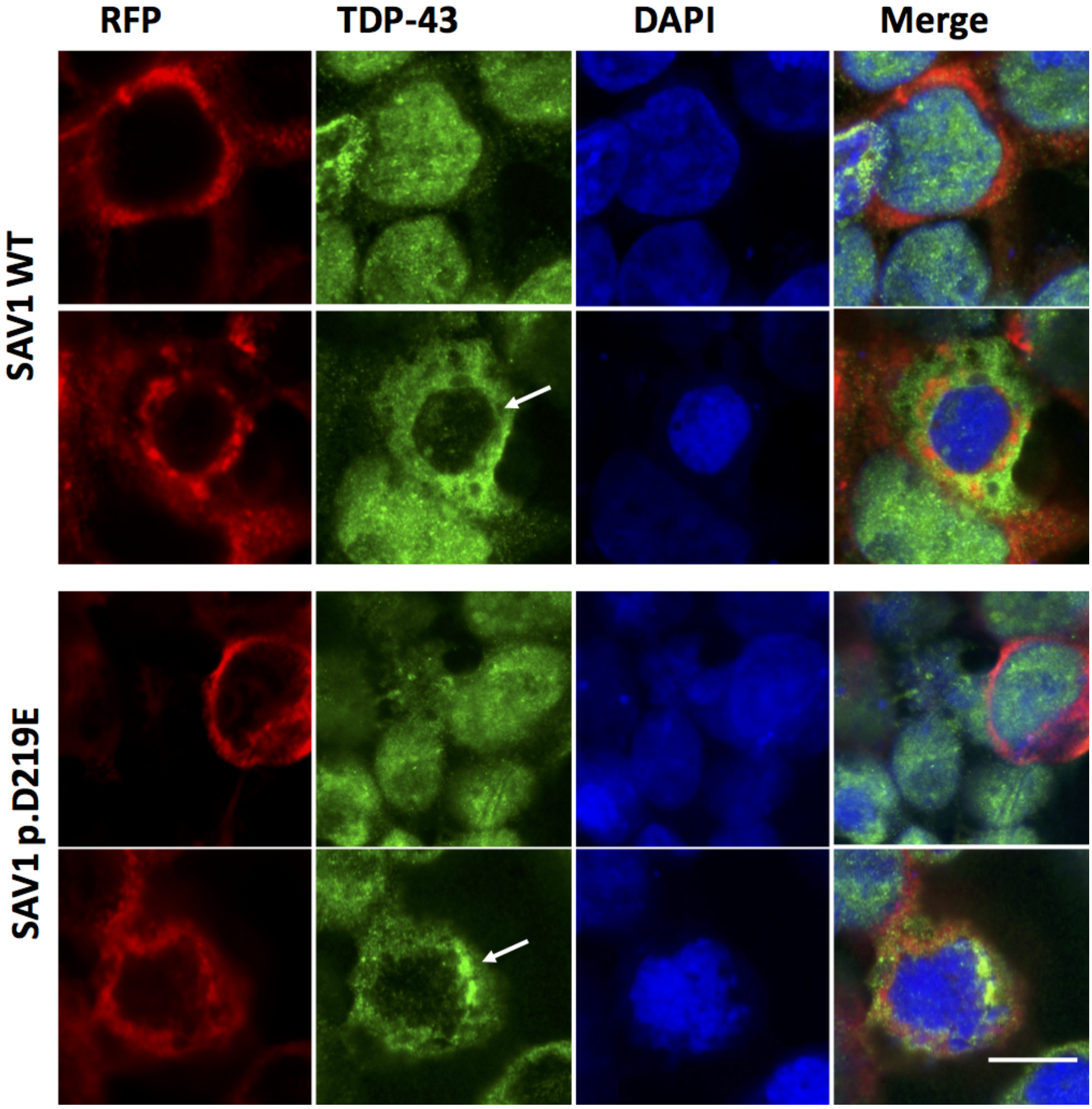


FIGURE 4.15: **Localisation of SAV1 in HEK293T cells**. HEK293T cells were transfected with pCMV6-AC-RFP wild type and mutant SAV1 constructs (red) and stained with nuclear marker DAPI (blue). Cells were also immuno-stained for endogenous TDP-43 (green). WT and mutant SAV1 recombinant proteins localised to the cytoplasm. TDP-43 was located in either the nucleus or cytoplasm (marked by arrow) and there was no clear co-localisation pattern between TDP-43 and SAV1. Scale bar= 10um.

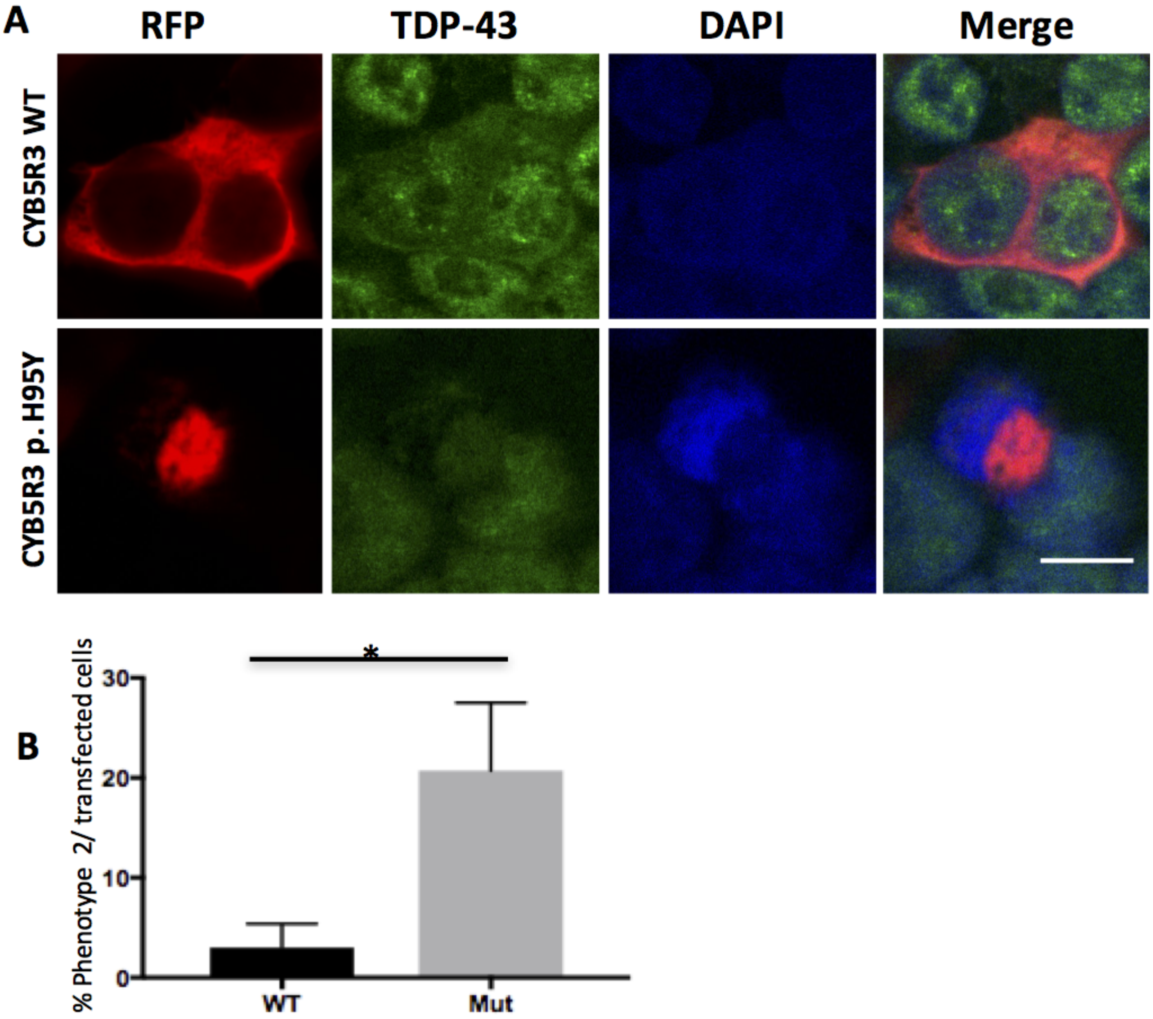


Figure 4.16: **Localisation of CYB5R3 in HEK293T cells**. HEK293T cells were transfected with pCMV6-AC-RFP wild type and mutant CYB5R3 constructs (red) and stained with nuclear marker DAPI (blue). Cells were also immuno-stained for endogenous TDP-43 (green). Both CYB5R3-WT and CYB5R3-H95Y were predominately located in the cytoplasm. CYB5R3-H95Y formed rounded inclusion-like structures. Scale bar = 10um (B) Counting was performed on a random sample of 100 CYB5R3-WT and CYB5R3- H95Y transfected cells, and a comparative t-test showed the cytoplasmic inclusion phenotype to be significantly more common in mutant than CYB5R3-WT transfected cells (*p<0.05.). Error bars show S.E.M.

4.4 Generation of a *VPS29* knockout cell line using CRISPR-Cas9

Another aim of this project was to develop a protocol to create a *VPS29* knockout cell line using CRISPR-Cas9 technology. Knockout cell lines will form part of the *in vitro* analysis pipeline to assess the pathogenicity of candidate genes in ongoing studies. Figure 4.18 shows the different stages, including transfection, flow sorting, cell maintenance, dot blotting, western blotting and Sanger sequencing, as well as the number of cell colonies that remained at each stage. Following transfection with CRISPR reagents and flow sorting, two 96-well plates (n=192 colonies) were obtained. Contaminated colonies, or those originating from multiple cells (determined by microscope visualisation) were discarded. Fifty-two cell colonies were passaged to 12-well plates, and protein lysates were collected for a total of 34 of these colonies. The rest were discarded due to poor growth. Only cell colonies that showed reduced or no VPS29 expression upon dot blotting were maintained (Figure 4.17). The first round of dot blotting resulted in 12 promising colonies that showed reduced or no VPS29 expression. These were selected for a second round of dot blotting to avoid false positives. As a result, only 4 were positive in the second round and these cell colonies underwent Western blotting (performed by other lab members), all of which showed strong VPS29 expression. Therefore, no cell colonies were sent for Sanger sequencing.

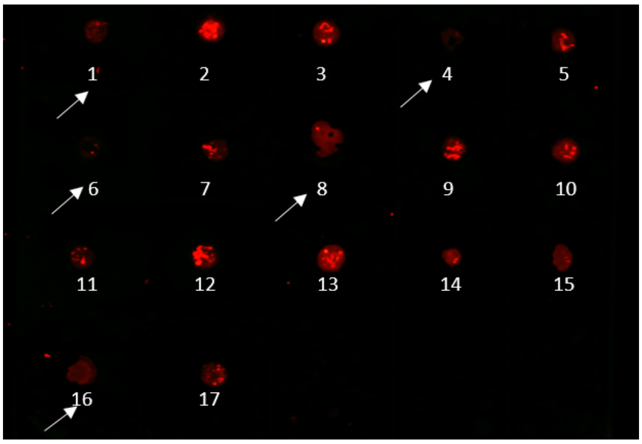


FIGURE 4.17: **Dot blot of VPS29 expression in CRISPR-Cas9 edited cell colonies**. HEK293T cells were transfected with CRISPR-Cas9 reagents in an effort to create a knockout VPS29 cell line. Cell colonies were collected, lysed and blotted on a nitrocellulose membrane with anti-VPS29 antibodies. Arrows show the colonies that demonstrated reduced/no VPS29 expression (potential gene knockdown/knockout) and that were maintained.

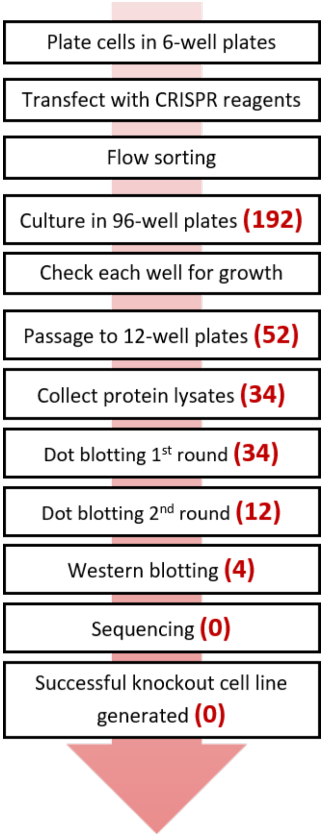


Plate cells in 6-well plates
Transfect with CRISPR reagents
Flow sorting
Culture in 96-well plates (192)
Check each well for growth
Passage to 12-well plates (52)
Collect protein lysates (34)
Dot blotting 1st round (34)
Dot blotting 2nd round (12)
Western blotting (4)
Sequencing (0)
Successful knockout cell line generated (0)

FIGURE 4.18: **Summary of the protocol and results of the *VPS29* knockout cell using CRISPR-Cas9**. The number of cell colonies (shown in parentheses) remaining after each stage of the CRISPR-Cas9 protocol.

Table 4.8: ***In silico* scores for candidate gene variants**

		Conservation					
Candidate	**Predicted deleterious**	*Conserved residue*	*Conserved flanking region*	*PhyloP & PhastCons*	**Natural variation**	**Expression brain/spinal cord**	**Function related to ALS**
Family FALS147 WES candidates (done prior)							
SAV1	8/10	1A	99.1%	3	low	yes	yes
DCAF7	1/10	1A	95.9%	3	medium	yes	possible
CYB5R3	9/10	1A	90.0%	0	medium	yes	yes
HERC2	3/10	1B	91.7%	3	low	yes	yes
PPP1R9A	2/10	3	64.6%	0	medium	yes	no
NDEL1	0/9	3	75.2%	3	medium	yes	no
FGD6	0/10	3	66.7%	0	low	low	no
AFF2	0/10	2	63.6%	3	low	low	yes
FBLN1	1/1	2	90.0%	3	high	low	no
ZZEF1	1/10	4	92.7%	3	high	yes	function unknown
NR1H4	2/9	4	76.0%	3	low	no	no
ZBTB8A	0/9	4	73.8%	3	low	no	no
ANKRD31	0/9	3	72.7%	0	medium	no	function unknown
DNAH8	8/10	1B	82.7%	3	high	no	no
OR2A25	5/10	1B	84.8%	0	high	no	no
ISM2	4/9	n/a*	n/a*	3	high	no	function unknown
TTN	2/9	2	81.8%	0	high	no	yes
PRAMEF12	0/10	2	63.6%	0	high	no	function unknown
Additional family FALS147 WGS candidate genes							
NUTM2A	1/10	2	62.9%	0	Medium	no	function unknown
C2CD4A	1/9	1B	100%	3	Medium	no	no
TERF2IP	1/11	3	57.8%	2	low	yes	no
ZNF580	7/11	1B	98.3%	0	High	yes	Maybe
Candidate gene screening in SALS							
DCAF7	3/8	3	72.7%	3	Medium	yes	yes
CYB5R3 (p.P73H)	11/11	1A	95%	3	High	yes	yes
CYB5R3 (p.S38N)	3/9	3	25%	0	High	yes	yes
CYB5R3 (p.Q28R)	2/9	3	0%	0	High	yes	yes
KIF5A	1/11	2	84.4%	2	Low	yes	yes
USP7	2/10	3	55%	3	Low	yes	yes
ABCD1	2/10	3	77%	2	Low	low	maybe
AUH	9/11	2	82%	3	Low	yes	yes
GBE1	10/11	3	97%	3	High	low	no
LAMA2	10/11	2	74.5%	3	High	low	maybe

5

Discussion

Overview

This thesis aimed to identify novel ALS genes from familial and sporadic cases. NGS analysis identified 22 candidate gene variants from family FALS147, among which the *CYB5R3*, *DCAF7*, and *SAV1* variants ranked highest using a comprehensive *in silico* pipeline. These three candidate gene variants were also assessed using general cell toxicity and cellular localisation analyses. While none of the mutant candidates showed increased toxicity compared to wild type, overexpression of CYB5R3-H95Y lead to the formation of large round cytoplasmic inclusions that were not seen in cells expressing CYB5R3-WT (Figure 4.16). Whole-genome sequencing data from 635 sporadic ALS patients was also interrogated for ALS candidate genes (n=21). Ten novel variants were identified, including a novel *CYB5R3* variant (p.P73H) which scored strongly as a potential pathogenic variant by *in silico* analyses. This discussion will first provide an overview of ALS gene discovery and NGS methods, followed by a discussion of the methods used during this project, the results obtained, and finally conclude with future work and directions. A discussion of the CRISPR-Cas9 protocol used and improvements for the future will also be discussed.

