Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

From single nucleotide polymorphisms to high-throughput sequencing in the complex genetics of amyotrophic lateral sclerosis

Thesis presented for the degree of Doctor of Philosophy

in Molecular Medicine

2012

Russell Lewis McLaughlin BSc (Hons)

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed 3.8.2012

Summary

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterized by progressive weakening of limb and bulbar muscles resulting in paralysis and death from respiratory failure within three to five years of symptom onset. The disease manifests as a consequence of sudden and rapid degeneration of upper and lower motor neurones, for which the causative biomolecular processes are still relatively unknown. There is no cure. A number of genes have been shown to cause the condition; however a substantial proportion of its heritability is still unexplained by genetics. In an attempt to address this, a number of genome-wide association studies (GWAS) have been attempted in recent years in ALS, in several populations including the Irish.

This thesis describes work that was carried out to investigate the contribution of genetic variation to the pathogenesis of ALS. Four separate bodies of work are represented in the thesis. The first investigated the contribution of genetic variation across the *ANG* locus (a known ALS gene) to serum, plasma and cerebrospinal fluid levels of angiogenin in ALS patients and controls from Ireland, Sweden and Poland. This study confirmed previously reported associations of *ANG* polymorphisms with ALS. Furthermore, angiogenin levels were observed to be lower in ALS patients than in controls and there was a tissue-differentiated dysregulation of the protein observed in ALS.

Secondly, the genome-wide single nucleotide polymorphism (SNP) dataset from the 2008 Irish ALS GWAS was augmented by further genotyping and this larger dataset was

used in several analyses to identify regions of the genome implicated in ALS aetiology. Association testing, copy number variation mapping and mapping of recurrent, overlapping, ALS-specific runs of homozygosity were carried out using these data, revealing several genomic intervals that may harbour genes that play a role in the pathogenesis of ALS. Through these analyses, several genes were also identified as candidates for follow-up work.

The third body of work was aimed at assaying the contribution of rare variation to ALS aetiology through next-generation sequencing of the exons of candidate genes that overlapped with regions identified in the work on genome-wide SNPs. A large number of rare variants were identified in the dataset, for which control sequencing will be required so that Irish population variants can be ruled out. Nevertheless, many interesting findings are presented from this work, including the observation of a possible burden of rare variants in *HYDIN*, a gene already implicated in neurological development and function, as well as rare variants in *UNC13A*, a gene previously associated with ALS through GWAS.

Finally, both the augmented genome-wide SNP dataset and the rare variant data generated in the NGS project were used to investigate the optimal design of future studies involving exome sequencing. This was primarily achieved through assessment of identity-by-descent (IBD) in the Irish population, using a British dataset as a comparison. IBD was found to be higher within the Irish population than within the British, and individuals showed some clustering of inter-relatedness as well geographical clustering. Exome sequencing of as few as three or four inter-related cases and two hypernormal controls could reveal novel variants associated with ALS.

In summary, the work presented in this thesis has attempted to describe the contribution of genetic variation to ALS aetiology, primarily in the Irish population. In doing so, several avenues for future research have been indicated. It is hoped that this work will contribute to a better understanding of ALS aetiology and help towards the future development of a cure.

Contents

List of figures

List of tables

List of abbreviations

aCGH Array comparative genomic hybridization

ALS Amyotrophic lateral sclerosis

AMD Age-related macular degeneration

ANOVA Analysis of variance

BAF B-allele frequency

BSA Bovine serum albumin

CNV Copy number variant

CSF Cerebrospinal fluid

DNA Deoxyribonucleic acid

EBI European Bioinformatics Institute

EDTA Ethylenediaminetetraacetic acid

ELISA Enzyme-linked immunosorbent assay

EM Expectation maximization

ENCODE Encyclopedia of DNA Elements

EWAS Epigenome-wide association study

FTD Frontotemporal dementia

GRCh37 Human Genome Reference Consortium genome build '37'

