A comprehensive analysis of known and candidate amyotrophic lateral sclerosis genes

Emily McCann

42111072

Australian School of Advanced Medicine, Macquarie University

A thesis submitted for the partial fulfillment of the requirements for the degree of Master of Research in Advanced Medicine

Supervisors

Associate Professor Ian Blair[1]

Dr Kelly Williams[1]

[1]Motor Neuron Disease Research Centre, Australian School of Advanced Medicine, Faculty of Human Sciences, Macquarie University, Sydney, NSW, Australia

Keywords: amyotrophic lateral sclerosis, motor neuron disease, gene discovery, epigenetics, DNA methylation

Main text word count: 17 899

Abstract character count: 1710

Number of figures: 58

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Declaration

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

The code in find_my_gene.sh script and awk scripts were previously written by Kelly Williams.

Primer sets for CCNF_Ex1, CCNF_gDNA_Ex1, CCNF_gDNA_Ex3, FUS_Ex15, TARDBP_Ex6 and rs3849942 were designed and optimised by Kelly Williams and Jennifer Fifita.

Primer sets for *SOD1* exons 1-5, CCNF_Ex3, OPTN_Ex10, microsatellite markers and SNP markers were designed by Kelly Williams and Jennifer Fifita.

VCP primer design, optimisation and mutation screening was performed by Alison Hogan.

Cluster analysis was performed by Aidan OBrien at CSIRO, North Ryde, NSW.

DNA sequencing was carried out by Macrogen Sequencing Korea.

All other research described in this project is my own original work.

Acknowledgements

Firstly, I wish to express my gratitude to my supervisor Dr. Ian Blair, for his constant support, guidance and advice. Thank you for this amazing opportunity to embark on my scientific journey. Your immense knowledge and brilliance has inspired me to pursue my love of genetics.

I must of course sincerely thank Dr. Kelly Williams. Kelly, there really are no words to describe how grateful I am for all you have done for me throughout this year. Your dedication to providing me with a learning experience of the highest standard is unmatched and you have taught me more in the last year than I ever could have imagined possible. You are truly the most incredible mentor I could ask for and if I can ever amount to half the scientist you are I will count myself lucky. On top of all that, your friendship has been an added bonus.

To the rest of our lab group Jenn Fifita, Katharine Zhang, Sadaf Warraich, the absent yet ever present Shu Yang and my fellow MRes comrade Alison Hogan, the numerous morning teas, lunches, walks and impromptu chats have been instrumental to the maintenance of my sanity throughout this year. Jenn, I must make special note of the wisdom and great friendship you have offered me all year. You have had a great impact on my experience and it would not have been half as much fun without your influence.

Finally, to my family, I must thank you for your never ending love, support and most importantly your incredible patience for my ever changing moods, not just this year but over the past four years of my academic journey. I would not be who I am today without you and I am so very grateful for all that you do.

Emily

Manuscripts arising from this candidature

K.L. Williams, S.Yang, X.Hu, J.A. Fifita, S.T. Warraich, K.Y. Zhang, N.Farrawell, B. Smith, S. Topp, C. Vance, A. Chesi, C.S. Leblond, V. Sundaramoorthy, C. Dobson-Stone, A. Lee, S.L. Rayner, M.P. Molloy, M.van Blitterswijk, D.W. Dickson, R.C. Petersen, N.R. Graff-Radford, B.F. Boeve, M.E. Murray, C. Pottier, E. Don, C. Winnick, A.P. Badrock, **E.P. McCann**, A. Hogan, H. Daoud, A. Levert, P.A. Dion, J. Mitsui, H. Ishiura, Y. Takahashi, J. Goto, J. Kost, C. Gellera, A. Soragia Gkazi, J. Miller, J. Stockton, W.S. Brooks, K. Boundy, M. Polak, J.L. Muoz-Blanco, J. Esteban-Prez, A. Rbano, O. Hardiman, K.E. Morrison, N. Ticozzi, V. Silani, J.D. Glass, J.B.J. Kwok, G.J. Guillemin, R.S. Chung, S. Tsuji, R.H. Brown Jr, A. Garca-Redondo, R. Rademakers, J.E. Landers, A.D. Gitler, G.A. Rouleau, N.J. Cole, J.J. Yerbury, J.D. Atkin, C.E. Shaw, G.A. Nicholson and I.P. Blair. *CCNF* mutations in amyotrophic lateral sclerosis and frontotemporal dementia. [Manuscript submitted to Nature Genetics]

V. Sundaramoorthy, A.K. Walker, V. Tan, J.A. Fifita, **E.P. McCann**, K.L. Williams, I.P. Blair, G.J. Guillemin, M.A. Farg and J.D. Atkin. Defects in optineurin and myosin VI mediated cellular trafficking in amyotrophic lateral sclerosis. [Manuscript under review at Acta Neuropathologica]

Abstract

Amyotrophic lateral sclerosis (ALS), also known as motor neuron disease (MND), is a debilitating and ultimately fatal neurodegenerative disease affecting 2.74/100000 Australians. ALS is caused by progressive degeneration and elimination of both upper and lower motor neurons. Muscles become progressively atrophic and weak leading to spasticity and fasciculations such that dexterity and gait are adversely affected to the point that every day activities become impossible. Voluntary muscle paralysis occurs gradually as motor neurons die, generally leading to respiratory failure causing death within three to five years of diagnosis. A family history of ALS is observed in roughly 10% of cases, while the remaining 90% are classified as sporadic. Gene mutations are the only known cause of ALS. To date, over twenty genes have been implicated as ALS genes, with hexanucleotide repeat expansions in *C9ORF72* accounting for one third of familial cases and a further 20% being attributable to *SOD1* mutations. As the only known cause of ALS, a thorough understanding of these causative gene mutations is imperative for an appreciation of disease pathogenesis. Here we present a comprehensive analysis of genetic investigations into ALS. Optimised protocols for known gene screening through new patients have been established. Bioinformatic scripts to interrogate patient exome sequencing data for known ALS genes and cryptic relatedness between individuals have been developed. Furthermore, detailed mutation validation and control cohort screening has been performed for several ALS candidate genes. Finally, preliminary efforts towards methylation analysis of the major ALS genes *C9ORF72* and *SOD1* are also underway.

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.

Emily McCann

Associate Professor Ian Blair

Dr Kelly Williams

Contents

List of Figures

List of Tables

1

Introduction

Amyotrophic lateral sclerosis (ALS, also known as motor neuron disease, MND) is a debilitating neurodegenerative disease. The clinical features of ALS are caused by the progressive degeneration and elimination of both the upper motor neurons in the cerebral cortex of the brain and the lower motor neurons in the brain stem and spinal cord (**de Carvalho and Swash**, 2011; **Dion et al.**, 2009). As the anterior horn cells of the spinal cord degenerate, symptoms are triggered by the failure of axonal connections as the axons retract, leading to muscle denervation (**Al-Chalabi et al.**, 2012; **Robberecht and Philips**, 2013). Muscles become progressively atrophic and weak with piecemeal development of spasticity and fasciculations that adversely affect manual dexterity and gait (**Robberecht and Philips**, 2013; **Worms**, 2001). Paralysis of voluntary muscles occurs gradually as motor neurons die, eventually leading to respiratory failure causing death, generally within 3–5 years of initial symptom onset (**de Carvalho and Swash**, 2011; **Robberecht and Philips**, 2013; **Worms**, 2001). Significant variation is observed amongst ALS patients in terms of both site and age of onset, progression rate, prognosis and benefits gained from treatment (**Tremolizzo et al.**, 2014). Men are more commonly affected than women, with the average age of disease onset being between 55–65 years of age, however this varies widely, with an Australian incidence of 2.74 per 100,000 people (**Australian Institute of Health and Welfare**, 2011). A proportion of ALS cases also develop

clinical or subclinical frontotemporal dementia (FTD) (**Robberecht and Philips**, 2013). FTD is the fourth most common form of dementia and the most common type of presenile dementia after Alzheimers disease (AD). Clinical and pathologic overlap is well recognised for ALS and FTD, indicating that they represent a spectrum of disease (**Fecto and Siddique**, 2011).

The underlying pathological process of ALS is poorly understood, yet many pathogenic mechanisms of ALS have been proposed including oxidative stress, glutamate excitotoxicity, impaired axonal transport, neurotrophic deprivation, neuroinflammation, apoptosis, altered protein turnover, and mitochondrial dysfunction (**de Carvalho and Swash**, 2011). A pathological hallmark feature of ALS is the presence of ubiquinated misfolded protein inclusions within affected motor neurons. In 2006, it was shown that the TDP–43 protein was a principal component of these inclusions in non–*SOD1* ALS (**Neumann et al.**, 2006). Subsequently, the *TARDBP* gene that encodes the TDP–43 protein was investigated as an ALS candidate gene, and was shown to harbour familial and sporadic ALS causative mutations (**Sreedharan et al.**, 2008). Similarly, after identification as disease genes in ALS, the protein products of *FUS*, *UBQLN2*, *OPTN*, *MATR3* and *p62* were shown to be variably immunopositive in these protein inclusions (**Deng et al.**, 2011b, 2010; **Hortobagyi et al.**, 2011; **Williams et al.**, 2012). It is believed that this protein aggregation may, at least in part, be triggered by mutations in disease genes, which may cause protein misfolding (**Al-Chalabi et al.**, 2012; **Robberecht and Philips**, 2013).

To date, the only known cause of ALS are gene mutations (Table 1.1). Around 10% of ALS cases are hereditary and are therefore classified as familial ALS (FALS) (**Rowland and Shneider**, 2001). The remaining 90% of ALS cases are apparently sporadic (SALS), however gene mutations identified in FALS cases have also been found in a small proportion of SALS (**Martin and Wong**, 2013). FALS cases can potentially be misclassified as sporadic owing to a lack of extensive family history and incomplete disease penetrance, a phenomenon frequently observed in ALS (**Andersen and Al-Chalabi**, 2011; **Al-Chalabi et al.**, 2012; **Wijesekera and Leigh**, 2009). Furthermore, familial and sporadic ALS cases are clinically and pathologically indistinguishable (**Andersen and Al-Chalabi**, 2011; **Martin and Wong**, 2013).

The first gene mutation identified as causing ALS was found in the superoxide dismutase 1 gene (*SOD1*, OMIM105400) (**Rosen**, 1993), and remained the only gene implicated as causing ALS for 15 years. This ALS locus was identified on chromosome

TABLE 1.1: **Known ALS gene mutations**. GWAS; Genome–wide association study. *Recessive for p.D90A in Scandinavian populations. #May be recessive in Japanese populations.

	Gene Name	Gene locus	Mode of inheritance	Estimated % of FALS	Method of discovery	Pathway involvement	Reference
Angiogenin	*ANG*	14q11.2	Dominant	<1%	Association studies	RNA–binding and/or processing protein dysfunctions	(Greenway et al., 2006)
Ataxin 2	*ATXN2*	12q24	Dominant	<1%	Candidate Gene Analysis	Repeat expansion	(Elden et al., 2010)
Chromosome 9 open reading frame 72	*C9ORF72*	9p21.3p13.3	Dominant	40–50%	GWAS	Repeat expansion	(DeJesus-Hernandez et al., 2011; Renton et al., 2011)
Chromatin modifying protein 2B	*CHMP2B*	3p11	Dominant	Unkown	Linkage Analysis	Proteostatic proteins	(Parkinson et al., 2006)
D–amino–acid oxidase	*DAO*	12q24	Dominant	<1%	Linkage Analysis	Excitotoxicity	(Mitchell et al., 2010)
V–Erb–B2 avian erythroblastic leukemia viral oncogene homolog 4	*ERBB4*	2q34	Dominant	<1%	Linkage Analysis and NGS	Growth and differentiation	(Takahashi et al., 2013)
Ewing sarcoma breakpoint region 1	*EWSR1*	22q12.2	Unkown	Unkown	Candidate Gene Analysis	RNA–binding and/or processing protein dysfunctions	(Couthouis et al., 2012)
Phosphatidylinositol 3,5bisphosphate5phosphatase	*FIG4*	6q21	Dominant	Unkown	Candidate Gene Analysis	Proteostatic proteins	(Chow et al., 2009)
Fused in sarcoma	*FUS*	16p11.2	Dominant	1–5%	Linkage Analysis	RNA–binding and/or processing protein dysfunctions	(Kwiatkowski et al., 2009; Vance et al., 2009)
Optineurin	*OPTN*	10p15p14	Dominant#	<1%	Homozygosity mapping	Proteostatic proteins	(Maruyama et al., 2010)
Prolyl 4–hydroxylase, beta polypeptide	*P4HB*	17q25.3	Susceptibility gene	Unkown	Association studies	Enzyme	(Kwok et al., 2013)
Profilin 1	*PFN1*	17p13.2	Dominant	Unkown	NGS	Cytoskeleton/cellular transport deficits	(Wu et al., 2012)
Senataxin	*SETX*	9q34	Dominant	<1%	Linkage Analysis	RNA–binding and/or processing protein dysfunctions	(Chen et al., 2004)
Survival of motor neuron 1, telomeric	*SMN1*	5q13.2	Unkown	Unkown	Candidate Gene Analysis	Copy number variation	(Corcia et al., 2002)
Survival Of Motor Neuron 2, Centromeric	*SMN2*	5q13.2	Unkown	Unkown	Candidate Gene Analysis	Copy number variation	(Moulard et al., 1998)
Superoxide dismutase 1	*SOD1*	21q22.1	Dominant*	20%	Linkage Analysis	Enzyme	(Rosen, 1993)
Sequestosome	*SQSTM1*	5q35	Dominant	Unkown	Candidate Gene Analysis	Proteostatic proteins	(Fecto et al., 2011)
Synovial Sarcoma Translocation Gene On Chromosome 18–Like	*SS18L1*	20q13.33	Unkown	Unkown	Exome sequencing	RNA–binding and/or processing protein dysfunctions	(Chesi et al., 2013)
TATA–binding protein associated factor 15	*TAF15*	17q11.1q11.2	Unkown	Unkown	Candidate Gene Analysis	RNA–binding and/or processing protein dysfunctions	(Couthouis et al., 2011)
TAR DNA–binding protein 43 (TDP–43)	*TARDBP*	1p36.2	Dominant	1–5%	Linkage Analysis	RNA–binding and/or processing protein dysfunctions	(Sreedharan et al., 2008)
Ubiquilin 2	*UBQLN2*	Xp11	Dominant	<1%	Exome sequencing	Proteostatic proteins	(Deng et al., 2011b; Williams et al., 2012)
Vesicle–associated membrane protein–associated protein B and C	*VAPB*	20q13.3	Dominant	Unkown	Candidate Gene Analysis	Cytoskeleton/cellular transport deficits	(Nishimura et al., 2004)
Valosin–containing protein	*VCP*	9p13	Dominant	<1%	Candidate Gene Analysis	Proteostatic proteins	(Johnson et al., 2010)

21q22.1–22.2 by a collaborative effort using 18 ALS pedigrees for linkage analysis (**Siddique et al.**, 1991). Screening the *SOD1* gene located in the linkage region revealed 11 missense mutations in 13 ALS families (**Rosen**, 1993). Approximately 20% of FALS cases are accounted for by the 160+ identified *SOD1* mutations (**Dion et al.**, 2009; **Robberecht and Philips**, 2013), which are almost all missense, inherited in a dominant fashion and affect the active site or protein structure (**Chio et al.**, 2008; **Dion et al.**, 2009; **Sreedharan and Brown**, 2013). Despite the dominance of these mutations, some do show reduced penetrance (i.e. mutation carriers do not develop disease). Especially interesting are those which show variable penetrance between different populations (**Dion et al.**, 2009). For example, the p.D90A mutation shows recessive inheritance in Scandinavian populations, and was previously considered a benign polymorphism in heterozygotes who showed no signs of ALS (**Andersen et al.**, 1996), however heterozygote carriers in other populations develop rapidly progressing ALS (**Andersen et al.**, 1995). *SOD1* mutations have also been identified in 1% of SALS cases (**Chio et al.**, 2008).

In 2006, linkage analysis in families with comorbid ALS/FTD implicated a locus on the short arm of chromosome 9 (**Morita et al.**, 2006; **Vance et al.**, 2006). Over the next five years, major linkage efforts identified several more ALS/FTD families linked to chromosome 9p21.2 and reduced the linked locus to a 3.7Mb region containing just five known genes (**Boxer et al.**, 2011; **Gijselinck et al.**, 2010; **Le Ber et al.**, 2009; **Luty et al.**, 2008; **Valdmanis et al.**, 2007). In addition, large genome–wide association studies (GWAS) using either SALS, FALS or FTD samples showed association to the 9p21.2 ALS/FTD locus (**Laaksovirta et al.**, 2010; **Shatunov et al.**, 2010; **van Es et al.**, 2009; **Van Deerlin et al.**, 2010).

In 2011, two groups independently identified a GGGGCC hexanucleotide repeat expansion in an intronic region of the chromosome 9 open reading frame 72 gene (*C9ORF72*, OMIM105550) as the cause of ALS/FTD in families linked to this region (**DeJesus-Hernandez et al.**, 2011; **Renton et al.**, 2011). The expansion site is located between two non–coding exons, 1a and 1b (**Al-Chalabi et al.**, 2013). In control populations, between 2 and 23 hexanucleotide repeats are present at this genomic location, however expansions of hundreds of repeat units have been found in ALS/FTD patients linked to this region. It has been demonstrated that patients carrying the repeat expansion also possess a full or partial common founder haplotype on 9p21.2, containing up to 42 SNPs (**Majounie et al.**, 2012). The repeat is also unstable, especially in the expanded form and is susceptible to somatic mutation

so that repeat–length mosaicism is observed within the same tissue (**Robberecht and Philips**, 2013). Such expansions account for 38.5% of FALS cases and 3.5% of sporadic cases in Australian cohorts, making this mutation the most common known cause of ALS to date. Further, *C9ORF72*–linked Australian ALS cases show penetrance levels of 50% at 58 years in males and 63 years in females (**Williams et al.**, 2013). Phenotypic variability is observed in *C9ORF72* expansion carriers in that some present with ALS or FTD, others with comorbid conditions and that the age of onset, severity and progression can be quite variable (**Majounie et al.**, 2012; **Renton et al.**, 2014; **Williams et al.**, 2013).

Questions have been raised as to the influence of repeat length and homozygosity on disease severity, however no definitive answer has so far been obtained (**Robberecht and Philips**, 2013). There have been reports of a second ALS gene being present in *C9ORF72* expansion patients, including *TARDBP*, *FUS* and *ANG* (**Chio et al.**, 2012; **van Blitterswijk et al.**, 2012a,b; **Williams et al.**, 2013). This hints at a possible oligogenic basis of disease in some families, and may also explain some of the phenotypic variability observed (**Robberecht and Philips**, 2013; **Williams et al.**, 2013).

Combined, mutations in *SOD1* and *C9ORF72* account for more than half of familial ALS cases and have even been identified in approximately 8% of SALS cases, highlighting the genetic component of SALS (**Andersen and Al-Chalabi**, 2011; **Dion et al.**, 2009; **Renton et al.**, 2014) which is solidified by heritability estimates of 0.61 using ALS disease discordant twin cohorts (**Al-Chalabi et al.**, 2010).

In 2008 and 2009, ALS–causing mutations were identified in the TAR DNA binding protein (*TARDBP*, OMIM612069) (**Sreedharan et al.**, 2008) and Fused in sarcoma (*FUS*, OMIM608030) genes respectively (**Kwiatkowski et al.**, 2009; **Vance et al.**, 2009). The proteins encoded by these genes, TDP–43 and FUS, are both heterogeneous nuclear ribonucleoproteins (hnRNPs) involved in RNA processing (**Al-Chalabi et al.**, 2012; **Andersen and Al-Chalabi**, 2011). These were seminal discoveries as they implicated defective RNA metabolism as a new pathogenic mechanism in ALS (**Kim et al.**, 2013), as both protein products are involved in pre–mRNA splicing, RNA transport and RNA translation (**Robberecht and Philips**, 2013). *TARDBP* was investigated as a candidate gene owing to its encoded protein's presence in the neuronal inclusions found in the majority of ALS patients. It was discovered that exon 6, encoding a C–terminal glycine–rich domain, contained many ALS mutations, mostly

missense but some deletions resulting in truncations (**Robberecht and Philips**, 2013). Many dominant missense mutations have been identified in *FUS* in FALS patients (**Kwiatkowski et al.**, 2009; **Vance et al.**, 2009). Similar to *TARDBP*, *FUS* contains a glycine–rich domain imperative for translational functions, with many ALS mutations identified in this region (**Robberecht and Philips**, 2013). Further, both *TARDBP* and *FUS* contain prion–like domains (**Kim et al.**, 2013). These domains may promote aggregation as they act as templates to induce natively folded proteins to convert, causing entrapment in the aggregate (**Robberecht and Philips**, 2013; **Kim et al.**, 2013).

Several other genes have been implicated in ALS, and many of these encode proteins present in the ubiquitinated inclusions found in the affected motor neurons of ALS patients. These are also generally involved in proteostasis and include optineurin (*OPTN*, OMIM613435) (**Maruyama et al.**, 2010; **Maruyama and Kawakami**, 2013), valosin–containing protein (*VCP*, OMIM613954) (**Johnson et al.**, 2010), ubiquilin 2 (*UBQLN2*, OMIM300857) (**Deng et al.**, 2011b; **Williams et al.**, 2012) and sequestosome 1 (*SQSTM1*, OMIM601530) (**Fecto et al.**, 2011), and support a role for the failure of protein degradation pathways in ALS pathogenesis, but occur in <1% of familial ALS cases (**Robberecht and Philips**, 2013). Recently, our laboratory identified mutations in the cyclin F gene, *CCNF*, causing ALS and FTD in a large Australian ALS/FTD family. A large worldwide collaborative patient screening effort identified multiple mutations in *CCNF* in familial and sporadic ALS and FTD cases at frequencies ranging from 0.6–3% (unpublished, manuscript submitted).

Recently, there has been a surge in interest in the link between neurodegenerative diseases such as ALS and epigenetic mechanisms. Epigenetics encompasses the molecular modifications that result in gene expression changes without alteration of the DNA sequence (**Bell and Spector**, 2011). Most epigenetic modifications are centred around the modulation of chromatin structure (Figure 1.1) (**Bell and Spector**, 2011), such as DNA methylation, histone modifications including acetylation and phosphorylation, nucleosome and chromatin remodelling, and noncoding RNA actions (**Al-Chalabi et al.**, 2013; **Handel et al.**, 2010; **Teperino et al.**, 2013). The complement of epigenetic modifications that are present in the genome is referred to as the epigenome. These mechanisms act both in isolation and in combination to control gene expression, working in feedback loops to either reinforce or disable each other (**Al-Chalabi et al.**, 2013; **Handel et al.**, 2010; **Jaenisch and Bird**, 2003). To date, DNA methylation is the most studied and best understood epigenetic mechanism.

DNA methyltransferase proteins (Dnmts) can add methyl groups to Cytosine residues which precede a Guanine residue, deemed a CpG site or CG dinucleotide (**Teperino et al.**, 2013; **Zhang et al.**, 2012). The presence of DNA methylation acts to hinder the accessibility of the DNA by the transcription machinery, so as to repress gene expression, as seen in Figure 1.2 (**Jaenisch and Bird**, 2003; **Zhang and Pradhan**, 2014). This most commonly occurs in gene promoter regions, which contain GC dense regions, known as CpG islands, which is not surprising given the role promoters play in regulating gene expression (**Teperino et al.**, 2013).

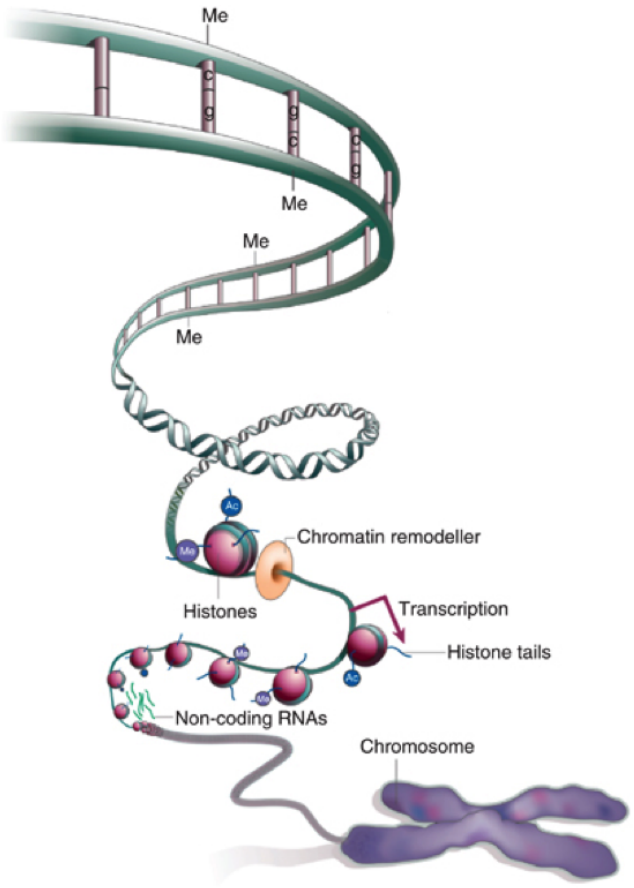

FIGURE 1.1: **Diagramatic representation of the different levels of epigenetic modifications acting on chromatin structure to regulate gene expression**. The vast majority of DNA methylation occurs at sites where the Cytosine residue is followed by a Guanine residue, commonly referred to as CpG sites or CG dinucleotides where a methyl group is added to the C5 carbon of the Cytosine by DNA methyltransferase proteins. Histone modifications can take the form of acetylation and phosphorylation, while various non–coding RNAs such as miRNA can also interact with DNA. Figure adapted with permission from (**Jones et al.**, 2008).

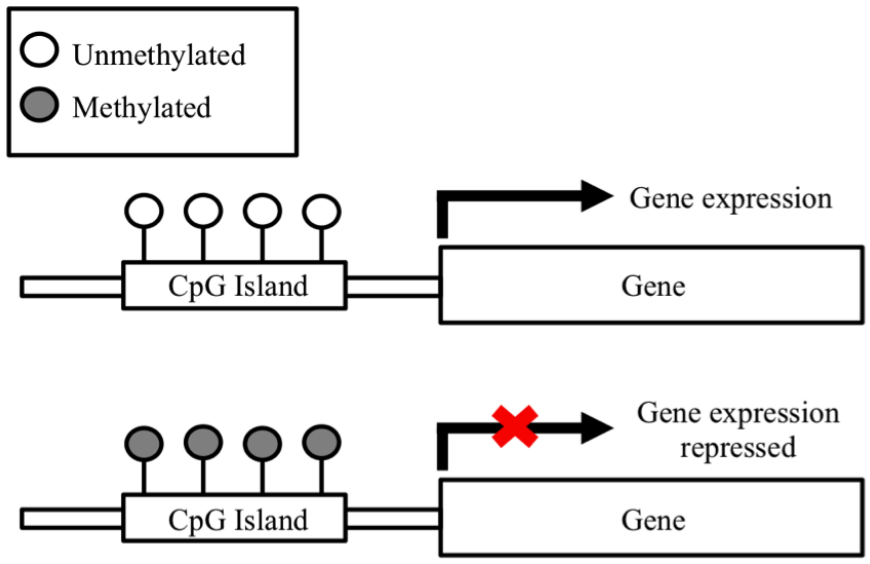

FIGURE 1.2: **Effect of DNA methylation on gene expression.** The presence of DNA methylation acts to block the binding of the transcription machinery so that the gene is not expressed.

Epigenetic modifications are not stable over time and undergo precise changes in order to control tissue specific gene expression throughout different stages of development and also in response to environmental pressures (**Al-Chalabi et al.**, 2013; **Feil and Fraga**, 2011). This instability is likely the result of influences from the environment, developmental programs, hormones and stochastic events (**Kaminsky et al.**, 2009). Further, evidence suggests changes to DNA methylation patterns as people age (**Fraga et al.**, 2005), or between generations (**Handel et al.**, 2010). The dynamic nature of epigenetic marks is an important way in which organisms may adapt rapidly to environmental influences, and thus can be seen as a reflection of environmental exposures (**Handel et al.**, 2010). As such, and owing to the heritable potential of epigenetic marks, these changes to gene expression are of immense importance to evolution (**Handel et al.**, 2010; **Zhang et al.**, 2012). There is still much to learn about how epigenetic marks act, become established and are regulated (**Handel et al.**, 2010).

It has been strongly suggested that epigenetic modifications play an important role in gene expression during cancer (**Zhang et al.**, 2012), and recent studies have laid the foundation for an epigenetic role in neurodegenerative diseases including ALS (**Belzil et al.**, 2014; **Callaghan et al.**, 2011; **Chestnut et al.**, 2011; **Figueroa-Romero et al.**, 2012; **Morahan et al.**, 2007, 2009; **Oates and Pamphlett**, 2007; **Tremolizzo**

et al., 2014; Xi et al., 2013; Yang et al., 2010b), Alzheimer's disease (Bakulski et al., 2012; Chen et al., 2009; Chouliaras et al., 2013; Mastroeni et al., 2009, 2010; Sanchez-Mut et al., 2014; Tohgi et al., 1999; Urdinguio et al., 2009), and Parkinsons disease (PD) (Jowaed et al., 2010; Kontopoulos et al., 2006; Masliah et al., 2013; Matsumoto et al., 2010; Voutsinas et al., 2010). All of these neurodegenerative diseases are highly complex disorders with poor understanding of the disease aetiology.

Despite recent gene discoveries, there clearly remains much to be learnt about the underlying aetiology and pathogenesis of ALS (Al-Chalabi et al., 2012; Deng et al., 2011b). In order to take steps to achieve this, this thesis presents a comprehensive collection of genetic investigations into ALS, which aim to contribute to uncovering the genetic basis of the disease. Optimised protocols have been established for mutation analysis of *C9ORF72* and *SOD1* in new patients. Bioinformatic scripts have been written to interrogate patient exome sequencing data for known ALS genes and cryptic relatedness between individuals. Detailed mutation validation and control cohort screening has been performed for several ALS candidate genes. Finally, preliminary efforts towards methylation analysis of the major ALS genes *C9ORF72* and *SOD1* through large disease discordant cohorts are also underway.

2

Subjects and materials

2.1 Subjects

Patients, family members of patients and unrelated control individuals were ascertained through the Macquarie University Hospital Neurology Clinic and recruited under informed written consent as approved by the Macquarie University Human Research Ethics Committee (HREC) under the approval number 5201300333. Patients were clinically diagnosed with definite or probable ALS according to El Escorial criteria (**Brooks et al.**, 2000) with most participants being of European descent. Whole blood samples were collected by staff at the Macquarie University Hospital Neurology Clinic in 10mL Vacutainer EDTA collection tubes (Becton, Dickinson and Company, NJ, USA) which were subsequently stored at –30°C. Genomic DNA (gDNA) was extracted from 3mL of whole blood using the QIAsymphony DSP DNA midi kit on the QIAsymphony Sample Preparation System (Qiagen, CA, USA). DNA samples were then quantitated using either the NanoDrop 2000/200c Spectrophotometer (Thermo Fisher Scientific, DE, USA) or the QIAxpert microfluidic UV/VIS spectrophotometer (Qiagen).