5.1 Gene discovery in ALS

To-date, gene mutations remain the only known cause of ALS. The pathological mechanisms underlying the disease are still poorly understood. Ongoing research to identify novel ALS genes is essential to provide insight and improve our understanding of the molecular mechanisms of the disease. Since the discovery of *SOD1* as the first ALS gene identified in 1993 (**Rosen et al.**, 1993), the development of sequencing technologies including recent NGS and bioinformatics pipelines, have largely facilitated gene discovery in ALS. For example, our laboratory recently used NGS data, genetic linkage and segregation analysis in a large ALS-FTD family to identify mutations in *CCNF* as a genetic cause of FALS (**Williams et al.**, 2016). Candidate gene screening was then carried out to determine whether *CCNF* mutations were also present in other ALS-FTD patients from different ethnic populations. This led to the identification of numerous novel *CCNF* mutations in both FALS and SALS, providing strong genetic evidence of the pathogenic nature of *CCNF* mutations. Despite great progress in the gene discovery field, known mutations only account for about 60% of FALS and 10% of SALS patients **McCann et al.** (2017). This shows the great need for ongoing research, using both novel gene discovery and candidate gene screening approaches, to find and validate novel ALS genes.

Novel gene discovery

Previously, most novel FALS genes were discovered using a combination of genetic linkage analysis, and candidate gene screening, mainly from large multi-generational families. Gene discovery methods have been very successful to date. However, remaining families are likely to have very rare and less penetrant variants, which makes novel gene discovery more difficult. Moreover, limited samples is available for these families. A late age of disease onset can also limit the availability of patient samples because individuals may already have died or might not show disease symptoms at time of collection. These factors can inhibit the ascertainment of accurate family histories and can lead to classification of familial cases as apparently sporadic. Therefore, the power for informative linkage and segregation analysis is restricted. NGS based analysis offers the best hope for gene discovery in smaller families, such as family FALS147 described in this thesis. However, NGS analysis from small families typically implicates many candidate genes, only one of which is truly disease causing. Therefore, this project made use of a combination of genetic, *in silico* and *in vitro* tools to help identify the disease causing gene by reducing the number of candidate genes through elimination of false-positive findings that arise from genetic analysis alone (**MacArthur et al.**, 2014).

In silico tools are crucial to rank and prioritise variants, while an *In vitro* pipeline can provide further experimental evidence of pathogenicity for candidate mutations.

The successful identification of the genetic cause of ALS in a family can benefit both the patients and their relatives. It provides an opportunity to prevent the disease gene mutation from being inherited by subsequent generations, through the use of genetic testing of embryos following *in vitro* fertilisation (IVF) before pregnancy. Moreover, genetic screening can provide early diagnosis if a known ALS mutation is identified within a patient, which is beneficial to improve patients quality of life. It can also allow for the selection of patients for specific clinical trials based on genetic cause. and allow for a more informed prognosis to be given while also impacting treatment decisions. Without genetic screening, diagnosis is currently difficult and mainly relies on observing disease progression.

Candidate gene screening

The recruitment and study of ALS families have been vital for the discovery of novel gene variants causing both FALS and SALS. For example, the *C9ORF72* locus was first mapped in FALS, and has since been recognised to be the most common cause of ALS, contributing to about 40% of FALS as well as about 6% of SALS in populations of European descent (**DeJesus-Hernandez et al.**, 2011; **Majounie et al.**, 2012; **Renton et al.**, 2011). Similarly, other ALS genes identified from the study of ALS families, including *SOD1* and *TADRBP* (**Chio et al.**, 2008; **Kabashi et al.**, 2008), have also been reported in SALS cases. It is possible that some of these SALS cases are FALS that have been wrongly classified due to the high variability in disease penetrance. Thus, the need for a candidate gene screening approach that examines both FALS and SALS cases.

This project screened 21 candidate genes acquired from three sources (shared variant analysis of family FALS147, recent publications and collaborators) in the Australian SALS cohort. Screening of FALS candidate genes seeked to provide more evidence for the pathogenic nature of disease. If mutations in these candidate genes exist in SALS patients, this provides further genetic evidence of a pathogenic role. Screening of newly published genes may validate those genes, particularly if the same mutations are found, and determine the incidence of mutations in Australian SALS. A list of candidate genes that are biologically relevant to ALS were generated from collaborators, and screened for the presence of additional mutations in our Australian SALS cohort. Finding mutations in those genes can help validate their role and assist in understanding associated disease mechanisms. NGS provides researchers with a cost-effective way to unravel remaining disease-causing genes. Finding new genes and their function will lead to a better understanding of ALS disease mechanisms and provide clues to the functional consequences of the mutations in ALS to develop drug therapies.

Whole-exome or whole-genome sequencing?

Choosing between WES and WGS depends on several factors including cost, aim of the study, data storage, data analysis and data quality (summarised in Table 5.1). Whole-exome sequencing encompasses exons of protein coding genes (about 2%), whereas WGS covers the whole genome. With the price of sequencing falling rapidly in recent years, the cost of WGS is approaching that of WES (**Majewski et al.**, 2011). Prior to this project, WES data from family FALS147 was analysed. For this project, newly acquired WGS data was also analysed to identify novel variants in coding genes that were missing from WES. As such, four novel gene variants were identified from WGS data analysis.

WGS offers a more consistent coverage of the genome, providing more accurate detection of structural variants, copy number variations, and genomic rearrangements. However, WGS provides about one hundred times more data than WES, necessitating more data storage, management and analysis (**Majewski et al.**, 2011; **Warr et al.**, 2015). WGS produces about 200 times more variants than WES, not only because it is sequencing more genetic material, but also because regions outside the exome are much less conserved (**Mu et al.**, 2011). This also results in the identification of many more benign variants. Furthermore, as there has been less research on non-coding regions, they are not well understood so determining the effect of a variant in these regions is difficult (**Mu et al.**, 2011). Only coding variants have been analysed during this project due to time constraints and a lack of information and tools to analyse non-coding variants. Nevertheless, non-coding regions will be examined for potentially pathogenic copy number variants (CNVs) and repeat expansions by our laboratory as a future project. Repeat expansions may be found using WGS data through tools that are increasingly in development (**Bahlo et al.**, 2018; **Tankard et al.**, 2017). *C9orf72* hexanucleotide repeat expansions are currently the most common cause of ALS (**DeJesus-Hernandez et al.**, 2011; **Renton et al.**, 2011) and analysing repeat expansions from WGS data can therefore be informative to find novel ALS genes.

For the moment, WES remains an effective approach for initial screening during novel gene discovery as most disease causing variants have been found in coding regions

TABLE 5.1: **Comparison of whole-exome and whole-genome sequencing**

Factors to consider	**Whole-exome sequencing**	**Whole-genome sequencing**
Sequencing data	2% of genome	Full genome
Estimated number of variants	20,000 variants	3.5million variants
Cost	Cheaper	Costly, but decreasing price
Coverage	Non-uniform	Uniform
Data storage	5-6GB storage	90GB storage
Data analysis	Easier	More challenging due to non-coding data
Sequencing time	Faster	Slower
Sequencing quality	Good	Better

(**Botstein and Risch**, 2003). However, WGS will ultimately replace WES as data storage costs drop and analysis pipelines improve so that disease-causing mutations in coding and non-coding regions can be discovered.

5.2 *In vitro* models of ALS

In situations where the power of NGS analysis is limited, functional data is crucial to assess the functional effect of putative novel gene mutations identified using NGS and *in silico* analysis. Indeed, most recent reports of novel ALS gene mutations have always been supported by cell biology and biochemical changes, including cellular mislocalisation and aggregation, abnormal ubiquitination of proteins, interaction with TDP-43, and induction of ER stress. For example, following the identification of *CCNF* mutations as a likely cause of ALS, our laboratory performed *in vitro* studies using neuronal cells to show that the expression of mutant CCNF caused abnormal ubiquitination and accumulation of ubiquitinated proteins including TDP-43 (**Williams et al.**, 2016). This provided *in vitro* evidence for the involvement of mutant *CCNF* in ALS pathogenesis and supported the role of abnormal protein homeostasis as the cause of disease.

The mouse NSC-34 (neuroblastoma/motor neuron-enriched primary spinal cord hybrid), neuro-2a (mouse neuroblastoma), SH-SY5Y (human neuroblastoma) and HEK293T (embryonic human kidney fibroblast) cell lines have commonly been used to study ALS-associated gene function and pathways (**Elden et al.**, 2010; **Williams et al.**, 2016). While these simple disease models do not mimic the complexity of human disease, they provide a relatively simple but powerful tool to assess whether candidate gene variants induce ALS-relevant phenotypes. They are also a powerful first step for assessing ALS disease mechanisms. Nevertheless, while *In vitro* analysis can be useful, especially in cases where an *in vitro* assay directly tests an established ALS disease mechanism, it has limitations. As mentioned, it does not accurately represent biological environment and alone, cannot directly prove causation of disease (**Duzkale et al.**, 2013). Moreover, overexpression models, as used during this project, may not reflect the effects of the endogenous gene mutation as protein overexpression can have unexpected side effects. For example, the greater abundance of the protein may be toxic or cells may gain functions they previously lacked, such as downregulation of DNA repair mechanisms, abnormal cell division, and compromised metabolism and growth **Krämer et al.** (2010).

To address these issues, genome editing techniques, such as CRISPR-Cas9, are developing and increasingly adopted for creating cell lines that endogenously express a target gene mutation. Since motor neurons cannot be collected and cultured directly from ALS patients, CRISPR-Cas9 has been useful to develop models of neurological diseases such as Huntington's disease **Xu et al.** (2017), frontotemporal dementia (FTD) **Nimsanor et al.** (2016), and ALS **Wang et al.** (2017). This project sought to develop a strategy using CRISPR-Cas9 to generate a *VPS29* (a candidate gene identified by our laboratory) knockout cell line for use in future functional studies.

5.3 Evaluation of methods used during this project

5.3.1 NGS analysis and bioinformatics

NGS data from family FALS147 were analysed using an established bioinformatics pipeline. Since most ALS genes found so far have an autosomal dominant pattern of inheritance, homozygous variants were also removed. This reduced the number of exonic variants by more than half (from 3214 to 1452), improving the ease of data analysis. Variants were filtered further using publicly available control databases, however, these databases have limitations. For example, the control databases contain a huge number of variants from many healthy control individuals, but medical information is often lacking. Control populations may include presymptomatic individuals with late onset diseases, asymptomatic individuals with low penetrance diseases, heterozygous carriers of recessive traits, and variants should not be assumed to be benign due to their presence in these control cohorts. In addition, many variants may be false-positives as the quality of variants called from NGS data can often vary and few have been validated using traditional PCR and Sanger sequencing. Indeed, known ALS mutations are present in ExAC and variants should therefore be filtered carefully.

Additionally, genetic variation, including genetic causes of ALS, is extremely diverse across populations and between different ethnic groups (**Cronin et al.**, 2007). Since family FALS147 is of European descent, variants identified in European (Non-Finnish) individuals were prioritised for filtering over variants found from other ethnic groups. To prevent the removal of pathogenic variants, singletons (variants present in a single individual) were not excluded, a minor allele frequency (MAF) threshold was set based on sample size (typically 0.0001), and each candidate gene variant was considered on a case-by-case basis. Furthermore, publicly available control databases are constantly being updated as NGS data is rapidly generated. Additional filtering carried out on family FALS147 WES-derived candidate genes using updated databases (gnomAD, MGRB, and ProjectMinE) resulted in the removal of three additional candidate gene variants (Section 4.1.5).

Uncertain Sanger sequencing results

Candidate NGS-derived variants underwent direct DNA sequencing for validation. Two variants, textitDCAF7 (p. c.G457C) and *NUTM2A* (c.C483A), gave uncertain results following Sanger sequencing (data not shown). For *DCAF7*, a second peak with the correct mutant allele (C) was present in the SALS patient, but the peak was not strong enough for a confident validation. Amplification of the same region in a healthy individual (control) only showed one clear peak (which was slightly higher than in the affected individual). For *NUTM2A*, a small second peak with the mutant allele (A) was present in the two affected individuals, but the second peak was not strong enough to fully validate the variant. This second peak was not present in the control married-in parent. Sequencing was repeated, with no changes in results.

Since traditional direct sequencing was inconclusive, a cloning step using pGEM-T vector could be added to determine the presence or absence of these variants. pGEM-T contains 3′-T overhangs at both ends and therefore allows for the easy cloning of any DNA fragment amplified with Taq polymerase (which leaves 3′-A overhangs at the end of the amplified DNA) (**Yao et al.**, 2016). As a result, ligation of the PCR product to pGEM-T vector readily occurs and the vector with separate gene inserts from both alleles can be sent for Sanger sequencing. Cloning makes it possible to distinguish different variants of a gene during sequencing, because only one allele of the gene is present in each clone. Sufficient clones should be sequenced to determine whether the two different alleles are present, as in theory, there is a 50% chance of obtaining either of the two alleles.

5.3.2 *In silico* functional analysis and variant prioritisation

The vast majority of the huge numbers of variants derived from NGS have unknown clinical significance. In the absence of sufficient genetic power, it is currently difficult to differentiate pathogenic variants from those that are benign. *In silico* tools can help to rank and assess potential pathogenicity of different variants. Yet, *in silico* analyses are predictive in nature, they can only assess variants based on currently available information while using algorithmic models. Considering multiple computational predictions and factors in combination, can improve the capacity to rank different candidate gene variants for potential pathogenicity. This is why this project made use of a combination of multiple *in silico* tools. However, these strategies provide supporting data as *in silico* analysis alone is not enough to conclude that a candidate gene variant is definitely pathogenic (**Duzkale et al.**, 2013).

The *in silico* analysis pipeline used in this project included protein prediction tools, conservation, gene natural variation, expression in brain/spinal cord and gene function. Eleven different protein prediction tools, which utilise different algorithms and factors (such as sequence conservation, homology information, structural and functional annotation of proteins) to predict potential protein pathogenicity were used. With different but related algorithms, protein prediction programs are neither directly comparable nor can they be regarded as independent evidence (**Duzkale et al.**, 2013). To address this issue, this project used the Combined Annotation Dependent Depletion (CADD) score, a novel tool that integrates results of many known prediction tools into one metric by comparing variants that survived natural selection with simulated mutations (**Kircher et al.**, 2014). CADD directly measures deleteriousness, in contrast to pathogenicity of molecular functionality. As such, CADD compensates for the incompleteness and bias of other existing methods, and provides a tool that allows users to have information combined from several factors. Nevertheless, it is important to keep in mind that hundreds to thousands of coding variants in an individual are usually labelled as potentially deleterious by these programs (**MacArthur et al.**, 2014).

Natural variation in humans may be a better indicator of pathogenicity in human diseases than protein conservation. Evolutionary conservation patterns are useful to prioritise candidate variants but many deleterious variants do not present with a strong conservation score, especially if the gene has undergone rapid evolution in humans. Genes that are intolerant to mutations (that is, have low natural variation in human control populations) indicate a vital role and most Mendelian disorders and ALS disease are caused by intolerant genes (**Petrovski et al.**, 2013). Assessing whether a candidate gene has a function that is related to a known ALS gene can also provide clues to the potential pathogenicity. However, it is important to note that a mutation may confer a gain of function to a candidate gene whose otherwise normal function may not be considered relevant to ALS.

During this project, candidate gene variants were prioritised by giving greater weight to natural variation, expression in brain/spinal cord and protein function, and less to protein prediction programs and conservation. A strength of this pipeline is the inclusion of numerous tools that can collectively implicate the potential pathogenicity of variants and reduce the chance of false positives. Using a comprehensive *in silico* pipeline remains the best way to rank candidate genes in situations where genetic power has been exhausted. To illustrate, our laboratory used this pipeline to assess candidate variants in family MQ1 and prioritise *VPS29* as the strongest candidate gene. *In vitro* analysis further supported this finding. Thus, the pipeline has been validated and is a strong strategy to assess, rank and filter candidate genes.

5.3.3 *In vitro* functional analysis of family FALS147 candidate genes

As discussed above, *in vitro* analyses are necessary to further assess the candidate gene variants identified from NGS and *in silico* analysis. Using cell studies is faster, allows for easy replication, and makes interpretation and deduction of mechanisms simpler as compared to the use of whole organisms (**Gruber and Hartung**, 2004). As such, *in vitro* analyses is an ideal starting point for testing the functional consequences of such candidates. In this project, a general toxicity assay (SYTOX blue) was used to determine cytotoxicity of candidate genes in HEK293T cells. While it is a useful assay, here it used transfection and therefore overexpression of candidates. It is possible that protein overexpression impacted cytotoxicity, including during overexpression of wild type candidate genes. Indeed, wild type candidate genes were significantly more toxic than non-transfected samples (Figure 4.13). Although non-significant, cytotoxicity was also higher in CYB5R3-WT and DCAF7-WT as compared to their mutant groups. Protein overexpression on its own can therefore have cytotoxicity effects.