GWAS Genome-wide association study

HGP Human genome project

HMM Hidden Markov model

IBD Identity by descent

IBS Identity by state

Indel Insertion/deletion

LD Linkage disequilibrium

LRR Log-R ratio

MAF Minor allele frequency

mRNA Messenger ribonucleic acid

NCBI36 National Centre for Biotechnology Information human genome build 36

ncRNA Non-coding ribonucleic acid

NGS Next-generation sequencing

OB Objective Bayes

PCA Principal components analysis

PCR Polymerase chain reaction

qPCR Quantitative polymerase chain reaction

RNA Ribonucleic acid

ROH Run of homozygosity

RPM Revolutions per minute

SD Standard deviation

SNP Single nucleotide polymorphism

SNV Single nucleotide variant

TE Tris-ethylenediaminetetraacetic acid

UCSC University of South California, Santa Cruz

WTCCC Wellcome Trust Case-Control Consortium

Acknowledgements

The research detailed herein was carried out between October 2008 and December 2011 with the support of the Irish Health Research Board and Research Motor Neurone. It concerns the disease Amyotrophic Lateral Sclerosis, and has made extensive use of DNA, blood and CSF samples contributed by patients and controls from Ireland. Sweden and Poland. I would primarily like to thank all individuals, especially patients, who have generously donated specimens throughout the years, for making this research possible.

As with any PhD thesis, one author takes credit for work which would not have been possible without the support of a legion of others. Foremost thanks go to my supervisors, **Professor Dan Bradley** and **Professor Orla Hardiman**, for conceiving the project and for taking me onboard. **Simon Cronin** helped me find my feet in the early days of the project, and for discussions, arguments and general support throughout its duration I am greatly indebted to my past and present colleagues **Sarah Connell**, **Tim Downing**, **Ceiridwen Edwards**, **Emma Finlay**, **Yonas Hirutu**, **Eppie Jones**, **Kevin Kenna**, **Lilian Lau**, **Valeria Mattiangeli**, **Ian Richardson**, **Frauke Stock** and **Matthew Teasdale** of the Bradley Lab, and **Peter Bede**, **Susan Byrne**, **Marwa Elamin** and **Catherine Lynch** of Prof. Hardiman's group, without which the project would have taken much longer.

Finally, I owe unending thanks to my highly supportive network of friends and family, especially **Alice Vajda**, who kept me relatively human throughout the whole endeavour.

Chapter 1

Introduction

The completion of the human genome project (HGP) in 2001 heralded a paradigm shift for biomedical research [1,2]. Over the course of ten years, at a cost of around 2.7 billion dollars, thousands of research scientists at twenty sequencing centres across the planet painstakingly pieced together the three billion bases of instructions for making a human. Ten years on from the initial release of the draft human genome sequence, it is possible for a single laboratory to sequence an entire human genome within weeks at a cost approaching one millionth of that of the HGP. This is attributable to the many scientific and technological advances that were made as a direct consequence of the HGP, but it is also only possible thanks to the immediate product of the project itself: the provision of a reference sequence for the human genome which could be used in any subsequent project. The various efforts downstream of the HGP have been diverse, allowing humankind to take a close look at the variations between individuals and populations, the differences that cause disease in some individuals, the specificities that make humans different to other species, and the historical events that shaped our evolution.

One of the major derivatives of the HGP was the International SNP Map Working Group, which assessed the prevalence and distribution of single nucleotide polymorphisms (SNPs) across the human genome in a panel of ethnically diverse individuals [3]. A SNP

is a site in the genome where an alternate allele exists at appreciable frequency for one base in the genetic sequence. SNPs are typically biallelic and their frequencies within populations are usually represented in terms of the minor (less frequent) allele; thus minor allele frequency (MAF) is a commonly-used term.

The major finding of the International SNP Map Working Group was that SNPs are frequent across the human genome; at the time the authors reported an average density of one SNP every 1.9 kilobases (kb). These findings directly fuelled the first iteration of the International HapMap Project [4, 5] which sought to investigate the patterns of genomic variation between individuals by analyzing the co-occurrences of SNPs on haplotypes within and between individuals derived from, at first, four different populations [5,6], later to be extended to 11 populations [7]. This also provided a snapshot of the global variation in genetic diversity as the allele frequencies of all SNPs assayed could be interrogated and compared between populations.