2.2 Materials

2.2.1 Growth media, reagents and buffers

1×TBE (Tris–Borate–EDTA) buffer
89mM Tris; 89mM boric acid; 2mM EDTA (pH 8.3).

T4 DNA ligase (Promega, WI, USA)
T4 DNA ligase in buffer containing 10mM Tris–HCl (pH 7.4 at 25°C), 50mM KCl, 1mM DTT, 0.1mM EDTA and 50% glycerol.

2×Ligation buffer (Promega, WI, USA)
60mM Tris–HCl (pH 7.8); 20mM $MgCl_2$; 20mM DTT; 2mM ATP; and 10% polyethylene glycerol.

SOC medium (Life Technologies, CA, USA)
2% tryptone; 0.5% yeast extract; 10mM NaCl; 2.5mM KCl; 10 mM $MgCl_2$; 10 mM $MgSO_4$; and 20mM glucose.

Luria Broth (LB)
10g/L Bacto–tryptone; 5g/L Bacto–yeast extract; and 5g/L NaCl. LB was subjected to autoclaving prior to use.

LB agar
15g/L Davis Agar dissolved in LB.

IPTG (Isopropyl–1–thio–B–D–galactopyranoside)
1.2g IPTG dissolved in 50mL sterile water. Solution was filter–sterilized and stored at 4°C.

Ampicillin (Bioline, NSW, Australia)
100mg/mL in H_2O.

X–Gal (5–Bromo–5–chloro–3–indolyl–B–D–galactopyranoside)
100mg/mL 5–bromo–4–chloro–3–indolyl–B–D–galactoside in N,N'–dimethyl–formamide. X–Gal solution was stored in the dark at –20°C.

2.2.2 Vector

pGEM–T (Promega, WI, USA) (Figure 2.1)

The pGEM–T vector system allows for rapid screening of colonies. The vector was prepared by the manufacturers through linearization of the pGEM–5Zf(+) vector with EcoRV at base 51 and adding a Thymine to both 3' ends. This overhang is compatible with PCR products generated by thermostable polymerases and increases ligation efficiency. The multiple cloning site of the pGEM–T vector is contained within a coding region for the β–galactosidase enzyme. As such, when DNA is inserted into this multiple cloning site, the α–peptide of the *lacZ* gene is disrupted and not expressed, so that functional β–galactosidase is not produced. This results in the production of white colonies containing insert DNA. In cases where no DNA has been inserted, the gene stays intact and thus β–galactosidase is expressed and a blue colony results.

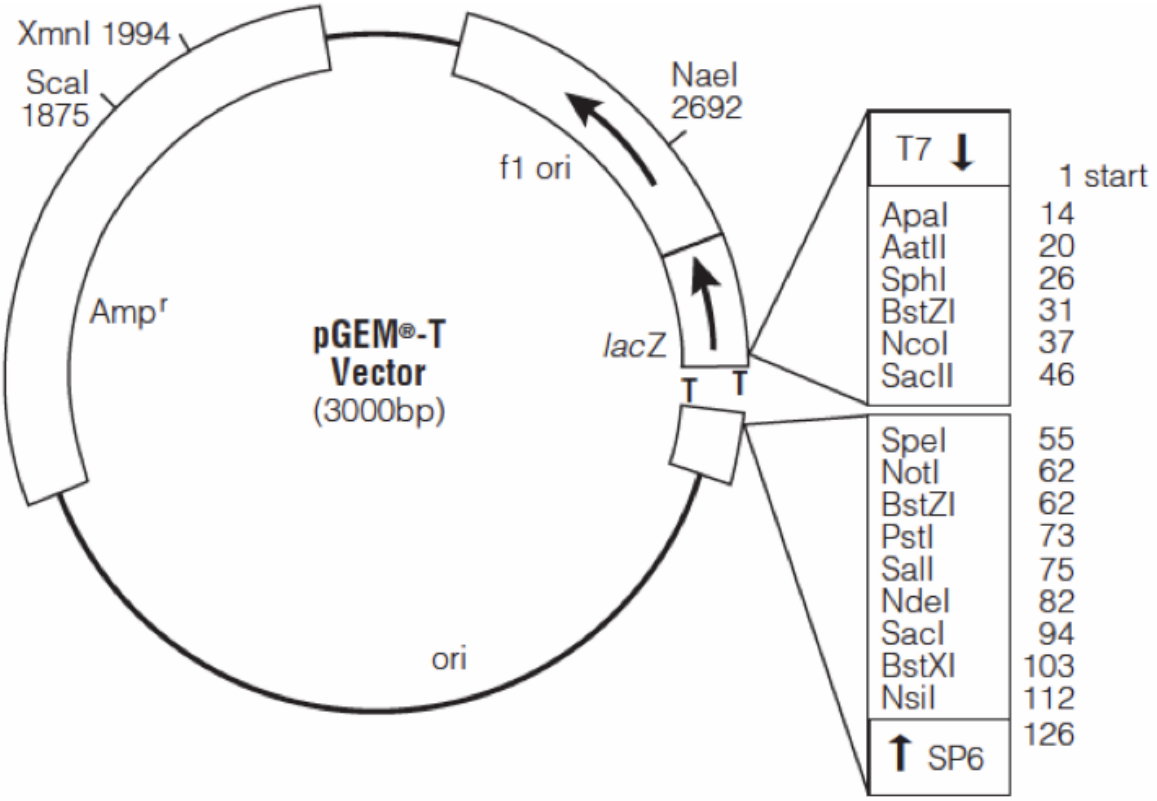

Figure 2.1: **Diagram of pGEM–T vector system from Promega.**

3 Methods

3.1 Pedigrees

Pedigrees are vital genetic tools, especially for determining the mode of inheritance for a trait, providing an easily discernable visual representation of relationships between members of a family and the occurrence of phenotypes. An ALS affected proband presented at the Macquarie University Hospital Neurology Clinic, with four deceased family members also known to have been affected by ALS. As such, a pedigree for this family, named Mq1, was drawn using the PedDraw program (http://www.pedigree-draw.com/) and compiled using information obtained from patient clinical files, death certificates and ancestry.com.

3.2 Bioinformatics and custom script development

Exome capture and sequencing was conducted prior to this project at Axeq (Korea) using patient gDNA samples. Bioinformatic analysis provided by Axeq comprised of alignment to the hg19 reference genome using BWA (**Li and Durbin**, 2010), variant identification using SAMtools (**Li et al.**, 2009) and annotation of variants with ANNOVAR (**Wang et al.**, 2010). The resultant variant files were used in further

bioinformatic analysis during this project.

3.3 Search_my_exome.sh

A custom shell script was written using Xcode software (v5.1.1) and was developed for use with the Unix programming language. This shell script utilises the awk command to interrogate individual patient exome data for specified ALS genes. The list of genes is provided in Table 3.1.

3.4 Relatedness testing

Custom shell scripts were written using the PLINK/SEQ library for working with human genetic variation data (http://atgu.mgh.harvard.edu/plinkseq/), to ascertain relatedness between two individuals based on exome sequencing data variant call format (VCF) files which were generated using vcf tools (**Danecek et al.**, 2011).

Patients identified as distantly related by more complex relatedness testing underwent common variant identification. The Compare two Datasets tool on NBIC Galaxy (http://galaxy.nbic.nl/) was used to identify common variants between the distantly related individuals. Additional control exome data was also interrogated for the identified common variants, with those present being removed from further analysis. Filtering was conducted in Microsoft Excel to remove known SNPs based on dbSNP135 as well as synonymous mutations. Variant filtering was performed using the previously developed custom shell scripts shown in Appendices A.1.1 and A.1.2, based on dbSNP (releases 137 and 141; https://www.ncbi.nlm.nih.gov/SNP/) and the Exome Variant Server (data release ESP6500SI–V2; http://evs.gs.washington.edu/EVS/), with any further identified SNPs removed from analysis. The Integrative Genomics Viewer (IGV) (**Thorvaldsdottir et al.**, 2013) was then used to visualise the position at which the variant occurred to confirm it was only present in the two individuals. The genomic position at which insertions or deletions occurred was confirmed in this way, and if a discrepancy was encountered, the confirmed position was investigated using the UCSC (University of California, Santa Cruz) Genome browser (http://genome.ucsc.edu/) to determine if it was in fact a known variant. *In silico* protein predictions were conducted using MutationTaster2 (**Schwarz et al.**, 2014) (http://www.mutationtaster.org/), PolyPhen–2 (**Adzhubei et al.**, 2010)

Table 3.1: **Known ALS genes screened through exome sequencing files using the Search_my_exome.sh script.**

ALS Genes
ANG
ATXN2
C9ORF72
CCNF
CELF4
CHMP2B
CREST
DAO
ERBB4
EWSR1
FBXO
FIG4
FUS
HNRNP
OPTN
P4HB
PDIA2
PDIA3
PFN1
RRM2
SETX
SMN1
SMN2
SOD1
SQSTM1
SRRM2
SS18L1
TAF15
TARDBP
UBQLN2
VAPB
VCP

(http://genetics.bwh.harvard.edu/pph2/), Pon_P2 (http://structure.bmc.lu.se/PON-P2/) and NetPhos 2.0 (**Blom et al.**, 1999) (http://www.cbs.dtu.dk/services/NetPhos/). MutationTaster2 (**Schwarz et al.**, 2014) was also utilised to determine conservation of individual genomic sites across species. Domain localisation information concerning mutations of interest was determined using InterPro (http://www.ebi.ac.uk/interpro/).

3.5 Candidate gene analysis

When candidate ALS genes were identified (by either our laboratory or a collaborator), patient exome sequence data was searched for the candidate genes, listed in Table 3.2, using a previously written shell script (Appendix A.1.3). Filtering was conducted in Microsoft Excel to remove known SNPs based on dbSNP135 as well as mutations that did not alter the protein sequence. Variant filtering was performed using the previously developed custom shell scripts shown in Appendices A.1.1 and A.1.2, based on dbSNP (releases 137 and 141; https://www.ncbi.nlm.nih.gov/SNP/) and the Exome Variant Server (data release ESP6500SI–V2; http://evs.gs.washington.edu/EVS/), with any further identified SNPs removed from analysis.

TABLE 3.2: **Candidate genes screened through exome sequencing files.**

Candidate Genes
RBM14
DAZAP1
PSPC1
SFPQ
NONO
RBM4
FAM98A
FIGN
CIRBP
HNRNPK
RBMX
CPSF6

3.6 Conservation of protein sequence

To determine the species conservation of mutated residues, mutiple sequence alignment using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/) was employed using sequences obtained from NCBI Nucleotide (http://www.ncbi.nlm.nih.gov/nucleotide/).

3.7 Polymerase Chain Reaction

Polymerase Chain Reaction (PCR) was carried out using the Mastercycler Pro S (Eppendorf, NY) and was used for mutation validation, known gene screening and genotyping. Standard PCR conditions are outlined in Table 3.3. All new primer sets used in this project were designed using either ExonPrimer (http://ihg.gsf.de/ihg/ExonPrimer.html) or Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/) and were synthesised by Sigma Aldrich (NSW, Australia). Optimal PCR conditions for each new primer set were determined using standard protocols utilising My*Taq* HS Red Mix (Bioline) and an annealing temperature gradient thermocycling program according to Table 3.4. When optimal conditions could not be determined by this method, a refined annealing temperature gradient was employed if appropriate or 10× PCR enhancer (Life Technologies) was added to the reaction. In particularly difficult cases a nested PCR approach was used. One PCR reaction was performed and the resultant amplified product used as the template in a subsequent PCR reaction, either diluted or undiluted. In such cases, the high fidelity MyFi DNA polymerase (Bioline) was utilised in conjunction with 10× PCR enhancer (Life Technologies). Previously designed primer sets which underwent optimisation are listed in Table 3.5. Additional primer sets previously designed and optimised were also utilised (Table 3.6).

3.8 Agarose gel electrophoresis

In order to determine if PCR amplification was successful, resultant products were gel electrophoresed to see if a product was amplified. Gels were prepared using 1.5%

TABLE 3.3: **Standard PCR reaction mixture and standard thermocycling conditions.** T_A, annealing temperature.

Reagent	1× Volume (μl)
dH_2O	7.4
My*Taq* HS red mix	10.0
10mM F primer	0.8
10mM R primer	0.8
20ng DNA	1.0
Total	20.0

Temperature	Time	Cycles
95°C	10 min	1×
94°C	15 sec	
T_A°C	30 sec	30×
72°C	30 sec	
72°C	7 min	1×
15°C	Hold	

TABLE 3.4: **Standard PCR reaction mixture and thermocycling conditions used for optimising PCR conditions for new primer sets.**

Reagent	1× Volume (μl)
dH_2O	3.2
My*Taq* HS red mix	5.0
10mM F primer	0.4
10mM R primer	0.4
20ng DNA	1.0
Total	10.0

Temperature	Time	Cycles	T_A Gradient
95°C	10 min	1×	53.4
			55.7
94°C	15 sec		58.3
50–72°C	30 sec	30×	61.0
72°C	30 sec		63.7
			66.1
72°C	7 min	1×	68.0
15°C	Hold		69.4

TABLE 3.5: **Primer sequences and product sizes for various gene primer sets designed prior to candidature.**

Primer Name	5'–3' Sequence	PCR product size
SOD1_Ex1_NewF	ATTGGTTTGGGGCCAGAG	408
SOD1_Ex1_NewR	TGACTCAGCACTTGGGCAC	
SOD1_Ex2_NewF	GTCAGCCTGGGATTTGGAC	355
SOD1_Ex2_NewR	CGACAGAGCAAGACCCTTTC	
SOD1_Ex3_NewF	CAGAAGTCGTGATGCAGGTC	313
SOD1_Ex3_NewR	CAGCAAGTTCAAAAGCAAAGG	
SOD1_Ex4_NewF	GACGTGAAGCCTTGTTTGAAG	418
SOD1_Ex4_NewR	AATTGTCCAATAAAATTGCTTTT	
SOD1_Ex5_NewF	TTCATTTAGACAGCAACACTTACC	572
SOD1_Ex5_NewR	CAAAATACAGGTCATTGAAACAGAC	
CCNF_Ex3_F	AGGTGTGGGGCTTTTGG	231
CCNF_Ex3_R	CAGACTGGCACATAGGGAGG	
OPTN_Ex10_F	TGGTTCAGCCTGTTTTCTCC	372
OPTN_Ex10_R	TTCATGCTCACACATTAACTGG	

Table 3.6: **Primer sequences, product sizes and optimal PCR reaction conditions and annealing temperatures for various gene primer sets.**

Primer Name	5'–3' Sequence	PCR product size	Optimised PCR conditions	T_A
CCNF_gDNA_Ex1_F	gacgcccacaaacccctg	839bp	Nested PCR to CCNF_Ex1 using MyFi+PCR enhancer	59°C
CCNF_gDNA_Ex1 R	ctttccacagagctaggtcca			
CCNF_Ex1_F	GGTAAATGAGGCGAGCACAG	348bp	Nested PCR from CCNF_gDNA Ex1 using MyFi+PCR enhancer	55°C
CCNF_Ex1_R	GGTTAAAGAGCTCGAGCCAG			
CCNF_gDNA_Ex3_F	ggggaagcacaaaacacatttc	837bp	MyFi + enhancer	60°C
CCNF_gDNA_Ex3_R	gccagtgagtgaaacgctatag			
VCP_Ex5_F	TGATGAGTTCTCACTTTGTCTTG	266bp	My*Taq*	66°C
VCP_Ex5_R	GACAGTTACCACATGATGCCAC			
VCP_Ex14_F	ccaccacgTTTGCCTAGAG	441bp	My*Taq*	66°C
VCP_Ex14_R	CTAAAGAGCACTCCGTACCAGC			
FUS_Ex15_F	TACTCGCTGGGTTAGGTAGGAG	218bp	My*Taq*	62°C
FUS_Ex15_R	GGAAGGTTACAAAATAACGAG			
TARDBP_Ex6_F	TGCTTGTAATCTAAGTTTTGT	751bp	My*Taq*	64°C
TARDBP Ex6 R	TGCTGAATATACTCCACACTG			
rs3849942_F	TGGCTCTCCAACACTTACAGAA	203bp	My*Taq*	64°C
rs3849942_R	ATGCTAGGCACTGAGACACAAA			

w/v of agarose powder (Bioline) to 1×TBE with the addition of 1×SYBRSafe DNA gel stain (Life Technologies) to visualise DNA. Gel lanes were loaded with 3μl of each reaction in horizontal submarine gel tanks and electrophoresis performed at 100V for 45 minutes. EasyLadder I (Bioline) size marker was also run for determination of PCR product size and concentration. SYBRSafe does not require the use of UV light for visualisation, therefore gels were visualised using a blue filter under white light and the image captured using Gel Doc EZ imager and Image Lab software (Bio–Rad, NSW, Australia).

3.9 PCR cleanup

ExoSAP treatment was applied to PCR products prior to sequencing to remove excess primers and dNTPs. To 17μl of PCR product, a master mix containing 0.2μl of each exonuclease I (New England Biolabs, MA, USA) and Thermosensitive Alkaline Phosphatase (TSAP, Promega), and 4.6μl of distilled water was added. This was followed by incubation at 37°C for 40 minutes and an enzyme denaturing step at 80°C for 20 minutes using the Mastercycler Pro S (Eppendorf).

3.10 Sequencing and Analysis

Sanger sequencing of amplified products was conducted using Big–Dye terminator sequencing on an ABI 3730XL sequencer (Macrogen) and analysed using Sequencher v5.1 software (Gene Codes, MI, USA). Sequencing chromatograms were analysed by direct comparison to reference sequences obtained from the UCSC Genome browser (http://genome.ucsc.edu/). Heterozygous SNPs were indicated by double sequencing peaks and a reduced peak height compared to wild type alleles, indicating the presence of two different bases at the SNP location. Confirmation of potential SNPs was conducted by repeating PCR reactions and sequence analysis in both forward and reverse directions. When interpreting SNP results, dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) was consulted to determine if the SNP had been previously identified. The Amyotrophic Lateral Sclerosis Online Genetic Database (ALSoD v6.0, http://alsod.iop.kcl.ac.uk/) (**Radunovic and Leigh**, 1999) was also consulted to determine if the SNP had been previously reported as linked to ALS.

3.11 TaqMan SNP genotyping

Custom TaqMan SNP genotyping assays (Life Technologies) were designed using gDNA sequences obtained from the UCSC Genome browser (http://genome.ucsc.edu/) and the Custom TaqMan Assay Design Tool (https://www.lifetechnologies.com/order/custom-genomic-products/tools/genotyping/). PCR products were generated using TaqMan mastermix according to manufacturers instructions. Briefly, in a 10μl reaction for each 1μl DNA sample, 2.5μl of TaqMan SNP Genotyping Mastermix and 0.25μl of the custom TaqMan SNP Genotyping primer were used. A plate set up as shown in Figure 3.1 was utilised. Subsequent analysis on the Viia 7 real time PCR system (Life Technologies) using the thermocycling conditions listed in Table 3.7 revealed genotypes for the SNP of interest.

NTC	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
NTC	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
NTC	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
NTC	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
neg	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
neg	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
pos	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
pos	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12

FIGURE 3.1: **Plate setup for customised TaqMan SNP genotyping assays.** For column one: NTC; no template control, neg; negative control DNA, pos; positive control DNA and columns two to twelve indicate the well of the non related control plate from which DNA was used for the assay.

TABLE 3.7: **Thermocycling program used for TaqMan SNP genotyping assays.**

TaqMan SNP Genotyping Thermocycling Program	
60°C for 30 sec	
95°C for 10 min	
95°C for 15 sec 60°C for 1 min	× 40 cycles
60°C for 30 sec	

3.12 *C9ORF72* genotyping

3.12.1 Patient Screening

As the *C9ORF72* hexanucleotide repeat expansion is the most frequently occurring gene mutation causing ALS, each new presenting familial and sporadic ALS case at Macquarie University Hospital Neurology Clinic was screened for the expansion using a combination of repeat primed PCR (Section 3.12.2) and rs3849942 genotyping (Section 3.12.3).

3.12.2 Repeat primed PCR

Classical PCR approaches are inadequate for analysis of the *C9ORF72* hexanucleotide repeat expansion owing to its GC density and repetitive nature. An alternative PCR technique, repeat–primed PCR is used for screening patients for the *C9ORF72* hexanucleotide repeat expansion. The commonly used method described by **Renton et al.** (2011) required further optimisation. Initially, DNA concentrations of 50ng, 100ng and 150ng were trialled according to Table 3.8. The use of two different *Taq* polymerases, My*Taq* HS Red Mix (Bioline) and One*Taq* (New England Biolabs, MA, USA) were also trialled in conjunction with an ExoSAP PCR cleanup protocol in which 0.3μl each of exonuclease I (New England Biolabs, MA, USA) and TSAP (Promega) and 7.5μl of distilled water was added to each reaction followed by incubation and enzyme denaturation as described above (Section 3.9). Fragment analysis of the *C9ORF72* repeat primed PCR products was performed on an ABI 3730XL sequencer (Macrogen, Korea) with subsequent analysis using Peak Scanner (v1.0) (Life Technologies).

3.12.3 rs3849942 genotyping

The rs3849942 G>A SNP has been shown to be frequently inherited as part of the common founder haplotype that segregates with the pathogenic *C9ORF72* hexanucleotide repeat expansion. The presence of the SNP does not guarantee that a patient necessarily carries an expansion, however expansion carriers will almost always present with the SNP. A TaqMan assay was designed and performed for this SNP as previously described (Section 3.11). Samples identified as harbouring the rs3849942 G>A SNP were validated by PCR and Sanger sequencing (Tables 3.3 and 3.6).

TABLE 3.8: **PCR reaction conditions trialled for optimising the *C9ORF72* repeat primed PCR reaction.**

	50ng DNA 1×	100ng DNA 1×	150ng DNA 1×
My*Taq* HS Red Mix	12.50	12.50	12.50
0.2mM deazoGTP	1.00	1.00	1.00
1M Betaine	5.00	5.00	5.00
DMSO	1.75	1.75	1.75
FAM forward primer	1.75	1.75	1.75
Anchor Reverse primer	1.75	1.75	1.75
Reverse Primer	0.88	0.88	0.88
50ng/μl DNA	1.00	2.00	3.00
Total	25.63	26.63	27.63

3.13 *SOD1* patient screening

Primer sets for each of the five *SOD1* exons were optimised using the standard conditions outlined in Table 3.4. As the second most frequently occurring gene mutation causing familial ALS, each new presenting familial ALS case at Macquarie University Hospital Neurology Clinic was screened for mutations in all *SOD1* exons using PCR and Sanger sequencing.

3.14 DNA cloning

3.14.1 Ligation

For ligation of an insert into the pGEM–T vector (Promega), the region of interest first underwent PCR amplification (Tables 3.3 and 3.6), followed by gel electrophoresis as previously described. The remaining 17μl of amplified product was also electrophoresed and gel extracted using the Bioline Isolate II PCR and Gel Kit according to manufacturers instructions. The sample was then quantitated using the NanoDrop 2000/200c Spectrophotometer (Thermo Fisher Scientific). A 10μl ligation reaction was performed using 1 unit of T4 DNA ligase (Promega), 2× ligation buffer (Promega), 50ng pGEM–T vector and the appropriate amount of DNA template as calculated according to manufacturers instructions. A control ligation was also conducted, using the provided control insert DNA in place of the PCR template.

All ligation reactions were thoroughly mixed by pipetting and incubated at room temperature for 6 hours.

3.14.2 Transformation

Resultant constructs were transformed into competent *E. coli* cells (Alpha–Select Gold efficiency, Bioline). Two microlitres of the ligation reaction was added to 25μl of pre–chilled competent cells, followed by gentle stirring and chilling on ice for 30 minutes. The cell suspension was heat–shocked at 42°C for 45 seconds, followed by cooling on ice for 2 minutes. One hundred and twenty five microlitres of SOC medium (Life Technologies) was then added and the mixture incubated at 37°C for 60 minutes with shaking at 200rpm. Ten microlitres and 100μl of the transformation mixture was then plated on separate pre–warmed LB agar plates supplemented with 0.5mM IPTG, 100μg/mL ampicillin and 50μg/mL X–Gal, inverted and incubated at 37°C overnight.

3.14.3 Selection of recombinant clones

Recombinant clones were identified using blue/white colony selection. A white colour indicated a recombinant colony containing the insert DNA, owing to insertional disruption of the β–galactosidase coding region located within the pGEM–T vector. The largest and most isolated white colonies were selected and screened for the insert by PCR. Half the colony was added to a standard PCR reaction utilising T7 and SP6 primers with thermocycling conditions as provided in Table 3.9.

TABLE 3.9: PCR reaction mixture and thermocycling conditions to check transformed *E.coli* colonies for target insert.

Reagent	**1× Volume (μl)**
dH_2O	4.2
My*Taq* HS red mix	5.0
20mM T7 Primer	0.4
20mM Sp6 primer	0.4
Colony	+
Total	10.0

Temperature	**Time**	**Cycles**
95°C	3 min	1×
95°C	15 sec	
55°C	15 sec	30×
72°C	1 min	1×
15°C	Hold	

3.14.4 Plasmid DNA Purification

Colonies that were confirmed to contain the DNA insert by PCR were subjected to plasmid DNA isolation. Recombinant clones were grown overnight in 5mL of LB medium supplemented with 100μg/mL ampicillin at 37°C with shaking at 200rpm. Glycerol stocks were prepared by adding 750μl of overnight culture to 250μl glycerol and stored at –80°C.The Isolate II nucleic acid isolation kit (Bioline) was used to purify plasmid DNA, which was eluted into 50μl and quantitated using the NanoDrop 2000/200c Spectrophotometer (Thermo Fisher Scientific).

3.15 Epigenetic Analysis

3.15.1 Zygosity testing of disease discordant twins and triplets

3.15.1.1 Microsatellite marker genotyping

Rutgers combined linkage–physical maps (http://compgen.rutgers.edu/rutgers_maps.shtml) (**Matise et al.**, 2007) were used to obtain heterozygosity and informative meiosis values to determine which microsatellite primer sets, fluorescently labelled with FAM and available in the laboratory, were most informative. After PCR optimisation for primer sets in Table 3.10 according to Table 3.4, the conditions outlined in Table 3.3 were used for PCR amplification of these microsatellite markers in the triplet and twin sets. Fragment analysis of PCR products was performed on an ABI 3730XL sequencer (Macrogen, Korea) and size analysis was completed using GeneMarker V2.6.3 software (SoftGenetics, PA, USA).

3.15.1.2 SNP genotyping

Minor allele frequencies (MAF) for the SNPs with primer sets available in laboratory stock inventory were obtained using a dbSNP batch query search (http://www.ncbi.nlm.nih.gov/SNP/dbSNP.cgi?list=rslist) and were used to determine the SNPs that would be most informative. After PCR optimisation for primer sets (Table 3.11) according to Table 3.4, the twin and triplet sets were subjected to genotyping for these SNPs, using the conditions outlined in Table 3.3.

TABLE 3.10: **Primer sequences, fluorescent label, product sizes and size range for microsatellite genotyping markers designed prior to candidature.**

Primer Name	5'–3' Sequence	Forward label	Range	PCR product size
D9S268_F	CACCAGGATATACATGACTCCATT	FAM	240–250	250
D9S268_R	GGCCTAACTTTTTAAGACAGCA			
D15S1038_F	ATTCTTGCTCACTCTGCTTC	FAM	121–147	145
D15S1038_R	CCCAGCCTTCCAGTGT			
D6S1651_F	CCAGGTATGTGGGATTGC	FAM	174–216	204
D6S1651_R	ACAAGCTCTGGAGTTGGAAG			
D3S3705_F	AGCTACTGGAGGATGACTG	FAM	228–236	234
D3S3705_R	AACCTGGGCATGGATA			
DXS8022_F	CTGTCACAGAAGTCCCATTTTA	FAM	160–188	168
DXS8022_R	GGAAACTAATGCAGCATGTC			
DXS991_F	ACTTCAACCACAGAAGCCTC	FAM	256–290	166
DXS991_R	ATCATTTGAGCCAATTCTCC			
DXS1190_F	ATCACCAGACAGAATCACC	FAM	210–212	218
DXS1190_R	TTTTATCCATTCAGCCAC			
DXS8032_F	CATTTTATTTTGCTTTGTATTTGGC	FAM	160–200	190
DXS8032_R	CTCCTAGAACAGTACCTGACACG			
D9S1674_F	GGCCTACCCTGTAGACTGA	FAM	216–232	218
D9S1674_R	TTAGAAGTGAGCCAAACTCAAA			
D16S411_F	TCATCTCCAAAGGAGTTTC	FAM	209–229	228
D16S411_R	GTGCATGTGTTCGTATCAA			
D17S802_F	GCCACCTGCCCCTCAA	FAM	166–188	178
D17S802_R	CTGCCAGCAGAGGCCA			
D8S601_F	TTGGCAATCACATTTCAGC	FAM	223–236	225
D8S601_R	GCACAGTTGGATCTTGTGTC			
D11S875_F	ACTGTCCTCTCATCCTACT	FAM	103–125	113
D11S875_R	TACAGAGCTGAGTTTGTAGC			
D20S892_F	CATGTAACTTCCTTAGAGGCACTG	FAM	177–223	209
D20S892_R	TATCTTGACTGCACTGTGGG			
D3S1555_F	GAAGGTGATGAAGCCAAGAG	FAM	221–237	221
D3S1555_R	CGGGAGTGACATGACATGAT			
D9S1869_F	CCTAACTTGAAGTGTAAATGGT	FAM	269–293	271
D9S1869_R	TTAGGGAAGACTGTCCTTAAT			

3.15.2 Methylation analysis

3.15.2.1 EpiTYPER assay design

CpG islands were identified for *C9ORF72* and *SOD1* using the UCSC genome browser (http://genome.ucsc.edu/). The CpG island track indicated which regions represented CpG islands, and the UCSF Brain DNA Methylation track was used to verify the quality of the CpG islands by providing DNA methylation data based on genome wide methylation analyses generated from postmortem human frontal cortex gray matter of a 57 year–old male using Methylated DNA immunoprecipitation (MeDIP–seq) and Methylation–sensitive Restriction Enzyme Sequencing (MRE–seq) (**Maunakea et al.**, 2010).