The impact of candidate variants on potential ubiquitination, and interaction with endogenous TDP-43 were also assessed in HEK293T cells. Confocal microscopy can provide co-localisation information and is also the best way to examine protein aggregation. In the future, to further examine co-aggregation, immunoprecipitation (IP) with the candidate gene followed by western blotting staining with TDP-43 can be carried out. This project examined effects on endogenous TDP-43. In future, co-transfections of TDP-43 and candidate gene proteins could be performed to confirm any co-localisation of proteins when both are overexpressed.

Selecting a cell line for *in vitro* studies

HEK293T cell line was chosen to study protein localisation and cytotoxicity by overexpression of wild-type and mutant candidate proteins for this project. HEK293T is an embryonic human kidney fibroblast cell line that is easy to transfect, and which has widely been used in ALS research (**Elden et al.**, 2010; **Freischmidt et al.**, 2015). Due to time constraints, only HEK293T was used in this project to evaluate candidate genes. In practice, findings should be replicated in more than one cell type to provide confidence. NSC-34 and SH-SY5Y cell lines, which has been widely used as a cellular model system for motor neuron studies (**Cashman et al.**, 1992; **Matusica et al.**, 2008; **Zhou et al.**, 2013), could be used in future replication studies. However, the NSC-34 cell line is derived from mouse cells, which may influence the phenotype seen in cell culture, while SH-SY5Y originates from a human neuroblastoma cell line, but is harder to transfect and would require further optimisation. Fibroblasts offer a great alternative, are readily available from ALS patients and have been shown to display some ALS-associated characteristics, such as abnormal TDP-43 aggregate formation, protein mislocalisation and protein co-localisation (**Sabatelli et al.**, 2015; **Yang et al.**, 2015). A better alternative for model development would be the use of iPSC (induced pluripotent stem cells)-derived motor neurons. These could be obtained from the exact patient that genetic analysis is being performed on, and would provide a means to study candidate ALS genes in the correct patient and cell type. However, reprogramming and differentiation can be difficult and time consuming and may not represent adult neurons. Human primary neurons are also available in our laboratory and are a future option but typically prove difficult to transfect or sustain expression for significant lengths of time. Human primary neurons are also a very precious resource. Considering all the factors, HEK293T cell line was chosen as the cheapest, easiest and fastest strategy for initial *in vitro* screening of candidate genes.

Filtering out variants based on *in silico* and *in vitro* analysis

In silico and *in vitro* data can be used for prioritising purposes but it is often unclear when a candidate gene variant can be removed. Previous application of these pipelines to another ALS family in the gene discovery pipeline in our laboratory showed that *in* *silico* results can reflect in *in vitro* results (unpublished results). *In silico* analysis of the five candidate variants identified in family MQ1 found that only *VPS29* was predicted deleterious by 7/10 protein prediction programs, was highly conserved between species, had low natural variation in humans, was highly expressed in brain and spinal cord, and was functionally related to ALS. Therefore, this variant was ranked as the most likely to be pathogenic from five candidates. This was then reflected in results from the *in vitro* pipeline, where mutant VPS29 was significantly more toxic than wild type in cells, and produced aggregates when co-expressed with TDP-43. None of the other four candidates generated this ALS-like pathology in cells. Thus, the gene discovery pipeline used has been successful in the past to discover, rank, and test new candidate genes.

5.4 Evaluation of results obtained during this project

5.4.1 Family FALS147 WGS analysis

Family FALS147 WES analysis, completed prior to this project, led to the identification of 18 candidate gene variants. WGS data recently became available for the same family and this project aimed to find additional candidate gene variants. Following WGS analysis, 15 additional novel candidate gene variants were identified. However, all but four (*NUTM2A, CDC24A, TERF2IP*, and *ZNF580*) had extremely high allele counts as they were called in more than 90% of samples. Therefore, they were recognised as false-positives and artefacts of WGS analysis, and were removed from analysis.

5.4.2 Family FALS147 *In silico* results

NUTM2A, CDC24A, TERF2IP, and *ZNF580* were ranked using the *in silico* pipeline to explore potential pathogenicity and likelihood of being disease causing. While none of these four candidates ranked highly in all aspects of the *in silico* pipeline, the *ZNF580* variant was the most promising, predicted to be pathogenic by six (out of ten) *in silico* tools, with high sequence homology across many species, and high expression in brain/spinal cord. However, *ZNF580* was found to have high natural variation in humans (and is therefore tolerant to variation), which suggests the variant may be benign. However, this gene product is involved in cell proliferation and migration, as well as regulation of inflammation. Neuroinflammation can impair neuronal survival, increasing the progression and propagation of the degenerative process seen in ALS (**Papadimitriou et al.**, 2010). For example, mutation in the ALS gene (*TBK1*), which belongs to the IKK-kinase family of kinases involved in immunity signalling pathways, autophagy and mitophagy, increased neuroinflammation induced damage due to a loss of protective regulation by T cells (**Oakes et al.**, 2017). Thus, pathways that lead to a loss of neuroinflammation control may contribute to ALS. The other three candidate genes (*NUTM2A, C2CD4A, TERF2IP*) ranked low in most steps of the *in silico* analysis and rank low in the list of candidate genes for family FALS147. The remaining top ranked candidate genes identified from WES and WGS can now be considered for *in vitro* for functional studies. They should also be screened in SALS cohorts to determine whether additional candidate gene variants are present, supporting a pathogenic role in ALS.

Three candidate gene variants (*CYB5R3, DCAF7*, and *SAV1*) were identified as the most likely to cause ALS in family FALS147. Among these three candidates, the *SAV1* p.D219E variant was ranked the strongest. This variant was predicted to be pathogenic by 8/10 protein prediction tools, had high sequence homology across many species, low variation in humans, was ubiquitously expressed in both the brain and spinal cord, and had a function that is related to known ALS disease mechanisms. It was more challenging to rank the remaining variants, as they did not score highly in all aspects of the *in silico* analysis pipeline. For instance, *CYB5R3* p.H95Y also scored highly for most aspects of the pipeline, as it was predicted to be deleterious by 9/10 *in silico* tools, was ubiquitously expressed and had a function related to ALS disease mechanisms. However, *CYB5R3* had medium variation in humans and was not conserved. Similarly, *DCAF7* p.V252I ranked positively in conservation, expression and function, but had a medium natural variation score in humans and was predicted to be deleterious by only one prediction tool. In comparison, the candidate genes that ranked lowest were predicted pathogenic by few protein prediction programs, with low conservation scores, low expression in brain and spinal cord, high natural variation, and did not have a function related to known ALS genes.

5.4.3 Family FALS147 *In vitro* results

In vitro analysis was performed to assess the functional effects of the top FALS147 candidate gene variants, and determine if there was *in vitro* support for a role in disease by validation of strong *in silico* prediction scores. This analysis showed that the three candidate gene variants did not increase cytotoxicity when expressed in HEK293T cells. Moreover, there were no significant changes in protein localisation for SAV1 and DCAF7, while overexpression of CYB5R3-H95Y formed large round cytoplasmic inclusions in a subpopulation of cells (Figure 4.16). Additional functional studies are now required to further assess the *SAV1* and *DCAF7* variants, but due to time constraints assays such as immunoprecipitation and western blotting to investigate protein co-aggregation and protein solubility were not carried out at this stage. The known functions of SAV1, DCAF7 and CYB5R3 are described below.

CYB5R3

CYB5R3 encodes for cytochrome b5 reductase protein, which plays a vital role in desaturation of fatty acids in the body (**Siendones et al.**, 2014). It also contributes to regulation of cytosolic NAD+/NADH ratio, redox homeostasis, cellular senescence and aerobic metabolism (**Siendones et al.**, 2014). *CYB5R3* has different isoforms that can be transcribed into either a soluble form found in erythrocytes (isoforms 1 and 2), or a membrane-bound form in somatic cells anchored in endoplasmic reticulum, mitochondria and other membranes (isoform 3). These isoforms arise from tissue-specific splicing for alternative transcripts that differ in the first exon. Mutations in *CYB5R3* have been linked to autosomal recessive methaemoglobinaemia in humans, a severe disease characterised by decreased oxygen carrying capacity of the blood, mental deficiency and other neurologic symptoms (**Kugler et al.**, 2001; **Lorenzo et al.**, 2011). Type I methaemoglobinaemia is characterised by a deficiency in the soluble isoform while type II is caused by defects in both isoforms. CYB5R3 is the main effector of oxidative stress and in the desaturation of fatty acids. These pathways have commonly been associated with ALS (**Barber and Shaw**, 2010; **Schmitt et al.**, 2014).

SAV1

The salvador family WW domain containing protein 1 (*SAV1*) is part of the Hippo pathway, which also includes STE20-like 1/2 (MST1/2) and large tumor suppressor 1/2 (LATS1/2) (**Park and Lee**, 2011). MST interacts with SAV1 to enhance MST kinase activity and control cell proliferation, apoptosis, regeneration and differentiation (**Zhao et al.**, 2011). Dysregulation in the Hippo pathway has been implicated in several human cancers, as it usually has a tumor suppressor role by preventing cell proliferation and promoting apoptosis (**Park and Lee**, 2011). Interestingly, the Hippo pathway components have also been implicated in ALS, with increased activity of MST1 in motor neurons from SOD1 mice (**Lee et al.**, 2013). Moreover, depletion of SAV1/MST complex can promote ciliogenesis in cultured cells and induce ciliopathy phenotypes in zebrafish (**Kim et al.**, 2014). Cilia disruption has been liked to neurological defects such as hydrocephalus, brain dysgenesis and intellectual disability (**Kenna et al.**, 2016). The recently identified ALS susceptibility gene, *NEK1*, also plays a role in cilia formation (**Kenna et al.**, 2016; **Shalom et al.**, 2008).

DCAF7

DDB1 and CUL4 associated factors (DCAFs), including DCAF7, form part of the CUL4-DDB1 ubiquitin ligase, which regulates cell proliferation, survival, DNA repair and genomic integrity by ubiquitination of key regulators (**Lee and Zhou**, 2007). The main role of DCAFs is to help CUL4-DDB1 to recruit substrates for ubiquitination and are necessary for proper folding and assembly of a functional CUL4-DDB1-DCAF ubiquilin ligase complex (**Lee and Zhou**, 2007). In ALS, protein aggregates found in motor neurons are ubiquitinated, a major hallmark of the disease. Dysregulation in the ubiquitin proteosome system (UPS) is now widely recognised in ALS. Genes involved in the UPS have been found to be mutated in ALS, including *CCNF* and *UBQLN2* (**Deng et al.**, 2011; **Williams et al.**, 2016). Indeed, CCNF encodes a ubiquitin ligase, part of a ubiquitin ligase complex, that when mutated in ALS leads to UPS dysfunction (**Galper et al.**, 2017). *DCAF7* function is conserved in animals from invertebrates to vertebrates (**Morriss et al.**, 2013). The *DCAF7* zebrafish homologue, wdr68, is essential for craniofacial development, (**Nissen et al.**, 2006) and the *Drosophila* homologue, *wap*, is essential in muscle development (**Morriss et al.**, 2013).

5.4.4 Candidate gene screening in SALS WGS data

Screening of candidate genes obtained from family FALS147

As part of the gene screening strategy, *CYB5R3, DCAF7* and *SAV1* were screened in the 850VCF SALS cohort to find identical or additional candidate variants. Four novel variants were found, one in *DCAF7* and three in *CYB5R3*, each in different patients. Three of the four variants scored poorly after *in silico* analysis. The exception was the *CYB5R3* p.P73H variant, which was predicted pathogenic by 10/10 prediction programs, and was highly conserved between many species. However, as mentioned, natural variation of *CYB5R3* in humans was predicted to be medium, suggesting that novel nonsynonymous variants in this gene may be benign. However, natural variation was assessed for the gene as a whole, it remains possible that natural variation is low in the protein domain in which the variation is located.

The *CYB5R3* p.P73H variant identified in the SALS patient is located close (22 amino acid residues) to the *CYB5R3* p.H95Y variant found in family FALS147. As such, the protein Interpro database (`https://www.ebi.ac.uk/interpro/`) was used to inspect whether they are located within the same protein domain. Both variants are in the ferredoxin reductase-type FAD-binding domain (17-129aa), which is a member of the flavin adenine dinucleotide (FAD)-binding domain family. Flavoenzymes catalyse different biochemical reactions, such as metabolites, electron transfer to and from redox centres, and in the activation of oxidation and hydroxylation reactions. Work can now commence, including *in vitro* analysis, to assess whether this new variant (p.P73H) has functional significance to ALS.

Candidate gene screening *in silico* results

From the 18 candidate genes screened in the SALS whole-genome dataset, six novel candidate gene variants were identified and scored using the *in silico* pipeline (Table 4.8). Of these variants, the *AUH*, *GBE1* and *LAMA2* variants scored highest from protein prediction programs. These three candidate genes were obtained from a collaborator, as they are involved with white-matter diseases. However, only the *AUH* variant had low natural variation in humans, is expressed ubiquitously in the brain and spinal cord, and has a function related to ALS. *GBE1* and *LAMA2* variants both had a high natural variation in humans, suggesting that novel nonsynonymous variants in this gene are more likely to be benign. A novel variant was also identified in the white matter disease gene *ABCD1*, however this variant was ranked poorly by the *in silico* pipeline.

Novel variants were also identified in *KIF5A* and *USP7* candidate genes. They were predicted to be deleterious by few protein prediction tools, and were poorly conserved across species. However, natural variation in humans was predicted to be "low", and both *KIF5A* and *USP7* have functions related to ALS. *KIF5A* was recently identified as an ALS-associated gene following a large-scale GWAS study (**Nicolas et al.**, 2018). *KIF5A* encodes a member of the kinesin family of proteins, which play an essential role in axonal transport. Defects in axonal transport have been observed in ALS patients and are known to directly contribute to motor neuron degeneration (**Chevalier-Larsen and Holzbaur**, 2006; **Hirokawa et al.**, 2010). *USP7* candidate gene was obtained from a collaborator and of interest to ALS research due to its interaction with TDP-43 protein. *USP7* encodes for a ubiquitin specific protease, which cleaves ubiquitin from its substrates. Although more work needs to be done on these candidates, including *in vitro* analysis, these results show that candidate gene screening and *in silico* analysis pipelines are promising for the identification of additional gene variants that cause or are associated with ALS.

5.5 Generation of a *VPS29* knockout cell line using CRISPR-Cas9

The majority of the *in vitro* assays used in this project used overexpression techniques to study the effects of candidate gene mutations in cells. Overexpression has limitations as discussed previously, but these can be avoided by generating cell lines, using genome editing techniques, that carry the target gene mutation in the endogenous gene. Several gene silencing and gene editing tools exist including, ASOs, siRNA, meganucleases, ZFNs, TALENs, and more recently, CRISPR-Cas9. Each method has its own advantages and drawbacks in terms of specificity, simplicity, efficiency and toxicity, that need to be considered when designing an ALS disease model (Table 5.5). This project used the CRISPR-Cas9 technique for the first time in our laboratory, with the aim of developing a protocol that could be implemented to assess candidate genes in the future. Although a VPS29 knockout cell line was not generated in this project, numerous issues and improvements to the current protocol have been identified, as described below, that can be addressed in ongoing CRISPR-cell line development.

Despite starting with 192 flow sorted cell colonies, only 52 colonies were apt for dot blotting. Culturing from a single cell was difficult, many cells were unhealthy, dead, or stopped growing. A confounding factor of this was that the HEK293T cells had reached a high passage of ~35, which may have contributed to the poor health of cells noticed around the end of the CRISPR experiment. Cells grown at higher passage numbers are less viable and may develop mutations that alter the original functional characteristics and gene expression of the line, as documented on several occasions (**Sambuy et al.**, 2005; **Wenger et al.**, 2004; **Yu et al.**, 1997). A lower passage number of HEK293T cells can be used for future CRISPR-Cas9 experiments.