The findings of the International HapMap Project were numerous, but one of its major outcomes was a detailed description of the extent of linkage disequilibrium (LD) between SNPs across the genome. LD is a statistical measure of the co-occurrence of particular alleles of neighbouring SNPs, and exists because genetic recombination affects large chunks of the genome at a time, meaning that genetic variation is passed on in chunks. When the presence of one allele of a SNP has no predictive effect on which allele occurs at another SNP, this is called linkage equilibrium, and any departure from this is LD, measured on a scale of 0 to 1. LD is often quantified by one of two statistical metrics, D' or r^2, where, for a 2-SNP haplotype comprising SNPs A and B,

$$D' = \frac{x_{11} - p_1 q_1}{\min(p_1 q_1, p_2 q_2)}, \tag{1.1}$$

and

$$r^2 = \frac{(x_{11} - p_1 q_1)^2}{p_1 q_1 p_2 q_2}, \tag{1.2}$$

where x_{ij} is the observed frequency of the haplotype $A_i B_j$, p_i is the frequency of A_i and q_i is the frequency of B_i. In words, D' describes the normalised deviation of the haplotype frequency from linkage equilibrium, and r^2 is a measure of the statistical correlation between a pair of SNPs.

Using these measures of LD between the millions of SNPs characterized in the HapMap project, a reduced marker set could be ascertained that captured the majority of human genetic variation in Europeans [8]. This permitted the development of genome-wide SNP arrays that could genotype hundreds of thousands of SNPs in a single experiment. Two companies emerged as the main competitors for provision of the technology, Illumina and Affymetrix [9], each of which used subtly different strategies to decide their marker sets. While Affymetrix included SNPs so that the whole genome was covered relatively regularly, Illumina based their marker set design on LD patterns revealed by HapMap to maximise the amount of common genetic variation captured [10]. This resulted in demonstrably better coverage of the genome with a similar number of markers [11], and improved power to detect associations with disease [12]. Consequently, the Illumina SNP array has since been the tool of choice for many research groups in the design of genome-wide association studies (GWAS).

1.1 SNP chips and genome-wide association studies

The essential principles of the Illumina bead chip and the Infinium assay, which forms the biochemical basis for genotyping on Illumina SNP arrays, are summarised in figure 1.1. DNA is added to the SNP chip and, following the single base extension process, fluorescence signals are read from the chip by an array scanner. This way, in a single

straightforward assay, most of the common variation (MAF $>$ 1%) in an individual's genome can be determined. When this is carried out on many individuals across many SNP chips, a dataset can be generated for use in a GWAS.

The common disease-common variant hypothesis, on which GWAS are based in principle, is an idea that began to emerge in the early days of the HGP [13–15]. Scientists became increasingly aware that common variation in the human genome may contribute to the genetic risk for diseases that are common in the population, and that assaying the common variation across the genome in a panel of individuals selected by disease phenotype may yield associations of these common variants with disease when compared with a panel of controls assayed for the same variation. Risch and Merikangas issued a "charge to the molecular technologists to develop the tools to meet this challenge," [13] which would come to be realised by the likes of Illumina and Affymetrix in the aftermath of the HapMap project.

The potential of GWAS was popularized in a widely-cited early example, a study conducted in age-related macular degeneration (AMD). This research compared genotypes for 105,980 SNPs in 96 cases and 50 controls [16]. Although successful in identifying strongly associated SNPs with AMD, this was a small study by modern standards and for most complex genetic diseases this sample size would be at least one order of magnitude too small to have the power to detect the more modest effects expected (see section 3.1.1 and [12]). Nevertheless, it acheived success in its goals and paved the way for many subsequent study designs in a variety of conditions and traits.

The use of GWAS as an approach for studying the genetics of human diseases has generated an abundance of novel data that will take many years of research to disentangle fully. Indeed, for the second quarter of 2011, the National Human Genome Research Institute's GWAS catalogue [17] cited 1,449 published GWAS that demonstrate significant associations with disease at $p \leq 5 \times 10^{-8}$ for 237 traits, indicating that the method