Table 3.11: **Primer sequences and product sizes for single nucleotide polymorphism genotyping markers designed prior to candidature.**

Primer Name	5'–3' Sequence	PCR product size
rs1474891_F	TTGTGCTTAGGTTCGGCATT	456
rs1474891_R	CCACTACGCCCAGCTAATGT	
rs763183_F	GCTTCTATGGAGCTTCCCTCTT	498
rs763183_R	AACCCTGACCAAATGACTTCAC	
rs1565948_F	GGTCAGTGAAGAAAAGTGCAAA	174
rs1565948_R	TGTCCCTTCTAGGGAGCTATTG	
rs2235580_F	TCCTCCGTCTGAGAGCACTT	242
rs2235580_R	GACATATGAGCAGCGTGGAG	
rs2282241_F	TCAGAATACTTGAAATACGGTGTCA	185
rs2282241_R	GAGCAGGTAAATGCTGGCTTAG	
rs2294605_F	AGAACAGCGATCCCATTCTG	220
rs2294605_R	CTGCATCGTAGGTGAGTCCA	
rs2473057_F	CCAATCTCCAGTTCTTTCTGGT	500
rs2473057_R	TCCCCGTGAGACAGTAGTACCT	
rs6600147_F	GCCGTAAAACAGCACACACA	250
rs6600147_R	AGGTGTGCCTGTAGGATGC	

4 Results

4.1 Genomic DNA extractions

Genomic DNA was successfully extracted from 3mL of whole blood from 206 new patients, family members and unrelated controls. Results are summarised in Table 4.1. The 260:280nm absorption ratio was used as an indication of DNA purity. All but one sample had a ratio between 1.8 and 1.9, indicating non–contaminated DNA.

TABLE 4.1: **Range and mean values for the quantity, concentration and 260:280 of 206 genomic DNA samples extracted from 3mL of whole blood.**

	μg per sample	**Concentration per sample (ng/μl per sample)**	**260:280 ratio**
Range	13.16–106.00	65.8–530.4	1.73–1.97
Mean	57.14	285.7±61.05	1.87±0.02

4.2 ALS gene primer set optimisation

Primer sequences and optimal PCR conditions were determined for new gene primer sets as shown in Table 4.2. Optimal PCR conditions were determined for previously designed gene primer sets as described in Table 4.3.

TABLE 4.2: **Primer sequences, product sizes and empirically determined optimal PCR reaction conditions and annealing temperatures for gene primer sets designed and optimised during candidature.**

Primer Name	5'–3' Sequence	PCR product size	Optimised PCR conditions	T_A
RBM14_Ex3_F	ACCTGGAGTCCTCCCCTC	329	My *Taq*+enhancer	61°C
RBM14_Ex3_R	GATCTGTAACCCTACCCACCTC			
SFPQ_Ex5_F	AAACCCCTCCTCGTTTTGC	365	My *Taq*+enhancer	58°C
SFPQ_Ex5_R	TGCATTTCTTTACGTTTCTGCA			
KDM2A_S104G_F	ccattgtctttcccattctttgt	231	My *Taq*	62°C
KDM2A_S104G_R	TCAATGCCTTTCTGTGTGTTCA			

TABLE 4.3: **Empirically determined optimal PCR reaction conditions and annealing temperatures for various gene primer sets designed prior to candidature.**

Primer Set	Optimised PCR conditions	T_A
SOD1_Ex1_New	My *Taq*+enhancer	64°C
SOD1_Ex2_New	My *Taq*+enhancer	64°C
SOD1_Ex3_New	My *Taq*	61°C
SOD1_Ex4_New	My *Taq*	61°C
SOD1_Ex5_New	My *Taq*	60°C
CCNF_Ex3	My *Taq*	67°C
OPTN_Ex10	My *Taq*	63°C

4.3 Analysis of known ALS genes

Over 20 ALS genes have been implicated in disease pathogenesis and together account for roughly two thirds of FALS cases. Of these, *C9ORF72* expansions account for one third of FALS while mutations in *SOD1* account for a further 20%. As such, it is important to establish whether these major ALS genes are the cause of ALS in each new patient before entering the gene discovery pipeline.

4.3.1 Optimised methods to detect the *C9ORF72* hexanucleotide repeat expansion

In order to identify ALS patients carrying a pathogenic hexanucleotide repeat expansion in *C9ORF72*, a combination of the repeat primed PCR and rs3849942 TaqMan genotyping was performed (Figure 4.1). To validate that the TaqMan

genotyping was accurate, selected samples that were positive for the rs3849942 SNP (as indicated in green and red in Figure 4.1) underwent PCR using SNP–specific primers (Table 3.6).

Optimal repeat primed PCR conditions were determined to include 100ng of DNA and My*Taq* HS Red Mix DNA polymerase according to Table 4.4. ExoSAP treatment which removes excess dNTPs and primers was trialled to remove small fragments from analysis however did not appear to improve the resolution of repeat fragment analysis. This was determined based on the clarity of the trace obtained after fragment analysis, as illustrated in Figure 4.2. Unfortunately, high background noise is still present in the resultant electropherogram (Figure 4.3). A positive rs3849942 result (i.e. G>A SNP) indicated the sample could potentially be positive for the *C9ORF72* hexanucleotide repeat expansion, while a negative rs3849942 result (i.e. GG homozygous) suggested a sample was negative for the expansion. As such, the SNP result aids in calling positive or negative expansion results in cases where an inconclusive result was obtained for the repeat primed PCR.

Table 4.4: **PCR reaction mix and touchdown thermocycling program used in repeat primed PCR for amplification of the *C9ORF72* hexanucleotide repeat expansion.**

Reagent	**1×**	**Volume (μl)**
My*Taq* HS Red Mix		12.5
0.2mM deazoGTP		1.0
1M Betaine		5.0
DMSO		1.75
20mM FAM Forward		1.75
20mM Anchor		1.75
20mM Reverse		0.88
50ng/μl DNA		2.0
Total		26.63

Temperature	**Time**	**Cycles**
95°C	10 min	1×
95°C	30 sec	
70°C *	45 sec	8×
72°C	3 min	
95°C	30 sec	
56°C	45 sec	32×
72°C	3 min	
Hold 15°C		

*annealing temperature decreases by 2°C each cycle

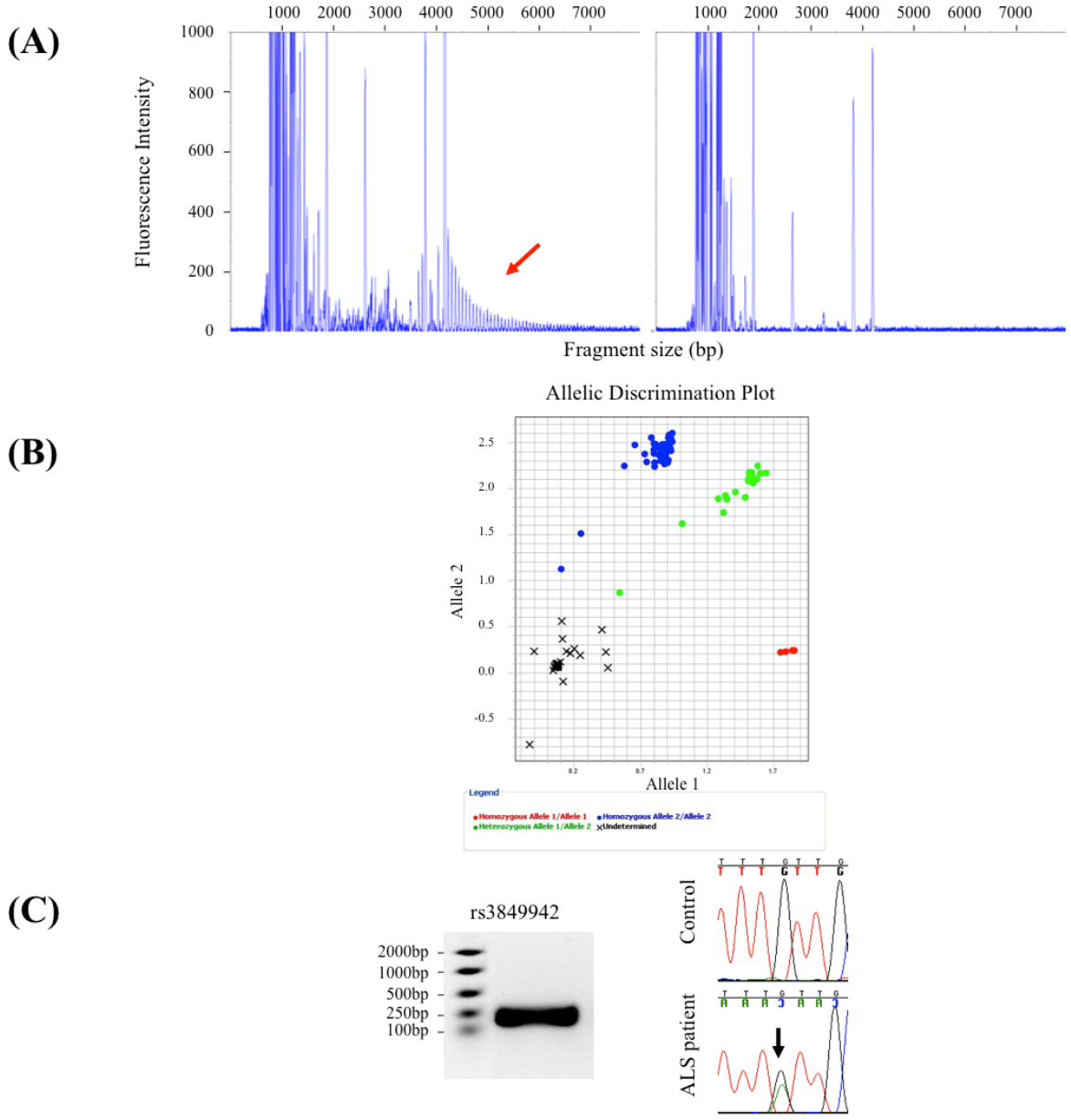

FIGURE 4.1: **Screening for the *C9ORF72* hexanucleotide repeat expansion.** A combination of repeat–primed PCR (image on the left is that of an expansion positive patient exhibiting a sawtooth pattern indicated by the red arrow, while that on the right is of a expansion negative patient) (A), rs3849942 genotyping by TaqMan analysis, where allele 1 is wildtype and allele 2 is the rare SNP (B) and validation of rs3849942 TaqMan positive samples by PCR sequencing (C) was determined to be the ideal strategy for *C9ORF72* screening.

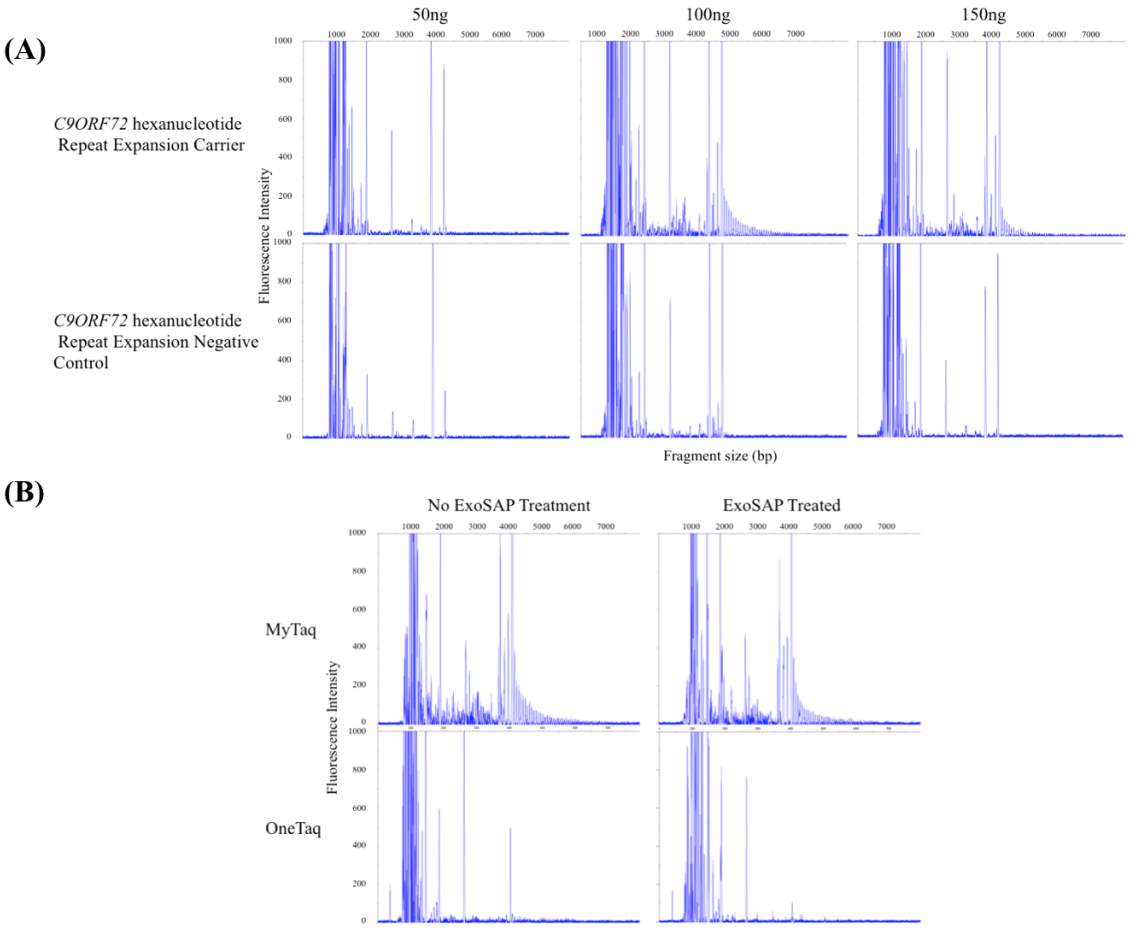

Figure 4.2: **Optimal *C9ORF72* repeat primed PCR conditions.** Optimal conditions for the *C9ORF72* repeat primed PCR were determined based on the clarity of resultant fragment analysis traces. Three different concentrations of DNA were trialled using My *Taq* and DNA samples known to be positive and negative for the *C9ORF72* hexanucleotide repeat expansion (A) and two *Taq* polymerases and the use of ExoSAP PCR cleanup were then trialled using 100ng of DNA known to be positive for the *C9ORF72* hexanucleotide repeat expansion (B). Red arrows indicate the characteristic sawtooth pattern associated with a repeat expansion.

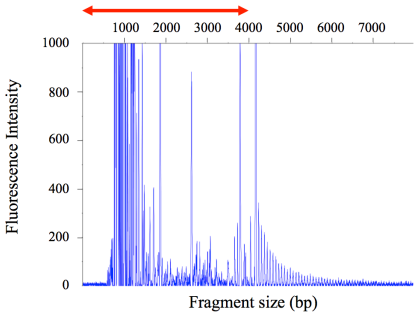

FIGURE 4.3: **Electropherogram resulting from *C9ORF72* repeat primed PCR.** Red arrow indicates the background noise of small fragments of no relevance, which ideally should be removed from analysis.

4.3.2 SALS and FALS patient screening for *C9ORF72*

Unfortunately, repeat primed PCR analysis of *C9ORF72* and rs3849942 genotyping by custom TaqMan analysis was not possible for newly presenting FALS and SALS patients within the time constraints of this project.

4.3.3 *SOD1* sequencing in FALS cases

Optimal PCR conditions for all five *SOD1* exons were determined as shown in Table 4.5 (Figure 4.4). All *SOD1* exons were successfully PCR amplified using these conditions in 21 FALS samples and a positive control sample. A negative control was included for each exon to confirm that no DNA contamination was present. Resultant sequencing chromatograms revealed that four patients carried *SOD1* (NM_000454) mutations as listed in Table 4.6 and shown in Figure 4.5. These samples were subsequently added to the cohort for methylation analysis using the Sequenom EpiTYPER.

TABLE 4.5: **PCR reaction mix and thermocycling conditions for all five *SOD1* exons.**

Reagent	1× Volume (µl)	
	Exons 1–2	Exons3–5
dH_2O	5.4	7.4
My*Taq* HS red mix	10.0	10.0
10x PCR Enhancer	2.0	0.0
10mM F primer	0.8	0.8
10mM R primer	0.8	0.8
20ng DNA	1.0	1.0
Total	20.0	20.0

Temperature	Time	Cycles	Optimised T_A	
95°C	10 min	1×		
94°C	15 sec		Exon 1	64°C
T_A°C	30 sec	30×	Exon 2	64°C
72°C	30 sec		Exon 3	61°C
			Exon 4	61°C
72°C	7 min		Exon 5	60°C
15°C	Hold			

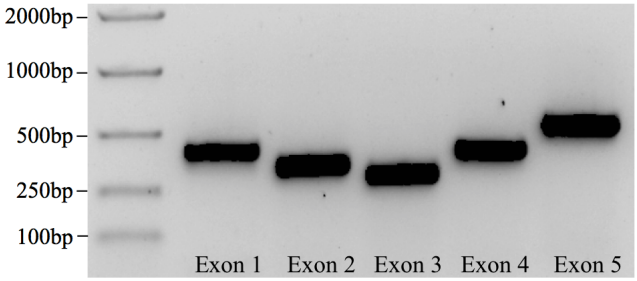

Figure 4.4: ***SOD1* exons 1–5 PCR optimisation.** PCR conditions for *SOD1* exons 1–5 (from left to right) which were subsequently used for familial patient screening for *SOD1* mutations.

Table 4.6: **PCR sequencing results for familial ALS patients identified as carrying mutations in *SOD1*.**

Patient /DNA Sample	Exon	Nucleotide Change	Amino Acid Change	Reference SNP ID	Original Report of Mutation
MQ130018	4	c.272A>C	p.D91A	rs80265967	(Andersen et al., 1995)
MQ130091	4	c.341T>C	p.I114T	rs121912441	(Rosen, 1993)
MQ140054	4	c.302A>G	p.E101G	rs121912439	(Rosen, 1993)
MQ130082	5	c.446T>G	p.V149G	NA	(Deng et al., 1993)

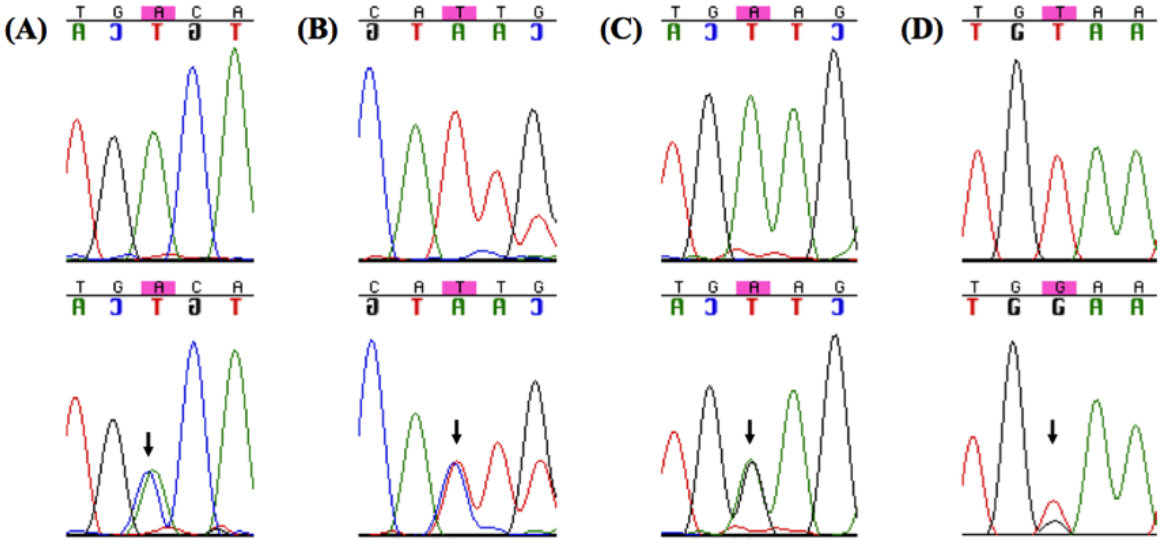

Figure 4.5: ***SOD1* mutations identified in the familial ALS cohort.** Top panels are control sequences and bottom panels are FALS patients. MQ130018 exhibited a p.D91A, c.272A>C mutation (A), MQ130091 exhibited a p.I114T, c.341T>C mutation (B), MQ140054 exhibited a p.E101G, c.302A>G mutation (C) and MQ130082 exhibited a p.V149G, c.446T>G mutation (D).

4.3.4 Targeted SALS patient screening for *SOD1*, *FUS* exon 15, *TARDBP* exon 6 and *C9ORF72*

As part of a collaborator's cell biology study, ten sporadic ALS patients required genotyping for the most frequently mutated ALS genes. DNA extracted from brain tissue was screened for gene mutations in *SOD1*, *TARDBP* exon 6 and *FUS* exon 15 using previously optimised PCR conditions (Table 3.6) and Sanger sequencing. A previously reported *TARDBP* (NM_007375) missense mutation (c.880G>C, p.G294A, rs80356721) (**Williams et al.**, 2009) was identified in one of the screened sporadic patients (Figure 4.6). No *SOD1* or *FUS* exon 15 mutations were found in the remaining patients.

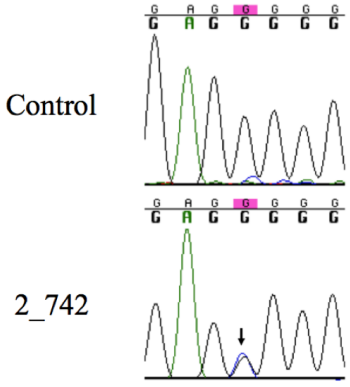

FIGURE 4.6: **One *TARDBP* exon 6 mutation was identified amongst ten sporadic ALS patients.** Patient 2_742 appears to carry the known ALS variant p.G294A, c.880G>C (rs80356721). This mutation has previously been described in an Australian sporadic ALS patient (**Williams et al.**, 2009).

Initial genotyping of the *C9ORF72* expansion associated SNP (rs3849942, section 3.12.3) by TaqMan and PCR validation revealed that three of the ten samples were heterozygous for the SNP (Table 4.7). Therefore, the samples potentially carry the *C9ORF72* hexanucleotide repeat expansion associated founder haplotype. The optimised repeat–primed PCR (Table 4.4) was performed for all ten samples. Fragment analysis did not identify the characteristic sawtooth trace (Figure 4.2) indicative of a repeat expansion in any sample (Table 4.7). An inconclusive result was obtained for one sample positive for the rs3849942 SNP. However, no definitive conclusion can be drawn as to whether this patient carries the *C9ORF72* hexanucleotide repeat expansion and should be confirmed with diagnostic Southern Blot analysis.

Table 4.7: ***C9ORF72* screening results for SALS samples used in cell biology studies.** Numbers 1–4 in the header row indicate the repeat primed PCR replicate number.

Sample	**RP PCR (1)**	**RP PCR (2)**	**RP PCR (3)**	**RP PCR (4)**	**rs3849942 TaqMan**	**rs3849942 PCR**	
						Initial PCR	Validation PCR
2_742	negative (6 repeats)	negative (6 repeats)			negative	GG	
3_550	negative (4 repeats)	repeat			negative	GG	
4_343	negative (5 repeats)	negative (5 repeats)			**heterozygous**	**GA**	**GA**
5_743	repeat (negative 5 repeats?)	negative (5 repeats)	inconclusive?	inconclusive?	**heterozygous**	**GA**	**GA**
6_076	negative (5 repeats)	negative (3 repeats)			failed	GG	
7_346	negative (7 repeats)	repeat negative (3 repeats)			negative	GG	
8_849	negative (5 repeats)	negative (5 repeats)			**heterozygous**	**GA**	**GA**
9_263	failed	repeat (negative 9 repeats?)			negative	GG	
10_441	negative (7 repeats)	negative (6 repeats)			negative	GG	

4.4 Discovery of genetic variants underlying ALS and candidate gene analysis

4.4.1 Mq1 Pedigree

To facilitate ALS gene discovery, the pedigree shown in Figure 4.7 was constructed for ALS family Mq1. Names, birth dates as well as dates and causes of death have been removed for anonymity. Information for all family members was obtained from personal communications. Information for family members in generations I–IV was found using ancestry.com. Additional information for family members marked with a blue asterisk was ascertained from death certificates, while clinical information was available and utilised for family members marked with an orange asterisk. Individuals underlined in purple provided blood samples and DNA has since been extracted. The green boxes indicate the same individual who was involved in two unions resulting in offspring. Further investigation revealed a consanguineous union which is indicated by a double line joining two individuals in generation VII.

4.4.2 Search_my_exome.sh

Programming code was written to produce a custom shell script shown in Appendix A.2.1 which is used to interrogate an individuals exome sequence data for the known ALS genes in Table 3.1, utilising variant report files obtained from the bioinformatics pipeline (alignment, variant calling, quality filtering and annotation). The resultant text file output lists all genomic variants present in that sample for all queried ALS genes, and was formatted to be opened in Microsoft Excel for further analysis. An example of the output file obtained is shown in Appendix A.3.1.

4.4.3 Relatedness testing

In some instances, apparently unrelated ALS families may share a recent common ancestor, and thereby carry the same mutation that causes ALS. In order to determine the level of relatedness between two individuals from separate families, the custom shell script shown in Appendix A.2.2 was designed using the PLINK/SEQ library for working with human genetic variation data (`http://atgu.mgh.harvard.edu/plinkseq/`), to use individual patient VCF files from exome sequencing data to determine the

Figure 4.7: **Nine generation pedigree for Mq1 constructed using personal communication, death certificates and investigation on ancestry.com.au.** Female family members are indicated by circles while male family members are indicated by squares, and diamond shapes are used in cases where multiple progeny of both genders are present. A black shape indicates an individual with ALS. Obligate carriers are indicated by dots. Deceased family members are represented by a diagonal line through the middle of the square/circle. Information for all family members was obtained from personal communications. Information for family members in generations I–IV was found using ancestry.com.au. Additional information for family marked with a blue asterisk was ascertained from death certificates, while clinical information was available and utilised for family members marked with an orange asterisk. Individuals underlined in purple provided blood samples and DNA has since been extracted. The green boxes indicate the same individual who was involved in two unions resulting in offspring. Further, investigation revealed a consanguineous union which is indicated by a double line joining two individuals in generation VII.

number of shared genotypes between each pair of individuals. This includes their total number of non–missing genotypes, i.e. the total number of genetic sites at which a genotype was found for both individuals. An example of the output text file is shown in Appendix A.3.2 and was designed to be easily opened in Microsoft Excel for subsequent calculation of the ratio between these two parameters, with a higher ratio indicating a closer level of relatedness. This aimed to determine if the ratio of shared alleles could be used as a measure of relatedness. Unfortunately no formula for relatedness could be established between this ratio and the number of meiosis separating two individuals. As such, bioinformatics collaborators at CSIRO carried out further investigations of relatedness testing by way of cluster analysis. This revealed that two ALS patients, 158–030516 and 304–100786 are likely to be distantly related (Figure 4.8). The common variants between the two were identified through NCBI Galaxy (`http://galaxy.nbic.nl/`) and underwent further bioinformatic filtering steps (Section 3.4). The *KDM2A* (NM_012308) p.S104G, c.310A>G variant was identified as an interesting candidate. Subsequent PCR (Tables 3.7 and 4.3) and sequencing suggested it was absent from both patients (Figure 4.9) and was thus confirmed as a false positive. Visual observation of the nucleotide position in IGV revealed the *KDM2A* c.310A>G variant was present in exome sequencing reads in almost every individual (Figure 4.10A), strongly suggesting it was a sequencing artefact. Careful inspection of the sequence reads across several individuals showed that the G nucleotide was present at the end of low quality sequencing reads (Figure 4.10B). A bioinformatic analysis pipeline should ideally remove low quality bases at the start or end of a read, however this was not the case for these two individuals. *In silico* protein predictions suggested that this variant would be highly damaging, is contained in an unintegrated signature domain, and is highly conserved across species (Table 4.8).

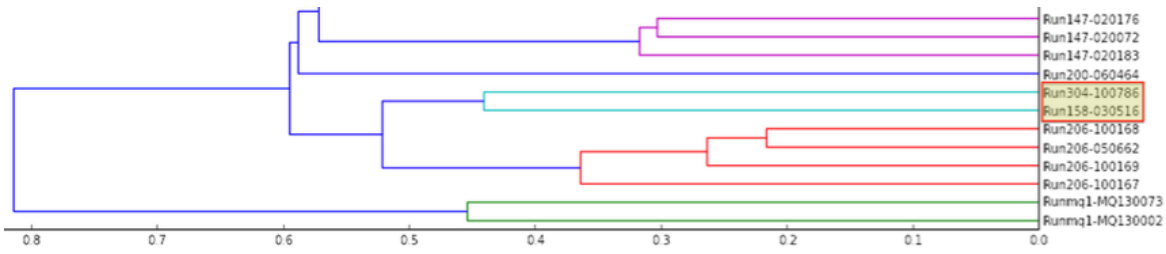

FIGURE 4.8: **Screenshot of cluster analysis performed by collaborators on exome sequencing data for individual patients.** The red shaded box indicated the two individuals identified as distantly related.

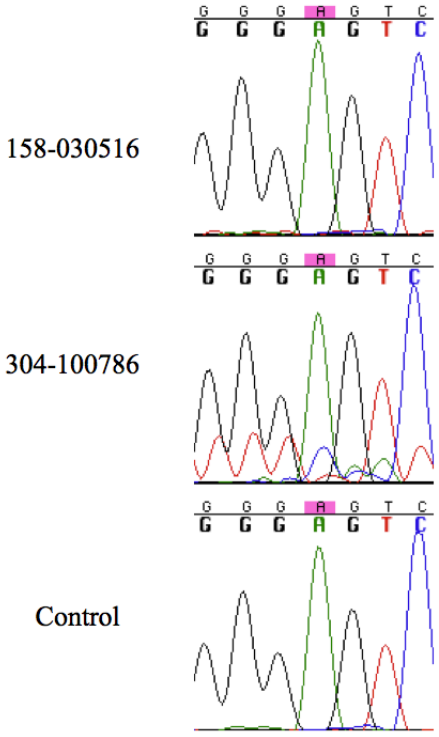

Figure 4.9: **The *KDM2A* p.S104G, c.310A>G mutation identified by exome sequencing was absent from two distantly related ALS patients.** PCR and Snager sequencing of 158–030516 and 304–100786 revealed that the c.310A>G variant was absent from these two individuals and was a false positive identification by exome sequencing.