A more specific VPS29 antibody will also be required. The antibody used during this study appeared to lack specificity as shown by the dot blotting images (Figure 4.17) and western analysis (performed by other lab members), leading to false-positives. The screening of cells using the Atto488 fluorescent tag required a flow sorting instrument, which was time and labour intensive. Alternative methods can be tested, such as the use of an antibiotic resistance gene for selection of cells growing in that antibiotic. VPS29 knockout lines have previously been created using this method (**Kvainickas et al.**, 2017). Alternatively, if CRISPR-Cas9 editing disrupts a restriction site, digestion patterns using restriction enzymes would be altered, thus allowing for cell screening based

Table 5.2: **Comparison of various genome-editing tools**

Methods	**Key features**	**Pros**	**Cons**	**References**
Meganucleases	• DNA binding domain • Target length >14bp	• Allows gene knockout and editing • Occur naturally as nucleases • Low chances of off-target effects	• Protein overexpression • Very difficult to engineer • Time consuming	(**Nelson and Gersbach**, 2016; **Silva et al.**, 2011)
ZNF	• DNA binding domain • FokI endonucleases fused to zinc fingers interacts with DNA • Target length 18-36bp	• Allows gene knockout and editing	• Time consuming • Protein overexpression • Cytotoxic effects • Difficult to engineer • Efficiency varies • Costly • High chances of off-target effects	(**Nelson and Gersbach**, 2016; **Porteus and Baltimore**, 2003)
TALENs	• DNA binding domain • FokI endonucleases fused to TALE domains interacts with DNA • Target length 24-38bp	• Allows gene knockout and editing	• Time consuming • Costly • Protein overexpression • Cytotoxic effects • Efficiency varies • High chances of off-target effects	(**Chen et al.**, 2013; **Porteus and Baltimore**, 2003)
CRISPR-Cas9	• Guide RNA (gRNA) • Cas9 endonuclease recruited to DNA via gRNA • Target length about 22bp	• Allows gene knockout and editing • Occurs naturally as nucleases • Allows for easy multiplex gene targeting • Very easy to engineer • Fast • No protein overexpression • Long term expression • Low toxicity effects • Higher efficiency • Low cost	• High chances of off-target effects	(**Cong et al.**, 2013; **Fu et al.**, 2013; **Hsu et al.**, 2013)
CRISPR-Cas9 nickase	• Guide RNA • Cas9 nickase with paired sgRNAs bind to DNA	• Allows gene knockout and editing • Occurs naturally as nucleases • Allows for easy multiplex gene targeting • Very easy to engineer • No protein overexpression • Long term expression • Low toxicity effects • Higher efficiency • Low cost	• More time-consuming to engineer than CRISPR-Cas9 • Reduced efficiency compared to CRISPR-Cas9	(**Ren et al.**, 2014; **Shen et al.**, 2014)

on unique band patterns following PCR-restriction digestion and gel electrophoresis.

This project aimed to develop a gene knockout line as a preliminary step to development of a specific gene mutation knock-in model. Although creating gene knockouts is feasible using the developed protocol, making precise single-base changes or substitutions is more challenging. This is because non-homologous end joining (creation of insertion and deletions) occurs much more readily that homology-directed repair (precise and accurate repair) across eukaryotes, even with the introduction of a donor template (**Eid et al.**, 2018). Specificity of CRISPR-Cas9 gene editing tool can be improved by inactivating either the RuvC or NHN nuclease domains (Cas9 nucleases responsible for DNA cleavage) so that only one DNA strand is nicked at a time. Double-stranded break will only occur if both Cas9 nicking events take place on opposite strands at their specific target locations, leading to improved specificity up to 1,500X greater than wild-type Cas9 **Ran et al.** (2013); **Ren et al.** (2014); **Shen et al.** (2014). This system still makes use of the cells endogenous repair pathway upon encountering double-stranded breaks but with improved specificity.

Recently, gene editing without creating double-stranded DNA breaks have been developed, known as CRISPR base editors (**Eid et al.**, 2018), such as cytidine deaminase-based DNA base editors (**Komor et al.**, 2016; **Shimatani et al.**, 2017) and adenine deaminase-based DNA editors (**Gaudelli et al.**, 2017). Base editors are more efficient and have less off target effects. Moreover, because they do not use double stranded breaks, resulting indels and subsequent gene expression knockouts are avoided.

5.6 Future directions

This research utilised a novel disease gene discovery pipeline to find the genetic cause of ALS in family FALS147, as well as a candidate gene discovery approach to screen for gene mutations in apparently SALS patients. This project used an *in silico* pipeline to prioritise NGS-derived candidate gene variants for *in vitro* analysis. Due to time constraints, only the top three candidate genes were assessed in *in vitro* studies, with the *CYB5R3* p.P73H variant showing some support for a role in disease. Future work using additional *in silico*, *in vitro* and *in vivo* strategies are required to further assess the effects of these candidate genes. Moreover, the remaining 19 candidate genes discovered from WES and WGS analysis of family FALS147 should also be investigated. Future additional experiments are explored below.

Investigation of protein binding partners and changes in protein solubility

To further evaluate whether a candidate protein interacts with TDP-43 or other known ALS proteins, immunoprecipitation (IP) followed by western blotting could be performed. IP is used to isolate a target protein from a sample as well as proteins which are directly bound to the target protein. During IP, the targeted protein (i.e. the candidate protein) is collected using an antibody that specifically binds to that particular protein. This complex is then precipitated using beads, analysed by SDS-PAGE (to confirm target protein) and by western blotting to probe for TDP-43 or other ALS associated proteins of interest. The presence of TDP-43 for example, would prove a direct protein-protein interaction between the candidate protein and TDP-43. This technique would allow us to first determine if candidate proteins interact with known ALS proteins, and secondly whether there are any changes between the interaction of the wild type or mutant candidate protein with ALS associated proteins such as TDP-43. Western blotting could also be used to assess whether expression of the candidate variant leads to any shift in protein solubility by examining soluble and insoluble fractions of the cell. Previous research have demonstrated that ALS proteins often become increasingly insoluble and aggregate (**Basso et al.**, 2009; **Walker et al.**, 2015).

A proteomics approach could also be carried out to determine the protein composition of cell lysates transfected with wild type or mutant candidate genes to investigate any global or specific protein changes. Protein identification and quantification would be measured using mass spectrometry techniques (**Han et al.**, 2008). This approach was successfully used to identify changes in protein expression associated with wild type and mutant CCNF, as well as other cellular pathways that are disrupted by mutant CCNF (**Hogan et al.**, 2017). A proteomics approach would provide further insight regarding the biological functions and processes that are up- and/or down-regulated by the expression of mutant candidate genes.

Animal models

Animal models are essential to validate *in vitro* effects and to help elucidate the pathological mechanisms associated with new ALS genes. Multiple species are available for rapid assessment, however, the zebrafish are particularly useful for rapid assessment of motor neurons and motor phenotypes in a vertebrate species. Zebrafish that express motor neuron reporter constructs are readily available. Transient overexpression-based models can quickly be developed in these zebrafish by injecting candidate wild type and mutant mRNA into fish embryos. The Centre for MND Research has an established zebrafish facility, and our laboratory have validated protocols for this analysis.

For example, the ALS gene *CCNF* was successfully assessed in the zebrafish, showing axonopathy and reduced motor response in CCNF-mutant fish (**Hogan et al.**, 2017). Strong candidate genes from family FALS147, prioritised following *in vitro* analysis, can therefore be further assessed in future zebrafish studies.

CRISPR-Cas9 cell line models

The trial of genome editing using CRISPR-Cas9 during this project has identified issues with the current protocol and identified necessary improvements such as the use of more specific VPS29 antibodies, lower cell line passage number, and better flow sorting. Beyond the protocol development stage, work can commence to generate CRISPR-Cas9 edited cell lines that carry the candidate gene variants, rather than just knockout cell lines. The protocols would be similar, but with the addition of a template DNA containing the desired mutation to allow for homologous repair in the cells. This would allow for the generation of stable cell lines with desired mutations that could be used to study cytotoxicity, physiological and long-term effects in cells and limit the drawbacks associated with overexpression studies.

Investigation of *CYB5R3* candidate gene

As discussed, *CYB5R3* was flagged as a strong candidate gene obtained from NGS analysis in family FALS147. CYB5R3 is involved in redox homeostasis, oxidative stress and stress protection. *CYB5R3* can be further investigated genetically using burden analysis and association studies in our ALS cohort. Additional *in vitro* analyses including IP and Western blotting should also be performed, as described above, to investigate its binding partners (including known ALS proteins) and the impact of the candidate variants. Since oxidative stress is an important pathway in ALS, future experiments could also assess differential oxidative stress between wild type and mutant CYB5R3.

Additionally, mutation screening of *CYB5R3* in apparently SALS patients identified a strong *in silico* candidate variant (p.P73H). A detailed family history can be sought for this patient to determine whether they are an unrecognised familial case. As described previously, past experience shows us that such cases are not uncommon within apparently SALS cohorts. Also, the effects of this variant can now be scrutinised using *in vitro* studies. Q5 mutagenesis can be performed to introduce the desired variant in a pCMV6-Entry-CYB5R3 construct (purchased from OriGene during this project), following which subcloning into pCMV6-AC-RFP for assessment using the *in vitro* pipeline.

5.7 Conclusion

With improvements in NGS and bioinformatics, ALS gene discovery has accelerated in the past few years, contributing to the identification of causative gene mutations that now account for around 60% of Australian FALS cases. This study has demonstrated a novel gene discovery pipeline, using a combination of computational and functional analyses, to help to prioritise the numerous candidate genes that arise from NGS analysis of small ALS families. The *in vitro* strategies described here can provide further evidence to support causality for candidate genes and in the longer term, help determine the molecular mechanisms underlying disease and provide models for future development of treatments. It is important that new ALS genes are screened in apparently SALS patients, as this will help elucidate the genetic cause of those ALS cases that carry disease genes with reduced penetrance or are otherwise unrecognised familial cases. There is great potential for this pipeline to solve the genetic cause of disease in family FALS147 and other Australian ALS families with unknown gene mutations. Discovering new ALS genes will also be of vital clinical importance as genetic diagnostic screening can include these new genes so that early support and care can be provided to early- or even pre-symptomatic patients. Diagnosis will also have significance for other at-risk family members who may choose to utlilise preimplantation genetic diagnosis (PGD) following IVF to prevent their children from carrying pathogenic mutations. There is currently no cure for this devastating disease, but the development and testing of new treatments is underway thanks to the increasing knowledge derived from research into ALS genetics, pathology and molecular mechanisms.

A

Appendix

A.1 Unix scripts

A.1.1 gene_search.sh

The following shell script was used to find all variants present in candidate genes of interest from the entire cohort of genome sequenced samples (850VCF). Running this script resulted in an output text file that can be opened in Microsoft excel or R studio for further analysis. This script was useful to search for varaints in a single gene.

```
#!/bin/sh   e

# 850VCF_gene_search.sh (Emily McCann)
# Adapted by Sandrine Chan on 13/10/17.
#
#
   ########################################################################

#  this code is for looking a candidate gene variant in the
   850VCF
```

```
#
    ##################################################################################



# define gene name to search for
GENE="Gene_name"

# define output file names
OUT1="$GENE"_"variants.vcf"
OUT2="$GENE"_"variants_headed.vcf"
OUT3="$GENE"_"variants_headed.txt"

# perform gene search
awk  v NAME="$GENE" '$8 ~ NAME { print $0 }' $file > $OUT1
```

A.1.2 gene_search_pipe.sh

The following shell script had the same aim as in Script A.1.1, however, it used a pipe command to allow for several candidate genes to be interrogated simultaneously. The output of each gene was then saved into a single output file. Below is an example used to find candidate genes obtained from a collaborator: *AARS2, ABCD1, ARSA, AUH, CLN6, COQ2, CSF1R, CYP27A1, EIF2B5, GBE1, HTRA1, LAMA2, LMNB1, PLP1*, and *TREX1.*

```
#!/bin/sh  e
#  Written by Emily McCann
#  Adapted by Sandrine Chan on June2018.
################################################################
# perform gene search for multiple genes simultaneously, print
    the whole rows and save output
################################################################

awk '$8 ~ /AARS2|ABCD1|ARSA|AUH|CLN6|COQ2|CSF1R|CYP27A1|EIF2B5
    |GBE1|HTRA1|LAMA2|LMNB1|PLP1|TREX1/ { print $0 }' $file >
    multiple_candidate_genes.vcf
```

A.1.3 gene_variant_search.sh

This script was developed to find specific gene variants present in the 850VCF dataset based on chromosomal position. Specific gene variants that utilised this script included candidate gene variants identified in family FALS147 exome data. The table header and output results were combined using **cat** (concatenate) command and saved as a .txt file.

```
#!/bin/sh  e

#  filter.sh and 850_ALSOD_search (Emily McCann)
#  Adapted by Sandrine Chan on 28/02/18

#############################################
To search for gene variant by chromosome number and location
#############################################

# navigate to directory, open and denote the file
cd /"file"
file= "file"

#SAV1 variant search
awk '($1 ~ 14 && $2 ~ 51111611) {print $0}' $file > 850
   VCF_SAV1.txt

#DCAF7 variant search
awk '($1 ~ 17 && $2 ~ 61662588) {print $0}' $file > 850
   VCF_DCAF7.txt

#CYB5R3 variant search
awk '($1 ~ 22 && $2 ~ 43024269) {print $0}' $file > 850
   VCF_CYB5R3.txt

#PRAMEF12 variant search
awk '($1 ~ 1 && $2 ~ 12837663) {print $0}' $file > 850
   VCF_PRAMEF12.txt

#ZBTB8A variant search
```

```
awk '($1 ~ 1 && $2 ~ 33059103) {print $0}' $file > 850
   VCF_ZBTB8A.txt

#TTN variant search
awk '($1 ~ 2 && $2 ~ 179546450) {print $0}' $file > 850
   VCF_TTN.txt

#ANKRD31 variant search
awk '($1 ~ 5 && $2 ~ 74491500) {print $0}' $file > 850
   VCF_ANKRD31.txt

#DNAH8 variant search
awk '($1 ~ 6 && $2 ~ 38879296) {print $0}' $file > 850
   VCF_DNAH8.txt

#OR2A25 variant search
awk '($1 ~ 7 && $2 ~ 143771959) {print $0}' $file > 850
   VCF_OR2A25.txt

#PPP1R9A variant search
awk '($1 ~ 7 && $2 ~ 94540079) {print $0}' $file > 850
   VCF_PPP1R9A.txt

#FGD6 variant search
awk '($1 ~ 12 && $2 ~ 95604810) {print $0}' $file > 850
   VCF_FGD6.txt

#NR1H4 variant search
awk '($1 ~ 12 && $2 ~ 100904749) {print $0}' $file > 850
   VCF_NR1H4.txt

#ISM2 variant search
awk '($1 ~ 14 && $2 ~ 77950679) {print $0}' $file > 850
   VCF_ISM2.txt

#HERC2 variant search
```

```
awk '($1 ~ 15 && $2 ~ 28474716) {print $0}' $file > 850
   VCF_HERC2.txt

#NDEL1 variant search
awk '($1 ~ 17 && $2 ~ 8347611) {print $0}' $file > 850
   VCF_NDEL1.txt

#ZZEF1 variant search
awk '($1 ~ 17 && $2 ~ 3961306) {print $0}' $file > 850
   VCF_ZZEF1.txt

#FBLN1 variant search
awk '($1 ~ 22 && $2 ~ 45928969) {print $0}' $file > 850
   VCF_FBLN1.txt

#AFF2 variant search
awk '($1 ~ X && $2 ~ 147743972) {print $0}' $file > 850
   VCF_AFF2.txt

#The header and all variant lines into the one text file using
    the concatenate (cat) function
cat header.txt 850VCF_CYB5R3.txt 850VCF_PRAMEF12.txt 850
   VCF_ZBTB8A.txt 850VCF_TTN.txt 850VCF_ANKRD31.txt 850
   VCF_DNAH8.txt 850VCF_OR2A25.txt 850VCF_PPP1R9A.txt 850
   VCF_FGD6.txt 850VCF_NR1H4.txt 850VCF_ISM2.txt 850VCF_HERC2.
   txt 850VCF_NDEL1.txt 850VCF_ZZEF1.txt 850VCF_FBLN1.txt 850
   VCF_AFF2.txt > 850VCF_FALS147_all_candidates_headed.txt
```

A.1.4 gene_variant_search_pipe.sh

Alternatively, rather than inputting each candidate gene variant into an individual line, the pipe command was used to search for multiple genes. Below is the main script used for this purpose.

```
#!/bin/sh  e

#  Adapted by SandrineChan on 26/02/18.
# I have used the script created by Emily McCann
```

```
#####################################################
#  this code is for looking for FALS147 candidate gene
   variants in the 850VCF using a piping system (faster than
   doing one by one gene)
######################################################

# perform gene search for FALS147 candidate genes, print the
   whole rows and save as new vcf
awk '$8 ~ /FBLN1|SAV1|CYB5R3|NDEL1|DNAH8|NR1H4|DCAF7|OR2A25|
   AFF2|ZBTB8A|ISM2|HERC2|ZZEF1|PRAMEF12|PPP1R9A|FGD6|ANKRD31/
    { print $0 }' $file > FALS147_candidate_genes.vcf

#Now that I have pulled out all the candidate genes from the
   850VCF file, I need to find the specific variants using the
    position of the variant from the subset.
#I am using the filter.sh script from Emily McCann

#Find variants by position
awk ' ($2~
   /12837663|33059103|74491500|38879296|143771959|94540079|95604810|

100904749|77950679|51111611|28474716|61662588|8347611|3961306|43024269|

45928969|147743972/
{ print $0 }' FALS147_candidate_genes.vcf >
   FALS147_candidates_genes_variants.vcf
```