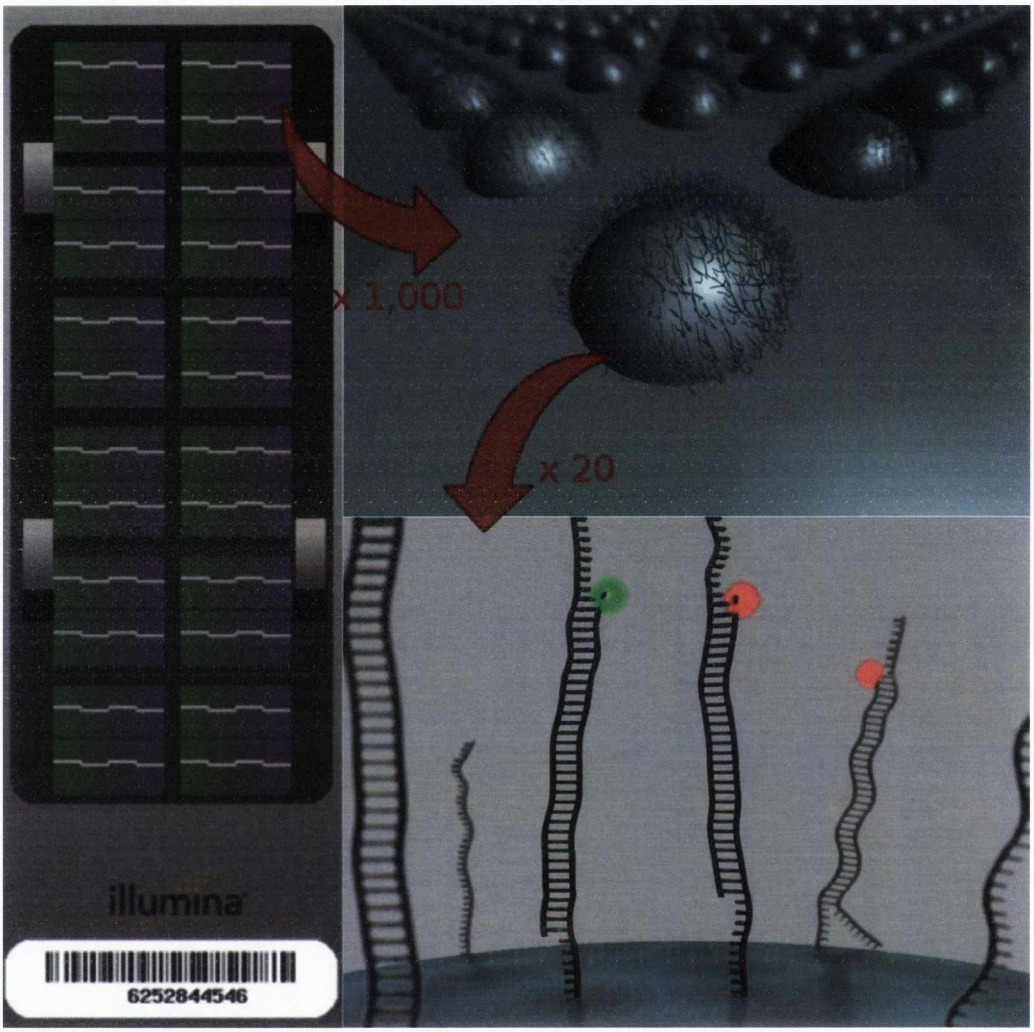


Figure 1.1: The Illumina Infinium assay-based genome-wide SNP array. Each 'chip' has hundreds of thousands of beads with 40-mer oligonucleotide sequences immobilized on their surfaces. These oligonucleotides are complementary to the sequences prior to SNPs of interest, and each bead represents one SNP. When fragmented DNA is added, it hybridizes to the immobilized DNA and a single base extension is carried out with fluorescently-labelled nucleotides. This way, the two alleles of the SNP of interest can be characterized by reading an overall fluorescence signal from the bead, where red or green means that the individual is only carrying allele A or B (interpreted as a homozygote AA or BB) and a composite yellow signal means that the individual is heterozygous, carrying both alleles A and B.

is succeeding triumphantly in generating new hypotheses. Nevertheless, as long as the GWAS method assays only common variation in the human genome, it will stay bound by that limitation, and some of the heritability of complex genetic diseases will remain enigmatic. In many cases, some of this heritability is likely to be driven by rare variants that are either not in strong LD with the common SNPs interrogated in GWAS, or by multiple rare variants within a locus that have arisen on different haplotype backgrounds, thus being tagged differently. While it has been argued that multiple rare variants can drive synthetic association of common variants with disease [18], it has been elegantly demonstrated by Wray and colleagues that this is unlikely for the majority of GWAS [19].

The only solution that GWAS methodology has to addressing the contribution of rare variants to genetically heterogenous diseases is the genotyping of huge numbers of individuals. This is potentially problematic for rare diseases for which it could take several years to build a sample cohort large enough to meet such needs. Furthermore, multiplicity of rare disease-causing variants at a particular locus has the potential to quench GWAS signals. For example, if a locus has two possible haplotypes of equal frequency, each tagged by opposite alleles of the same SNP, and one disease-causing mutation arises on each haplotype simultaneously, a GWAS conducted on a future generation would not detect the locus (assuming no genetic drift or selection on a very large, randomly-breeding population). While purely hypothetical, this situation illustrates how multiple rare variants at a disease locus could remain undetected by GWAS.