Table 4.8: ***In silico* predictions of the pathogenic likelihood of the *KDM2A* p.S104G , c.310A>G mutation.**

	***In silico* Prediction**
Mutation Taster	Disease Causing; 0.999
PolyPhen	Probably Damaging;0.987
Pon_P2	0.242
NetPhos	No change
Domain	Unintegrated signature
Species conservation	Drosophila, Chicken, Mouse, Pufferfish, Chimp, Macaque

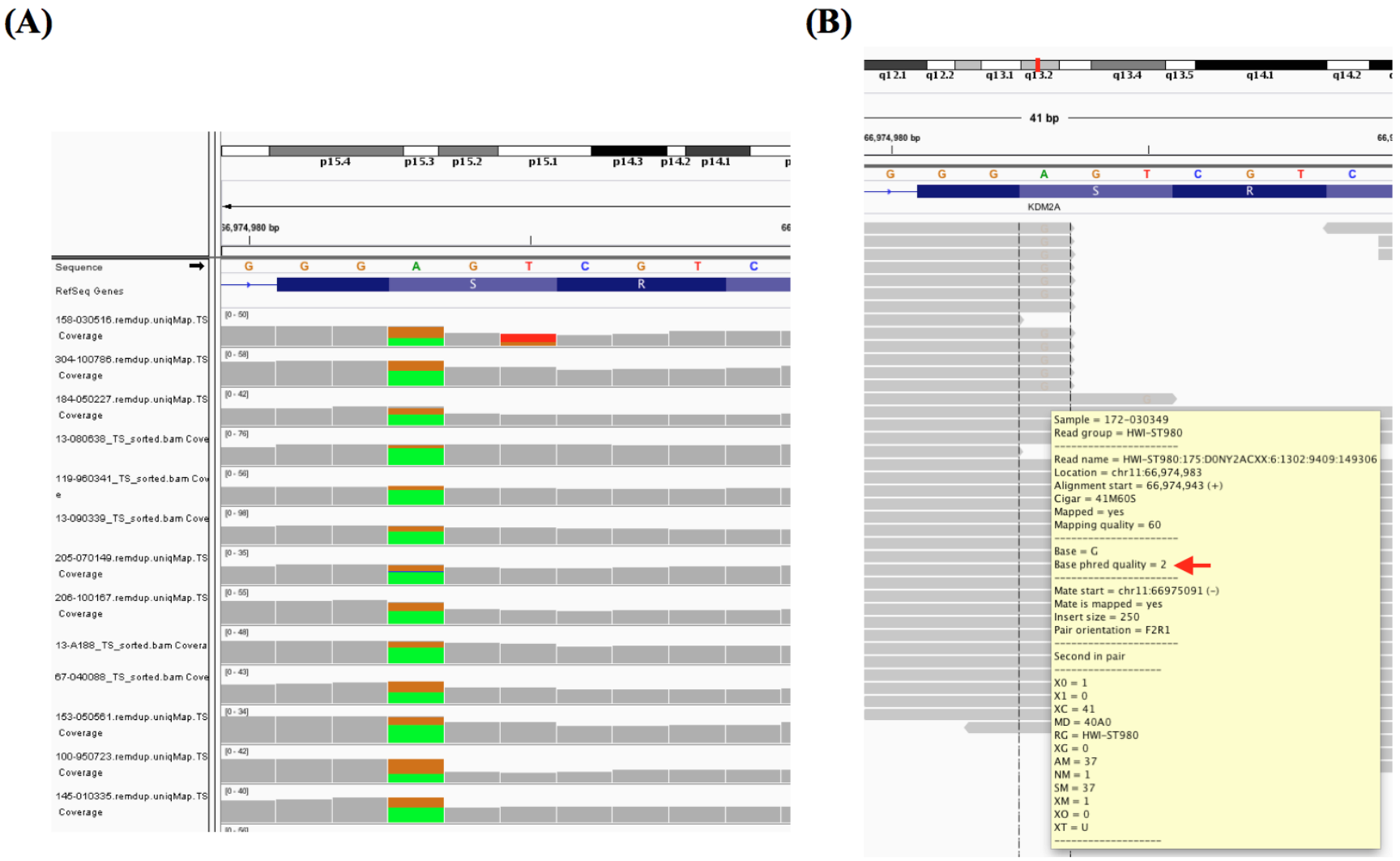

FIGURE 4.10: **Screenshots of visualisation of the *KDM2A* variant p. S104G , c. 310A>G using the Intergrative Genomics Viewer in the exome sequenced cohort.** It is apparent that most individuals do display the Guanine variant (indicated by brown blocks) in addition to the reference Adenine (indicated by green blocks), not only the apparently distantly related ALS patients 158–030516 and 304–100786 (A). This base generally occurs at the end of sequencing reads and is of poor quality, as demonstrated by the poor quality score indicated by the red arrow (B).

4.4.4 Analysis of candidate variants identified by exome sequencing and custom scripts

4.4.4.1 *SFPQ* and *RBM14*

After rigorous bioinformatic testing (Section 3.5) for the candidate genes listed in Table 3.2, a total of 84 variants withstood analysis, and of these *SFPQ* (NM_005066) p.N533H, c.1597A>C and *RBM14* (NM_006328) p.S620X, c.1859C>A were identified as interesting ALS–linked candidates based on protein structure and function knowledge.

After optimising PCR primer sets for *SFPQ* exon 5 and *RBM14* exon 3 (Table 4.2), sequencing validated the presence of the *SFPQ* p.N533H, c.1597A>C mutation in patient 294–100203 (Figure 4.11A). Further, to confirm the absence of the mutation in control samples, TaqMan SNP genotyping (Section 3.11) revealed that the *SFPQ* p.N533H, c.1597A>C mutation was absent from 250 non–related control samples. Interrogation of control exome sequence data indicated the absence of *SFPQ* p.N533H in a further 1000 Australian control samples, supporting its role as a ALS mutation. In contrast, patient 117–960663 was found to be negative for the *RBM14* p. S620X, c.1859C>A mutation (Figure 4.11B). This result, coupled with low read depth from exome sequencing, suggests that the *RBM14* mutation was a false positive.

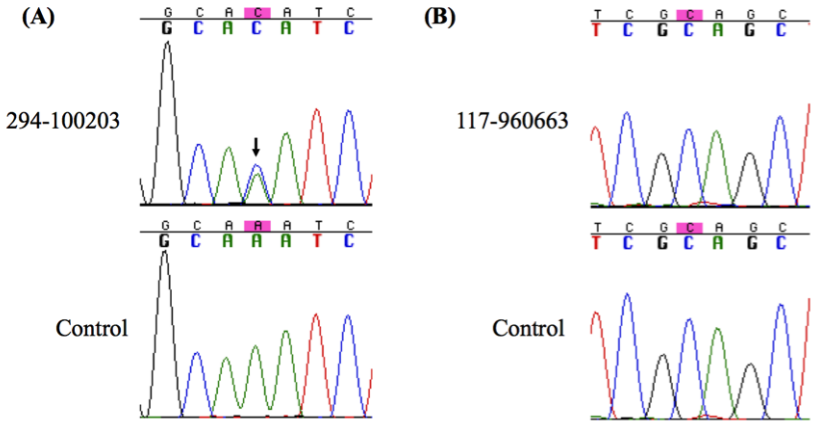

Figure 4.11: **Candidate ALS gene validation.** The candidate mutations in *SFPQ* exon 5 and *RBM14* exon 3 that were identified in patient exome sequence data were examined by Sanger sequencing in patient DNA for validation purposes. (A) The *SFPQ* p.N533H, c.1597A¿C mutation was confirmed by Sanger sequencing. (B) The putative *RBM14* p.S620X, c.1859C¿A mutation was not present following Sanger sequencing, indicating that it was a false positive.

4.4.4.2 *CCNF*

Because *CCNF* exon 1 was not captured by TruSeq Exome capture and was therefore not covered in Australian patient exome sequence data, this exon was screened in 27 FALS probands by Sanger sequencing (Table 3.6), with no mutations identified. TaqMan SNP genotyping (Section 3.11) of non–related control samples was performed for the *CCNF* (NM_001761) variants p.S3G (c.123A>T) and p.H69Y (c.205C>T), which had previously been identified in ALS cohorts by collaborators. No p.S3G or p.H69Y mutations were identified in 649 and 587 non–related control samples respectively. Four control samples appeared to carry the p.H69Y mutant allele, however upon PCR and direct sequencing using the CCNF_Ex3 primer set (Table 4.3), three were shown to be false positives. The fourth sample, NRCO19 appeared to carry the mutation (Figure 4.12A) and was repeated. Unfortunately, the result was unclear and thus the PCR product was used as the template in a new nested PCR reaction, which resulted in a homozygous mutant upon sequencing (Figure 4.12B). Using a new genomic DNA primer set (Table 3.6), the reaction was repeated but failed, suggesting the patient DNA was of poor quality. At this point, there was insufficient patient DNA to repurify, and insufficient time to source a kit for whole genome amplification. As such, an attempt was made to clone the remnant PCR product into pGEM–T, however transformation of the product failed to produce any clones containing the DNA insert. Therefore the presence of the H69Y mutation in NRCO19 could not be confirmed in the time–frame of this study.

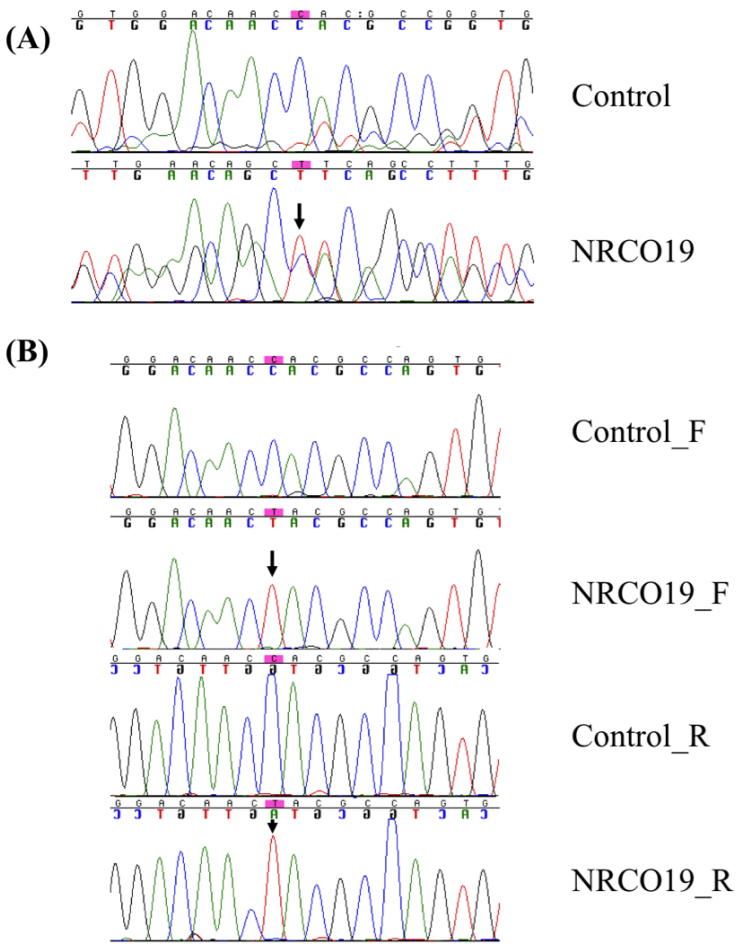

Figure 4.12: **Investigation of a *CCNF* p.H69Y, c.205C>T mutation in NRCO19.** After a positive result for *CCNF* H69Y TaqMan SNP genotyping (not shown), a heterozygous *CCNF* p.H69Y, c.205C>T was identified (A) and after a nested PCR an apparently homozygous mutant was identified (B).

4.4.4.3 *VCP*

A patient presenting with features of ALS, FTD, and Inclusion Body Myopathy (IBM), that are often associated with patients harbouring a VCP mutation, was initially screened for mutations in all 17 *VCP* exons by a member of our laboratory. Sequence analysis revealed that the patient carried a p.R159C, c.475C>T mutation in exon 5 (NM_007126), which was validated using the PCR conditions in Table 3.6 and Sanger sequencing (Figure 4.13A). Clustal Omega analysis revealed that the Arginine residue at position 159 of VCP is highly conserved across species (Figure 4.13B). The patient also appeared to carry a variation in the length of a polyA region residing within exon 14 at c.1839–1847 (Figure 4.13C). After PCR amplification of this region (Table 3.6) and subsequent cloning (Section 3.14), three alleles appeared to be present,

one carrying each of eight, nine and ten Adenine bases (Figure 4.13D).

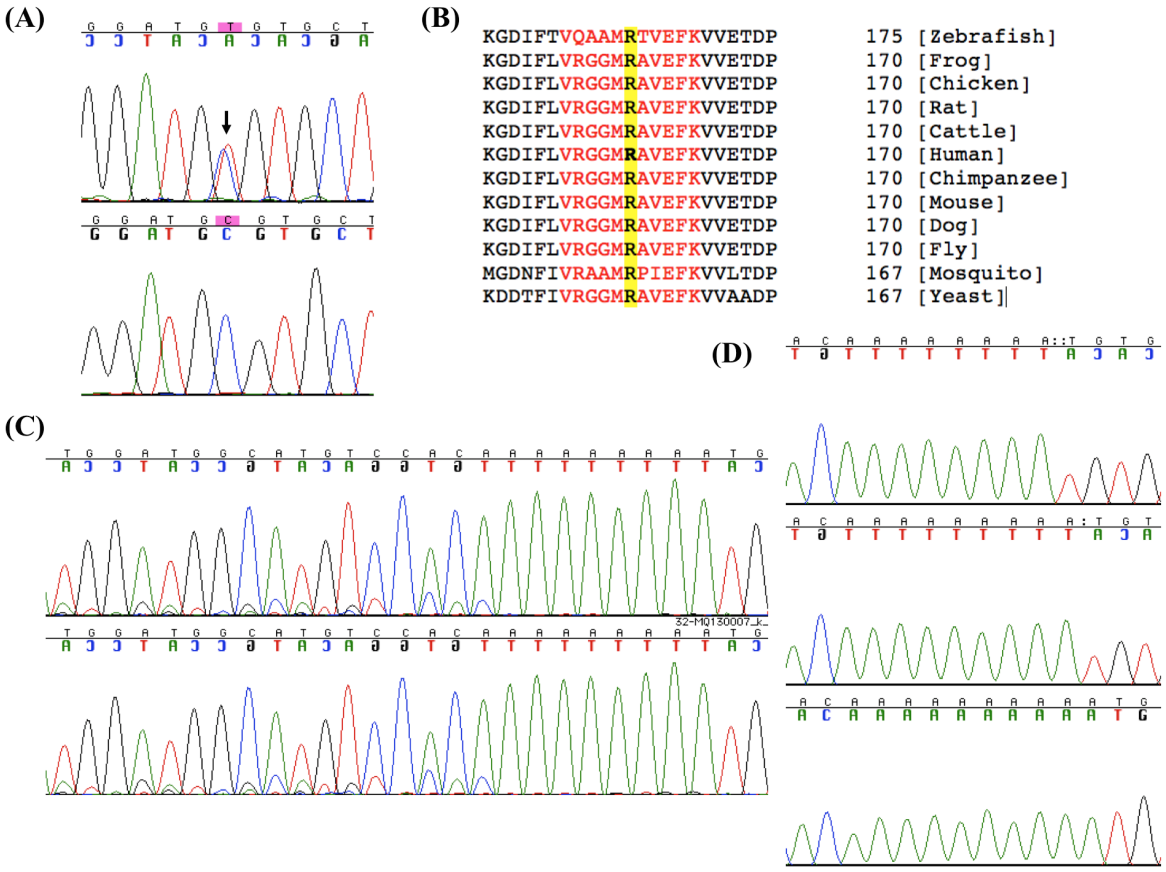

FIGURE 4.13: **Mutation analysis of *VCP* in a patient presenting with features of ALS, FTD and IBM that are often assoctaied with VCP mutations.** PCR sequencing identified a p.R159C, c.475C>T mutation in *VCP* exon 5 (A). Clustal Omega analysis revealed that the mutated Arginine residue is highly conserved between species (B). A possible insertion mutation was identified in a polyA region of exon 14 (C) and upon pGEM–T cloning, three different alleles containing either eight, nine or ten Adenine bases were identified (D).

4.4.4.4 *OPTN*

Using optimised PCR conditions (Table 4.3) the presence of the *OPTN* (NM_001008213) p.V295F, c.883G>T mutation in the index sample was confirmed (Figure 4.14). TaqMan SNP genotyping analysis (Section 3.11) was performed to confirm the absence of the *OPTN* variant p.V295F in control samples. No non–related control samples carried this mutation amongst 263 samples, nor was it present in exome sequence from a further 967 controls.

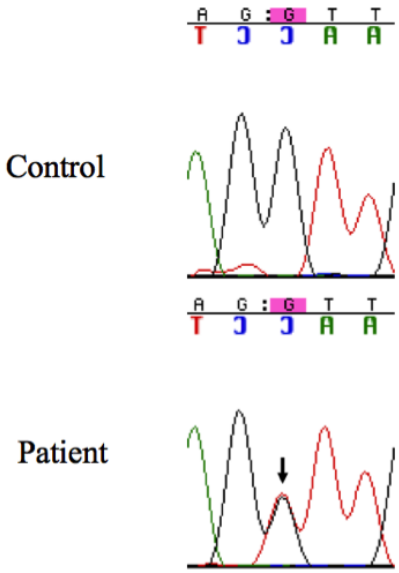

Figure 4.14: **The *OPTN* p.V295F, c.883G>T mutation was confirmed in Mutation analysis of *VCP* in a patient presenting with features of ALS, FTD and IBM that are often assoctaied with VCP mutations, by PCR and Sanger sequencing.**

4.5 Epigenetic Analysis

4.5.1 Zygosity testing

Disease discordant twin and triplet sets where one individual has the disease and the other does not are powerful tools for identifying epigenetic regulators of disease. However, before proceeding towards expensive twin studies, it is vital to confirm that the twin and triplet sets are monozygotic (i.e. identical) and not dizygotic. Microsatellites are short tandem repeats of 2–5 base pairs which are variable in number and highly polymorphic, thus serving as valuable genetic markers. SNPs that occur commonly in a population can also serve as informative genetic markers. Thus here both microsatellites and SNPs have been used for zygosity testing of twin and triplet sets.

The number of informative meiosis and heterozygosity values for the available microsatellite markers were extracted from Rutgers maps (Appendix A.4.1) and a combination of the highest values for both parameters were utilised to determine that

those highlighted would be most informative for zygosity testing. Similarly, MAF values for the available SNPs were found using a dbSNP batch query (Appendix A.4.2). A combination of the highest MAF values and spread across chromosomes was employed to determine that those highlighted would be most informative for zygosity testing.

Optimal PCR conditions were determined for both microsatellite (Table 4.9) and SNP (Table 4.10) markers. A set of reportedly MZ triplets comprising one ALS affected individual and two unaffected siblings (Figure 4.15), as well as a twin set comprising one ALS affected individual and one unaffected sibling (Figure 4.16), underwent zygosity testing using a range of microsatellite and SNP markers (using respective optimal PCR conditions, Tables 4.9 and 4.10). For the triplet set, nine microsatellites across seven different chromosomes and seven SNPs across 4 chromosomes were genotyped while twelve microsatellites across nine different chromosomes and two SNPs over two chromosomes were genotyped for the twin set. Zygosity testing (Table 4.11) suggested that within in the triplet set, the affected person and one of their unaffected siblings are monozygotic, with identical genotypes for nine microsatellite markers and two SNP markers. However, the second unaffected triplet appears to be dizygotic when compared with the other two, as they carry different genotypes for two microsatellite markers. Table 4.12 indicates that the twin set is also monozygotic, sharing identical genotypes for all twelve microsatellite markers and seven SNP markers.

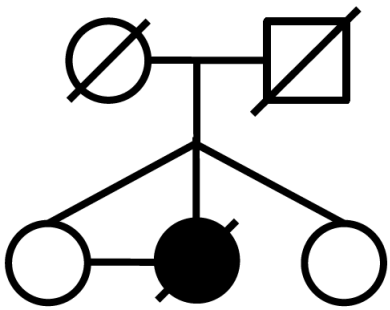

FIGURE 4.15: **Pedigree of the triplet set.**

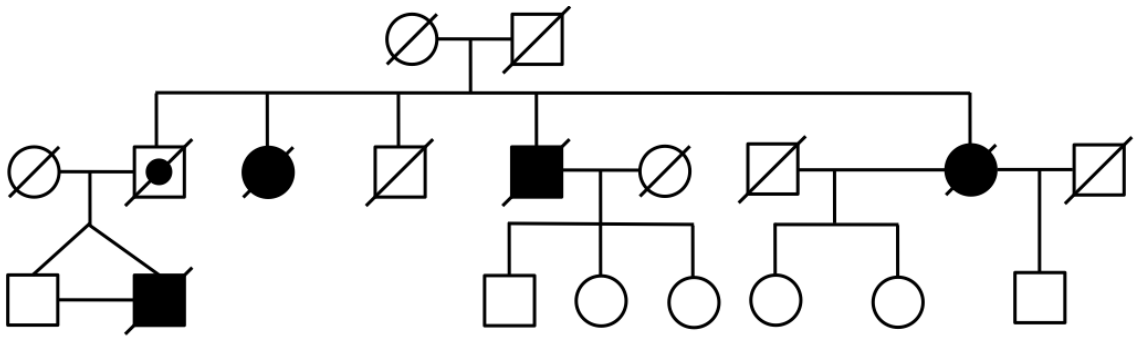

Figure 4.16: **Pedigree of the twin set.**

Table 4.9: **Empirically determined optimal PCR reaction conditions and annealing temperatures for genotyping with microsatellite markers.**

Microsatellite marker	Optimised PCR conditions	T_A
D9S268	My *Taq*	62°C
D15S1038	My *Taq*+enhancer	61°C
D6S1651	My *Taq*	62°C
D3S3705	My *Taq*	60°C
DXS8022	My *Taq*	62°C
DXS991	My *Taq*	66°C
DXS1190	My *Taq*+enhancer	61°C
DXS8032	My *Taq*	61°C
D9S1674	My *Taq*	63°C
D16S411	My *Taq*	61°C
D17S802	My *Taq*	68°C
D8S601	My *Taq*	61°C
D11S875	My *Taq*	59°C
D20S892	My *Taq*	61°C
D3S1555	My *Taq*	64°C
D9S1869	My *Taq*	58°C

TABLE 4.10: **Empirically determined optimal PCR reaction conditions and annealing temperatures for genotyping selected single nucleotide polymorphic markers.**

SNP primer set	Optimised PCR conditions	T_A
rs1474891	My *Taq*	61°C
rs763183	My *Taq*	66°C
rs1565948	My *Taq*	58°C
rs2235580	My *Taq*	64°C
rs2282241	My *Taq*	60°C
rs2294605	My *Taq*	61°C
rs2473057	My *Taq*	64°C
rs6600147	My *Taq*	66°C

TABLE 4.11: **Genotype data for zygosity testing of a triplet set using microsatellite and single nucleotide polymorphic markers.** Individual 990306 is affected by ALS while siblings 990370 and 990705 are unaffected.

	990306	**990370**	**990705**
D3S3705	191, 193	191, 193	191, 193
D6S1651	181, 181	**145, 145**	181, 181
D8S601	188, 186	188, 186	188, 186
D9S268	196, 198	**216, 218**	196, 198
D9S1674	186, 186	186, 186	186, 186
D16S411	189, 191	189, 191	189, 191
D20S892	184, 188	184, 188	184, 188
DXS8022	166, 168	166, 168	166, 168
DXS8032	175, 175	175, 175	175, 175
rs1565948	TT	TT	TT
rs2235580	GG	GG	GG

Table 4.12: **Genotype data for zygosity testing of a twin set using microsatellite and single nucleotide polymorphic markers.** Individual MQ140066 is affected by ALS while sibling MQ140115 is unaffected.

	MQ140066	**MQ140115**
D3S3705	191, 193	191, 193
D3S1555	188,188	188,188
D6S1651	181, 185	181, 185
D8S601	192, 194	192, 194
D9S268	196, 200	196, 200
D9S1674	187, 191	187, 191
D9S1869	210, 212	210, 212
D11S875	133, 137	133, 137
D15S1038	146, 148	146, 148
D16S411	259, 259	259, 259
D17S802	167, 167	167, 167
D20S892	186, 186	186, 186
rs763183	TT	TT
rs1474891	TT	TT
rs1565948	TC	TC
rs2235580	GA	GA
rs2282241	GG	GG
rs2294605	TT	TT
rs2473057	CC	CC

4.5.2 Methylation Analysis

Not only are twins/triplets a valuable research asset, large disease discordant cohorts that harbour an identical gene mutation (in the case of *C9ORF72* positive individuals) or containing mutations in the same gene (including *SOD1* families) are also a powerful tool for identifying genetic and epigenetic regulators of disease.

The large disease discordant, mutation known DNA sample cohorts that are available in the laboratory, will undergo specific quantitation of CpG site methylation within the promoter region of either the *C9ORF72* gene or *SOD1* gene, respective to the mutation in the individual using the Sequenom EpiTYPER system in future studies. The cohort and experimental design are the current focus.

4.5.2.1 EpiTYPER Design

To facilitate ongoing epigenetic studies of disease discordance in *C9ORF72* and *SOD1* positive ALS patients including the MZ twins identified above, CpG islands were identified for both *C9ORF72* and *SOD1* using the UCSC genome browser (Section 3.15.2.1). CpG islands were identified for *C9ORF72* at chr:27573760–27573987 (227bp) and chr9:27572967–27573553 (586bp) (Figure 4.17). The CpG islands overlap the promoter regions of exons 1a and 1b for all three *C9ORF72* transcripts (NM_018325.3, NM_001256054.1 and NM_145005.5). The *SOD1* (NM_000454) CpG island was identified as chr21:33031735–33032657 (922bp) at the 5' end of the gene and overlapping the promoter region and the first exon (Figure 4.18).

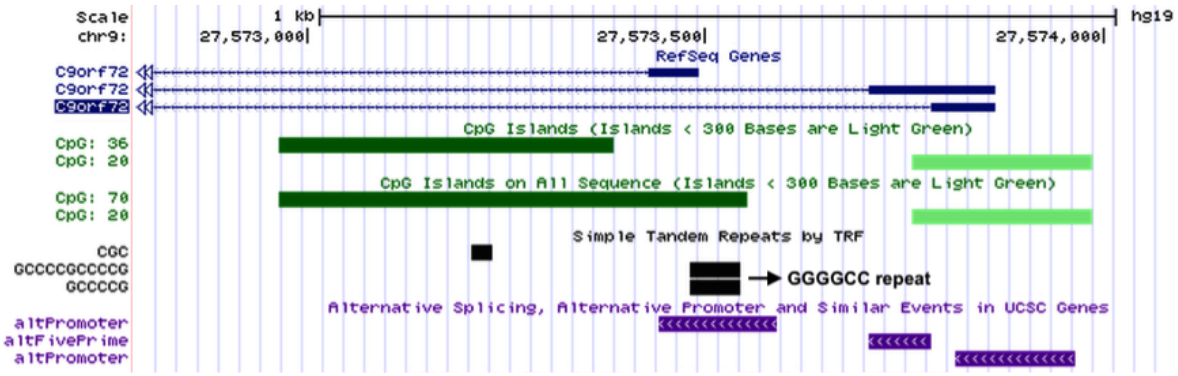

Figure 4.17: **CpG islands flanking the *C9ORF72* hexanucleotide repeat expansion.** Using the UCSC database (http://genome.ucsc.edu) two CpG islands were predicted to surround the *C9ORF72* hexanucleotide repeat region. The CpG islands overlap the promoter regions of exons 1a and 1b for all three *C9ORF72* transcripts (NM_018325.3,NM_001256054.1 and NM_145005.5).

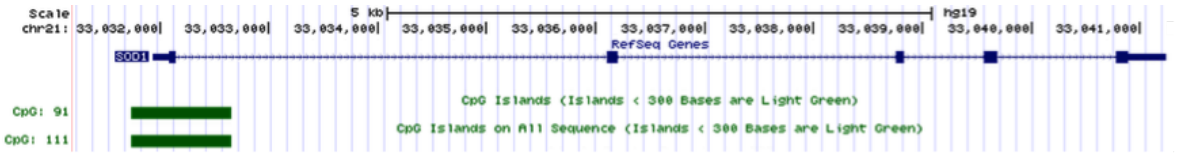

Figure 4.18: C**pG islands flanking the *SOD1* gene.** Using the UCSC database (CpG islands flanking the *SOD1* gene. Using the UCSC database (http://genome.ucsc.edu) a single 5 CpG island overlaps the promoter region of the *SOD1* gene (NM_000454)).

5

Discussion

Genetic analyses of ALS lay the foundation on which downstream *in vitro* and *in vivo* studies of this devastating disease are built, in the hope that one day these efforts will lead to the development of effective diagnostic and treatment strategies to improve outcomes for patients. As can be seen in Figure 5.1, there is a complex ALS genetic analysis pipeline performed in our laboratory. The various studies performed in this project are part of this pipeline. Despite many genetic mutations implicating pathogenic roles in ALS, there is still a great need to identify the underlying genetic cause in newly presenting ALS families, whether it be to determine the presence of a previously known ALS gene or to discover new ALS variants.