A.1.5 shared_variant.sh

This script was used to output variants that are common between the two affected individuals, and different to the control individual. The script needed troubleshooting. The last line in the script shows the successful code which utilised a different type of **awk** command.

```
#!/bin/sh  e

#  Adapted from filter.sh (Emily McCann).
#  Re adapted by Sandrine Chan on 27/02/18.
##########################################################
#Shared variant analysis for family FALS147 using the 2
    affected ($126 and $127) and 1 control ($128).
#This script aims to find all the 0/1 (variant present) in
    affected individuals and 0/0 (WT) in control , and save the
    output to a new file name
##########################################################

# navigate to folder
cd /FALS147

# This was the original code used, I thought this was correct
    but no nonsynonymous variants in the output
##awk ' ($126 ~ /^0\/1:/ && $127 ~ /^0\/1:/ && $128 ~
    /^0\/0:/) {print $0}' FALS147.txt > FALS147_shared_variants
    .txt

## Other codes were hence tried to fix the code ##

# 1) Look for non synonymous first , then do the shared variant
       nonsynonymous worked, but together with shared variant
    comes up as zero
#awk '$9 ~ "nonsynonymous" {print $0}' FALS147_headed.txt >
    FALS147_filtering_nonsyn.txt

# 2) let's add a space and see if this fixes the problem  Does
     not change anything
#awk ' ($126 ~ /^0\/1:/ && $127 ~ /^0\/1:/ && $128 ~ /^0\/0:/)
     {print $0}  ' FALS147.txt > FALS147_shared_variants_NEW.txt

# 3) this should give us variants with genotype 0/1 in all 3
    individuals this did something different , so something is
    working  but still not correct
#awk ' ($126 ~ /^0\/1:/ && $127 ~ /^0\/1:/ && $128 ~ /^0\/1:/)
     {print $0}  ' FALS147_.txt > FALS147_shared_variants_test.
    txt
```

```
# 4) Will try this without the brackets and without space  no
     difference found
#awk '$126 ~ /^0\/1:/ && $127 ~ /^0\/1:/ && $128 ~ /^0\/0:/ {
    print $0}' FALS147.txt > FALS147_shared_variants_NEW.txt

# 5) try curly brackets  did not work, gave the full genome
    file (9GB)
#awk '{$126 ~ /^0\/1:/ && $127 ~ /^0\/1:/ && $128 ~ /^0\/0:/}
     {print $0}' FALS147_.txt > FALS147_hared_variants_NEW1.txt

# 6) try IF command  file size is a bit bigger  seems to have
     nonsynonymous variants  This is the version that finally
    worked!!!!
awk F "\t" '{ if(($126 ~ /^0\/1:/) && ($127 ~ /^0\/1:/) && (
    $128 ~ /^0\/0:/)) { print } }' FALS147_.txt >
    FALS147_shared_variants_NEW2.txt


## Coding version no. 6 worked.
```

A.2 R scripts

A.2.1 SALS_subset.sh

This script has been used in R studio to pull out SALS cohort data from 850VCF, as well as to provide the details for a specific candidate gene. *ARPP21* candidate gene was used as an example in the script below. The script also removed variants that are not present or not called in the SALS subset.

```
#This is an R script, provided by Ingrid Tarr. I have adapted
    it for specific gene searches.

##################################################
#This is an example of how the script was used to find all
    ARPP21 variants present in the SALS subset only.
##################################################
```

```
# this is the file that has the IDs and project code
library(gdata)

full < read.xls("/Volumes/data_FMHS/Restrict/Blair\␣Group/
   Genetics/Project\␣MiNE/Manifests/Master_manifest 850
   sequenced_14 12 17␣JF.xlsx", header = T, sheet = 1,
   stringsAsFactors = F, nrow = 1000)

# sample ID and experiment code are the columns of interest
table(full$Blair.experiment.code)

SALS < full[full$Blair.experiment.code == "SALS", "SampleID"]
FALS < full[full$Blair.experiment.code == "FALS", "SampleID"]
SOD1 < full[full$Blair.experiment.code == "SOD1", "SampleID"]
twin < full[full$Blair.experiment.code == "Twin", "SampleID"]
twin_sod1 < full[full$Blair.experiment.code == "Twin SOD1", "
   SampleID"]


# change file path to variant file
ARPP21_full_850VCF < read.delim("/Volumes/data_FMHS/Restrict/
   Blair␣Group/Genetics/WGS_gene_searches/raw_data/
   HPC_resultant_txt/ARPP21_variants_headed.txt", header = T,
   skip = 0, nrows = 14000, sep = "\t")

# if there are still some samples referred to as WIL..... then
    these two lines should replace them with the MQIDs
# if there aren't then it won't run these two lines
matchup < if(length(grep("WIL", colnames(ARPP21_full_850VCF))
   ) > 0){
  read.xls("/Volumes/data_FMHS/Restrict/Blair\␣Group/Genetics/
     Project\␣MiNE/WIL\␣ID\␣conversion.xlsx", header = T,
     stringsAsFactors = F, col.names = c("tube", "mq", "wil",
     "manifest", "full_tube_id", "fastQ"))
}
```

```
colnames(ARPP21_full_850VCF)[grep(”WIL”, colnames(
   ARPP21_full_850VCF))] < if(length(grep(”WIL”, colnames(
   ARPP21_full_850VCF))) > 0){
  as.character(matchup[match(colnames(ARPP21_full_850VCF)[grep
     (”WIL”, colnames(ARPP21_full_850VCF))], matchup$wil), ”mq
     ”])
}

# tidying of the MQIDs in the variant data for matching
colnames(ARPP21_full_850VCF)[grep(”MQ160198”, colnames(
   ARPP21_full_850VCF))] < ”12 MQ160198”
colnames(ARPP21_full_850VCF) < gsub(”[[:punct:]]”, ” ”,
   colnames(ARPP21_full_850VCF))
colnames(ARPP21_full_850VCF) < gsub(”^X”, ””, colnames(
   ARPP21_full_850VCF))

# create a new data frame with the variant info for the group
   of interest:
ARPP21_SALS_850VCF < ARPP21_full_850VCF[, c(rep(T, 9),
   colnames(ARPP21_full_850VCF)[10:ncol(ARPP21_full_850VCF)] %
   in% SALS)]

# when saving, change the filepath
write.csv(ARPP21_SALS_850VCF, ”./ARPP21_SALS_850VCF.csv”)



##########################################
###Removing boring variants####
# Removed any variants that are reference or not called for
   all samples
##########################################

# This analysis may start with a .csv.
Boring < read.csv(”ARPP21_SALS_850VCF.csv”)

ind < apply(
  Boring[ , (which(colnames(Boring) == ”FORMAT”)+1):ncol(
     Boring)], # for the sample columns
  1, # look at each row
  function(u) !all(grepl(”0/0|\\./\\.”, u)) # and say if
     either 0/0 or ./. is found in ALL of the samples (and
     invert so T means keep  is at least one sample not
     wildtype, F means remove  all samples are wildtype or
     not called)
)


write.table(variant_table[ind, ], ”./
   ARPP21_SALS_filtered_850VCF.txt”, sep = ”\t”)

write.csv(variant_table[ind, ], ”./ARPP21_SALS_filtered_850VCF
   .csv”) #Saving it as a .csv

ARPP21_filtered_850VCF < read.csv(”
   ARPP21_SALS_filtered_850VCF.csv”)
```

A.2.2 gene_analysis_filtering

The aim of this script was to further filter exome and/or genome data to remove gene variants that do not meet the ”PASS” criteria, that are not exonic, that are synonymous, and that are not novel based on dbSNP147 and ExaC. This script used tools available from the dplyr and tidyverse packages in R studio. This considerably reduced the number of candidate gene variants.

```
#!/bin/sh

# Written by Sandrine Chan Moi Fat on 6/3/18. uses tools
   available in dplyr and tidyverse packages
####This script is to filter and look for new variants in
   genes/gene variants obtained from 850VCF (already subsetted
    to SALS only)####

#Set the working directory
```

```
setwd("/Volumes/shares/42794471/Sandrine_gene_850VCF_SALS/
   ARPP21/")

#Install and open packages
library(dplyr)
library(tidyverse)


#Import the data and assign name
ARPP21_filtered_850VCF < read_csv("
   ARPP21_SALS_filtered_850VCF.csv")

#Select only the ones with Filter=PASS (using pipe command
   from tidyr)
ARPP21_filtered_PASS < ARPP21_filtered_850VCF %>%
  filter(FILTER=="PASS")

#To check whether only PASS have been included, use the
   following.
ARPP21_filtered_PASS %>% group_by(FILTER) %>% tally()


#Find exonic only (n=16)
ARPP21_filtered_PASS_exonic < ARPP21_filtered_PASS [grep("
   Func.refGene=exonic", ARPP21_filtered_PASS$INFO), ]


#Remove synonymous variants (n=8)
ARPP21_filtered_PASS_exonic_nonsyn <
   ARPP21_filtered_PASS_exonic [grep("ExonicFunc.refGene=
   nonsynonymous", ARPP21_filtered_PASS_exonic$INFO), ]


#Are they novel in dbSNP147? (filter for dbSNP=.) (n=0)
ARPP21_filtered_PASS_exonic_nonsynonymous_novel147 <
   ARPP21_filtered_PASS_exonic_nonsyn[grep("avsnp147=\\.",
   ARPP21_filtered_PASS_exonic_nonsyn$INFO), ]
```

```
#No need to filter by Exac since no variants left... (n=0)
#ARPP21_filtered_PASS_exonic_nonsyn_novel147_novelExac <
   ARPP21_filtered_PASS_exonic_nonsynonymous_novel147[grep("
   ExAC_ALL=\\.",
   ARPP21_filtered_PASS_exonic_nonsynonymous_novel147$INFO), ]
#Just saving an empty dataset for records

 #Saving results as a new .txt and .csv
write.table(ARPP21_filtered_PASS_exonic_nonsynonymous_novel147
   "./ARPP21_filtered_PASS_exonic_nonsynonymous_novel147.txt",
    sep = "\t")
write.csv(ARPP21_filtered_PASS_exonic_nonsynonymous_novel147"
   ./ARPP21_filtered_PASS_exonic_nonsynonymous_novel147.csv")
```

A.3 Additional tables

This section shows the supplementary tables generated during this project.

Table A.1: **FALS147 candidate gene variants previously identified from WES analysis**

Gene	**Location**	**Accession no**	**Variant (DNA)**	**Variant (Protein)**
PRAMEF12	chr01:12837663	NM_001080830	c.G1373A	p.R458H
ZBTB8A	chr01:33059103	NM_001040441	c.A571G	p.K191E
TTN	chr02:179546450	NM_133378	c.T29378A	p.I9793N
CCDC50	chr03:191107301	NM_174908	c.A811G	p.M271V
IL17RE	chr03:9944363	NM_153483	c.C64A	p.P22T
ANKRD31	chr05:74491500	NM_001164443	c.T973C	p.F325L
LRRC70	chr05:61876755	NM_181506	c.A1490T	p.E497V
DNAH8	chr06:38879296	NM_001206927	c.G9793T	p.G3265C
OR2A25	chr07:143771959	NM_001004488	c.C647G	p.S216C
PPP1R9A	chr07:94540079	NM_001166163	c.C654G	p.I218M
FGD6	chr12:95604810	NM_018351	c.C250G	p.Q84E
NR1H4	chr12:100904749	NM_005123	c.G273T	p.M91I
ISM2	chr14:77950679	NM_182509	c.A614G	p.N205S
SAV1	chr14:51111611	NM_021818	c.T657A	p.D219E
HERC2	chr15:28474716	NM_004667	c.A5010C	p.E1670D
DCAF7	chr17:61662588	NM_005828	c.G754A	p.V252I
NDEL1	chr17:8347611	NM_001025579	c.G22C	p.D8H
ZZEF1	chr17:3961306	NM_015113	c.T5147C	p.V1716A
CYB5R3	chr22:43024269	NM_007326	c.C283T	p.H95Y
FBLN1	chr22:45928969	NM_006487	c.C571T	p.R191X
AFF2	chrX:147743972	NM_001169124	c.T724C	p.S242P

TABLE A.2: **Novel variants identified from candidate gene screening**

Gene	**Location**	**Acession no**	**Variant(DNA)**	**Variant (Protein)**
DCAF7	ch17:61657233	NM_005828	c.G457C	p.G153R
CYB5R3	ch22:43026934	NM_007326	c.C218A	p.P73H
CYB5R3	ch22:43040403	NM_001171660	c.G113A	p.S38N
CYB5R3	ch22:43040433	NM_001171660	c.A83G	p.Q28R
ABCD1	chrX:152990929	NM_000033	c.G208C	p.V70L
AARS2	chr6:44272510	NM_020745	c.G1624T	p.D542Y
AUH	chr9:93976697	NM_001698	c.C953T	p.T318I
AUH	chr9:94123951	NM_001698	c.A221C	p.E74A
CYP27A1	chr2:219679313	NM_000784	c.G1309T	p.A437S
GBE1	chr3:81643080	NM_000158	c.C1087T	p. L363F
GBE1	chr3:81754764	NM_000158	c.G144T	p.R48S
GBE1	chr3:81810577	NM_000158	c.T92A	p.L31H
HTRA1	chr10:124248938	NM_002775	c.G573C	p.K191N
LAMA2	chr6:129371083	NM_000426	c.A133C	p.N45H
LAMA2	chr6:129511430	NM_000426	c.T1548A	p.D516E
LAMA2	chr6:129571323	NM_000426	c.A1849C	p.T617P
LAMA2	chr6:129591890	NM_000426	c.C2444T	p.S815F
LAMA2	chr6:129641769	NM_000426	c.G4145T	p.C1382F
LAMA2	chr6:129748968	NM_000426	c.C5937G	p.N1979K
LAMA2	chr6:129766925	NM_000426	c.A6388T	p.I2130L
LAMA2	chr6:129785500	NM_000426	c.G7058C	p.R2353P
LAMA2	chr6:129799911	NM_001079823	c.C7513T	p.L2505F
KIF5A	chr12:57969899	NM_004984	c.A2053G	p.K685E
USP7	chr16:8994445	NM_003470	c.G2251C	V751L

TABLE A.3: ***In silico* scores for protein prediction**

Candidate gene	**SIFT**	**Polyphen2_HVAR**	**MutationTaster**	**MutationAssessor**	**FATHMM**	**PROVEAN**	**PonP2**	**PhD SNP**	**SNP & Go**	**PANTHER**	**CADD**	**Score**
NUTM2A	Tolerated (0.234)	Probably damaging (0.984)	Polymorphism (1)	Low (1.755)	Tolerated (1.99)	N (-1.2)	Neutral (0.062)	Neutral (0.250)	Neutral (0.055)	NA	Medium (10.15)	1/10
C2CD4A	Tolerated (0.393)	Benign (0.013)	Disease causing (0.999)	Low (1.65)	Tolerated (0.85)	N (-2.4)	Neutral (0.105)	Neutral (0.406)	Neutral (0.097)	NA	Medium (19.38)	1/9
TERF2IP	Tolerated (0.085)	Benign (0.005)	Polymorphism (1)	Medium (2.2)	Tolerated (0.87)	N (-0.43)	Neutral (0.199)	Neutral (0.069)	Neutral (0.037)	Neutral (0.357)	Low (2.191)	1/11
ZNF580	Deleterious (0.19)	Deleterious	Disease causing (1)	Neutral (0.745)	Tolerated (-1.13)	D (-3.79)	Pathogenic (0.883)	Disease (0.544)	Neutral (0.121)	Neutral (0.353)	High (34)	7/11
DCAF7	NA	Benign (0.029)	Disease causing (1)	Medium (2.905)	NA	NA	Unknown (0.673)	Disease (0.576)	Neutral (0.254)	Neutral (0.239)	Low (8.322)	3/8
CYB5R3 (p.P73H)	Deleterious (0)	Probably damaging (1)	Disease causing (1)	High (4.74)	Deleterious (-2.25)	Deleterious (-8.79)	Pathogenic (0.961)	Disease (0.925)	Disease (0.842)	Disease (0.962)	High (24.4)	11/11
CYB5R3 (p.S38N)	Tolerated (0.203)	Benign (0.043)	Polymorphism (1)	NA	Deleterious (-2.26)	Neutral (-0.26)	Neutral (0.120)	Disease (0.452)	Disease (0.154)	NA	Medium (11.14)	3/9
CYB5R3 (p.Q28R)	Deleterious (0.014)	Benign (0.047)	Polymorphism (1)	NA	Deleterious (-2.19)	neutral (-0.1)	Neutral (0.136)	Neutral (0.214)	Neutral (0.077)	NA	Low (0.001)	2/9
KIF5A	Tolerated (0.113	Benign (0.005	Disease causing (0.991)	Neutral (0.55)	Tolerated (-1.02)	Neutral (-1.86)	Unknown (0.432)	Neutral (0.420)	Neutral (0.334)	Neutral (0.278)	Medium (15.08)	1/11
USP7	Tolerated (0.321)	Benign (0.001)	Disease causing (1)	Neutral (-1.06)	Tolerated (3.34)	Neutral (-0.15)	Unknown (0.540)	Neutral (0.157)	Neutral (0.082)	NA	High(23.1)	2/10
ABCD1	Tolerated (0.155)	Benign (0.136)	Deleterious (0.991)	Neutral (-0.11)	Deleterious (-6.28)	Neutral (-0.7)	NA	Neutral (0.425)	Neutral (0.137)	Neutral (0.184)	Medium (9.999)	2/10
AUH	Deleterious (0)	Deleterious (0.998)	Deleterious (1)	Medium (3.335)	Tolerated (-1.23)	Deleterious (-5.26)	Pathogenic (0.851)	Disease (0.750)	Neutral (0.367)	Disease (0.812)	High (32)	9/11
GBE1	Deleterious (0.001)	Deleterious (0.988)	Deleterious (1)	High (4.415)	Deleterious (-2.19)	Deleterious (-3.31)	Unknown (0.588)	Disease (0.808)	Disease (0.624)	Disease (0.606)	High (24.4)	10/11
LAMA2	Deleterious (0)	Deleterious (0.999)	Deteterious (1)	High (4.06)	Tolerated (-1.29)	Deleterious (-4.51)	Pathogenic (0.910)	Disease (0.789)	Disease (0.744)	Disease (0.806)	High (25.7)	10/11