Therefore, the inability of GWAS to identify disease-associated loci in such cases highlights the need to assay rare variation in many complex genetic diseases (complex diseases are diseases that are likely to have several contributory elements, many of which are genetic). A very large international project called The 1000 Genomes Project [20] has been underway for some time to investigate rare genetic variation in humans across the planet, making data available to the research community as it is generated. The prinicpal techno-

logical innovation that has made such a project possible is the advent of next-generation sequencing (NGS) technologies. As well as assessing population genetic phenomena in projects such as The 1000 Genomes Project, NGS has potential in the design of experiments assaying causative variants in complex genetic diseases.

1.2 Next-generation sequencing

For over three decades, Sanger sequencing [21] has been the most widely-used method for determining the sequence of bases in a DNA molecule. However, the upper limits of throughput have been reached with current technologies despite a continually growing interest in large scale genetic variation. This has necessitated the development of technologies that can sequence DNA in ultra-high throughput, a challenge which has been met by a few NGS methods, including those of Roche/454, Applied Biosystems and Illumina. The underlying principles of all three technologies are similar: DNA is highly fragmented, these fragments are sequenced in parallel in an ultra-high throughput manner (tens of millions of molecules at a time), then the resulting sequence reads are aligned to the reference genome, allowing some variation in the sequence. The resulting alignments can then be assessed for sequence variants. Such methods have revolutionized the field of genomics research on a scale similar to GWAS, with many studies now harnessing the technologies in the assessment of genetic variation in a variety of fields.

Technical details of NGS are discussed extensively in chapter 4. One disease that is a suitable candidate for both NGS and GWAS is amyotrophic lateral sclerosis (ALS). ALS is a complex genetic disease, the aetiology of which has only partially been explained by genetic factors. The utility of genomic studies such as GWAS and NGS in ALS is the focus of this thesis.

1.3 Amyotrophic lateral sclerosis

ALS is a fatal neurodegenerative condition characterised by progressive loss of motor neurones, resulting in death from respiratory failure typically within three to five years of disease onset. The lifetime risk for adults developing the condition is roughly 1 in 400 [22–24], but prevalence at any given time is fairly low (around 4-6 per hundred thousand person-years [25]) due to its poor prognosis. With a few exceptions, there is little geographic variation in the incidence; in Ireland the incidence has been estimated to be around 2.8 per hundred thousand person-years for adults [25]. The geographical outliers for incidence include the Pacific Island of Guam and the Kii Peninsula in Japan, where an agressive form of ALS/Parkinson-dementia complex is more common [26], possibly due to the biomagnification of the neurotoxin BMAA in the diets of the inhabitants of these areas [27].

The diagnosis of ALS is made by a combination of neurological examination, electrophysiological testing and in some cases, family history. A requisite for its diagnosis is the exclusion of other clinically similar conditions which can mimic the symptoms of ALS. In order for ALS to be clinically 'definite' or 'probable', signs of both upper and lower motor neurone damage must be present (figure 1.2, [28]). In around 75% of cases, the disease manifests first in an extremity such as an arm or leg (commonly termed limb onset or spinal onset); in the remainder of cases the site of onset is in the muscles of the face, head and neck (bulbar onset). Approximately 5% of ALS patients also have frontotemporal dementia (FTD), and as much as 30-50% may have milder cognitive impairment, the identification of which requires detailed neuropsychological testing [29, 30].

The consequence of motor neurone death is progressive muscle weakness, as decreasing efferent innervation leads to muscle atrophy. This effect spreads from the site of onset eventually to involve the majority of skeletal muscles, although generally autonomic function is spared, along with muscles that control eye movement and bladder and bowel function.