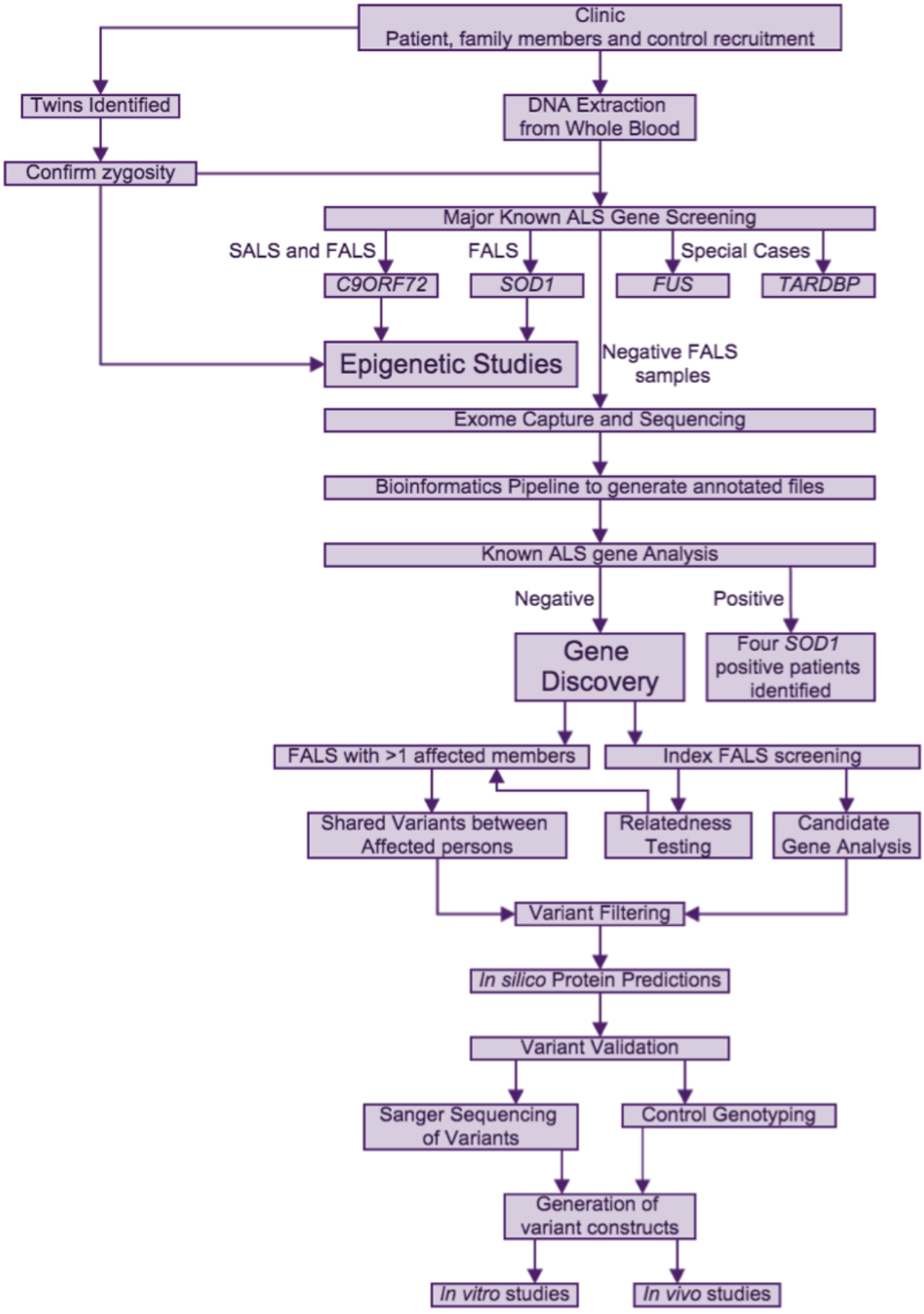

FIGURE 5.1: **Graphical representation of the ALS genetic analysis pipeline employed here.**

5.1 Genomic DNA extractions

Using an automated protocol for whole blood DNA extractions, high quality and quantity DNA has been extracted from whole blood for use in downstream analyses. The use of the automated protocol using the QIAsymphony DSP DNA midi kit on the QIAsymphony Sample Preparation System (Qiagen) consistently produces DNA of of high quality, exemplified by an average 260:280 ratio of 1.87, and concentration of 285.7ng/μl obtained from each 2μl elution equating to an average of 57.14μg of DNA produced from each 3mL of whole blood. This is in stark contrast to lower purity DNA with significantly lower concentrations of generally just 150-300ng/μl obtained from manual extraction protocols, which were naturally more laborious, time consuming and thus prone to contamination than the automated procedure. Thus the current automated procedure is superior to manual protocols in not only quality and quantity but also efficiency. As illustrated in Figure 5.1, DNA extractions are central to the genetic pipeline, thus it is imperative that DNA of such a high standard as produced by the current methodology, is extracted from peripheral whole blood samples as this high quality and concentration DNA is required for use in a wide variety of downstream genetic assays, most notably PCR reactions from which products are frequently obtained for further analysis such as direct sequencing or construct generation for *in vitro* and *in vivo* studies. If DNA samples are of substandard quality, downstream analyses utilising such samples will be imperatively difficult to interpret and any results obtained will be prone to inaccuracies and thus any conclusions drawn will be unreliable.

5.2 Known ALS gene analysis

When new ALS patient samples are obtained from the clinic, the first step is to determine whether the patient carries a mutation in a known ALS gene, as indicated in Figure 5.1. If this is the case, the patient and/or family will be excluded from further gene discovery pipelines as the causative mutation is apparent. However, identifying cases of known ALS genes is still of vital importance as healthcare management, disease prognosis and family planning may be dependent on the underlying genetic mutation (**Katsanis and Katsanis**, 2013). Inclusion of mutation-known patients and/or families in downstream research efforts aimed at further characterisation of disease modifiers, epidemiology, aetiology or pathology specific to a particular ALS gene are also dependent on accurate classification of the underlying ALS variant.

5.2.1 *C9ORF72*

Since the discovery that a GGGGCC hexanucleotide repeat expansion in the intronic region of *C9ORF72* accounted for roughly one-third of familial ALS and approximately 6% of SALS cases (**DeJesus-Hernandez et al.**, 2011; **Renton et al.**, 2014, 2011), there have been extensive efforts to further characterise this major ALS gene. The GGGGCC repeat expansion in *C9ORF72* is the only known copy number variation (CNV) ALS mutation, and requires a unique form of detection. Here an optimised methodology was developed for classifying ALS patient DNA samples as positive or negative for the repeat expansion, which includes repeat primed PCR and rs3849942 SNP genotyping.

To optimise the repeat primed PCR assay, different DNA concentrations were trialled using previously optimised conditions. This revealed that 50ng of DNA was insufficient to produce a clear sawtooth pattern for known expansion carriers; however, both 100ng and 150ng of DNA were in fact sufficient for this purpose (Figure 4.2). An increase in DNA concentration from 100ng to 150ng produced a negligible improvement in the quality of sawtooth pattern obtained from fragment analysis. As such, the use of extra DNA was seen as wasteful, and the 100ng DNA concentration condition as favourable. This is particularly relevant, as these reactions need to be carried out in duplicate, again owing to the erratic nature of the assay. The trial of two different *Taq* polymerase mixes clearly showed that My*Taq* (Bioline) produced a much clearer trace than did One*Taq* (New England Biolabs). This result was surprising as One*Taq* was developed and marketed for use with high GC content regions. The only downfall of the current conditions is the high background noise seen for short genomic regions in the electropherogram (Figure 4.3). ExoSAP treatment was used in an effort to remove this background noise, but was unsuccessful. A further attempt at removing these small noise fragments may be trialled by performing a gel extraction of the PCR product at the required size.

The repeat-primed PCR is necessary to amplify this expanded repeat as classical PCR approaches are not amenable to these repeat expansions as they are far too large, repetitive and GC rich to be faithfully amplified using simple forward and reverse primers coupled with Sanger or Next Generation sequencing (NGS) (**Chen et al.**, 2010). As such, the repeat primed PCR consists of three primers as seen in Figure 5.2, that is two traditional gene specific forward and reverse primers flanking the repeat

region, with a fluorescent tag on the forward primer for subsequent fragment length analysis by capillary electrophoresis, as well as an anchor primer complementary to the repeat region (**Chen et al.**, 2010; **Renton et al.**, 2011). Additionally, dGTP is substituted for 7-deaza-2-deoxy GTP as signal intensity is increased under such conditions (**Hantash et al.**, 2010). A touchdown thermocycling program with gradually decreasing annealing temperatures is utilised in order to increase specificity for the repeat region. The theory underlying this technique is that the anchor primer may bind at different points throughout the repeat expansion so that amplicons of incrementally differing sizes are produced. The reverse primer is present at a lower concentration so that once it is exhausted the anchor primer is preferentially used in its place. The fragment lengths produced are then analysed using capillary electrophoresis, with samples carrying the expansion displaying an elongated sawtooth pattern of fluorescence intensity versus base pair length relative to that of controls (**Renton et al.**, 2011), as is observed in Figure 4.1A. This assay design has proven highly robust, with concordance rates of 98.6% and 100% (**Hantash et al.**, 2010) with Southern blot analysis for the classification of the full range of expanded alleles observed in Fragile X syndrome.

Traditionally, detection of large repeat expansions has been achieved using Southern Blot analysis. However, despite this approach giving accurate quantitative results, it is highly labourious. Southern Blotting typically requires over 10μg of DNA and is immensely time consuming, in addition to requiring expensive membranes and affinity reagents including radioactive protocols, potentially leading to toxic exposure. Fluorescence *in situ* hybridisation (FISH) has also been used to confirm repeat length (**Renton et al.**, 2011), but it too requires large amounts of DNA as well as technically challenging karyotyping analysis using fluorescence microscopy at the chromosome level. The alternative approach of repeat-primed PCR is far simpler and less technically challenging.

By nature, the repeat-primed PCR can only detect repeat expansions of approximately 60 repeats (**Renton et al.**, 2011), in contrast to the more accurate quantification provided by Southern Blotting. However, as a classification tool this resolution is sufficient. Despite some controversy as to the repeat size distinction between pathogenic and normal alleles, the benchmark is set at more than 30 repeat units indicating a pathogenic allele, while wildtype is considered to contain 23 or less (**DeJesus-Hernandez et al.**, 2011). Therefore, using the repeat-primed PCR method, samples may be classified as *C9ORF72* repeat expansion positive or negative.

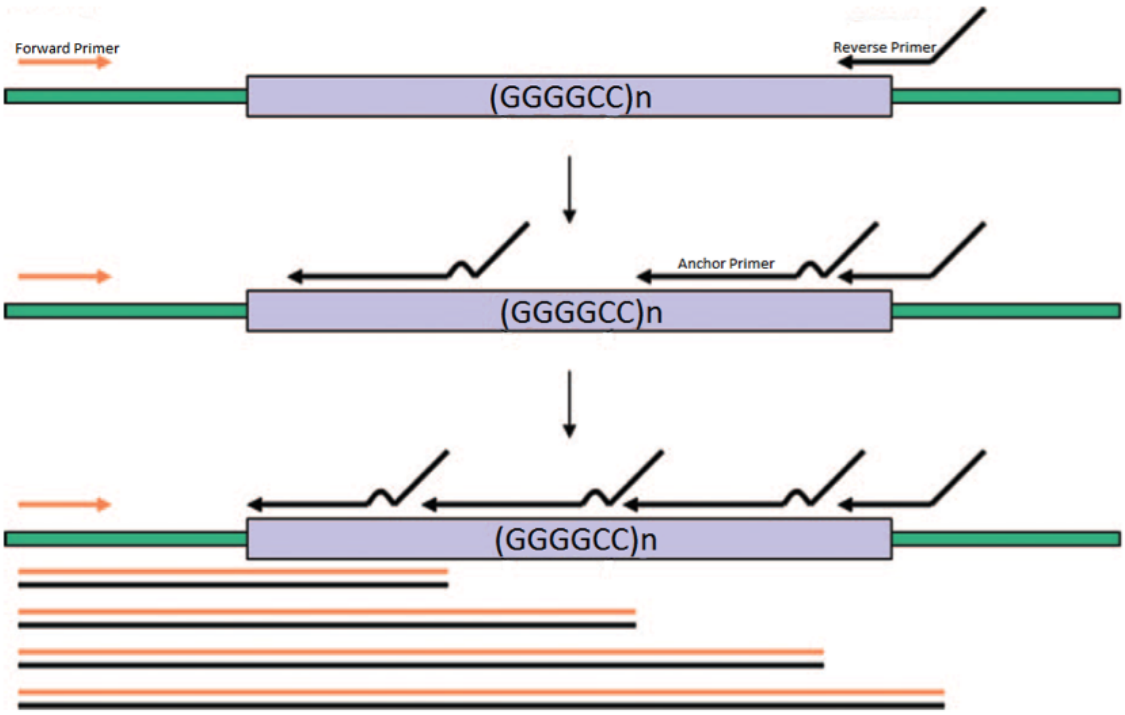

FIGURE 5.2: **Schematic representation of the repeat-primed PCR.** The forward and reverse primers are shown. The reverse primer binds the 3 end of the repeat region but may also bind randomly across the repeat. After initial rounds of PCR when the reverse primer is exhausted, the anchor primer binds in place of the reverse primer. This results in a stutter pattern of the PCR product sizes when subjected to capillary electrophoresis. Figure adapted with permission from (**Hantash et al.**, 2010).

The repeat-primed PCR method is notoriously temperamental, an attribute exhibited by the need for numerous additional reagents, including 7-deaza-2-deoxy GTP, DMSO and magnesium, as well as its high DNA content demands compared to standard PCR reactions. This nature also lends itself to varying results as demonstrated in Table 4.7, as the assay is extremely sensitive. As such, it is not uncommon to obtain an inconclusive result from this assay. The concurrent use of SNP genotyping based on linkage disequilibrium provides a useful additional classifying tool. Multiple studies have demonstrated that a common founder haplotype is inherited with the repeat expansion (**Laaksovirta et al.**, 2010; **Majounie et al.**, 2012; **Shatunov et al.**, 2010; **van Es et al.**, 2009). That is, all persons carrying the *C9ORF72* repeat expansion will also carry the founder haplotype, however individuals who have the founder haplotype will not necessarily carry the repeat expansion. Although, being an indirect measure, a definitive result cannot be obtained by this method. Thus

the rs3849942 SNP, located within the founder haplotype region, has been used as a complimentary tool for classification of individuals as positive or negative for the *C9ORF72* hexanucleotide repeat expansion. This SNP has been consistently shown to be in linkage disequilibrium with ALS caused by the *C9ORF72* expansion in multiple GWAS studies, showing significant association exceeding genome wide significance (**Laaksovirta et al.**, 2010; **Shatunov et al.**, 2010; **van Es et al.**, 2009). The combination of the two genotyping assays is imperative as they provide complimentary evidence to support or refute the presence of the expansion. That is, if an inconclusive result is repeatedly obtained from a patient's repeat primed PCR, the presence of a wildtype SNP allele can be used instead to infer the absence of an expansion. However, if the alternative SNP allele is present, it is likely, though not certain, that the patient does carry the expansion, and further diagnostic testing by Southern Blot is warranted.

Identification of pathogenic *C9ORF72* expansions has multifaceted significance. Patients identified as positive for the expansion in a research setting may be referred for diagnostic testing. This information may be utilised by at risk family members to undergo screening and/or the use of embryonic screening and IVF to prevent offspring from carrying this pathogenic mutation. After performing both assays on all newly presenting FALS and SALS cases and according to Figure 5.1, those FALS patients determined as negative for the *C9ORF72* expansion, are subsequently screened for *SOD1* mutations, while SALS patients are subjected to relatedness testing pipelines as detailed in section 3.4. Patients determined as carriers of the expansion are also assigned to the epigenetic analysis cohort to be described later.

5.2.2 *SOD1*

Standard gene primer sets were previously used for gene screening of the five *SOD1* exons. However, these primer sets were increasingly producing unsatisfactory sequence chromatograms for certain exonic regions, thus new primer sets were designed to encompass additional flanking sequences of each exon in order to improve sequence quality. The new primer sets required optimisation for PCR conditions, which was achieved in a straight forward manner, using standard PCR and thermocycler conditions with various annealing temperatures adequately amplifying exons 3, 4 and 5 while exons 1 and 2 required the simple addition of PCR enhancer reagent (Life Technologies) to the reaction mix (Table 4.5). The addition of PCR enhancer improves the thermostabilisation of the DNA polymerase, which broadens the range of effective annealing temperatures, allowing a cleaner product to be produced (**Spiess**

et al., 2004).

Screening of the Macquarie familial ALS cohort revealed the presence of four *SOD1* mutations (p.D91A, p.E101G, p.V149G and p.I114T) amongst 21 individuals (Figure 4.5 and Table 4.6). As such, validation was performed and the mutations were confirmed. All of these *SOD1* mutations have been previously implicated in ALS (**Andersen et al.**, 1996; **Deng et al.**, 1993; **Rosen**, 1993). This equates to roughly 19% of FALS cases in our cohort carrying a *SOD1* mutation, which is as expected.

As the first gene implicated in ALS aetiology by **Rosen** (1993), approximately 20% of FALS cases are accounted for by the 160+ identified *SOD1* mutations. Almost all are autosomal dominant missense mutations affecting the active site or protein structure. All five exons of the *SOD1* gene are known to harbour mutations linked to ALS. Thus there exists substantial reasoning to interrogate the entire *SOD1* gene in newly presenting FALS patients for causative mutations. However, as mutations in *SOD1* account for just 1% of SALS cases (**Chio et al.**, 2008), the likelihood of encountering such a causative mutation is not substantial enough to warrant screening of SALS patients. Additionally, gene discovery using SALS patients is not routinely conducted, as appropriate comparisons with affected and unaffected family members are not possible to elucidate causative variants.

The identification of these mutations is crucial for both clinical and research purposes. Clinically, this allows the patient to be referred for diagnostic testing for an official diagnosis of *SOD1* aetiology which may, if desired, lead to at risk family members undergoing screening and/or the use of embryonic screening and IVF to prevent offspring from carrying a pathogenic mutation. In a research context, these identifications allow patients to be assigned to appropriate research cohorts for further analyses. Samples positive for *SOD1* mutations are often utilised in downstream analyses to assess properties associated with specific mutation types and to ascertain differences, or the lack thereof, to that of wild-type samples. Most relevantly for the current study, as shown in Figure 5.1, *SOD1* positive patients are assigned to the cohort for epigenetic analysis to be described later.

5.2.3 *TARDBP* and *FUS*

As part of a collaboration, it was necessary to determine the mutation status of ten sporadic ALS patients from whom tissues were derived for cell biology studies. In

order to determine whether causative mutations could be a contributing factor to the findings of this study, the genotype of all participants was determined for the most frequently occurring ALS genes. A SALS patient carrying a p.G294A, c.880G>C mutation in *TARDBP* was identified. This result is inherently rare as a miniscule <1% of SALS cases are attributable to *TARDBP* mutations.

Mutations in *TARDBP* and *FUS* account for roughly just 1-5% of FALS each (**Robberecht and Philips**, 2013), thus being the third and fourth most common known causes of FALS. However, with six and fifteen exons respectively, screening the entirety of *TARDBP* and *FUS* in all newly presenting FALS cases is comparatively expensive and time consuming, given the small likelihood of a positive result. As such, generally only those exons most likely to harbour an ALS mutation, exon 6 of *TARDBP* and exon 15 of *FUS* (**Ling et al.**, 2013), are screened for research purposes such as that in this collaboration. Ordinarily, in a gene discovery pipeline using ALS patients presenting at a clinic, the labour, monetary and time costs weighed against the likelihood of a positive result deem it inefficient to screen for mutations in *TARDBP* and *FUS* by PCR approaches, thus samples generally proceed directly to exome sequencing without undergoing this step.

5.3 Gene Discovery and Candidate Gene Analysis in ALS

With only 10% of ALS patients having a family history and two thirds of those cases being accounted for by known gene mutations, sophisticated approaches are required to simplify the process to uncover new, rare genetic variants underlying ALS. The discovery of new ALS genes is dependent on such further studies.

5.3.1 Mq1 Pedigree

Construction of the Mq1 pedigree revealed that ALS followed an autosomal dominant mode of inheritance, though with incomplete penetrance, within this family. Further, it appeared that family members in earlier generations (i.e. pre 1900) likely suffered from ALS and/or FTD, but were misdiagnosed as ALS was not desribed until 1869 (**Goetz and Charcot**, 2000). Through construction of the pedigree, a consanguineous union was also found to be present, which may complicate how the disease is traced

back to the common ancestor. Family pedigrees are vital tools for use in all gene discovery pipelines. Pedigrees assist in determining the mode of inheritance a trait follows, and can provide clues as to the degree of penetrance a trait exhibits and which family members will provide the most informative genetic information. Unfortunately due to time constraints of this project, a full genetic analysis pipeline for this single family was not possible.

5.3.2 Bioinformatic analysis of exome sequence data

Bioinformatic scripts have been developed as part of this project to interrogate large volumes of sequence data generated by NGS exome sequencing. The wealth of data produced by NGS strategies is of little use without the availability of appropriate analysis strategies (**Katsanis and Katsanis**, 2013). In fact, the rate limiting step of NGS studies lies in the bioinformatic analysis rather than data acquisition, and clearly such analysis will be easier for exome sequencing for which less data is output compared with whole genome sequencing (WGS) (**Wang et al.**, 2013). If too much irrelevant data is generated there is a risk of data fatigue, in that researchers' ability to decipher the information presented for millions of positions will not be consistent throughout (**Teer and Mullikin**, 2010). With so many different reads obtained with wide ranging variants, downstream bioinformatic pipelines are vital components of data interpretation and require steps for alignment, variant calling, quality filtering and SNP annotation (**Wang et al.**, 2013). The criteria for the variant calling step can be made more stringent to decrease false positive variants (**Ku et al.**, 2012). Subsequently, application of filtering steps is required to reduce the possible disease causing variants, such as to remove those falling outside the coding region, being synonymous or present in controls or the reference SNP databases (**Gilissen et al.**, 2012). After all these steps there is still the possibility that false negatives have occurred and the actual disease causing variant has been discarded (**Ku et al.**, 2012). There is thus a great need for the generation of bioinformatic pipelines that may be applied to sequencing data to derive meaningful biological insights.

Previous pipelines have been established to deal with this output in terms of alignment, variant calling, quality filtering and annotation. Additional bioinformatic scripts had also been generated in order to search the resultant annotated exome variant file for genes of interest, according to user input (Appendix A.1.3).

5.3.2.1 Search_my_exome.sh

The script in Appendix A.2.1 was developed to extract all SNPs associated with ALS genes from a single patient exome data file. The resultant data file (example shown in Appendix A.3.1) also contains associated information such as chromosome position, reference and alternative bases, quality information, mutation type and SNP reference if applicable and can be used to quickly and easily assess whether a known ALS mutation is present or whether the patient carries a novel ALS gene variant. A patient who is positive for an ALS gene may be assigned to a cohort for further genetic, epigenetic or cell biology investigations, or a patient who is negative may undergo further filtering steps in an effort to identify novel ALS candidate genes.

The need to quickly and seamlessly search ALS patient exome data files for genetic variants in known ALS genes arises from a multifaceted background. As roughly two thirds of the genetic basis underlying FALS has already been elucidated, the likelihood that a patient with a family history of ALS carries a mutation in a known ALS gene is substantial. However, screening all patients for mutations in the more than 20 known ALS genes, totalling over 100 exons, would be incredibly laborious and expensive (**Ku et al.**, 2012). The different thermocycling and reagent conditions required for each exon would require extended periods of time and consume vast amounts of costly reagents. When we consider that sequencing the entire exome of a person now costs just $1000, it is far more cost- and time-effective to proceed directly to exome sequencing after the two major ALS genes, *C9ORF72* and *SOD1* have been excluded. All that is required is DNA extraction and sample preparation of one DNA sample to be subjected to an exome enrichment technique followed by NGS (**Boycott et al.**, 2013) which takes far less time and reagents, and outputs a wealth of variant information. The added advantage of using exome sequencing is that all variant information is obtained, thus a variant in a known ALS gene will be identified (except *C9ORF72*), or other exonic variants in those cases that carry a novel mutation. So effectively, exome sequencing has a dual role for identification of known mutations as well as use in novel mutation discovery.

The script in Appendix A.2.1 was written using the Unix coding language, which is widely used in the bioinformatic landscape as well as among the broader coding community. This coding language was chosen owing to its simplicity compared with its counterparts such as Perl. It was deemed necessary that a novel script was required as alternative approaches, such as using the search tool in Microsoft Excel to parse the document of roughly 80 000 lines for each ALS gene, are highly inefficient, taking

hours with the consideration for the need to copy and paste each line to a new document. Using the previously developed script (Appendix A.1.3), each gene had to be input by the user to create separate result files for each, which then needed to be combined. This approach required roughly half an hour to execute, and whilst a significant improvement compared to the mundane excel approach, still required some improvement for efficiency. Using the novel script, less than a minute is required to compute the search of the annotated exome data file for each ALS gene and output the corresponding lines of information to a new result file, which conveniently contains all relevant variation data to determine whether an ALS-linked mutation is present in that patient. Additionally, no commercial software is available for such a task, and any such software developed would inevitably be outdated very quickly given the rate of gene discovery in ALS. This is in contrast to this novel script that may be customised whenever desired, to include all currently known ALS genes.

5.3.2.2 Relatedness Testing

Bioinformatic analysis to identify related individuals was attempted using the script shown in Appendix A.2.2 utilising the PLINK/SEQ library for working with human genetic variation data (`http://atgu.mgh.harvard.edu/plinkseq/`). The output file (Appendix A.3.2)reports the number of shared genotypes between two individuals as well as the number of non-missing genotypes, that is the total number of genetic sites at which a genotype was found for both individuals. Using these outputs, a ratio can be easily computed using Microsoft Excel. A general trend implicating a higher ratio with a lower number of meioses between the pair was observed, indicating a closer level of relatedness. However, no direct formula could be determined to link this ratio with the number of meioses separating two individuals. As a result, collaborators at the CSIRO with advanced bioinformatic expertise were tasked with generating relatedness testing pipelines.

Our collaborators were able to perform a cluster analysis on the exome-sequenced cohort and identified two individuals whom are distantly related (Figure 4.8). After extensive bioinformatic analysis, *KDM2A* p.S104G, c.310A>G appeared to be a likely disease-causing variant. Howvever, was found to be absent in the two patients after PCR and Sanger sequencing (Figure 4.9), and as such represents a false positive identification by exome sequencing. Upon consultation of IGV, it was discovered that this variant appeared to occur in multiple individuals (Figure 4.10A). This variant was thus discarded as a potential disease-causing variant and deemed a false positive

result. It is likely that the variant was not called in many other individuals as it generally occurred at the end of low quality sequencing reads (Figure 4.10B). This is interesting considering that this residue is highly conserved amongst various species indicating the integral role, and in that the protein change was predicted to have a significant impact on the protein structure and function with the potential damaging properties (Table 4.8). Though, owing to the fact that the change occurs within an unintegrated protein domain suggests that the result of the mutation may not have a large influence on protein structure or function.

As outlined above, a large proportion of the genetic basis of ALS has already been elucidated, with most genes of large effect having been identified. Thus the discovery of novel ALS genes is becoming increasingly difficult. The majority of known ALS genes were identified through investigations of families carrying dominant mutations with relatively large effects. As the remaining variants have failed to be identified by classical genetic approaches such as linkage analysis, it is likely that they are relatively rare and of smaller effect. Rare variants are inherently difficult to identify owing to their uncommon occurrence, and are most easily identified in large families with many affected individuals. Owing to the rarity of these variants, it is likely that many people who carry them are indeed related and have inherited the variant from a common ancestor. Likewise, owing to the relative rarity of ALS, it is likely that many ALS patients are distantly related. As such, testing for relatedness is an important step in genetic investigations of ALS. Distant relatedness can serve as an important tool in candidate gene identification, in that variants shared between two related affected patients yet absent in unaffected relatives are incredibly strong candidate disease genes. As such, establishing relatedness between cohort members is an imperative tool in ALS gene discovery that has been underutilised to date.

5.3.3 Identifying candidate genes based on functional knowledge

In the era of NGS and exome sequencing, interrogation of the human genome is now possible on an unprecedented level. As such, investigating candidate genes identified based on functional knowledge can be achieved quickly and simply with the use of bioinformatics. Once exome sequencing data has undergone standard pipelines for alignment, variant calling, quality filtering and SNP annotation, bioinformatic tools can be utilised. Utilising the script in Appendix A.1.3, the entire cohort of exome

datasets was inspected for a list of candidate genes (Table 3.2), which collaborators suggested might play a causative role in ALS pathogenesis based on knowledge of protein function. Further functional and structural protein knowledge was applied by our collaborators to determine which candidate gene variants amongst the ALS patient cohort were most likely to contribute to ALS pathogenesis. The variants identified as such were *SFPQ* p.N533H, c.1597A>C and *RBM14* p.S620X, c.1859C>A.

SFPQ encodes the Splicing Factor Proline/Glutamine-Rich (SFPQ) protein, which is a DNA and RNA binding protein with extensive roles mRNA splicing (**Gozani et al.**, 1994). It has been postulated that the SFPQ protein forms a heterotetramer with the NONO protein, each constituting two subunits (**Sewer et al.**, 2002). Residue 533 falls in a critical long coiled/coil oligomerisation domain which acts to target the protein to paraspeckles (**Bond and Fox**, 2009), thus it is likely that this mutation may adversely affect the localisation of the protein. The structure of this domain is likely to be affected by the amino acid change from Asparagine to Histidine. This change represents a shift from a neutral polar residue to a positively charged basic residue. While both have hydrophilic properties and a propensity to form hydrogen bonds, the Histidine side chain is significantly bulkier than that of Asparagine, which could cause steric hindrance. This is particularly important as Asparagine residues commonly have roles in secondary structure formation in that they commonly occur at the beginning of alpha helices and in beta sheet turns (**Fujiwara et al.**, 2012). As such, steric hindrance in these regions could severely impact on the tertiary structure of the monomer protein unit which could also lead to difficulties in the formation of quaternary structure in the heterotetramer complex with NONO subunits. Combined with these potential structural and functional impacts, the validation of this mutant by Sanger sequencing and its absence from control samples (Section 4.4.4.1), a strong case for a link between this *SFPQ* mutant and the pathogenesis of ALS is apparent.

Encoding a hnRNP, the *RBM14* gene appeared to be an interesting ALS candidate. However, as the c.1859C>A mutation identified by exome sequencing was not replicated by PCR and Sanger sequencing (4.4.4.1), this mutant represented a false positive mutation. This was not unexpected as a low read depth was initially obtained for this variant during exome sequencing. Read depth is of paramount importance to exome sequencing, in that it must be as high as possible so as to minimise opportunities for relevant mutations to be overlooked (**Teer and Mullikin**, 2010). A high read depth is required for heterozygote identification and should also compensate for the raw base-calling errors encountered in NGS technologies which are

higher than that of Sanger sequencing, which is subsequently used to confirm variants identified by exome sequencing (**Ku et al.**, 2012). This high coverage is also vital to avoid biases inflicted from inherent uneven sequencing of different fragments as well as the bias in exome capture imposed by the capture method of choice (**Ku et al.**, 2012). This result demonstrates the importance of both obtaining adequate read depth during NGS sequencing to obtain valid results as well as that of PCR and Sanger sequencing validation of exome sequencing results to avoid false positive identifications.

5.3.4 *CCNF*

Using a combination of linkage analysis, exome sequencing, bioinformatic analysis and Sanger sequencing validation, mutations in the *CCNF* gene were implicated in ALS by our laboratory (unpublished, manuscript submitted). *CCNF* comprises 17 exons and encodes the cyclin F protein spanning 786 amino acids. Cyclin F is a component of an E3 ubiquitin-protein ligase complex. E3s mediate the ubiquitination and proteasomal degradation of target proteins and are an integral component of the ubiquitin proteasome system (UPS) (**D'Angiolella et al.**, 2010), which has been implicated extensively in ALS (**Robberecht and Philips**, 2013).

Firstly, in an attempt to identify new *CCNF* variants and/or multiple occurrences of *CCNF* variants, screening of *CCNF* exon 1 in proband ALS cases was performed. This was required as previous TruSeq exome capture did not cover this exon, thus manual interrogation was required. Owing to the small size, and GC density of this exon, a nested PCR approach was required using a high fidelity *Taq* polymerase. Subsequently, no mutations were identified amongst the 27 probands.

In order to support the previously identified mutations as causative of ALS, their absence in control samples required confirmation. Screening of non related control samples was thus conducted for the *CCNF* mutants p.S3G and p.H69Y.

Being absent from 649 control samples, the p.S3G mutation withstood control screening and appears to be a valid ALS variant. Interestingly, the p.S3G variant falls outside the functional domains of the cyclin F protein, however is in relatively close proximity to the catalytic F-box domain which is essential for the protein to bind interacting partners and thus for its function in the ubiquitin proteasome system. The change from Serine, a polar residue to the non-polar Glycine has the potential to inflict substantial structural modifications to the tertiary structure of the cyclin F protein.