Table A.4: ***In silico* scores for natural variation**

Gene	**z score**	**RVIS score**	**Natural variation**
NUTM2A	3.95	NA (NA)	Medium
C2CD4A	3.51	NA (NA)	Medium
TERF2IP	1.4	- 0.16 (41.64%)	low
ZNF580	2.24	NA (NA)	High
DCAF7 (p.G153R)	3.16	NA (NA)	Medium
CYB5R3 (p.P73H)	-1.2	0.51 (80.2%)	High
CYB5R3 (p.S38N)	-1.2	0.51 (80.2%)	High
CYB5R3 (p.Q28R)	-1.2	0.51 (80.2%)	High
KIF5A	4.8	- 1.31 (4.88%)	Low
USP7	6.18	- 1.33 (4.67%)	Low
ABCD1	2.41	-0.54 (20.54%)	Low
AUH	0.57	-0.38 (27.40%)	Low
GBE1	-1.83	0.62 (83.53%)	High
LAMA2	-1.58	0.5 (79.64%)	High

Table A.5: ***In silico* scores for brain and spinal cord expression**

Gene	**HBT**	**TPM score**	**Expression score**
NUTM2A	NA	None	no
C2CD4A	NA	None	no
TERF2IP	High	High	yes
ZNF580	High	Medium	yes
DCAF7 (p.G153R)	High	Medium	yes
CYB5R3 (p.P73H)	High	High	yes
CYB5R3 (p.S38N)	High	High	yes
CYB5R3 (p.Q28R)	High	High	yes
KIF5A	High	Medium	yes
USP7	High	Medium	yes
ABCD1	Low	Low	low
AUH	High	Medium	yes
GBE1	High	Low	low
LAMA2	Low	Low	low

Table A.6: ***In silico* scores for conservation**

Gene	**Conservation across species**	**Conserved flanking region (%)**	**Conservation Score**	**phyloP100**	**phastCons**	**PhyloP & PhastCons Score**
NUTM2A	5 of 7	62.9	2	-0.328	0	0
C2CD4A	5 of 5	100	1B	5.652	1	3
TERF2IP	6 of 9	57.8	3	0.516	0.005	2
ZNF580	6 of 6	98.3	1B	-0.538	0	0
DCAF7	11 of 15	72.7	3	9.775	1	3
CYB5R3	10 of 10	75	1A	9.972	1	3
CYB5R3	0 of 10	24	3	-1.335	0	0
CYB5R3	0 of 10	0	3	-0.058	0	0
KIF5A	8 of 9	84.4	2	0.942	0.962	2
USP7	11 of 20	55	3	5.768	1	3
ABCD1	7 of 11	77	3	6.179	0.995	2
AUH	12 of 14	82	2	8.652	1	3
GBE1	17 of 20	97	3	4.613	1	3
LAMA2	9 of 11	74.5	2	8.889	1	3

Table A.7: **Complete list of all gene variants found during this project**

Gene	Location	Accession no	Variant (DNA)	Variant (Protein)	Obtained from
PRAMEF12	chr01:12837663	NM_001080830	c.G1373A	p.R458H	FALS147 whole-exome analysis
ZBTB8A	chr01:33059103	NM_001040441	c.A571G	p.K191E	FALS147 whole-exome analysis
TTN	chr02:179546450	NM_133378	c.T29378A	p.I9793N	FALS147 whole-exome analysis
CCDC50	chr03:191107301	NM_174908	c.A811G	p.M271V	FALS147 whole-exome analysis
IL17RE	chr03:9944363	NM_153483	c.C64A	p.P22T,	FALS147 whole-exome analysis
ANKRD31	chr05:74491500	NM_001164443	c.T973C	p.F325L	FALS147 whole-exome analysis
LRRC70	chr05:61876755	NM_181506	c.A1490T	p.E497V	FALS147 whole-exome analysis
DNAH8	chr06:38879296	NM_001206927	c.G9793T	p.G3265C	FALS147 whole-exome analysis
OR2A25	chr07:143771959	NM_001004488	c.C647G	p.S216C	FALS147 whole-exome analysis
PPP1R9A	chr07:94540079	NM_001166163	c.C654G	p.I218M	FALS147 whole-exome analysis
FGD6	chr12:95604810	NM_018351	c.C250G	p.Q84E	FALS147 whole-exome analysis
NR1H4	chr12:100904749	NM_005123	c.G273T	p.M91I	FALS147 whole-exome analysis
ISM2	chr14:77950679	NM_182509	c.A614G	p.N205S	FALS147 whole-exome analysis
SAV1	chr14:51111611	NM_021818	c.T657A	p.D219E	FALS147 whole-exome analysis
HERC2	chr15:28474716	NM_004667	c.A5010C	p.E1670D	FALS147 whole-exome analysis
DCAF7	chr17:61662588	NM_005828	c.G754A	p.V252I	FALS147 whole-exome analysis
NDEL1	chr17:8347611	NM_001025579	c.G22C	p.D8H	FALS147 whole-exome analysis
ZZEF1	chr17:3961306	NM_015113	c.T5147C	p.V1716A	FALS147 whole-exome analysis
CYB5R3	chr22:43024269	NM_007326	c.C283T	p.H95Y	FALS147 whole-exome analysis
FBLN1	chr22:45928969	NM_006487	c.C571T	p.R191X	FALS147 whole-exome analysis
AFF2	chrX:147743972	NM_001169124	c.T724C	p.S242P	FALS147 whole-exome analysis
MST1L	1:17085995	NM_001271733	c.C902G	p.A301G	FALS147 whole-genome analysis
NBPF19	1:145366193	NM_001351365	c.C8548A	p.Q2850K	FALS147 whole-genome analysis
OR2T33	1:248436972	NM_001004695	c.T145G	p.W49G	FALS147 whole-genome analysis
OR2T12	1:248458736	NM_001004692	c.T145G	p.W49G	FALS147 whole-genome analysis
FAM8A1	6:17601035	NM_016255	c.A395G	p.H132R	FALS147 whole-genome analysis
FAM8A1	6:17601040	NM_016255	c.G400A	p.G134S	FALS147 whole-genome analysis
FAM8A1	6:17601041	NM_016255	c.G401T	p.G134V	FALS147 whole-genome analysis
FAM8A1	6:17601044	NM_016255	c.T404C	p.L135P	FALS147 whole-genome analysis
FAM8A1	6:17601058	NM_016255	c.G418A	p.A140T	FALS147 whole-genome analysis
NUTM2A	10:88988120	NM_001099338	c.C483A	p.H161Q	FALS147 whole-genome analysis
CDHR5	11:618998	NM_001171968	c.C1543G	p.P515A	FALS147 whole-genome analysis
OR6S1	14:21109726	NM_001001968	c.C125G	p.T42R	FALS147 whole-genome analysis
C2CD4A	15:62359942	NM_207322	c.G130A	p.D44N	FALS147 whole-genome analysis
TERF2IP	16:75682129	NM_018975	c.G349A	p.A117T	FALS147 whole-genome analysis
ZNF580	19:56154346	NM_001163423	c.C472T	p.R158C	FALS147 whole-genome analysis
DCAF7	ch17:61657233	NM_005828	c.G457C	p.G153R	FALS147 candidate gene screening
CYB5R3	ch22:43026934	NM_007326	c.C218A	p.P73H	FALS147 candidate gene screening
CYB5R3	ch22:43040403	NM_001171660	c.G113A	p.S38N	FALS147 candidate gene screening
CYB5R3	ch22:43040433	NM_001171660	c.A83G	p.Q28R	FALS147 candidate gene screening
ABCD1	chrX:152990929	NM_000033	c.G208C	p.V70L	Candidate gene screening
AARS2	chr6:44272510	NM_020745	c.G1624T	p.D542Y	Candidate gene screening
AUH	chr9:93976697	NM_001698	c.C953T	p.T318I	Candidate gene screening
AUH	chr9:94123951	NM_001698	c.A221C	p.E74A	Candidate gene screening
CYP27A1	chr2:219679313	NM_000784	c.G1309T	p.A437S	Candidate gene screening
GBE1	chr3:81643080	NM_000158	c.C1087T	p. L363F	Candidate gene screening
GBE1	chr3:81754764	NM_000158	c.G144T	p.R48S	Candidate gene screening
GBE1	chr3:81810577	NM_000158	c.T92A	p.L31H	Candidate gene screening
HTRA1	chr10:124248938	NM_002775	c.G573C	p.K191N	Candidate gene screening
LAMA2	chr6:129371083	NM_000426	c.A133C	p.N45H	Candidate gene screening
LAMA2	chr6:129511430	NM_000426	c.T1548A	p.D516E	Candidate gene screening
LAMA2	chr6:129571323	NM_000426	c.A1849C	p.T617P	Candidate gene screening
LAMA2	chr6:129591890	NM_000426	c.C2444T	p.S815F	Candidate gene screening
LAMA2	chr6:129641769	NM_000426	c.G4145T	p.C1382F	Candidate gene screening
LAMA2	chr6:129748968	NM_000426	c.C5937G	p.N1979K	Candidate gene screening
LAMA2	chr6:129766925	NM_000426	c.A6388T	p.I2130L	Candidate gene screening
LAMA2	chr6:129785500	NM_000426	c.G7058C	p.R2353P	Candidate gene screening
LAMA2	chr6:129799911	NM_001079823	c.C7513T	p.L2505F	Candidate gene screening
KIF5A	chr12:57969899	NM_004984	c.A2053G	p.K685E	Candidate gene screening
USP7	chr16:8994445	NM_003470	c.G2251C	V751L	Candidate gene screening

Table A.8: **Complete list of all PCR validated gene variants found during this project**

Gene	**Location**	**Accession no**	**Variant (DNA)**	Variant (Protein)	**Obtained from**
PRAMEF12	chr01:12837663	NM_001080830	c.G1373A	p.R458H	FALS147 whole-exome analysis
ZBTB8A	chr01:33059103	NM_001040441	c.A571G	p.K191E	FALS147 whole-exome analysis
TTN	chr02:179546450	NM_133378	c.T29378A	p.I9793N	FALS147 whole-exome analysis
ANKRD31	chr05:74491500	NM_001164443	c.T973C	p.F325L	FALS147 whole-exome analysis
DNAH8	chr06:38879296	NM_001206927	c.G9793T	p.G3265C	FALS147 whole-exome analysis
OR2A25	chr07:143771959	NM_001004488	c.C647G	p.S216C	FALS147 whole-exome analysis
PPP1R9A	chr07:94540079	NM_001166163	c.C654G	p.I218M	FALS147 whole-exome analysis
FGD6	chr12:95604810	NM_018351	c.C250G	p.Q84E	FALS147 whole-exome analysis
NR1H4	chr12:100904749	NM_005123	c.G273T	p.M91I	FALS147 whole-exome analysis
ISM2	chr14:77950679	NM_182509	c.A614G	p.N205S	FALS147 whole-exome analysis
SAV1	chr14:51111611	NM_021818	c.T657A	p.D219E	FALS147 whole-exome analysis
HERC2	chr15:28474716	NM_004667	c.A5010C	p.E1670D	FALS147 whole-exome analysis
DCAF7	chr17:61662588	NM_005828	c.G754A	p.V252I	FALS147 whole-exome analysis
NDEL1	chr17:8347611	NM_001025579	c.G22C	p.D8H	FALS147 whole-exome analysis
ZZEF1	chr17:3961306	NM_015113	c.T5147C	p.V1716A	FALS147 whole-exome analysis
CYB5R3	chr22:43024269	NM_007326	c.C283T	p.H95Y	FALS147 whole-exome analysis
FBLN1	chr22:45928969	NM_006487	c.C571T	p.R191X	FALS147 whole-exome analysis
AFF2	chrX:147743972	NM_001169124	c.T724C	p.S242P	FALS147 whole-exome analysis
NUTM2A	10:88988120	NM_001099338	c.C483A	p.H161Q	FALS147 whole-genome analysis
C2CD4A	15:62359942	NM_207322	c.G130A	p.D44N	FALS147 whole-genome analysis
TERF2IP	16:75682129	NM_018975	c.G349A	p.A117T	FALS147 whole-genome analysis
ZNF580	19:56154346	NM_001163423	c.C472T	p.R158C	FALS147 whole-genome analysis
ABCD1	chrX:152990929	NM_000033	c.G208C	p.V70L	Collaborator candidate gene screening
AUH	chr9:93976697	NM_001698	c.C953T	p.T318I	Collaborator candidate gene screening
GBE1	chr3:81643080	NM_000158	c.C1087T	p. L363F	Collaborator candidate gene screening
LAMA2	chr6:129371083	NM_000426	c.A133C	p.N45H	Collaborator candidate gene screening
KIF5A	chr12:57969899	NM_004984	c.A2053G	p.K685E	Collaborator candidate gene screening
USP7	chr16:8994445	NM_003470	c.G2251C	V751L	Collaborator candidate gene screening
DCAF7	ch17:61657233	NM_005828	c.G457C	p.G153R	FALS147 candidate gene screening
CYB5R3	ch22:43026934	NM_007326	c.C218A	p.P73H	FALS147 candidate gene screening
CYB5R3	ch22:43040403	NM_001171660	c.G113A	p.S38N	FALS147 candidate gene screening
CYB5R3	ch22:43040433	NM_001171660	c.A83G	p.Q28R	FALS147 candidate gene screening

Ethics approval

Office of the Deputy Vice-Chancellor
(Research)

Research Office
Research Hub, Building C5C East
Macquarie University
NSW 2109 Australia
T: +61 (2) 9850 4459
http://www.research.mq.edu.au/
ABN 90 952 801 237

MACQUARIE
University
SYDNEY·AUSTRALIA

30 June 2016

Dear Associate Professor Blair

Reference No: 5201600387

Title: *Macquarie University Neurodegenerative Disease Biobank*

Thank you for submitting the above application for ethical and scientific review. Your application was considered by the Macquarie University Human Research Ethics Committee (HREC (Medical Sciences)).

I am pleased to advise that ethical and scientific approval has been granted for this project to be conducted at:

- Macquarie University

This approval is subject to the following conditions as determined by the HREC (Medical Sciences) Executive:

- For Macquarie University Researchers (internal) accessing samples from the Biobank to conduct research that fits within the scope of current HREC approval, approval will be sought from the Biobank Committee. The Biobank committee will provide quarterly reports to the HREC on the research being conducted using the samples and the investigators involved.
- For Macquarie University Researchers (internal) accessing samples from the Biobank to conduct research that does not fit within the scope of current HREC approval, ethics approval must be sought from the MDS HREC.
- For External researchers wishing to access samples from the Biobank, Biobank approval will be sought. The Biobank committee will provide the MDS HREC with:
    - Information on the research being conducted using the samples
    - An MTA covering transfer of samples
    - A copy of the external institutions ethics application and ethics approval covering use of samples.
- This will be done for each request made by an external researcher and research will not commence using the samples until approval by the MDS HREC has been finalised.

This research meets the requirements set out in the *National Statement on Ethical Conduct in Human Research* (2007 – Updated May 2015) (the *National Statement*).

Standard Conditions of Approval:

1. Continuing compliance with the requirements of the *National Statement,* which is available at the following website:

http://www.nhmrc.gov.au/book/national-statement-ethical-conduct-human-research

2. This approval is valid for five (5) years, subject to the submission of annual reports. Please submit your reports on the anniversary of the approval for this protocol.

3. All adverse events, including events which might affect the continued ethical and scientific acceptability of the project, must be reported to the HREC within 72 hours.

4. Proposed changes to the protocol and associated documents must be submitted to the Committee for approval before implementation.

It is the responsibility of the Chief investigator to retain a copy of all documentation related to this project and to forward a copy of this approval letter to all personnel listed on the project.

Should you have any queries regarding your project, please contact the Ethics Secretariat on 9850 4194 or by email ethics.secretariat@mq.edu.au

The HREC (Medical Sciences) Terms of Reference and Standard Operating Procedures are available from the Research Office website at:

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics

The HREC (Medical Sciences) wishes you every success in your research.

Yours sincerely

Professor Tony Eyers
Chair, Macquarie University Human Research Ethics Committee (Medical Sciences)

This HREC is constituted and operates in accordance with the National Health and Medical Research Council's (NHMRC) *National Statement on Ethical Conduct in Human Research* (2007) and the *CPMP/ICH Note for Guidance on Good Clinical Practice*.

Abbreviations

ALS Amyotrophic lateral sclerosis
ALS-FTD Amyotrophic lateral sclerosis -frontotemporal dementia
Bp Base pair
CRISPR Clustered regularly interspaced short palindromic repeats
CrRNA CRISPR RNA
DMEM Dulbeccos modified Eagle medium
DMEM/F12 Dulbeccos modified Eagle medium: Nutrient mixture F-12
DNA deoxyribonucleic acid
ER endoplasmic reticulum
ExAC Exome Aggregation Consortium
FALS Familial amyotrophic lateral sclerosis
FALS147 Family with amyotrophic lateral sclerosis number 147
FSC Forward scatter
FTD Frontotemporal dementia
GnomAD Genome Aggregation Database
IP Immunoprecipitation
iPSC Induced pluripotent stem cell
IVF *In vitro* fertilisation
LB Luria broth
LMN Lower motor neuron
MAF Minor allele frequency
MGRB Medical Genome Reference Bank
MND Motor neuron disease
NGS Next generation sequencing
NSC-34 Mouse motor neuron-like hybrid cell line
PAM Protospacer adjacent motif
PCR Polymerase chain reaction
PDG Preimplantation genetic diagnosis

RCF Relative centrifugal force
RFP Red fluorescent protein
RPM Revolutions per minute
RVIS Residual variation intolerance score
SALS Sporadic amyotrophic lateral sclerosis
SDS-PAGE Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEM Standard error of the mean
SH-SY5Y Human neuroblastoma cell line
SNP Single-nucleotide polymorphism
$\mathbf{T^A}$ Annealing temperature
TALENS Transcription activator-like effector nucleases
TBE Tris-borate EDTA
tracrRNA Trans-activating crRNA
TSAP Thermosensitive alkaline phosphate
UMN Upper motor neuron
UPS Ubiquitin-proteosome system
VCF Variant call format
WES Whole-exome sequencing
WGS Whole-genome sequencing
ZFN Zinc finger nuclease

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010), A method and server for predicting damaging missense mutations, *Nature methods*, 7, 4, 248

Al-Chalabi, A., Jones, A., Troakes, C., King, A., Al-Sarraj, S., and Van Den Berg, L. H. (2012), The genetics and neuropathology of amyotrophic lateral sclerosis, *Acta neuropathologica*, 124, 3, 339–352

Atkin, J. D., Farg, M. A., Walker, A. K., McLean, C., Tomas, D., and Horne, M. K. (2008), Endoplasmic reticulum stress and induction of the unfolded protein response in human sporadic amyotrophic lateral sclerosis, *Neurobiol. Dis.*, 30, 3, 400–407

Bahlo, M., Bennett, M. F., Degorski, P., Tankard, R. M., Delatycki, M. B., and Lockhart, P. J. (2018), Recent advances in the detection of repeat expansions with short-read next-generation sequencing, *F1000Res*, 7

Banerjee, R., Beal, M. F., and Thomas, B. (2010), Autophagy in neurodegenerative disorders: pathogenic roles and therapeutic implications, *Trends in neurosciences*, 33, 12, 541–549

Bannwarth, S., Ait-El-Mkadem, S., Chaussenot, A., Genin, E. C., Lacas-Gervais, S., Fragaki, K., et al. (2014), A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement, *Brain*, 137, Pt 8, 2329–2345

Barber, S. C. and Shaw, P. J. (2010), Oxidative stress in als: key role in motor neuron injury and therapeutic target, *Free Radical Biology and Medicine*, 48, 5, 629–641

Basso, M., Massignan, T., Samengo, G., Cheroni, C., De Biasi, S., Salmona, M., et al. (2006), Insoluble mutant SOD1 is partly oligoubiquitinated in amyotrophic lateral sclerosis mice, *J. Biol. Chem.*, 281, 44, 33325–33335

Basso, M., Samengo, G., Nardo, G., Massignan, T., D'Alessandro, G., Tartari, S., et al. (2009), Characterization of detergent-insoluble proteins in ALS indicates a causal link between nitrative stress and aggregation in pathogenesis, *PLoS ONE*, 4, 12, e8130

Boillée, S., Velde, C. V., and Cleveland, D. W. (2006), ALS: a disease of motor neurons and their nonneuronal neighbors, *Neuron*, 52, 1, 39–59

Botstein, D. and Risch, N. (2003), Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease, *Nat. Genet.*, 33 Suppl, 228–237

Boylan, K. (2015), Familial amyotrophic lateral sclerosis, *Neurologic clinics*, 33, 4, 807–830

Bozzoni, V., Pansarasa, O., Diamanti, L., Nosari, G., Cereda, C., and Ceroni, M. (2016), Amyotrophic lateral sclerosis and environmental factors, *Functional neurology*, 31, 1, 7

Brooks, B. R., Miller, R. G., Swash, M., and Munsat, T. L. (2000), El escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis, *Amyotrophic lateral sclerosis and other motor neuron disorders*, 1, 5, 293–299

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., and Casadio, R. (2009), Functional annotations improve the predictive score of human disease-related mutations in proteins, *Hum. Mutat.*, 30, 8, 1237–1244

Capriotti, E., Calabrese, R., and Casadio, R. (2006), Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information, *Bioinformatics*, 22, 22, 2729–2734

Cashman, N. R., Durham, H. D., Blusztajn, J. K., Oda, K., Tabira, T., Shaw, I. T., et al. (1992), Neuroblastoma x spinal cord (NSC) hybrid cell lines resemble developing motor neurons, *Dev. Dyn.*, 194, 3, 209–221

Chen, S., Oikonomou, G., Chiu, C. N., Niles, B. J., Liu, J., Lee, D. A., et al. (2013), A large-scale in vivo analysis reveals that talens are significantly more mutagenic than zfns generated using context-dependent assembly, *Nucleic acids research*, 41, 4, 2769–2778

Chevalier-Larsen, E. and Holzbaur, E. L. (2006), Axonal transport and neurodegenerative disease, *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1762, 11-12, 1094–1108

Chio, A., Traynor, B. J., Lombardo, F., Fimognari, M., Calvo, A., Ghiglione, P., et al. (2008), Prevalence of SOD1 mutations in the Italian ALS population, *Neurology*, 70, 7, 533–537

Cho, S. W., Kim, S., Kim, J. M., and Kim, J.-S. (2013), Targeted genome engineering in human cells with the cas9 rna-guided endonuclease, *Nature biotechnology*, 31, 3, 230

Choi, Y. and Chan, A. P. (2015), PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels, *Bioinformatics*, 31, 16, 2745–2747

Cirulli, E. T., Lasseigne, B. N., Petrovski, S., Sapp, P. C., Dion, P. A., Leblond, C. S., et al. (2015), Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways, *Science*, 347, 6229, 1436–1441

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., et al. (2013), Multiplex genome engineering using CRISPR/Cas systems, *Science*, 1231143

Cox, P. A., Kostrzewa, R. M., and Guillemin, G. J. (2018), BMAA and neurodegenerative illness, *Neurotoxicity research*, 33, 1, 178–183

Cronin, S., Hardiman, O., and Traynor, B. J. (2007), Ethnic variation in the incidence of ALS: A systematic review, *Neurology*, 68, 13, 1002–1007

Cruz, M. P. (2018), Edaravone (radicava): a novel neuroprotective agent for the treatment of amyotrophic lateral sclerosis, *Pharmacy and Therapeutics*, 43, 1, 25

D'Angiolella, V., Donato, V., Vijayakumar, S., Saraf, A., Florens, L., Washburn, M. P., et al. (2010), SCF(Cyclin F) controls centrosome homeostasis and mitotic fidelity through CP110 degradation, *Nature*, 466, 7302, 138–142

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., et al. (2011), Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS, *Neuron*, 72, 2, 245–256

Deng, H. X., Chen, W., Hong, S. T., Boycott, K. M., Gorrie, G. H., Siddique, N., et al. (2011), Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia, *Nature*, 477, 7363, 211–215

Ding, W.-X., Ni, H.-M., Gao, W., Yoshimori, T., Stolz, D. B., Ron, D., et al. (2007), Linking of autophagy to ubiquitin-proteasome system is important for the regulation of endoplasmic reticulum stress and cell viability, *The American journal of pathology*, 171, 2, 513–524

Duzkale, H., Shen, J., McLaughlin, H., Alfares, A., Kelly, M., Pugh, T., et al. (2013), A systematic approach to assessing the clinical significance of genetic variants, *Clinical genetics*, 84, 5, 453–463

Eid, A., Alshareef, S., and Mahfouz, M. M. (2018), Crispr base editors: genome editing without double-stranded breaks, *Biochemical Journal*, 475, 11, 1955–1964

Elden, A. C., Kim, H.-J., Hart, M. P., Chen-Plotkin, A. S., Johnson, B. S., Fang, X., et al. (2010), Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for als, *Nature*, 466, 7310, 1069

Englund, B., Brun, A., Gustafson, L., Passant, U., Mann, D., Neary, D., et al. (1994), Clinical and neuropathological criteria for frontotemporal dementia, *J Neurol Neurosurg Psychiatry*, 57, 4, 416–8

Farg, M. A., Konopka, A., Soo, K. Y., Ito, D., and Atkin, J. D. (2017), The DNA damage response (DDR) is induced by the C9orf72 repeat expansion in amyotrophic lateral sclerosis, *Hum. Mol. Genet.*, 26, 15, 2882–2896

Fecto, F. and Siddique, T. (2011), Making connections: pathology and genetics link amyotrophic lateral sclerosis with frontotemporal lobe dementia, *Journal of Molecular Neuroscience*, 45, 3, 663

Freischmidt, A., Wieland, T., Richter, B., Ruf, W., Schaeffer, V., Muller, K., et al. (2015), Haploinsufficiency of TBK1 causes familial ALS and fronto-temporal dementia, *Nat. Neurosci.*, 18, 5, 631–636

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., et al. (2013), High-frequency off-target mutagenesis induced by crispr-cas nucleases in human cells, *Nature biotechnology*, 31, 9, 822

Galper, J., Rayner, S. L., Hogan, A. L., Fifita, J. A., Lee, A., Chung, R. S., et al. (2017), Cyclin F: A component of an E3 ubiquitin ligase complex with roles in neurodegeneration and cancer, *Int. J. Biochem. Cell Biol.*, 89, 216–220

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., et al. (2017), Programmable base editing of AT to GC in genomic DNA without DNA cleavage, *Nature*, 551, 7681, 464

Greenway, M. J., Andersen, P. M., Russ, C., Ennis, S., Cashman, S., Donaghy, C., et al. (2006), ANG mutations segregate with familial and sporadic amyotrophic lateral sclerosis, *Nature genetics*, 38, 4, 411

Gruber, F. P. and Hartung, T. (2004), Alternatives to animal experimentation in basic research., *Altex*, 21, 3–31

Guo, W., Naujock, M., Fumagalli, L., Vandoorne, T., Baatsen, P., Boon, R., et al. (2017), HDAC6 inhibition reverses axonal transport defects in motor neurons derived from FUS-ALS patients, *Nature communications*, 8, 1, 861

Han, X., Aslanian, A., and Yates III, J. R. (2008), Mass spectrometry for proteomics, *Current opinion in chemical biology*, 12, 5, 483–490

Hirokawa, N., Niwa, S., and Tanaka, Y. (2010), Molecular motors in neurons: transport mechanisms and roles in brain function, development, and disease, *Neuron*, 68, 4, 610–638

Hitomi, J., Katayama, T., Eguchi, Y., Kudo, T., Taniguchi, M., Koyama, Y., et al. (2004), Involvement of caspase-4 in endoplasmic reticulum stress-induced apoptosis and Abeta-induced cell death, *J. Cell Biol.*, 165, 3, 347–356

Hogan, A. L., Don, E. K., Rayner, S. L., Lee, A., Laird, A. S., Watchon, M., et al. (2017), Expression of ALS/FTD-linked mutant CCNF in zebrafish leads to increased cell death in the spinal cord and an aberrant motor phenotype, *Hum. Mol. Genet.*, 26, 14, 2616–2626

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., et al. (2013), Dna targeting specificity of rna-guided cas9 nucleases, *Nature biotechnology*, 31, 9, 827

Hwang, W. Y., Fu, Y., Reyon, D., Maeder, M. L., Tsai, S. Q., Sander, J. D., et al. (2013), Efficient genome editing in zebrafish using a CRISPR-Cas system, *Nature biotechnology*, 31, 3, 227

Johnson, J. O., Mandrioli, J., Benatar, M., Abramzon, Y., Van Deerlin, V. M., Trojanowski, J. Q., et al. (2010), Exome sequencing reveals VCP mutations as a cause of familial ALS, *Neuron*, 68, 5, 857–864

Johnson, J. O., Pioro, E. P., Boehringer, A., Chia, R., Feit, H., Renton, A. E., et al. (2014), Mutations in the Matrin 3 gene cause familial amyotrophic lateral sclerosis, *Nat. Neurosci.*, 17, 5, 664–666

Kabashi, E., Valdmanis, P. N., Dion, P., Spiegelman, D., McConkey, B. J., Vande Velde, C., et al. (2008), TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis, *Nat. Genet.*, 40, 5, 572–574

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., et al. (2011), Spatio-temporal transcriptome of the human brain, *Nature*, 478, 7370, 483

Kaufman, R. J. (2002), Orchestrating the unfolded protein response in health and disease, *J. Clin. Invest.*, 110, 10, 1389–1398

Kenna, K. P., van Doormaal, P. T., Dekker, A. M., Ticozzi, N., Kenna, B. J., Diekstra, F. P., et al. (2016), NEK1 variants confer susceptibility to amyotrophic lateral sclerosis, *Nat. Genet.*, 48, 9, 1037–1042

Kieran, D., Woods, I., Villunger, A., Strasser, A., and Prehn, J. H. (2007), Deletion of the BH3-only protein puma protects motoneurons from ER stress-induced apoptosis and delays motoneuron loss in ALS mice, *Proc. Natl. Acad. Sci. U.S.A.*, 104, 51, 20606–20611

Kiernan, M. C., Vucic, S., Cheah, B. C., Turner, M. R., Eisen, A., Hardiman, O., et al. (2011), Amyotrophic lateral sclerosis, *The Lancet*, 377, 9769, 942–955

Kim, H. J., Kim, N. C., Wang, Y. D., Scarborough, E. A., Moore, J., Diaz, Z., et al. (2013), Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS, *Nature*, 495, 7442, 467–473

Kim, M., Kim, M., Lee, M. S., Kim, C. H., and Lim, D. S. (2014), The MST1/2-SAV1 complex of the Hippo pathway promotes ciliogenesis, *Nat Commun*, 5, 5370

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014), A general framework for estimating the relative pathogenicity of human genetic variants, *Nat. Genet.*, 46, 3, 310–315

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., and Liu, D. R. (2016), Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage, *Nature*, 533, 7603, 420

Krämer, O., Klausing, S., and Noll, T. (2010), Methods in mammalian cell line engineering: from random mutagenesis to sequence-specific approaches, *Applied microbiology and biotechnology*, 88, 2, 425–436

Kugler, W., Pekrun, A., Laspe, P., Erdlenbruch, B., and Lakomek, M. (2001), Molecular basis of recessive congenital methemoglobinemia, types I and II: Exon skipping and three novel missense mutations in the NADH-cytochrome b5 reductase (diaphorase 1) gene, *Hum. Mutat.*, 17, 4, 348

Kvainickas, A., Jimenez-Orgaz, A., Nagele, H., Hu, Z., Dengjel, J., and Steinberg, F. (2017), Cargo-selective SNX-BAR proteins mediate retromer trimer independent retrograde transport, *J. Cell Biol.*, 216, 11, 3677–3693

Kwiatkowski, T. J., Bosco, D., Leclerc, A., Tamrazian, E., Vanderburg, C., Russ, C., et al. (2009), Mutations in the fus/tls gene on chromosome 16 cause familial amyotrophic lateral sclerosis, *Science*, 323, 5918, 1205–1208