Being polar, Serine is able to sit on the surface of proteins and interact with water. Conversely, as a non-polar residue, Glycine is hyprophobic, thus generally lies within protein structures to avoid contacts with water. Additionally, owing to its small size, Glycine offers minimal steric hindrance and often sits in secondary structure regions such as alpha helices and beta turns which require this to achieve such formations. As such, the presence of a Glycine residue opposed to a Serine residue would reduce the steric hindrance in the tertiary structure and may affect protein folding which could adversely affect the ability of the protein to bind substrates. Occurring so close to the catalytic domain of the protein, this is particularly likely to affect the function of the protein. Taken together, the *CCNF* p.S3G mutant has significant potential to be a disease causing variant.

However, though absent from 587 control samples, the p.H69Y mutation was detected by a TaqMan SNP genotyping assay in four control samples. Sanger sequencing validation confirmed that three of these occurrences were in fact false positive results, however the fourth appeared to be truly positive, though the quality of sequencing was inadequate and thus needed to be repeated (Figure 4.12A). Unfortunately, owing to its GC density exon 3 of *CCNF* proves difficult to amplify, with no clear band being obtained for the second attempt at PCR. So a nested PCR approach was adopted using the product generated as a template. As this template was already amplified, a bias was introduced for the mutant allele so that upon sequencing it appeared that a homozygous mutant was present (Figure 4.12B) opposed to the heterozygous mutant indicated by both TaqMan SNP genotyping and the first attempt at validation. A new primer set for genomic DNA was then trialled, however an insufficient number of thermocycling cycles were employed, thus no product was visible on the gel. At this point, the DNA sample was completely exhausted despite attempts at resuspension. In a final attempt, a pGEM-T cloning approach was adopted using the latter PCR product as a template, in the hope that a product would be sufficiently amplified. Unfortunately, transformation of the resultant ligation product failed to produce any clones containing the DNA insert. Due to time constraints, whole genome amplification was not possible, and with all other avenues of investigation being exhausted, no further action was possible to obtain a definitive result. Falling within the catalytic F-box domain of the cyclin F protein, the p.H69Y mutation has the potential to severely impact on the function of the protein. Though both the original Histidine and mutant Tyrosine residues are hydrophilic, the positive charge removed by the substitution for Tyrosine has the potential to adversely affect ionic interactions in addition to the increased steric hindrance imposed by the benzene ring of Tyrosine,

which is likely to affect protein folding. Thus the catalytic site of cyclin F is likely to undergo severe conformational modifications as a result of the p.H69Y mutation, which suggests the potential for a functional consequence. However, there is in fact a known SNP at this location, rs369245530 (p.H69D, c.205C>G), which strongly suggests that the p.H69Y, c.205C>T mutation is a benign SNP rather than a disease causing variant.

5.3.5 *VCP*

A patient with movement impairment presented to a collaborator with features of ALS, FTD an IBM that are often associated with mutations in the *VCP* gene. As such, it was requested that the patient be screened for mutations in *VCP*. After optimisation of PCR conditions for all 17 *VCP* exons and subsequent screening in the patient by another member of our laboratory, sequence analysis revealed that the patient carried a p.R159C, c.475C>T mutation in exon 5. Additionally, a rare stretch of Adenine bases was observed within exon 14 at c.1839-1847. These variants both underwent validation.

The Valosin Containing protein (VCP) is a member of the AAA+ protein family, meaning it is a multifunctional ATPase, of which six subunits interact to form a functional homohexamer (**Meerang et al.**, 2011). *VCP* was first implicated in IBM with early-onset Paget disease of the bone (PDB) and FTD (**Watts et al.**, 2004). Exome sequencing of two FALS patients revealed the same mutation identified by **Watts et al.** (2004) was present in these two patients, and further screening of the *VCP* gene in FALS patients identified a further three mutations in the gene (**Johnson et al.**, 2010). The link between *VCP* gene mutations and ALS is not surprising considering that *VCP* has functions in the maturation of ubiquitin-containing autophagosomes, which are utilised for protein degradation, a pathway commonly associated with ALS genes. Further, the toxic effects of mutant VCP act in part on the TDP-43 protein, which constitutes a significant proportion of the characteristic ubiquitin inclusions of ALS (**Johnson et al.**, 2010).

The p.R159C, c.475C>T mutation in exon 5 has been previously identified in a patient suffering from IBM and FTD (**Bersano et al.**, 2009). Further, this residue has also been implicated in ALS related conditions, though as different mutations. The first identification was of a p.R159H, c.476G>A mutation (rs121909335) in patients affected by IBM with early-onset PDB and FTD (**Haubenberger et al.**, 2005; **van der Zee et al.**, 2009). Subsequently, a p.R159G, c.475C>G (rs387906789) transversion mutation was found in ALS patients with or without FTD (**Johnson**

et al., 2010).

As being previously identified as pathogenic, this residue is clearly important for VCP protein function and this mutation potentially warrants further investigation as a disease causative variant. The functional importance likely stems from the fact that the mutation occurs at a highly conserved residue that lies within the CDC48 domain of the protein which is crucial for binding ubiquitin, thus the mutation identified here, and others at this residue, are likely to interfere with protein interactions between those involved in the ubiquitin proteasome and ER-associated degradation (**Bersano et al.**, 2009). Further, the change from an Arginine residue to a Cysteine residue induces the loss of a positive charge that may be involved in ionic bonds, which would be lost upon such a modification. Further, the introduction of a Cysteine residue has the potential to introduce a disulphite bridge into the tertiary structure of the protein, which could have drastic effects on the protein folding which could translate into adverse affects of quaternary structure of the fully functional homohexamer protein complex.

In addition to the point mutation identified in *VCP* exon 5 of the ALS patient under examination, a polyA region within exon 14 was identified and appeared to present an interesting insertion/deletion (indel) mutation. This was indicated by a sequence trace that appeared to be of high quality with little background noise, except that following the polyA region, double peaks were observed at all base positions, which appeared to represent sequence consistent with an indel mutation (Figure 4.13).

In order to determine whether this indel mutation was in fact present, a cloning approach was adopted, so that the length of the polyA stretch in each of the two alleles could be determined. However, upon cloning there appeared to be three alleles present, as stretches of eight, nine and ten Adenine bases were observed. This is of course puzzling as humans, being diploid organisms, carry two sets of chromosomes and hence only two alleles for each gene.

Long stretches of Adenine bases are rare in exonic regions, as they may introduce transcriptional slippage by RNA polymerase. This phenomenon has been reported for Adenine stretches of 11 or more (**Linton et al.**, 1997; **Wagner et al.**, 1990). As a result, long stretches of Adenine bases have been selected against in the genome and are quite a rare occurrence (**Linton et al.**, 1997; **Wagner et al.**, 1990).

Slippage also occurs during DNA replication by DNA polymerases. When a repetitive stretch of DNA is encountered during replication events, either the template DNA strand or the newly synthesised DNA strand may form a hairpin loop, resulting in a deletion or an insertion event respectively (**Streisinger et al.**, 1966). It is highly likely that by one process or another, DNA slippage underlies the heterogeneous stretch of Adeneine bases observed in *VCP* exon 14 in the patient sample under analysis.

As such, the presence of three different alleles could indicate one of two possibilities. Firstly, it may simply indicate the occurrence of a PCR error in which an extra Adenine base was added to or removed from the polyA stretch. This is quite plausible as repetitive stretches of DNA are difficult to faithfully replicate, and in conjunction with the use of a very basic *Taq* Polymerase PCR mastermix, such an event is not unlikely. Alternatively, the presence of the three alleles may represent a *de novo* somatic mutation in the patient, so that only relatively few cells carry the mutant allele compared to the parent alleles, which may account for the discrepancy in allele frequencies. The reference allele (NM_007126) contains nine Adenine bases, thus the presence of alternate eight and ten stretches of Adenine indicates both an insertion and a deletion mutation. As only two clones presented with the larger polyA region of ten Adenine bases while six clones contained the smaller eight Adenine bases in the polyA stretch, it is more likely that the eight Adenine allele represents a truly present variant in the patient. That is, the eight Adenine allele may have plausibly been produced by a *de novo* somatic mutation or is simply a naturally occurring variant, while it is highly probable that the ten Adenine allele is a PCR artefact, though it is still possible that it may represent a recent *de novo* mutation within the patient, causing its low frequency.

In order to further investigate the polyA region of *VCP* exon 14, the first step would likely be to either PCR the whole exon using a high fidelity *Taq* polymerase and/or to design primers specifically targeting the polyA region to more accurately represent this particular region. Subsequent cloning approaches would then yield more reliable allele representations. If this resulted in just two alleles being represented, it is most likely that a simple PCR error was at play. However, if three alleles remain to be observed, the likelihood of a *de novo* mutation is strengthened and further haplotyping would be appropriate. It is however anticipated that a PCR error is likely to underlie the presence of the three alleles owing to the presence of the p.R159C mutation in exon 5, especially taking into consideration the previously established

pathogenic involvement of this residue in ALS and related conditions. Though this indel mutation has the potential to cause a frameshift which could potentially change protein sequence downstream of this polyA region which may affect protein structure and function, particularly of the C-terminal of the VCP prtotein.

5.3.6 *OPTN*

Previous bioinformatic investigations using the script in Appendix A.1.3 to interrogate all available ALS patient exome sequencing data for variants in the *OPTN* gene revealed that a patient carried a p.V295F, c.883G>T variant. Validation by PCR and sequencing confirmed that this mutant was in fact present (Figure 4.14). The mutant was also found to be absent from 263 non related controls by TaqMan SNP genotyping and 1000 control exomes by bioinformatic analysis, strengthening the case for its involvement in ALS pathogenesis.

The *OPTN* gene was first implicated in ALS by (**Maruyama et al.**, 2010). Interestingly, three different *OPTN* mutation types were found to be associated with ALS. Two different null mutations of the gene being a homozygous deletion of exon 5 and a p.Q398X nonsense mutation, in addition to an p.E478G autosomal dominant missense mutation with incomplete penetrance (**Maruyama et al.**, 2010). The optineurin protein encoded for by the *OPTN* gene plays important roles in maintenance of the Golgi complex, membrane trafficking and exocytosis (**Sahlender et al.**, 2005) and has been identified in ALS protein inclusions (**Deng et al.**, 2011a).

The novel p.V295F, c.883G>T mutation occurs at a highly conserved residue within the proteins second coiled-coil domain, which is composed of alpha helices. The change from a Valine residue to that of Phenylalanine is significant in that a bulky benzene ring is introduced, which may introduce steric hindrance into the alpha helix, potentially disrupting this secondary structure which may induce conformational changes to the coiled coil domain and subsequently the tertiary protein structure. Thus it is highly likely that this is a disease causing variant.

5.4 Epigenetic investigation into ALS

Although gene mutations now account for two-thirds of familial ALS cases, phenotypic discordance amongst patients with identical gene mutations is highly prevalent. In particular, cases of MZ twins carrying identical mutations in either *C9ORF72* or *SOD1* with only one twin presenting with disease symptoms have been identified (**Al-Chalabi et al.**, 2010; **Dols-Icardo et al.**, 2014). As with other neurodegenerative diseases, such discordance has led to interest in evaluating epigenetic modifications in ALS.

We have in our possession large disease discordant cohorts, including twins and triplets. These individuals carry the same ALS mutation however present with various degrees of disease onset and progression. The disease variability observed between individuals with known mutations implicates factors other than genetic predisposition affecting the ALS phenotype, i.e. there is no genotype-phenotype correlation. Therefore, we are investigating additional modifying factors including epigenetic mechanisms. Indeed, there already exists a substantial foundation for an epigenetic role in ALS, as demonstrated by the following examples.

The first epigenetic analyses of ALS focused on locus-specific DNA methylation in sporadic patients (**Morahan et al.**, 2007). Methylation status of the gene promoters for *SOD1* and *VEGF* (Vascular endothelial growth factor) were analysed by bisulphite sequencing. Minimal promoter methylation with no discernable pattern for *VEGF* and *SOD1* in both SALS patients and controls was observed. Similarly, promoter methylation was analysed in metallothionein protein isoforms MT-Ia and MT-IIa by methylation sensitive restriction enzyme analysis and no DNA methylation was present in SALS patients or controls (**Morahan et al.**, 2007).

In ALS patients, the astroglial transporter *EAAT2/GLT1* is dramatically lost, contributing to astroglial dysfunction. The methylation status of the gene promoter was interrogated in a *EAAT2/GLT1* murine model by bisulphite sequencing. It was discovered that the *GLT1* promoter was selectively hypermethylated to repress activation, implicating a role of DNA methylation in ALS pathogenesis (**Yang et al.**, 2010b). Additionally, exposure to metals such as lead, mercury and selenium have been associated with ALS, and are suggested to induce epigenetic changes which may lead to symptom onset (**Callaghan et al.**, 2011).

In 2009, genome-wide methylation analysis was conducted on brain samples from ten SALS patients and ten neurologically normal individuals (**Morahan et al.**, 2009).

A total of 38 sites were identified as having differential methylation between SALS patients and controls, with 23 of these associated with genes, and pathway analysis revealing enrichment for those involved in calcium homeostasis, neurotransmission and oxidative stress (**Morahan et al.**, 2009). The Y chromosome showed significantly decreased methylation in SALS patients relative to controls, which is of interest considering the increased frequency of ALS amongst males when compared to females (**Morahan et al.**, 2009). Global 5mC content is increased in SALS spinal cord (**Figueroa-Romero et al.**, 2012), opposed to decreased global 5mC seen in AD and PD patients. This has recently been replicated in whole blood samples, which can have strong biomarker implications, as blood sample analysis is readily amenable to diagnostic and prognostic applications if the appropriate correlations are confirmed. Two DNA methyltransferase proteins, Dnmt1 and Dnmt3a, are upregulated in the motor cortex and spinal cord motor neurons of SALS patients, suggesting aberrant DNA methylation may be present in ALS (**Chestnut et al.**, 2011). Expression of these proteins in a cell model rose by five and two fold respectively while 5mC accumulated in the nuclei, eventually leading to neuronal apoptosis. This was mirrored in human ALS. This strongly suggests a strong role of DNA methylation in ALS, as these proteins are responsible for the creation and maintenance of DNA methylation modifications.

5.4.1 Epigenetics and the twin study design

Through zygosity testing, we identified that within the discordant triplet set, the affected individual and one unaffected individual are monozygotic (MZ), while the second unaffected individual is dizygotic (DZ). The discordant twin set, consisting of one affected and one unaffected individual, was also determined to be monozygotic.

Twin studies have long been used to determine the disease heritability. Comparisons between the phenotypic concordance rates between MZ and DZ twin sets allows estimation of the genetic contribution to a given trait. As MZ twins arise from a single zygote, they share an identical genetic makeup, whereas DZ twins share on average half their genetic identity just as any other sibship. Both types of twins are the same age and generally share an environment, therefore removing these as confounding factors in disease analysis.

A major result from twin heritability studies is the suggestion that factors other than genetics, including epigenetic modifications, play a role in disease discordance

(**Hanson et al.**, 1991). While having identical genetic makeups, MZ twins do not always exhibit identical phenotypes, suggesting phenotype is not only the result of genotype, but a combination of genotype with other factors such as the environment and epigenetics, with the latter likely to be influenced by the former (**Fraga et al.**, 2005; **Kaminsky et al.**, 2009; **Zhang and Pradhan**, 2014). As MZ twins do not possess the abundant confounding DNA sequence differences found in singleton studies, they represent the best opportunity to enhance our understanding of pathogenesis and phenotypic variability (**Bell and Spector**, 2011; **Ketelaar et al.**, 2012; **Petronis**, 2006). Additionally, the discordant MZ twin study design controls for genetic differences, sex, cohort effects, maternal effects and generally share a common environment (**Bell and Spector**, 2011). As such, direct comparison of discordant MZ twin sets represent an ideal study design for investigating the non-genetic factors influencing phenotypic expression (**Bell and Spector**, 2011; **Chatterjee and Morison**, 2011; **Ketelaar et al.**, 2012; **Petronis et al.**, 2003).

In order to draw accurate and relevant conclusions from such investigations, the determination of zygosity is imperative. Zygosity testing is often performed using SNP arrays as previously described, however the protocol requires specialised equipment which is prohibitively expensive, as well as extensive analysis techniques. There does of course exist the option to outsource the zygosity testing to a service provider, however this too is expensive. The alternative approach adopted here using both microsatellite and SNP genotyping was most appropriate for our purposes. In order to produce the most informative genotyping results, genetic markers should possess as great a degree of variability as possible, that is high heterozygosity values. Microsatellites are short tandem repeats of 2-5 base pairs which are highly variable in number and are thus very polymorphic. Similarly, SNPs that occur commonly throughout the population can serve as informative genetic markers. As such, both are well suited for use in zygosity testing. The most informative microsatellite markers were discerned based on heterozygosity values indicating the degree of variation observed at those loci as well as those with the highest number of informative meioses. SNPs were also chosen based on having the highest heterozygosity values, which correlate to MAF values close to 0.5, as well as being spread across different chromosomes to gain a representation of genotype across the genome.

As variable number tandem repeats, microsatellites may take the form of dozens of alleles, while SNPs generally only take two allele forms. As such, microsatellite markers are more informative than are SNPs, however microsatellite size analysis is

more challenging than SNP analysis owing to fluorescence attributes and fragment size discernment opposed to simple sequence analysis. Primer sets were readily available for a variety of microsatellite and SNP markers covering a range of chromosomes. Using these primer sets, simple PCR reactions were required for DNA amplification followed by sequence trace analysis. Both techniques are performed extensively and thus represent the opportunity for the most reliable results to be obtained.

By interrogating polymorphic markers across seven and eleven chromosomes for the triplet and twin sets respectively, confidence in zygosity results is high. With monzygotic, and therefore genetically identical individuals, one of which is affected by ALS, and a third dizygotic individual possessing an estimated 50% genetic identity with the other two individuals, a number of unique opportunities for epigenetic analyses are represented by the triplet set. Methylation profile differences between the two identical individuals may represent a gene expression regulatory event, which dictates the presence or absence of the disease state. This is also true of the ALS discordant monozygotic twin set. Additionally, by comparing the methylation pattern of the non-identical and non-affected triplet to that of the other individuals, the impact of genotype identity on methylation patterns may be evaluated.

DNA methylation profiles have been examined in MZ and DZ twin sets, and as would be expected, DZ twins exhibited far more divergent DNA methylation patterns than did MZ twins (**Chatterjee and Morison**, 2011; **Fraga et al.**, 2005; **Kaminsky et al.**, 2009). MZ twin pairs were found to show very similar methylation patterns at differentially methylated regions, with variation of just 3–4% depending on tissue type (**Ollikainen et al.**, 2010). It has also been revealed that tissue-specific variation exists between MZ twins, with close to 6000 unique metastable genomic regions identified between adult twins (**Kaminsky et al.**, 2009) and extensive differences observed also in newborns (**Ollikainen et al.**, 2010). However, evidence suggests that over time the DNA methylation patterns of MZ twins diverge, that is age-dependent epigenetic drift occurs, suggesting accumulation from influences both internal and external (**Fraga et al.**, 2005; **Kaminsky et al.**, 2009; **Ollikainen et al.**, 2010). MZ twins show variable degrees of discordance which is clearly a result of age, tissue type, site selection and sample size (**Czyz et al.**, 2012). It must also be noted that the external influences leading to epigenetic change have so far not been elucidated (**Steves et al.**, 2012).

MZ twins with a common genotype do in fact show varying penetrance for complex

disease such as neurological disorders (**Fraga et al.**, 2005). In MZ twins discordant for schizophrenia, the affected MZ twin was epigenetically more similar to affected concordant twin sets than to their own co-twin (**Petronis et al.**, 2003). Furthermore, twin studies have implicated an epigenetic role in other neurological conditions such as bipolar disorder (**Dempster et al.**, 2011), Rett syndrome, autism, and Alzheimers, Parkinsons and Huntingtons disease (**Urdinguio et al.**, 2009).

5.4.2 Locus specific DNA methylation analysis in ALS

Using the UCSC genome browser, CpG islands associated with both major ALS genes, *C9ORF72* and *SOD1* were identified. Accounting for the majority of FALS cases, *C9ORF72* and *SOD1* are obvious candidates to be investigated for DNA methylation differences. As such, future studies will elucidate the locus specific methylation status of the CpG islands associated with *C9ORF72* and *SOD1* in rare discordant MZ twins and other FALS and SALS cases carrying mutations in these respective genes compared to controls. Additionally, ALS patients identified as carrying either of these mutations in the previously described screening sections for both genes will form the respective cohorts for future methylation analyses.

The identified CpG islands are all located within GC rich gene promoters. This is not surprising as such regions harbour the majority of CpG sites within the genome and are further the most functionally relevant genetic regions as regulators of gene expression. The presence of DNA methylation within the promoter region is likely to turn off the promoter and hence downregulate gene expression whereas its absence will turn on the promoter, allowing the gene to be transcribed and hence its expression to be upregulated.

It has been hypothesised that repeats containing the CpG dinucleotide may be targets for aberrant DNA methylation, and that silencing machinery may play a role in the formation of all triplet repeat expansions, independent of sequence (**Robertson**, 2005). Considering the most common known cause of ALS are GGGGCC hexanucleotide repeat expansion in the *C9ORF72* gene, and each repeat introduces a new CpG site, it is logical to consider that aberrant DNA methylation likely occurs in patients with a *C9ORF72* repeat expansion. Such a role is strengthened by reports that another repeat expansion mutation associated with ALS, that in *ATXN2*, has been found with promoter hypermethylation in spinocerebellar ataxia type 2 (**Laffita-Mesa et al.**, 2012). Further, DNA methylation at repeat sequences

and their adjacent CpG islands have been shown to influence histone methylation which leads to severe gene silencing and loss of function (**Belzil et al.**, 2013).

Initial studies have revealed that in brain tissue the *C9ORF72* repeat expansion is repressed in affected ALS expansion carrier brains by the binding of mutant *C9ORF72* to trimethylated lysine residues within histones H3 and H4 (**Belzil et al.**, 2013) and also possibly by DNA hypermethylation of the upstream CpG island in approximately 40% of cases as determined using DNA extracted from whole blood . This seemingly low value of 40% may be misleading, as hypermethylation may still exist further upstream, or even within the repeat, at other CpG islands (**Belzil et al.**, 2013). However, there are discrepancies in that another study showed no hypermethylation at the *C9ORF72* locus for ALS patients using DNA extracted from brain tissue, though only four were included in the cohort (**Belzil et al.**, 2014). The work presented by **Xi et al.** (2013) is incredibly exciting in that hypermethylation of a 5' CpG island of the *C9ORF72* repeat was observed in ALS patients, and has also been shown to a lesser extent in FTD patients (**Xi et al.**, 2014). If this result can be replicated by DNA methylation studies such as that proposed here, the therapeutic and diagnostic potential for ALS would be huge.

As such, there is significant cause for investigation of the DNA methylation pattern associated with *C9ORF72*. As tempting as it is to analyse the methylation status of each CpG site in the expansion region, the GC density and repetitive nature of the region deem it unamenable to bisulphite conversion treatment followed by PCR, which lies at the core of all locus specific DNA methylation analysis techniques. However, we have identified *C9ORF72* associated CpG islands overlapping the promoter regions for the various isoforms of this gene, which are also incredibly relevant, with the potential of DNA methylation here having a role in regulating gene expression as previously described.

We have also identified a CpG rich region associated with the *SOD1* gene which extends from the promoter out to the first exon and intron, which has been preliminarily investigated in SALS patients with no known ALS mutation (**Oates and Pamphlett**, 2007). *SOD1* disease progression can vary significantly depending on the particular variant present, such that an p.A4V mutation causes rapid, aggressive degeneration leading to death within a year of symptom onset (**Cudkowicz et al.**, 1997) while a very slow and gradual disease course is experienced by carriers of the p.D90A variant. As such, investigating DNA methylation differences in this region

between patients with various *SOD1* mutations, and even with the same *SOD1* mutation will potentially yield interesting results of how *SOD1* mutants are expressed or repressed and the effect on phenotype. Further, the protein encoded by *SOD1* is involved in detoxifying oxygen free radicals, which have the potential to cause epigenetic damage (**Oates and Pamphlett**, 2007), suggesting a link between *SOD1* and DNA methylation in ALS is highly probable.

It is anticipated that the results uncovered here will demonstrate an epigenetic regulation of gene expression contributing to presentation of ALS. If this is shown and replicated, the potential use of epigenetic modifying drug treatments for ALS will be an attractive investment of time and resources for future studies to uncover the usefulness of such drugs for treating ALS. Further, the epigenetic marks may be used as biomarkers of ALS. This would be most useful for *C9ORF72* linked ALS cases, as described extensively, the expansion is notoriously difficult to amplify. This is particularly relevant, as methylation analysis will be conducted on DNA extracted from whole blood. Despite concerns raised about the conservation of epigenetic marks across different tissue types, high levels of epigenetic concordance have been observed between whole blood and brain tissue. Methylation ofthe Illumina 450K array probes has been shown to be relatively robust across tissues, with just 13% of CpG island sites showing differential expression across tissues (**Belzil et al.**, 2013). Considerable concordance has been observed between the methylation patterns of whole blood and neuronal tissue, including within ALS patients, suggesting that whole blood can be a valuable surrogate (**Davies et al.**, 2012; **Masliah et al.**, 2013; **Xi et al.**, 2013; **Yang et al.**, 2010a). This strengthens the potential use of distinctive epigenetics patterns as biomarkers, as blood samples are easily accessible and ideal for use in diagnostic or prognostic assays.

5.5 Conclusion

Despite the vast genetic basis already evident for ALS, the need still exists for further elucidation of the underlying variants in order to gain a detailed understanding of disease pathogenesis. The ALS genetic analysis pipeline presented here represents an efficient methodology for elucidating molecular modifications underlying the onset of ALS. The success of this approach has also been demonstrated at different stages of analysis, and the potential for this pipeline to provide insights into the underlying causes of ALS has been supported by the various mutations and insights discovered

throughout the course of this project.

A

Appendix

A.1 Previously developed bioinformatic scripts

A.1.1 dbSNP137/141 script example

The following shell script was written by Kelly Williams. This is an example of the command used to determine if a SNP on chromosome 11 at position 66, 436, 117 is present in the dbSNP137/141database.

```
awk ' { if($12=="66436117") print($7" "$12" rs"$1" build"$20"
"$23) } ' chr_11.txt
```

If the SNP is present, the output will describe the SNP by chromosome number, chromosomal location, SNP name, build and as follows:

```
11 66436117 rs145628127 build137 other
```

A.1.2 Exome Variant Server (EVS) script example

The following shell script was written by Kelly Williams. This is an example of the command used to determine if a SNP on chromosome 11 at position 66, 436, 117 is present in the Exome Variant Server (EVS) database.

```
awk ' { if($1=="11:66436117") print($1" "$2" "$3" "$4" "$5"
"$6" "$7" "$8" "$9" "$10" "$11" "$12" "$13" "$14" "$15"
"$16" "$17" "$22" "$25" "$28) } '
ESP6500SI-V2-SSA137.chr11.snps_indels.txt
```

If the SNP is present, the output will describe the SNP in detail.

A.1.3 Find_my_gene.sh

The following shell script was written by Kelly Williams, and is used to screen candidate genes through an entire cohort of exome sequenced samples. The resulting output text file can be easily opened in Microsoft excel for further analysis.

```
#!/bin/sh
#This is to look in .txt files for variants in genes
#by Kelly Williams 16/09/2011, edited 16/05/2013

echo Please input gene name you wish to look for
read GENE
OUT="$GENE"_"$(date '+%Y%m%d').txt"
echo "The output file is $OUT"
for file in *.txt
do
  awk -v NAME="$GENE" '$12 ~ NAME { print $0 }' $file >> $OUT
done
```

A.2 Novel Bioinformatic scripts

A.2.1 Search_my_exome.sh

The following example script is used to search annotated ANNOVAR files received as the result of exome sequencing analysis for single nucleotide polymorphisms in known ALS genes. The output text file can be easily opened in Microsoft excel for further analysis.

```
# Search_my_exome.sh
# by Emily McCann 29/01/2014
# This script is used to firstly convert an annotated ANNOVAR .xlsx
file generated from exome sequencing analysis into an uncorrupted
.txt file and then to subsequently search that .txt file for known
ALS genes in a single patient
# This script can be used on files generated from 2012 onwrads
# For this script to work, the user is required to have an annotated
ANNOVAR excel workwork. This workbook must then be saved as a .txt
(tab delimited text file). With the file open in excel, go to
File > Save as > Format > Tab Delimited text (.txt). This is the
file that must be input to the script.
# Must be in the right location! Make sure you are in the folder
that your file to search is in!!!
# Must ensure Header_Search_my_exome.txt file is in the same folder!

echo Please ensure you are working in the folder where the file you
wish to work with is located and that the Header_Search_my_exome.txt
file is also in this folder.
echo Please open the annotated ANNOVAR file for the exome you wish
to search in Excel and Save As a tab delimited text .txt file.
echo Have you done this? Y or N?
read response
if [ $response = "Y" ]
then
echo  Please input this patient exome file inwhich you wish to
look for ALS genes. Please save as a .txt and include the file
extension below.

read PATIENT
```

```
file="$PATIENT"
OUT="$PATIENT"_"$(date_'+%Y%m%d').txt"
echo "The_output_file_is_"RESULTS_$PATIENT"_"$(date '+%Y%m%d').txt""

for file in $PATIENT
do
cp $PATIENT $PATIENT.orig
perl -plne 's/\r/\n/g' $PATIENT > $PATIENT.new
mv $PATIENT.new $PATIENT
awk '{print FILENAME"\t"$0}' $PATIENT > $PATIENT.bk
mv $PATIENT.bk $PATIENT

awk ' { if($12=="ANG") print($0) } ' $file >> $OUT

awk ' { if($12=="ATXN2") print $0 } ' $file >> $OUT

awk ' { if($12=="C9orf72") print $0 } ' $file >> $OUT

awk ' { if($12=="CCNF") print $0 } ' $file >> $OUT

awk ' { if($12=="CELF4") print $0 } ' $file >> $OUT

awk ' { if($12=="CHMP2B") print $0 } ' $file >> $OUT

awk ' { if($12=="CREST") print $0 } ' $file >> $OUT

awk ' { if($12=="DAO") print $0 } ' $file >> $OUT

awk ' { if($12=="ERBB4") print $0 } ' $file >> $OUT

awk ' { if($12=="EWSR1") print $0 } ' $file >> $OUT

awk ' { if($12=="FBXO") print $0 } ' $file >> $OUT

awk ' { if($12=="FIG4") print $0 } ' $file >> $OUT

awk ' { if($12=="FUS") print $0 } ' $file >> $OUT
```

```
awk ' { if($12=="HNRNP") print $0 } ' $file >> $OUT

awk ' { if($12=="OPTN") print $0 } ' $file >> $OUT

awk ' { if($12=="P4HB") print $0 } ' $file >> $OUT

awk ' { if($12=="PDIA2") print $0 } ' $file >> $OUT

awk ' { if($12=="PDIA3") print $0 } ' $file >> $OUT

awk ' { if($12=="PFN1") print $0 } ' $file >> $OUT

awk ' { if($12=="RRM2") print $0 } ' $file >> $OUT

awk ' { if($12=="SETX") print $0 } ' $file >> $OUT

awk ' { if($12=="SMN1") print $0 } ' $file >> $OUT

awk ' { if($12=="SMN2") print $0 } ' $file >> $OUT

awk ' { if($12=="SOD1") print $0 } ' $file >> $OUT

awk ' { if($12=="SQSTM1") print $0 } ' $file >> $OUT

awk ' { if($12=="SRRM2") print $0 } ' $file >> $OUT

awk ' { if($12=="SS18L1") print $0 } ' $file >> $OUT

awk ' { if($12=="TAF15") print $0 } ' $file >> $OUT

awk ' { if($12=="TARDBP") print $0 } ' $file >> $OUT

awk ' { if($12=="UBQLN2") print $0 } ' $file >> $OUT

awk ' { if($12=="VAPB") print $0 } ' $file >> $OUT
```

```
awk ' { if($12=="VCP") print $0 } ' $file >> $OUT
done

cat Header_Search_my_exome.txt $OUT >>
"RESULTS_$PATIENT"_"$(date_'+%Y%m%d').txt"


echo You have used the Search my Exome script to search
for mutations in ANG, ATXN2, C9orf72, CCNF, CELF4, CHMP2B,
CREST, DAO, ERBB4, EWSR1, FBXO, FIG4, FUS, HNRNP, OPTN, P4HB,
PDIA2, PDIA3, PFN1, RRM2, SETX, SMN1, SMN2, SOD1, SQSTM1,
SRRM2, TAF15, TARDBP, UBQLN2, VAPB, VCP. This script was
updated with ALS genes as at 19/02/2014.

else
echo Please restart the program and follow instructions.
Fi
```

A.2.2 PLINKSEQ.sh

The following script is used to establish relatedness between individuals using non-basewise vcf files.

```
#!/bin/sh
#PLINKSEQ scripts for relatedness

echo start PLINKSEQ_test project creation

pseq family_187_test new-project --metameta
/Users/Emily/Documents/PLINKSEQ/meta.meta
--resources /Users/Emily/Documents/PLINKSEQ/hg19

echo finished PLINKSEQ_test project creation

echo start loading vcfs

pseq family_187_test load-vcf --vcf
/Users/Emily/Documents
/Family_vcf_files/Family_187/*.vcf

echo finished loading vcfs

echo start PLINKSEQ_test

pseq family_187_test ibs-matrix
--mask mac=2-5 --long-format
--two-counts >> result.txt

echo finished PLINKSEQ_test
```

A.3 Example output files from novel bioinformatic Scripts

A.3.1 Example of output text file from Search_my_exome.sh

This is an example of the output text file generated by running the Search_my_exome.sh script from A.2.1, listing all genomic variants present in that sample for all input ALS genes, and can be easily opened in Microsoft excel for further analysis.