Lagier-Tourenne, C., Polymenidou, M., and Cleveland, D. W. (2010), TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration, *Hum. Mol. Genet.*, 19, R1, 46–64

Lee, J. and Zhou, P. (2007), DCAFs, the missing link of the CUL4-DDB1 ubiquitin ligase, *Mol. Cell*, 26, 6, 775–780

Lee, J. K., Shin, J. H., Hwang, S. G., Gwag, B. J., McKee, A. C., Lee, J., et al. (2013), MST1 functions as a key modulator of neurodegeneration in a mouse model of ALS, *Proc. Natl. Acad. Sci. U.S.A.*, 110, 29, 12066–12071

Leigh, P., Dodson, A., Swash, M., Brion, J.-P., and Anderton, B. (1989), Cytoskeletal abnormalities in motor neuron disease: an immunocytochemical study, *Brain*, 112, 2, 521–535

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016), Analysis of protein-coding genetic variation in 60,706 humans, *Nature*, 536, 7616, 285–291

Levine, T. P., Daniels, R. D., Gatta, A. T., Wong, L. H., and Hayes, M. J. (2013), The product of C9orf72, a gene strongly implicated in neurodegeneration, is structurally related to DENN Rab-GEFs, *Bioinformatics*, 29, 4, 499–503

Lillo, P., Mioshi, E., Burrell, J. R., Kiernan, M. C., Hodges, J. R., and Hornberger, M. (2012), Grey and white matter changes across the amyotrophic lateral sclerosis-frontotemporal dementia continuum, *PLoS ONE*, 7, 8, e43993

Ling, S. C., Polymenidou, M., and Cleveland, D. W. (2013), Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis, *Neuron*, 79, 3, 416–438

Lomen-Hoerth, C., Murphy, J., Langmore, S., Kramer, J., Olney, R., and Miller, B. (2003), Are amyotrophic lateral sclerosis patients cognitively normal?, *Neurology*, 60, 7, 1094–1097

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013), The genotype-tissue expression (GTEx) project, *Nature genetics*, 45, 6, 580

Lorenzo, F. R., Phillips, J. D., Nussenzveig, R., Lingam, B., Koul, P. A., Schrier, S. L., et al. (2011), Molecular basis of two novel mutations found in type I methemoglobinemia, *Blood Cells Mol. Dis.*, 46, 4, 277–281

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., et al. (2014), Guidelines for investigating causality of sequence variants in human disease, *Nature*, 508, 7497, 469–476

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A., and Jabado, N. (2011), What can exome sequencing do for you?, *J. Med. Genet.*, 48, 9, 580–589

Majounie, E., Renton, A. E., Mok, K., Dopper, E. G., Waite, A., Rollinson, S., et al. (2012), Frequency of the c9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study, *The Lancet Neurology*, 11, 4, 323–330

Marangi, G. and Traynor, B. J. (2015), Genetic causes of amyotrophic lateral sclerosis: new genetic analysis methodologies entailing new opportunities and challenges, *Brain research*, 1607, 75–93

Maruyama, H., Morino, H., Ito, H., Izumi, Y., Kato, H., Watanabe, Y., et al. (2010), Mutations of optineurin in amyotrophic lateral sclerosis, *Nature*, 465, 7295, 223–226

Matusica, D., Fenech, M. P., Rogers, M.-L., and Rush, R. A. (2008), Characterization and use of the nsc-34 cell line for study of neurotrophin receptor trafficking, *Journal of neuroscience research*, 86, 3, 553–565

Maurel, C., Dangoumau, A., Marouillat, S., Brulard, C., Chami, A., Hergesheimer, R., et al. (2018), Causative genes in amyotrophic lateral sclerosis and protein degradation pathways: a link to neurodegeneration, *Molecular neurobiology*, 1–20

McCann, E. P., Williams, K. L., Fifita, J. A., Tarr, I. S., O'Connor, J., Rowe, D. B., et al. (2017), The genotype–phenotype landscape of familial amyotrophic lateral sclerosis in australia, *Clinical genetics*, 92, 3, 259–266

Miller, R. G., Mitchell, J., and Moore, D. H. (2012), Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND), *The Cochrane Library*

Morita, M., Al-Chalabi, A., Andersen, P., Hosler, B., Sapp, P., Englund, E., et al. (2006), A locus on chromosome 9p confers susceptibility to ALS and frontotemporal dementia, *Neurology*, 66, 6, 839–844

Morriss, G. R., Jaramillo, C. T., Mikolajczak, C. M., Duong, S., Jaramillo, M. S., and Cripps, R. M. (2013), The drosophila wings apart gene anchors a novel, evolutionarily conserved pathway of neuromuscular development, *Genetics*, 195, 3, 927–940

Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y., and Gerstein, M. B. (2011), Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project, *Nucleic Acids Res.*, 39, 16, 7058–7076

Nandi, D., Tahiliani, P., Kumar, A., and Chandu, D. (2006), The ubiquitin-proteasome system, *Journal of biosciences*, 31, 1, 137–155

Nelson, C. E. and Gersbach, C. A. (2016), Engineering delivery vehicles for genome editing, *Annual review of chemical and biomolecular engineering*, 7, 637–662

Neumann, M., Sampathu, D. M., Kwong, L. K., Truax, A. C., Micsenyi, M. C., Chou, T. T., et al. (2006), Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis, *Science*, 314, 5796, 130–133

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

Nicolas, A., Kenna, K. P., Renton, A. E., Ticozzi, N., Faghri, F., Chia, R., et al. (2018), Genome-wide analyses identify KIF5A as a novel ALS gene, *Neuron*, 97, 6, 1268–1283

Nimsanor, N., Poulsen, U., Rasmussen, M. A., Clausen, C., Mau-Holzmann, U. A., Nielsen, J. E., et al. (2016), Generation of an isogenic, gene-corrected ipsc line from a symptomatic 59-year-old female patient with frontotemporal dementia caused by an r406w mutation in the microtubule associated protein tau (mapt) gene, *Stem cell research*, 17, 3, 576–579

Niroula, A., Urolagin, S., and Vihinen, M. (2015), PON-P2: prediction method for fast and reliable identification of harmful variants, *PloS one*, 10, 2, e0117380

Nissen, R. M., Amsterdam, A., and Hopkins, N. (2006), A zebrafish screen for craniofacial mutants identifies wdr68 as a highly conserved gene required for endothelin-1 expression, *BMC Dev. Biol.*, 6, 28

Oakes, J. A., Davies, M. C., and Collins, M. O. (2017), TBK1: a new player in ALS linking autophagy and neuroinflammation, *Mol Brain*, 10, 1, 5

Papadimitriou, D., Le Verche, V., Jacquier, A., Ikiz, B., Przedborski, S., and Re, D. B. (2010), Inflammation in als and sma: sorting out the good from the evil, *Neurobiology of disease*, 37, 3, 493–502

Park, B. H. and Lee, Y. H. (2011), Phosphorylation of SAV1 by mammalian ste20-like kinase promotes cell death, *BMB Rep*, 44, 9, 584–589

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., and Goldstein, D. B. (2013), Genic intolerance to functional variation and the interpretation of personal genomes, *PLoS Genet.*, 9, 8, e1003709

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010), Detection of nonneutral substitution rates on mammalian phylogenies, *Genome research*, 20, 1, 110–121

Porteus, M. H. and Baltimore, D. (2003), Chimeric nucleases stimulate gene targeting in human cells, *Science*, 300, 5620, 763–763

Pourcel, C., Salvignol, G., and Vergnaud, G. (2005), CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies, *Microbiology (Reading, Engl.)*, 151, Pt 3, 653–663

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005), Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic acids research*, 33, suppl_1, D501–D504

Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., et al. (2013), Double nicking by rna-guided crispr cas9 for enhanced genome editing specificity, *Cell*, 154, 6, 1380–1389

Ren, X., Yang, Z., Mao, D., Chang, Z., Qiao, H.-H., Wang, X., et al. (2014), Performance of the cas9 nickase system in drosophila melanogaster, *G3: Genes, Genomes, Genetics*, 4, 10, 1955–1962

Renton, A. E., Chiò, A., and Traynor, B. J. (2014), State of play in amyotrophic lateral sclerosis genetics, *Nature neuroscience*, 17, 1, 17

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., et al. (2011), A hexanucleotide repeat expansion in C9orf72 is the cause of chromosome 9p21-linked ALS-FTD, *Neuron*, 72, 2, 257–268

Reva, B., Antipin, Y., and Sander, C. (2011), Predicting the functional impact of protein mutations: application to cancer genomics, *Nucleic acids research*, 39, 17, e118–e118

Rosen, D. R., Siddique, T., Patterson, D., Figlewicz, D. A., Sapp, P., Hentati, A., et al. (1993), Mutations in cu/zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis, *Nature*, 362, 6415, 59

Rowland, L. P. and Shneider, N. A. (2001), Amyotrophic lateral sclerosis, *New England Journal of Medicine*, 344, 22, 1688–1700

Sabatelli, M., Zollino, M., Conte, A., Del Grande, A., Marangi, G., Lucchini, M., et al. (2015), Primary fibroblasts cultures reveal TDP-43 abnormalities in amyotrophic lateral sclerosis patients with and without SOD1 mutations, *Neurobiol. Aging*, 36, 5, 5–2005

Sambuy, Y., De Angelis, I., Ranaldi, G., Scarino, M., Stammati, A., and Zucco, F. (2005), The Caco-2 cell line as a model of the intestinal barrier: influence of cell and culture-related factors on Caco-2 cell functional characteristics, *Cell biology and toxicology*, 21, 1, 1–26

Sasaki, S. (2010), Endoplasmic reticulum stress in motor neurons of the spinal cord in sporadic amyotrophic lateral sclerosis, *J. Neuropathol. Exp. Neurol.*, 69, 4, 346–355

Schmitt, F., Hussain, G., Dupuis, L., Loeffler, J. P., and Henriques, A. (2014), A plural role for lipids in motor neuron diseases: energy, signaling and structure, *Front Cell Neurosci*, 8, 25

Schwab, C., Arai, T., Hasegawa, M., Yu, S., and McGeer, P. L. (2008), Colocalization of transactivation-responsive DNA-binding protein 43 and huntingtin in inclusions of Huntington disease, *J. Neuropathol. Exp. Neurol.*, 67, 12, 1159–1165

Schwarz, J. M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010), Mutationtaster evaluates disease-causing potential of sequence alterations, *Nature methods*, 7, 8, 575

Shalom, O., Shalva, N., Altschuler, Y., and Motro, B. (2008), The mammalian Nek1 kinase is involved in primary cilium formation, *FEBS Lett.*, 582, 10, 1465–1470

Shen, B., Zhang, W., Zhang, J., Zhou, J., Wang, J., Chen, L., et al. (2014), Efficient genome modification by crispr-cas9 nickase with minimal off-target effects, *Nature methods*, 11, 4, 399

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., et al. (2013), Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models, *Hum. Mutat.*, 34, 1, 57–65

Shimatani, Z., Kashojiya, S., Takayama, M., Terada, R., Arazoe, T., Ishii, H., et al. (2017), Targeted base editing in rice and tomato using a CRISPR-Cas9 cytidine deaminase fusion, *Nature biotechnology*, 35, 5, 441

Siddique, T., Figlewigz, D. A., Pericak-Vance, M. A., Haines, J. L., Rouleau, G., Jeffers, A. J., et al. (1991), Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity, *New England Journal of Medicine*, 324, 20, 1381–1384

Siendones, E., SantaCruz-Calvo, S., Martin-Montalvo, A., Cascajo, M. V., Ariza, J., Lopez-Lluch, G., et al. (2014), Membrane-bound CYB5R3 is a common effector of nutritional and oxidative stress response through FOXO3a and Nrf2, *Antioxid. Redox Signal.*, 21, 12, 1708–1725

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005), Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, *Genome research*, 15, 8, 1034–1050

Silva, G., Poirot, L., Galetto, R., Smith, J., Montoya, G., Duchateau, P., et al. (2011), Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy, *Current gene therapy*, 11, 1, 11–27

Sreedharan, J., Blair, I. P., Tripathi, V. B., Hu, X., Vance, C., Rogelj, B., et al. (2008), TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis, *Science*, 319, 5870, 1668–1672

Talbot, K. (2009), Motor neuron disease: the bare essentials, *Practical neurology*, 9, 5, 303–309

Tankard, R. M., Delatycki, M. B., Lockhart, P. J., and Bahlo, M. (2017), Detecting known repeat expansions with standard protocol next generation sequencing, towards developing a single screening test for neurological repeat expansion disorders, *bioRxiv*, 157792

Taylor, J. P., Brown Jr, R. H., and Cleveland, D. W. (2016), Decoding ALS: from genes to mechanism, *Nature*, 539, 7628, 197

Thomas, P. D. and Kejariwal, A. (2004), Coding single-nucleotide polymorphisms associated with complex vs. mendelian disease: evolutionary evidence for differences in molecular effects, *Proceedings of the National Academy of Sciences*, 101, 43, 15398–15403

Turner, B. J. and Atkin, J. D. (2006), ER stress and UPR in familial amyotrophic lateral sclerosis, *Curr. Mol. Med.*, 6, 1, 79–86

Vance, C., Rogelj, B., Hortobágyi, T., De Vos, K. J., Nishimura, A. L., Sreedharan, J., et al. (2009), Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6, *Science*, 323, 5918, 1208–1211

Walker, A. K., Tripathy, K., Restrepo, C. R., Ge, G., Xu, Y., Kwong, L. K., et al. (2015), An insoluble frontotemporal lobar degeneration-associated tdp-43 c-terminal fragment causes neurodegeneration and hippocampus pathology in transgenic mice, *Human molecular genetics*, 24, 25, 7241–7254

Wang, H. Y., Wang, I. F., Bose, J., and Shen, C. K. (2004), Structural diversity and functional implications of the eukaryotic TDP gene family, *Genomics*, 83, 1, 130–139

Wang, L., Yi, F., Fu, L., Yang, J., Wang, S., Wang, Z., et al. (2017), Crispr/cas9-mediated targeted gene correction in amyotrophic lateral sclerosis patient ipscs, *Protein & cell*, 8, 5, 365–378

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., and Watson, M. (2015), Exome Sequencing: Current and Future Perspectives, *G3 (Bethesda)*, 5, 8, 1543–1550

Wenger, S. L., Senft, J. R., Sargent, L. M., Bamezai, R., Bairwa, N., and Grant, S. G. (2004), Comparison of established cell lines at different passages by karyotype and comparative genomic hybridization, *Biosci. Rep.*, 24, 6, 631–639

Williams, K. L., Topp, S., Yang, S., Smith, B., Fifita, J. A., Warraich, S. T., et al. (2016), CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia, *Nature communications*, 7, 11253

Williams, K. L., Warraich, S. T., Yang, S., Solski, J. A., Fernando, R., Rouleau, G. A., et al. (2012), UBQLN2/ubiquilin 2 mutation and pathology in familial amyotrophic lateral sclerosis, *Neurobiology of aging*, 33, 10, 2527–e3

Wilson, A. C., Dugger, B. N., Dickson, D. W., and Wang, D. S. (2011), TDP-43 in aging and Alzheimer's disease - a review, *Int J Clin Exp Pathol*, 4, 2, 147–155

Wu, C. H., Fallini, C., Ticozzi, N., Keagle, P. J., Sapp, P. C., Piotrowska, K., et al. (2012), Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis, *Nature*, 488, 7412, 499–503

Xu, X., Tay, Y., Sim, B., Yoon, S.-I., Huang, Y., Ooi, J., et al. (2017), Reversal of phenotypic abnormalities by crispr/cas9-mediated gene correction in huntington disease patient-derived induced pluripotent stem cells, *Stem cell reports*, 8, 3, 619–633

Yang, S., Zhang, K. Y., Kariawasam, R., Bax, M., Fifita, J. A., Ooi, L., et al. (2015), Evaluation of Skin Fibroblasts from Amyotrophic Lateral Sclerosis Patients for the Rapid Study of Pathological Features, *Neurotox Res*, 28, 2, 138–146

Yao, S., Hart, D. J., and An, Y. (2016), Recent advances in universal TA cloning methods for use in function studies, *Protein Eng. Des. Sel.*

Yu, H., Cook, T. J., and Sinko, P. J. (1997), Evidence for diminished functional expression of intestinal transporters in caco-2 cell monolayers at high passages, *Pharmaceutical research*, 14, 6, 757–762

Zhao, B., Tumaneng, K., and Guan, K. L. (2011), The Hippo pathway in organ size control, tissue regeneration and stem cell self-renewal, *Nat. Cell Biol.*, 13, 8, 877–883

Zhou, Y., Liu, S., Liu, G., Öztürk, A., and Hicks, G. G. (2013), ALS-associated FUS mutations result in compromised FUS alternative splicing and autoregulation, *PLoS genetics*, 9, 10, e1003895