Sample	#chr_name	chr_start	chr_end	ref_base	alt_base	hom_het	snp_quality	tot_depth	alt_depth	region	gene	change	annotation	dbSNP135_fi	dbSNP135_c	1000G_2010	1000G_2011	SCS	CLN	OMIM
mq1-MQ130	chr14	21162053	21162053	T	G	het	51	194	70	exonic	ANG	synonymous	ANG:NM_00	rs11701	rs11701	0.113	0.12	.	CLN	105850:611895
mq1-MQ130	chr12	111894072	111894072	C	T	het	58	10	4	intronic	ATXN2	.	.	rs2073950	rs2073950	0.221	0.23	.	.	.
mq1-MQ130	chr12	111895203	111895203	C	G	het	207	32	14	intronic	ATXN2	.	.	rs2301622	rs2301622	0.541	0.54	.	.	.
mq1-MQ130	chr12	111895272	111895272	C	T	het	225	26	16	intronic	ATXN2	.	.	rs2301621	rs2301621	0.225	0.23	.	.	.
mq1-MQ130	chr12	111923455	111923455	G	A	het	42	20	9	intronic	ATXN2	.	.	rs17805591	rs17805591	0.017	0.01	.	.	.
mq1-MQ130	chr12	112036753	112036823	GGCTGCTGC	GGCTGCTGC	hom	214	9	9	exonic	ATXN2	nonframeshi	ATXN2:NM_(	.	.	.	.	.	.	.
mq1-MQ130	chr12	112036929	112036929	G	A	hom	100	5	5	exonic	ATXN2	synonymous	ATXN2:NM_(	rs695872	rs695872	0.547	0.44	.	.	.
mq1-MQ130	chr12	112037000	112037000	G	C	hom	122	9	9	exonic	ATXN2	nonsynonym	ATXN2:NM_(	rs695871	rs695871	0.42	0.46	.	.	.
mq1-MQ130	chr09	27561468	27561468	-	TGT	hom	214	68	68	UTR3	C9orf72	.	.	rs3063748	.	.	.	.	.	.
mq1-MQ130	chr09	27562352	27562352	T	C	hom	222	20	20	intronic	C9orf72	.	.	rs2589050	.	1	1	.	.	.
mq1-MQ130	chr09	27567145	27567145	C	T	het	119	76	32	UTR5	C9orf72	.	.	rs10757668	rs10757668	0.174	0.18	.	.	.
mq1-MQ130	chr16	2479641	2479641	G	A	het	7.8	7	2	intronic	CCNF	.	.	.	.	.	.	.	.	.
mq1-MQ130	chr16	2480948	2480948	A	G	het	38	25	8	intronic	CCNF	.	.	rs12923030	rs12923030	0.105	0.11	.	.	.
mq1-MQ130	chr16	2485665	2485702	CACATATATA	CATATATATA	hom	14.4	2	2	intronic	CCNF	.	.	.	.	.	.	.	.	.
mq1-MQ130	chr16	2485896	2485896	A	G	het	97	19	7	intronic	CCNF	.	.	rs12928789	rs12928789	0.781	0.78	.	.	.
mq1-MQ130	chr16	2486998	2486998	A	T	het	35	6	5	intronic	CCNF	.	.	rs12921396	rs12921396	0.767	0.78	.	.	.
mq1-MQ130	chr16	2487012	2487012	-	T	het	115	13	8	intronic	CCNF	.	.	rs34456414	.	0.584287	.	.	.	.
mq1-MQ130	chr16	2488211	2488211	T	C	het	83	6	4	intronic	CCNF	.	.	rs12926008	rs12926008	0.763	0.78	.	.	.
mq1-MQ130	chr16	2489533	2489533	T	G	het	10.5	5	4	intronic	CCNF	.	.	rs28647184	rs28647184	0.751	0.78	.	.	.
mq1-MQ130	chr16	2498828	2498828	A	G	het	105	18	9	intronic	CCNF	.	.	rs28417759	rs28417759	0.755	0.78	.	.	.
mq1-MQ130	chr16	2498849	2498849	G	A	het	156	39	20	intronic	CCNF	.	.	rs28670436	rs28670436	0.709	0.75	.	.	.
mq1-MQ130	chr16	2499011	2499011	T	C	het	130	91	42	intronic	CCNF	.	.	rs8060813	rs8060813	0.755	0.78	.	.	.
mq1-MQ130	chr16	2499804	2499804	-	TCCCCTCCA	het	217	65	22	intronic	CCNF	.	.	.	.	.	.	.	.	.
mq1-MQ130	chr18	34850846	34850846	G	A	het	79	22	12	exonic	CELF4	synonymous	CELF4:NM_0	rs1443638	rs1443638	0.271	0.24	.	.	.
mq1-MQ130	chr18	34852895	34852895	C	T	het	107	29	13	intronic	CELF4	.	.	rs2241944	rs2241944	0.383	0.38	.	.	.
mq1-MQ130	chr18	35065396	35065396	T	C	hom	24	3	3	intronic	CELF4	.	.	rs4799451	rs4799451	.	.	.	.	.
mq1-MQ130	chr03	87295049	87295049	T	C	hom	168	25	25	exonic	CHMP2B	synonymous	CHMP2B:NM	rs11540913	rs11540913	0.816	0.84	.	CLN	.
mq1-MQ130	chr22	29687487	29687487	-	C	het	23.5	93	39	intronic	EWSR1	.	.	.	.	.	.	.	.	.
mq1-MQ130	chr22	29688632	29688632	T	G	het	124	155	82	intronic	EWSR1	.	.	rs3761426	rs3761426	0.149	0.15	.	.	.
mq1-MQ130	chr22	29692488	29692488	G	T	hom	86	15	15	intronic	EWSR1	.	.	rs140065	rs140065	0.85	0.86	.	.	.
mq1-MQ130	chr22	29692497	29692497	G	T	hom	74	10	10	intronic	EWSR1	.	.	rs131190	rs131190	0.881	0.89	.	.	.
mq1-MQ130	chr22	29693990	29693990	A	G	het	225	51	27	intronic	EWSR1	.	.	rs3747142	rs3747142	0.192	0.18	.	.	.
mq1-MQ130	chr06	110036479	110036479	A	T	het	12.3	9	4	intronic	FIG4	.	.	rs6924436	rs6924436	0.237	0.23	.	.	.
mq1-MQ130	chr06	110048500	110048500	-	T	hom	29.5	40	36	intronic	FIG4	.	.	rs11459279	.	.	.	.	.	.
mq1-MQ130	chr06	110059510	110059510	C	A	het	127	104	53	intronic	FIG4	.	.	rs2273752	rs2273752	0.312	0.35	.	.	.
mq1-MQ130	chr06	110064259	110064263	TCATT	-	het	217	23	11	intronic	FIG4	.	.	rs57291908	.	.	.	.	.	.
mq1-MQ130	chr06	110065052	110065054	CAA	-	het	121	10	4	intronic	FIG4	.	.	.	.	.	.	.	.	.
mq1-MQ130	chr06	110106234	110106234	A	G	hom	217	49	49	intronic	FIG4	.	.	rs10499054	rs10499054	0.4	0.43	.	.	.
mq1-MQ130	chr06	110110943	110110943	G	T	hom	109	23	23	intronic	FIG4	.	.	rs9384723	rs9384723	0.381	0.42	.	.	.
mq1-MQ130	chr06	110146303	110146303	G	A	het	146	145	65	exonic	FIG4	synonymous	FIG4:NM_01	rs9398218	rs9398218	0.384	0.39	.	.	.
mq1-MQ130	chr16	31191482	31191482	A	G	hom	222	149	149	UTR5	FUS	.	.	rs929867	rs929867	0.952	0.95	.	.	.
mq1-MQ130	chr16	31193942	31193942	C	A	hom	222	100	99	exonic	FUS	synonymous	FUS:NM_001	rs741810	rs741810	0.197	0.22	.	.	.
mq1-MQ130	chr16	31196148	31196158	TACTTTCTTT	-	hom	48	4	4	intronic	FUS	.	.	rs59633484	.	.	.	.	.	.
mq1-MQ130	chr10	13151224	13151224	G	A	het	225	44	20	exonic	OPTN	synonymous	OPTN:NM_0	rs2234968	rs2234968	0.147	0.18	.	.	.
mq1-MQ130	chr10	13152516	13152516	T	-	hom	46.5	54	54	intronic	OPTN	.	.	rs72043574	.	.	.	.	.	.
mq1-MQ130	chr10	13158262	13158262	C	T	hom	139	39	39	intronic	OPTN	.	.	rs2244380	rs2244380	0.786	0.8	.	.	.
mq1-MQ130	chr10	13164332	13164332	T	C	het	38	13	8	intronic	OPTN	.	.	rs765884	rs765884	0.18	0.19	.	.	.
mq1-MQ130	chr10	13164596	13164596	A	G	hom	222	33	33	intronic	OPTN	.	.	rs489040	rs489040	0.489	0.48	.	.	.
mq1-MQ130	chr10	13166076	13166076	A	G	hom	186	67	67	exonic	OPTN	nonsynonym	OPTN:NM_0	rs523747	.	0.992	0.99	.	.	.
mq1-MQ130	chr10	13167860	13167860	G	T	hom	166	64	64	intronic	OPTN	.	.	rs676302	rs676302	0.804	0.8	.	.	.
mq1-MQ130	chr10	13175682	13175682	A	C	het	111	83	41	intronic	OPTN	.	.	rs7086894	rs7086894	0.311	0.33	.	.	.
mq1-MQ130	chr16	333146	333146	C	T	het	31	22	7	UTR5	PDIA2	.	.	rs421195	rs421195	0.027	0.05	.	.	.
mq1-MQ130	chr16	333220	333220	G	A	het	139	29	8	exonic	PDIA2	synonymous	PDIA2:NM_0	.	.	.	.	.	.	.
mq1-MQ130	chr16	336660	336660	G	A	het	181	147	75	exonic	PDIA2	synonymous	PDIA2:NM_0	rs11647490	rs11647490	0.553	0.6	.	.	.
mq1-MQ130	chr15	44038899	44038899	C	T	het	109	56	25	exonic	PDIA3	synonymous	PDIA3:NM_0	rs2411284	rs2411284	0.518	0.57	.	.	.
mq1-MQ130	chr15	44053617	44053617	C	T	het	17.1	26	11	intronic	PDIA3	.	.	rs8040336	rs8040336	0.335	0.39	.	.	.
mq1-MQ130	chr15	44061802	44061802	C	T	het	159	90	44	exonic	PDIA3	synonymous	PDIA3:NM_0	rs1053492	rs1053492	0.444	0.53	.	.	.
mq1-MQ130	chr02	10262920	10262920	T	G	het	225	75	42	exonic	RRM2	nonsynonym	RRM2:NM_0	rs1130609	rs1130609	0.689	0.66	.	.	.
mq1-MQ130	chr02	10264057	10264057	C	C	hom	222	50	50	intronic	RRM2	.	.	rs6432065	rs6432065	0.694	0.66	.	.	.
mq1-MQ130	chr02	10267211	10267211	T	C	het	225	93	38	intronic	RRM2	.	.	rs61754180	rs61754180	0.052	0.05	.	.	.
mq1-MQ130	chr09	135202829	135202829	T	C	hom	222	105	105	exonic	SETX	nonsynonym	SETX:NM_01	rs543573	rs543573	0.534	0.59	.	.	.
mq1-MQ130	chr09	135203231	135203231	C	T	hom	126	87	87	exonic	SETX	nonsynonym	SETX:NM_01	rs1183768	rs1183768	0.52	0.59	.	.	.
mq1-MQ130	chr09	135203409	135203409	A	C	hom	222	70	70	exonic	SETX	nonsynonym	SETX:NM_01	rs1185193	rs1185193	0.626	0.66	.	.	.
mq1-MQ130	chr09	135206460	135206460	A	G	hom	222	60	60	exonic	SETX	synonymous	SETX:NM_01	rs9411449	rs9411449	0.608	0.65	.	.	.
mq1-MQ130	chr09	135221597	135221597	G	A	hom	108	24	24	intronic	SETX	.	.	rs497650	.	0.99	0.99	.	.	.
mq1-MQ130	chr16	2808363	2808363	C	T	hom	147	16	16	intronic	SRRM2	.	.	rs3094778	rs3094778	0.547	0.59	.	.	.
mq1-MQ130	chr16	2810623	2810623	T	G	hom	40.1	6	6	intronic	SRRM2	.	.	rs2285879	rs2285879	0.372	0.43	.	.	.
mq1-MQ130	chr16	2812890	2812890	A	G	hom	204	98	98	exonic	SRRM2	synonymous	SRRM2:NM_	rs2240141	rs2240141	0.643	0.68	.	.	.
mq1-MQ130	chr16	2812939	2812939	C	A	hom	186	117	117	exonic	SRRM2	nonsynonym	SRRM2:NM_	rs2240140	rs2240140	0.378	0.42	.	.	.
mq1-MQ130	chr16	2813517	2813517	C	T	het	82	210	82	exonic	SRRM2	synonymous	SRRM2:NM_	rs140191985	.	0.004	0.01	.	.	.
mq1-MQ130	chr16	2814162	2814162	G	A	hom	186	49	49	exonic	SRRM2	synonymous	SRRM2:NM_	rs3094775	rs3094775	0.615	0.65	.	.	.
mq1-MQ130	chr16	2815237	2815237	A	C	hom	222	204	204	exonic	SRRM2	synonymous	SRRM2:NM_	rs3094773	rs3094773	0.9	0.91	.	.	.
mq1-MQ130	chr16	2818161	2818161	T	C	hom	177	39	39	exonic	SRRM2	synonymous	SRRM2:NM_	rs2301802	rs2301802	0.654	0.67	.	.	.
mq1-MQ130	chr16	2818704	2818704	T	G	hom	222	96	96	intronic	SRRM2	.	.	rs3094792	rs3094792	0.632	0.68	.	.	.
mq1-MQ130	chr17	34171525	34171525	C	T	het	143	61	22	exonic	TAF15	nonsynonym	TAF15:NM_1	.	.	.	.	.	.	.
mq1-MQ130	chr01	11079077	11079077	A	G	het	49	8	3	intronic	TARDBP	.	.	rs2273348	rs2273348	0.576	0.63	.	.	.
mq1-MQ130	chr09	35060302	35060302	T	C	het	19.1	5	2	intronic	VCP	.	.	rs684562	rs684562	0.381	0.43	.	.	.
mq1-MQ130	chr09	35060955	35060955	T	C	het	176	164	83	intronic	VCP	.	.	rs2258240	rs2258240	0.69	0.7	.	.	.

A.3.2 Example of output text file from PLINKSEQ.sh

This is an example of the output text file generated by running the PLINKSEQ.sh script from A.2.2, listing the number of shared genotypes between each pair of individuals as well as their total number of non-missing genotypes, that is the total number of genetic

sites at which a genotype was found for both individuals, and can be easily opened in Microsoft excel for further analysis.

Sample 1	Sample 2	# of Genotypes shared	Total # of non-missing Genotypes for the pair	Ratio
206-100168	86-950164	9133	26085	0.35012459
206-100169	86-950164	9194	26007	0.35352021
206-100167	86-950164	9799	26266	0.37306784
206-100168	5-A124	10380	27880	0.3723099
206-100169	5-A124	10479	27910	0.37545683
206-100169	5-100248	10667	28759	0.37090998
206-100168	5-100248	10751	28824	0.37298779
206-100167	5-A124	10835	28250	0.38353982
5-970585	206-100168	10860	28626	0.37937539
5-970585	206-100169	11024	28641	0.38490276
206-100167	5-100248	11277	29062	0.38803248
206-100167	5-970585	11350	28646	0.39621588
5-100248	86-950164	12275	28876	0.4250935
5-970585	86-950164	12509	28766	0.43485365
5-A124	86-950164	13971	30210	0.46246276
206-100167	206-100169	15571	36675	0.42456714
206-100167	206-100168	16069	37306	0.430735
5-970585	5-100248	16695	34405	0.48524924
5-100248	5-A124	16976	35248	0.48161598
5-970585	5-A124	17173	34730	0.49447164
206-100168	206-100169	21804	42219	0.51644994

A.4 Parameters utilised to determine informative polymorphic markers for zygosity testing

A.4.1 Informative Meioses and Heterozygosity values for available microsatellites

By identifying the highest values for each of informative meioses and heterozygosity values and ensuring both sufficient, those microsatellites in bold were selected for use in zygosity testing.

Microsatellite	Informative meisosis	Heterozygosity
D3S1555	956	0.75
D3S3516	91	0.47
D3S3614	675	0.73
D3S3705	621	0.74
D6S1541	556	0.54
D6S1651	121	0.71
D8S509	741	0.63
D8S601	**724**	**0.8**
D9S166	**930**	**0.81**
D9S1674	**939**	**0.79**
D9S1684	812	0.71
D9S1852	89	0.64
D9S1869	589	0.74
D9S264	422	0.74
D9S268	755	0.64
D11S1911	435	0.72
D11S875	754	0.72
D15S1021	767	0.63
D15S1038	633	0.69
D15S122	158	0.85
D15S1514	809	0.75
D16S2619	143	0.77
D16S2622	654	0.58
D16S283	268	0.65
D16S3024	163	0.89
D16S3034	735	0.6
D16S3044	793	0.74

D16S3062	134	0.81
D16S3064	182	0.89
D16S3079	947	0.76
D16S3081	120	0.75
D16S3082	170	0.87
D16S3088	790	0.69
D16S3114	154	0.94
D16S3124	435	0.61
D16S3145	778	0.65
D16S401	**780**	**0.76**
D16S404	**911**	**0.79**
D16S411	**963**	**0.82**
D16S415	121	0.7
D16S423	**1141**	**0.76**
D16S475	**639**	**0.84**
D16S504	705	0.69
D16S521	784	0.76
D16S541	883	0.66
D16S682	86	0.66
D16S685	728	0.7
D16S690	221	0.76
D16S746	88	0.68
D16S753	**834**	**0.8**
D16S760	18	0.74
D16S768	44	0.83
D16S769	91	0.71
D16S770	741	0.65
D17S107	131	0.43
D17S1826	184	0.43
D17S1848	424	0.51
D17S802	**692**	**0.82**
D17S836	662	0.63
D17S927	736	0.68
D19S604	102	0.7
D20S103	983	0.68
D20S160	60	0.68
D20S172	143	0.69

D20S175	820	0.65
D20S177	643	0.57
D20S199	**376**	**0.86**
D20S27	986	0.73
D20S61	864	0.69
D20S873	105	0.63
D20S889	156	0.87
D20S892	504	0.81
D20S915	719	0.67
DXS1190	702	0.52
DXS8022	**1273**	**0.85**
DXS8032	1061	0.7
DXS991	**1202**	**0.8**

A.4.2 Minor allele frequencies for available SNPs

By identifying the highest values for minor allele frequencies, those SNPs in bold were selected for use in zygosity testing.

rs #	Chr	Minor Allele frequency
1474891	**1**	**0.371901**
2273348	1	0.370983
3765895	1	0.150138
7546123	**1**	**0.342057**
9430335	1	0.156566
774359	9	0.199725
1565948	9	0.421488
1948522	9	0.133609
2282241	**9**	**0.435721**
2814707	9	0.17539
3849942	9	0.191001
10122902	9	0.213039
10757665	9	0.172176
774359	9	0.199725
1565948	**9**	**0.421488**
1948522	9	0.133609
2814707	9	0.17539

3849942	9	0.191001
2235580	**16**	**0.385675**
2294605	**16**	**0.376033**
3794621	16	0.337466
6600147	16	0.355372
204141	X	0.448005
204165	X	0.442563
512119	X	0.423216
763183	**X**	**0.495768**
957721	X	0.451632
1536163	X	0.455865
1560514	X	0.498186
1927228	X	0.288996
2005463	X	0.368198
2473057	**X**	**0.331318**
3002630	X	0.270254

A.5 Ethics Approval

The work in this project was conducted in accordance with the below ethics approval.

MACQUARIE UNIVERSITY

Office of the Deputy Vice-Chancellor (Research)

Research Office
C5C Research HUB East, Level 3, Room 324
MACQUARIE UNIVERSITY NSW 2109 AUSTRALIA

Phone +61 (0)2 9850 4194
Fax +61 (0)2 9850 4465
Email ethics.secretariat@mq.edu.au

3 December 2013

Professor Gilles J Guillemin
Head of the Neuroinflammation group
Co-Director of the MND and
Neurodegenerative diseases Research Centre
Australian School of Advanced Medicine (ASAM)
Macquarie University

Dear Professor Guillemin,

RE: *Investigation of mechanisms involved in inflammatory diseases*

Thank you for your email dated 21 May 2013 responding to the issues raised by the Macquarie University Human Research Ethics Committee (HREC (Medical Sciences)).

Your reposnes and revised documents were reviewed by the HREC (Medical Sciences) at their meeting held on the 23 May 2013. This research meets the requirements set out in the *National Statement on Ethical Conduct in Human Research* (2007).

Details of this approval are as follows:

Reference No: 5201300333

Approval Date: 23 May 2013

This letter constitutes ethical and scientific approval only.

The following documentation has been reviewed and approved by the HREC (Medical Sciences):

Documents reviewed	Version no.	Date
Ethics and Privacy Application Form for Research Involving Humans	8	Nov 2005
Correspondence from Prof Gilles Guillemin addressing the HREC's feedback		Received 21/05/2013
MQ Participant Information and Consent Form	2	22/05/2013

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. Approval is for five (5) years, subject to the submission of annual reports.

First Annual Report Due: 1 June 2014

3. All adverse events must be reported to the HREC within 72 hours. Any issues which affect the continued ethical acceptability of the project must also be reported to the HREC.

4. Proposed changes to the protocol 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.

Please do not hesitate to contact the Ethics Secretariat should you have any questions regarding your ethics application.

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

Yours sincerely

Dr Karolyn White
Director, Research Ethics
Chair, 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) (the National Statement) and the CPMP/ICH Note for Guidance on Good Clinical Practice.

Abbreviations

5mC	5-methyl-cytosine
AD	Alzheimer's disease
ALS	amyotrophic lateral sclerosis
ALSoD	amyotrophic lateral sclerosis online genetic database
ANNOVAR	functional annotation of genetic variants from high-throughput sequencing data
ATXN2	ataxin-2
bp	base pair
BWA	burrows-wheeler alignment
C9ORF72	chromosome 9 open reading frame 72
CCNF	cyclin F
CG	cytosine and guanine
CNV	copy number variant
CpG	cytosine guanine dinucleotide
CSIRO	Commonwealth Scientific and Industrial Research Organisation
dbSNP	single nucleotide polymorphism database
dGTP	deoxyguanosine triphosphate
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid

Dnmt	DNA methyltransferase
dNTP	deoxyribonucleotide triphosphate
dsDNA	double stranded DNA
DZ	dizygotic
E. coli	*Escherichia coli*
EAAT2	excitatory amino-acid transporter 2
EDTA	ethylenediaminetetraacetic acid
ER	endoplasmic reticulum
EVS	exome variant server
FALS	familial ALS
FISH	fluorescent in situ hybridization
FTD	frontotemporal dementia
FUS	fused in sarcoma
GC content	guanine and cytosine content
gDNA	genomic DNA
GLT1	glial glutamate transporter 1
GTP	guanosine triphosphate
GWAS	genome-wide association studies
hnRNP	heterogeneousnuclear ribonucleoprotein
IBM	inclusion body myopathy
KDM2A	lysine (K)-specific demethylase 2A
LB	luria broth
MAF	minor allele frequency
MATR3	matrin 3

Mb	megabase
MeDIP-seq	methylated DNA immunoprecipitation sequencing
MND	motor neuron disease
MRE-Seq	methylation-sensitive restriction enzyme sequencing
mRNA	messenger RNA
MZ	monozygotic
NBIC	Netherlands bioinformatics c entre
NCBI	national center for biotechnology information
NGS	next-generation sequencing
NLS	nuclear localisation signal
OMIM	online mendelian inheritance in man
OPTN	optineurin
p62	nucleoporin p62
PCR	polymerase chain reaction
PD	Parkinson's disease
PDB	Paget disease of bone
RBM14	RNA binding motif protein 14
RNA	ribonucleic acid
SALS	sporadic ALS
SAM	sequence alignment/map
SFPQ	splicing factor proline/glutamine-rich
SNP	single nucleotide polymorphism
SOD1	superoxide dismutase 1
SQSTM1	sequestosome 1

TARDBP	TAR DNA binding protein gene
TBE	tris-borate EDTA
TDP-43	TAR DNA binding protein of 43 kDa
TSAP	thermosensitive alkaline phosphatase
UBQN2	ubiquilin 2
UCSC	university of California, Santa Cruz
UCSF	University of California, San Francisco
UPS	ubiquitin proteasome system
UPS	ubiquitin-proteasome system
UTR	untranslated region
UTR	untranslated region
VCF	variant call format
VCP	valosin containing protein
WGS	whole genome sequencing
VEGF	Vascular endothelial growth factor

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, *Nat. Methods*, 7, 4, 248–249 16

Al-Chalabi, A., Fang, F., Hanby, M. F., Leigh, P. N., Shaw, C. E., Ye, W., et al. (2010), An estimate of amyotrophic lateral sclerosis heritability using twin data, *J. Neurol. Neurosurg. Psychiatr.*, 81, 12, 1324–1326 5, 77

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 Neuropathol.*, 124, 3, 339–352 1, 2, 5, 9

Al-Chalabi, A., Kwak, S., Mehler, M., Rouleau, G., Siddique, T., Strong, M., et al. (2013), Genetic and epigenetic studies of amyotrophic lateral sclerosis, *Amyotroph Lateral Scler Frontotemporal Degener*, 14 Suppl 1, 44–52 4, 6, 8

Andersen, P. M. and Al-Chalabi, A. (2011), Clinical genetics of amyotrophic lateral sclerosis: what do we really know?, *Nat Rev Neurol*, 7, 11, 603–615 2, 5

Andersen, P. M., Forsgren, L., Binzer, M., Nilsson, P., Ala-Hurula, V., Keranen, M. L., et al. (1996), Autosomal recessive adult-onset amyotrophic lateral sclerosis associated with homozygosity for Asp90Ala CuZn-superoxide dismutase mutation. A clinical and genealogical study of 36 patients, *Brain*, 119 (Pt 4), 1153–1172 4, 64

Andersen, P. M., Nilsson, P., Ala-Hurula, V., Keranen, M. L., Tarvainen, I., Haltia, T., et al. (1995), Amyotrophic lateral sclerosis associated with homozygosity for an Asp90Ala mutation in CuZn-superoxide dismutase, *Nat. Genet.*, 10, 1, 61–66 4, 37

Australian Institute of Health and Welfare (2011), Analysis of Australian Institute of Health and Welfare National National Mortality Database, Technical report, Australian Institute of Health and Welfare, Canberra, NSW 1

Bakulski, K. M., Dolinoy, D. C., Sartor, M. A., Paulson, H. L., Konen, J. R., Lieberman, A. P., et al. (2012), Genome-wide DNA methylation differences between late-onset Alzheimer's disease and cognitively normal controls in human frontal cortex, *J. Alzheimers Dis.*, 29, 3, 571–588 9

Bell, J. T. and Spector, T. D. (2011), A twin approach to unraveling epigenetics, *Trends Genet.*, 27, 3, 116–125 6, 79

Belzil, V. V., Bauer, P. O., Gendron, T. F., Murray, M. E., Dickson, D., and Petrucelli, L. (2014), Characterization of DNA hypermethylation in the cerebellum of c9FTD/ALS patients, *Brain Res.* 8, 82

Belzil, V. V., Bauer, P. O., Prudencio, M., Gendron, T. F., Stetler, C. T., Yan, I. K., et al. (2013), Reduced C9orf72 gene expression in c9FTD/ALS is caused by histone trimethylation, an epigenetic event detectable in blood, *Acta Neuropathol.*, 126, 6, 895–905 82, 83

Bersano, A., Del Bo, R., Lamperti, C., Ghezzi, S., Fagiolari, G., Fortunato, F., et al. (2009), Inclusion body myopathy and frontotemporal dementia caused by a novel VCP mutation, *Neurobiol. Aging*, 30, 5, 752–758 73, 74

Blom, N., Gammeltoft, S., and Brunak, S. (1999), Sequence and structure-based prediction of eukaryotic protein phosphorylation sites, *J. Mol. Biol.*, 294, 5, 1351–1362 18

Bond, C. S. and Fox, A. H. (2009), Paraspeckles: nuclear bodies built on long noncoding RNA, *J. Cell Biol.*, 186, 5, 637–644 70

Boxer, A. L., Mackenzie, I. R., Boeve, B. F., Baker, M., Seeley, W. W., Crook, R., et al. (2011), Clinical, neuroimaging and neuropathological features of a new chromosome 9p-linked FTD-ALS family, *J. Neurol. Neurosurg. Psychiatr.*, 82, 2, 196–203 4

Boycott, K. M., Vanstone, M. R., Bulman, D. E., and MacKenzie, A. E. (2013), Rare-disease genetics in the era of next-generation sequencing: discovery to translation, *Nat. Rev. Genet.*, 14, 10, 681–691 67

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, *Amyotroph. Lateral Scler. Other Motor Neuron Disord.*, 1, 5, 293–299 11

Callaghan, B., Feldman, D., Gruis, K., and Feldman, E. (2011), The association of exposure to lead, mercury, and selenium and the development of amyotrophic lateral sclerosis and the epigenetic implications, *Neurodegener Dis*, 8, 1-2, 1–8 8, 77

Chatterjee, A. and Morison, I. M. (2011), Monozygotic twins: genes are not the destiny?, *Bioinformation*, 7, 7, 369–370 79, 80

Chen, K. L., Wang, S. S., Yang, Y. Y., Yuan, R. Y., Chen, R. M., and Hu, C. J. (2009), The epigenetic effects of amyloid-beta(1-40) on global DNA and neprilysin genes in murine cerebral endothelial cells, *Biochem. Biophys. Res. Commun.*, 378, 1, 57–61 9

Chen, L., Hadd, A., Sah, S., Filipovic-Sadic, S., Krosting, J., Sekinger, E., et al. (2010), An information-rich CGG repeat primed PCR that detects the full range of fragile X expanded alleles and minimizes the need for southern blot analysis, *J Mol Diagn*, 12, 5, 589–600 60, 61

Chen, Y. Z., Bennett, C. L., Huynh, H. M., Blair, I. P., Puls, I., Irobi, J., et al. (2004), DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4), *Am. J. Hum. Genet.*, 74, 6, 1128–1135 3

Chesi, A., Staahl, B. T., Jovi?i?, A., Couthouis, J., Fasolino, M., Raphael, A. R., et al. (2013), Exome sequencing to identify de novo mutations in sporadic ALS trios, *Nat. Neurosci.*, 16, 7, 851–855 3

Chestnut, B. A., Chang, Q., Price, A., Lesuisse, C., Wong, M., and Martin, L. J. (2011), Epigenetic regulation of motor neuron cell death through DNA methylation, *J. Neurosci.*, 31, 46, 16619–16636 8, 78

Chio, A., Restagno, G., Brunetti, M., Ossola, I., Calvo, A., Canosa, A., et al. (2012), ALS/FTD phenotype in two Sardinian families carrying both C9ORF72 and TARDBP mutations, *J. Neurol. Neurosurg. Psychiatr.*, 83, 7, 730–733 5

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 4, 64

Chouliaras, L., Mastroeni, D., Delvaux, E., Grover, A., Kenis, G., Hof, P. R., et al. (2013), Consistent decrease in global DNA methylation and hydroxymethylation in the hippocampus of Alzheimer's disease patients, *Neurobiol. Aging*, 34, 9, 2091–2099 9

Chow, C. Y., Landers, J. E., Bergren, S. K., Sapp, P. C., Grant, A. E., Jones, J. M., et al. (2009), Deleterious variants of FIG4, a phosphoinositide phosphatase, in patients with ALS, *Am. J. Hum. Genet.*, 84, 1, 85–88 3

Corcia, P., Mayeux-Portas, V., Khoris, J., de Toffol, B., Autret, A., Muh, J. P., et al. (2002), Abnormal SMN1 gene copy number is a susceptibility factor for amyotrophic lateral sclerosis, *Ann. Neurol.*, 51, 2, 243–246 3

Couthouis, J., Hart, M. P., Erion, R., King, O. D., Diaz, Z., Nakaya, T., et al. (2012), Evaluating the role of the FUS/TLS-related gene EWSR1 in amyotrophic lateral sclerosis, *Hum. Mol. Genet.*, 21, 13, 2899–2911 3

Couthouis, J., Hart, M. P., Shorter, J., DeJesus-Hernandez, M., Erion, R., Oristano, R., et al. (2011), A yeast functional screen predicts new candidate ALS disease genes, *Proc. Natl. Acad. Sci. U.S.A.*, 108, 52, 20881–20890 3

Cudkowicz, M. E., McKenna-Yasek, D., Sapp, P. E., Chin, W., Geller, B., Hayden, D. L., et al. (1997), Epidemiology of mutations in superoxide dismutase in amyotrophic lateral sclerosis, *Ann. Neurol.*, 41, 2, 210–221 82

Czyz, W., Morahan, J. M., Ebers, G. C., and Ramagopalan, S. V. (2012), Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences, *BMC Med*, 10, 93 80

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011), The variant call format and VCFtools, *Bioinformatics*, 27, 15, 2156–2158 16

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 71

Davies, M. N., Volta, M., Pidsley, R., Lunnon, K., Dixit, A., Lovestone, S., et al. (2012), Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood, *Genome Biol.*, 13, 6, R43 83

de Carvalho, M. and Swash, M. (2011), Amyotrophic lateral sclerosis: an update, *Curr. Opin. Neurol.*, 24, 5, 497–503 1, 2

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 3, 4, 60, 61

Dempster, E. L., Pidsley, R., Schalkwyk, L. C., Owens, S., Georgiades, A., Kane, F., et al. (2011), Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder, *Hum. Mol. Genet.*, 20, 24, 4786–4796 81

Deng, H. X., Bigio, E. H., Zhai, H., Fecto, F., Ajroud, K., Shi, Y., et al. (2011a), Differential involvement of optineurin in amyotrophic lateral sclerosis with or without SOD1 mutations, *Arch. Neurol.*, 68, 8, 1057–1061 76

Deng, H. X., Chen, W., Hong, S. T., Boycott, K. M., Gorrie, G. H., Siddique, N., et al. (2011b), Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia, *Nature*, 477, 7363, 211–215 2, 3, 6, 9

Deng, H. X., Hentati, A., Tainer, J. A., Iqbal, Z., Cayabyab, A., Hung, W. Y., et al. (1993), Amyotrophic lateral sclerosis and structural defects in Cu,Zn superoxide dismutase, *Science*, 261, 5124, 1047–1051 37, 64

Deng, H. X., Zhai, H., Bigio, E. H., Yan, J., Fecto, F., Ajroud, K., et al. (2010), FUS-immunoreactive inclusions are a common feature in sporadic and non-SOD1 familial amyotrophic lateral sclerosis, *Ann. Neurol.*, 67, 6, 739–748 2

Dion, P. A., Daoud, H., and Rouleau, G. A. (2009), Genetics of motor neuron disorders: new insights into pathogenic mechanisms, *Nat. Rev. Genet.*, 10, 11, 769–782 1, 4, 5

Dols-Icardo, O., Garcia-Redondo, A., Rojas-Garcia, R., Sanchez-Valle, R., Noguera, A., Gomez-Tortosa, E., et al. (2014), Characterization of the repeat expansion size in C9orf72 in amyotrophic lateral sclerosis and frontotemporal dementia, *Hum. Mol. Genet.*, 23, 3, 749–754 77

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, 1069–1075 3

Fecto, F. and Siddique, T. (2011), Making connections: pathology and genetics link amyotrophic lateral sclerosis with frontotemporal lobe dementia, *J. Mol. Neurosci.*, 45, 3, 663–675 2

Fecto, F., Yan, J., Vemula, S. P., Liu, E., Yang, Y., Chen, W., et al. (2011), SQSTM1 mutations in familial and sporadic amyotrophic lateral sclerosis, *Arch. Neurol.*, 68, 11, 1440–1446 3, 6

Feil, R. and Fraga, M. F. (2011), Epigenetics and the environment: emerging patterns and implications, *Nat. Rev. Genet.*, 13, 2, 97–109 8

Figueroa-Romero, C., Hur, J., Bender, D. E., Delaney, C. E., Cataldo, M. D., Smith, A. L., et al. (2012), Identification of epigenetically altered genes in sporadic amyotrophic lateral sclerosis, *PLoS ONE*, 7, 12, e52672 8, 78

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., et al. (2005), Epigenetic differences arise during the lifetime of monozygotic twins, *Proc. Natl. Acad. Sci. U.S.A.*, 102, 30, 10604–10609 8, 79, 80, 81

Fujiwara, K., Toda, H., and Ikeguchi, M. (2012), Dependence of -helical and -sheet amino acid propensities on the overall protein fold type, *BMC Struct. Biol.*, 12, 18 70

Gijselinck, I., Engelborghs, S., Maes, G., Cuijt, I., Peeters, K., Mattheijssens, M., et al. (2010), Identification of 2 Loci at chromosomes 9 and 14 in a multiplex family with frontotemporal lobar degeneration and amyotrophic lateral sclerosis, *Arch. Neurol.*, 67, 5, 606–616 4

Gilissen, C., Hoischen, A., Brunner, H. G., and Veltman, J. A. (2012), Disease gene identification strategies for exome sequencing, *Eur. J. Hum. Genet.*, 20, 5, 490–497 66

Goetz, C. G. and Charcot, J. M. (2000), Amyotrophic lateral sclerosis: early contributions of Jean-Martin Charcot, *Muscle Nerve*, 23, 3, 336–343 65

Gozani, O., Patton, J. G., and Reed, R. (1994), A novel set of spliceosome-associated proteins and the essential splicing factor PSF bind stably to pre-mRNA prior to catalytic step II of the splicing reaction, *EMBO J.*, 13, 14, 3356–3367 70

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, *Nat. Genet.*, 38, 4, 411–413 3

Handel, A. E., Ebers, G. C., and Ramagopalan, S. V. (2010), Epigenetics: molecular mechanisms and implications for disease, *Trends Mol Med*, 16, 1, 7–16 6, 8

Hanson, B., McGue, M., Roitman-Johnson, B., Segal, N. L., Bouchard, T. J., and Blumenthal, M. N. (1991), Atopic disease and immunoglobulin E in twins reared apart and together, *Am. J. Hum. Genet.*, 48, 5, 873–879 79

Hantash, F. M., Goos, D. G., Tsao, D., Quan, F., Buller-Burckle, A., Peng, M., et al. (2010), Qualitative assessment of FMR1 (CGG)n triplet repeat status in normal, intermediate, premutation, full mutation, and mosaic carriers in both sexes: implications for fragile X syndrome carrier and newborn screening, *Genet. Med.*, 12, 3, 162–173 61, 62

Haubenberger, D., Bittner, R. E., Rauch-Shorny, S., Zimprich, F., Mannhalter, C., Wagner, L., et al. (2005), Inclusion body myopathy and Paget disease is linked to a novel mutation in the VCP gene, *Neurology*, 65, 8, 1304–1305 73

Hortobagyi, T., Troakes, C., Nishimura, A. L., Vance, C., van Swieten, J. C., Seelaar, H., et al. (2011), Optineurin inclusions occur in a minority of TDP-43 positive ALS and FTLD-TDP cases and are rarely observed in other neurodegenerative disorders, *Acta Neuropathol.*, 121, 4, 519–527 2

Jaenisch, R. and Bird, A. (2003), Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals, *Nat. Genet.*, 33 Suppl, 245–254 6, 7

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 3, 6, 73

Jones, P. A., Archer, T. K., Baylin, S. B., Beck, S., Berger, S., Bernstein, B. E., et al. (2008), Moving AHEAD with an international human epigenome project, *Nature*, 454, 7205, 711–715 7

Jowaed, A., Schmitt, I., Kaut, O., and Wullner, U. (2010), Methylation regulates alpha-synuclein expression and is decreased in Parkinson's disease patients' brains, *J. Neurosci.*, 30, 18, 6355–6359 9

Kaminsky, Z. A., Tang, T., Wang, S. C., Ptak, C., Oh, G. H., Wong, A. H., et al. (2009), DNA methylation profiles in monozygotic and dizygotic twins, *Nat. Genet.*, 41, 2, 240–245 8, 79, 80

Katsanis, S. H. and Katsanis, N. (2013), Molecular genetic testing and the future of clinical genomics, *Nat. Rev. Genet.*, 14, 6, 415–426 59, 66

Ketelaar, M. E., Hofstra, E. M., and Hayden, M. R. (2012), What monozygotic twins discordant for phenotype illustrate about mechanisms influencing genetic forms of neurodegeneration, *Clin. Genet.*, 81, 4, 325–333 79

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 5, 6

Kontopoulos, E., Parvin, J. D., and Feany, M. B. (2006), Alpha-synuclein acts in the nucleus to inhibit histone acetylation and promote neurotoxicity, *Hum. Mol. Genet.*, 15, 20, 3012–3023 9

Ku, C. S., Cooper, D. N., Polychronakos, C., Naidoo, N., Wu, M., and Soong, R. (2012), Exome sequencing: dual role as a discovery and diagnostic tool, *Ann. Neurol.*, 71, 1, 5–14 66, 67, 71

Kwiatkowski, T. J., Bosco, D. A., Leclerc, A. L., Tamrazian, E., Vanderburg, C. R., Russ, C., et al. (2009), Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis, *Science*, 323, 5918, 1205–1208 3, 5, 6

Kwok, C. T., Morris, A. G., Frampton, J., Smith, B., Shaw, C. E., and de Belleroche, J. (2013), Association studies indicate that protein disulfide isomerase is a risk factor in amyotrophic lateral sclerosis, *Free Radic. Biol. Med.*, 58, 81–86 3

Laaksovirta, H., Peuralinna, T., Schymick, J. C., Scholz, S. W., Lai, S. L., Myllykangas, L., et al. (2010), Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study, *Lancet Neurol*, 9, 10, 978–985 4, 62, 63

Laffita-Mesa, J. M., Bauer, P. O., Kouri, V., Pena Serrano, L., Roskams, J., Almaguer Gotay, D., et al. (2012), Epigenetics DNA methylation in the core ataxin-2 gene promoter: novel physiological and pathological implications, *Hum. Genet.*, 131, 4, 625–638 81

Le Ber, I., Camuzat, A., Berger, E., Hannequin, D., Laquerriere, A., Golfier, V., et al. (2009), Chromosome 9p-linked families with frontotemporal dementia associated with motor neuron disease, *Neurology*, 72, 19, 1669–1676 4

Li, H. and Durbin, R. (2010), Fast and accurate long-read alignment with Burrows-Wheeler transform, *Bioinformatics*, 26, 5, 589–595 15

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009), The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, 25, 16, 2078–2079 15

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 65

Linton, M. F., Raabe, M., Pierotti, V., and Young, S. G. (1997), Reading-frame restoration by transcriptional slippage at long stretches of adenine residues in mammalian cells, *J. Biol. Chem.*, 272, 22, 14127–14132 74

Luty, A. A., Kwok, J. B., Thompson, E. M., Blumbergs, P., Brooks, W. S., Loy, C. T., et al. (2008), Pedigree with frontotemporal lobar degeneration–motor neuron

disease and Tar DNA binding protein-43 positive neuropathology: genetic linkage to chromosome 9, *BMC Neurol*, 8, 32 4

Majounie, E., Renton, A. E., Mok, K., Dopper, E., 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, *Lancet Neurol*, 11, 4, 323–330 4, 5, 62

Martin, L. J. and Wong, M. (2013), Aberrant regulation of DNA methylation in amyotrophic lateral sclerosis: a new target of disease mechanisms, *Neurotherapeutics*, 10, 4, 722–733 2

Maruyama, H. and Kawakami, H. (2013), Optineurin and amyotrophic lateral sclerosis, *Geriatr Gerontol Int*, 13, 3, 528–532 6

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 3, 6, 76

Masliah, E., Dumaop, W., Galasko, D., and Desplats, P. (2013), Distinctive patterns of DNA methylation associated with Parkinson disease: identification of concordant epigenetic changes in brain and peripheral blood leukocytes, *Epigenetics*, 8, 10, 1030–1038 9, 83

Mastroeni, D., Grover, A., Delvaux, E., Whiteside, C., Coleman, P. D., and Rogers, J. (2010), Epigenetic changes in Alzheimer's disease: decrements in DNA methylation, *Neurobiol. Aging*, 31, 12, 2025–2037 9

Mastroeni, D., McKee, A., Grover, A., Rogers, J., and Coleman, P. D. (2009), Epigenetic differences in cortical neurons from a pair of monozygotic twins discordant for Alzheimer's disease, *PLoS ONE*, 4, 8, e6617 9

Matise, T. C., Chen, F., Chen, W., De La Vega, F. M., Hansen, M., He, C., et al. (2007), A second-generation combined linkage physical map of the human genome, *Genome Res.*, 17, 12, 1783–1786 27

Matsumoto, L., Takuma, H., Tamaoka, A., Kurisaki, H., Date, H., Tsuji, S., et al. (2010), CpG demethylation enhances alpha-synuclein expression and affects the pathogenesis of Parkinson's disease, *PLoS ONE*, 5, 11, e15522 9

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D., et al. (2010), Conserved role of intragenic DNA methylation in regulating alternative promoters, *Nature*, 466, 7303, 253–257 28

Meerang, M., Ritz, D., Paliwal, S., Garajova, Z., Bosshard, M., Mailand, N., et al. (2011), The ubiquitin-selective segregase VCP/p97 orchestrates the response to DNA double-strand breaks, *Nat. Cell Biol.*, 13, 11, 1376–1382 73

Mitchell, J., Paul, P., Chen, H. J., Morris, A., Payling, M., Falchi, M., et al. (2010), Familial amyotrophic lateral sclerosis is associated with a mutation in D-amino acid oxidase, *Proc. Natl. Acad. Sci. U.S.A.*, 107, 16, 7556–7561 3

Morahan, J. M., Yu, B., Trent, R. J., and Pamphlett, R. (2007), Are metallothionein genes silenced in ALS?, *Toxicol. Lett.*, 168, 1, 83–87 8, 77

Morahan, J. M., Yu, B., Trent, R. J., and Pamphlett, R. (2009), A genome-wide analysis of brain DNA methylation identifies new candidate genes for sporadic amyotrophic lateral sclerosis, *Amyotroph Lateral Scler*, 10, 5-6, 418–429 8, 77, 78

Morita, M., Al-Chalabi, A., Andersen, P. M., 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 4

Moulard, B., Salachas, F., Chassande, B., Briolotti, V., Meininger, V., Malafosse, A., et al. (1998), Association between centromeric deletions of the SMN gene and sporadic adult-onset lower motor neuron disease, *Ann. Neurol.*, 43, 5, 640–644 3

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 2

Nishimura, A. L., Mitne-Neto, M., Silva, H. C., Richieri-Costa, A., Middleton, S., Cascio, D., et al. (2004), A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis, *Am. J. Hum. Genet.*, 75, 5, 822–831 3

Oates, N. and Pamphlett, R. (2007), An epigenetic analysis of SOD1 and VEGF in ALS, *Amyotroph Lateral Scler*, 8, 2, 83–86 8, 82, 83

Ollikainen, M., Smith, K. R., Joo, E. J., Ng, H. K., Andronikos, R., Novakovic, B., et al. (2010), DNA methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome, *Hum. Mol. Genet.*, 19, 21, 4176–4188 80

Parkinson, N., Ince, P. G., Smith, M. O., Highley, R., Skibinski, G., Andersen, P. M., et al. (2006), ALS phenotypes with mutations in CHMP2B (charged multivesicular body protein 2B), *Neurology*, 67, 6, 1074–1077 3

Petronis, A. (2006), Epigenetics and twins: three variations on the theme, *Trends Genet.*, 22, 7, 347–350 79

Petronis, A., Gottesman, I. I., Kan, P., Kennedy, J. L., Basile, V. S., Paterson, A. D., et al. (2003), Monozygotic twins exhibit numerous epigenetic differences: clues to twin discordance?, *Schizophr Bull*, 29, 1, 169–178 79, 81

Radunovic, A. and Leigh, P. N. (1999), ALSODatabase: database of SOD1 (and other) gene mutations in ALS on the Internet. European FALS Group and ALSOD Consortium, *Amyotroph. Lateral Scler. Other Motor Neuron Disord.*, 1, 1, 45–49 22

Renton, A. E., Chio, A., and Traynor, B. J. (2014), State of play in amyotrophic lateral sclerosis genetics, *Nat. Neurosci.*, 17, 1, 17–23 5, 60

Renton, A. E., Majounie, E., Waite, A., Simon-Sanchez, 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 3, 4, 24, 60, 61

Robberecht, W. and Philips, T. (2013), The changing scene of amyotrophic lateral sclerosis, *Nat. Rev. Neurosci.*, 14, 4, 248–264 1, 2, 4, 5, 6, 65, 71

Robertson, K. D. (2005), DNA methylation and human disease, *Nat. Rev. Genet.*, 6, 8, 597–610 81

Rosen, D. R. (1993), Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis, *Nature*, 364, 6435, 362 2, 3, 4, 37, 64

Rowland, L. P. and Shneider, N. A. (2001), Amyotrophic lateral sclerosis, *N. Engl. J. Med.*, 344, 22, 1688–1700 2

Sahlender, D. A., Roberts, R. C., Arden, S. D., Spudich, G., Taylor, M. J., Luzio, J. P., et al. (2005), Optineurin links myosin VI to the Golgi complex and is involved in Golgi organization and exocytosis, *J. Cell Biol.*, 169, 2, 285–295 76

Sanchez-Mut, J. V., Aso, E., Heyn, H., Matsuda, T., Bock, C., Ferrer, I., et al. (2014), Promoter hypermethylation of the phosphatase DUSP22 mediates PKA-dependent TAU phosphorylation and CREB activation in Alzheimer's disease, *Hippocampus*, 24, 4, 363–368 9

Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014), MutationTaster2: mutation prediction for the deep-sequencing age, *Nat. Methods*, 11, 4, 361–362 16, 18

Sewer, M. B., Nguyen, V. Q., Huang, C. J., Tucker, P. W., Kagawa, N., and Waterman, M. R. (2002), Transcriptional activation of human CYP17 in H295R adrenocortical cells depends on complex formation among p54(nrb)/NonO, protein-associated splicing factor, and SF-1, a complex that also participates in repression of transcription, *Endocrinology*, 143, 4, 1280–1290 70

Shatunov, A., Mok, K., Newhouse, S., Weale, M. E., Smith, B., Vance, C., et al. (2010), Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study, *Lancet Neurol*, 9, 10, 986–994 4, 62, 63

Siddique, T., Figlewicz, 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, *N. Engl. J. Med.*, 324, 20, 1381–1384 4

Spiess, A. N., Mueller, N., and Ivell, R. (2004), Trehalose is a potent PCR enhancer: lowering of DNA melting temperature and thermal stabilization of taq polymerase by the disaccharide trehalose, *Clin. Chem.*, 50, 7, 1256–1259 63

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 2, 3, 5

Sreedharan, J. and Brown, R. H. (2013), Amyotrophic lateral sclerosis: Problems and prospects, *Ann. Neurol.*, 74, 3, 309–316 4

Steves, C. J., Spector, T. D., and Jackson, S. H. (2012), Ageing, genes, environment and epigenetics: what twin studies tell us now, and in the future, *Age Ageing*, 41, 5, 581–586 80

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., et al. (1966), Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday, *Cold Spring Harb. Symp. Quant. Biol.*, 31, 77–84 75

Takahashi, Y., Fukuda, Y., Yoshimura, J., Toyoda, A., Kurppa, K., Moritoyo, H., et al. (2013), ERBB4 mutations that disrupt the neuregulin-ErbB4 pathway cause amyotrophic lateral sclerosis type 19, *Am. J. Hum. Genet.*, 93, 5, 900–905 3

Teer, J. K. and Mullikin, J. C. (2010), Exome sequencing: the sweet spot before whole genomes, *Hum. Mol. Genet.*, 19, R2, R145–151 66, 70

Teperino, R., Lempradl, A., and Pospisilik, J. A. (2013), Bridging epigenomics and complex disease: the basics, *Cell. Mol. Life Sci.*, 70, 9, 1609–1621 6, 7

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013), Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, *Brief. Bioinformatics*, 14, 2, 178–192 16

Tohgi, H., Utsugisawa, K., Nagane, Y., Yoshimura, M., Genda, Y., and Ukitsu, M. (1999), Reduction with age in methylcytosine in the promoter region -224 approximately -101 of the amyloid precursor protein gene in autopsy human cortex, *Brain Res. Mol. Brain Res.*, 70, 2, 288–292 9

Tremolizzo, L., Messina, P., Conti, E., Sala, G., Cecchi, M., Airoldi, L., et al. (2014), Whole-blood global DNA methylation is increased in amyotrophic lateral sclerosis independently of age of onset, *Amyotroph Lateral Scler Frontotemporal Degener*, 15, 1-2, 98–105 1, 8

Urdinguio, R. G., Sanchez-Mut, J. V., and Esteller, M. (2009), Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies, *Lancet Neurol*, 8, 11, 1056–1072 9, 81

Valdmanis, P. N., Dupre, N., Bouchard, J. P., Camu, W., Salachas, F., Meininger, V., et al. (2007), Three families with amyotrophic lateral sclerosis and frontotemporal dementia with evidence of linkage to chromosome 9p, *Arch. Neurol.*, 64, 2, 240–245 4

van Blitterswijk, M., van Es, M. A., Hennekam, E. A., Dooijes, D., van Rheenen, W., Medic, J., et al. (2012a), Evidence for an oligogenic basis of amyotrophic lateral sclerosis, *Hum. Mol. Genet.*, 21, 17, 3776–3784 5

van Blitterswijk, M., van Es, M. A., Koppers, M., van Rheenen, W., Medic, J., Schelhaas, H. J., et al. (2012b), VAPB and C9orf72 mutations in 1 familial amyotrophic lateral sclerosis patient, *Neurobiol. Aging*, 33, 12, 1–4 5

Van Deerlin, V. M., Sleiman, P. M., Martinez-Lage, M., Chen-Plotkin, A., Wang, L. S., Graff-Radford, N. R., et al. (2010), Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions, *Nat. Genet.*, 42, 3, 234–239 4

van der Zee, J., Pirici, D., Van Langenhove, T., Engelborghs, S., Vandenberghe, R., Hoffmann, M., et al. (2009), Clinical heterogeneity in 3 unrelated families linked to VCP p.Arg159His, *Neurology*, 73, 8, 626–632 73

van Es, M. A., Veldink, J. H., Saris, C. G., Blauw, H. M., van Vught, P. W., Birve, A., et al. (2009), Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis, *Nat. Genet.*, 41, 10, 1083–1087 4, 62, 63

Vance, C., Al-Chalabi, A., Ruddy, D., Smith, B. N., Hu, X., Sreedharan, J., et al. (2006), Familial amyotrophic lateral sclerosis with frontotemporal dementia is linked to a locus on chromosome 9p13.2-21.3, *Brain*, 129, Pt 4, 868–876 4

Vance, C., Rogelj, B., Hortobagyi, 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 3, 5, 6

Voutsinas, G. E., Stavrou, E. F., Karousos, G., Dasoula, A., Papachatzopoulou, A., Syrrou, M., et al. (2010), Allelic imbalance of expression and epigenetic regulation within the alpha-synuclein wild-type and p.Ala53Thr alleles in Parkinson disease, *Hum. Mutat.*, 31, 6, 685–691 9

Wagner, L. A., Weiss, R. B., Driscoll, R., Dunn, D. S., and Gesteland, R. F. (1990), Transcriptional slippage occurs during elongation at runs of adenine or thymine in Escherichia coli, *Nucleic Acids Res.*, 18, 12, 3529–3535 74

Wang, K., Li, M., and Hakonarson, H. (2010), ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, *Nucleic Acids Res.*, 38, 16, e164 15

Wang, Z., Liu, X., Yang, B. Z., and Gelernter, J. (2013), The role and challenges of exome sequencing in studies of human diseases, *Front Genet*, 4, 160 66

Watts, G. D., Wymer, J., Kovach, M. J., Mehta, S. G., Mumm, S., Darvish, D., et al. (2004), Inclusion body myopathy associated with Paget disease of bone and frontotemporal dementia is caused by mutant valosin-containing protein, *Nat. Genet.*, 36, 4, 377–381 73

Wijesekera, L. C. and Leigh, P. N. (2009), Amyotrophic lateral sclerosis, *Orphanet J Rare Dis*, 4, 3 2

Williams, K. L., Durnall, J. C., Thoeng, A. D., Warraich, S. T., Nicholson, G. A., and Blair, I. P. (2009), A novel TARDBP mutation in an Australian amyotrophic lateral sclerosis kindred, *J. Neurol. Neurosurg. Psychiatr.*, 80, 11, 1286–1288 38

Williams, K. L., Fifita, J. A., Vucic, S., Durnall, J. C., Kiernan, M. C., Blair, I. P., et al. (2013), Pathophysiological insights into ALS with C9ORF72 expansions, *J. Neurol. Neurosurg. Psychiatr.*, 84, 8, 931–935 5

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, *Neurobiol. Aging*, 33, 10, 3–10 2, 3, 6

Worms, P. M. (2001), The epidemiology of motor neuron diseases: a review of recent studies, *J. Neurol. Sci.*, 191, 1-2, 3–9 1

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 3

Xi, Z., Rainero, I., Rubino, E., Pinessi, L., Bruni, A. C., Maletta, R. G., et al. (2014), Hypermethylation of the CpG-island near the C9orf72 G4C2-repeat expansion in FTLD patients, *Hum. Mol. Genet.* 82

Xi, Z., Zinman, L., Moreno, D., Schymick, J., Liang, Y., Sato, C., et al. (2013), Hypermethylation of the CpG island near the G4C2 repeat in ALS with a C9orf72 expansion, *Am. J. Hum. Genet.*, 92, 6, 981–989 9, 82, 83

Yang, X., Lay, F., Han, H., and Jones, P. A. (2010a), Targeting DNA methylation for epigenetic therapy, *Trends Pharmacol. Sci.*, 31, 11, 536–546 83

Yang, Y., Gozen, O., Vidensky, S., Robinson, M. B., and Rothstein, J. D. (2010b), Epigenetic regulation of neuron-dependent induction of astroglial synaptic protein GLT1, *Glia*, 58, 3, 277–286 9, 77

Zhang, D., Li, S., Tan, Q., and Pang, Z. (2012), Twin-based DNA methylation analysis takes the center stage of studies of human complex diseases, *J Genet Genomics*, 39, 11, 581–586 7, 8

Zhang, G. and Pradhan, S. (2014), Mammalian epigenetic mechanisms, *IUBMB Life*, 66, 4, 240–256 7, 79