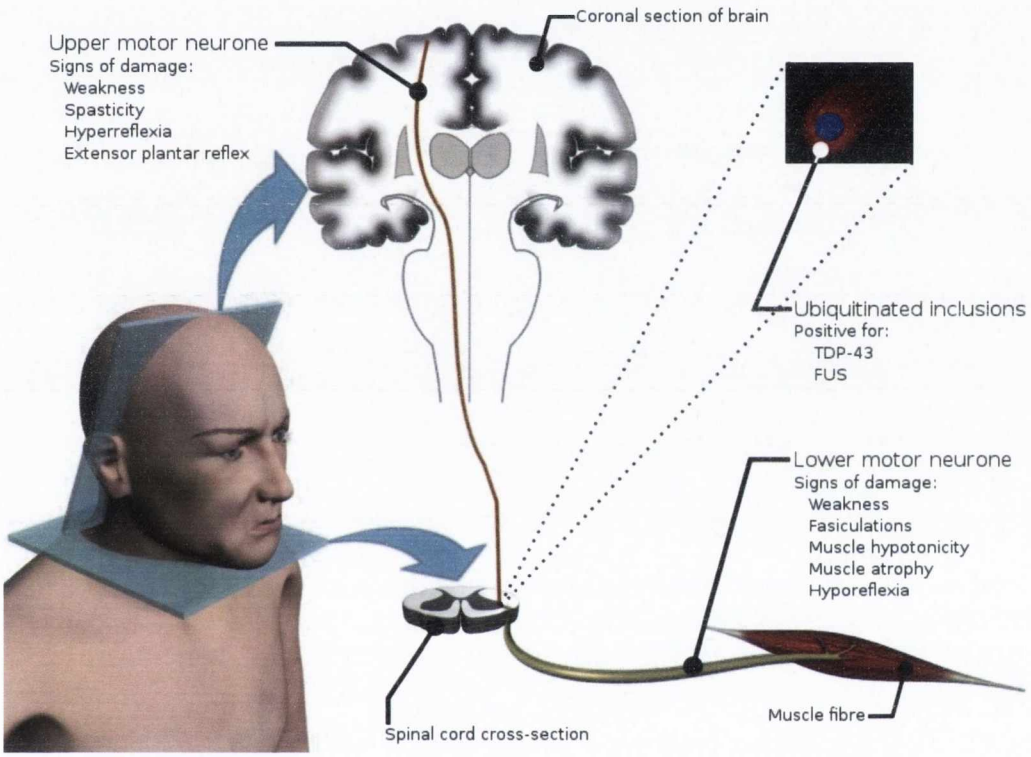


Figure 1.2: Features of ALS [31–34]. In order for ALS to be clinically definite or probable [28], signs of both upper and lower motor neurone damage must be evident. The biochemical hallmark of ALS is cytoplasmic ubiquitinated inclusions that immunostain positively for TDP-43 [33] or FUS [34].

Patients therefore end up profoundly disabled and completely dependent on care-givers, as well as being dysarthric (unable to speak) and dysphagic (unable to swallow). Progressive weakening of the respiratory muscles causes increasing difficulty in breathing, which leads to death from either respiratory failure or respiratory infection in the majority of cases [35]. At the time of writing, the only FDA-approved drug used to treat the progression of ALS is riluzole (Rilutek) [36], which prolongs survival by a just a few months [37]; otherwise the only treatment is palliative care [38]. The exact mechansim of action of riluzole in slowing ALS progression is unclear; however, it may act by increasing glutamate reuptake in the spinal cord, thus limiting the excitotoxic effects of the neurotransmitter [39].

In around 5% of the cases of ALS a family history of the condition is observed [40]. The remainder of ALS conditions are termed 'sporadic' ALS; however, it is widely accepted that a genetic component plays a role in the aetiology of the disease and therefore this terminology could be interpreted as somewhat of a misnomer. With the exception of the putatively environment-driven cases of ALS in Guam and the Kii Peninsula, there have been few, if any, conclusive causative links between environmental risk factors and ALS. Coupled with twin-based heritability estimates of 0.61 [41], this suggests a sizeable genetic component to the aetiology of the disease. Indeed, to date there have been several genes implicated in the pathogenesis of ALS. Generally, a pathophysiological hallmark of the disease is ubiquitinated inclusions in the cytoplasm of affected neurones; these inclusions have been shown to be positive for the proteins TDP-43 and FUS (figure 1.2, [33,34]).

The first genetic locus to be implicated in ALS was the long arm of chromosome 21, identified through familial linkage [42], which was later revealed to be due to mutations in *SOD1* [43]. In this discovery, Rosen *et al.* observed eleven heterozygous mutations in *SOD1* [43], and since then, over 150 mutations in this gene have been implicated in the aetiology of ALS, although not all of these variants are necessarily pathogenic [44]. *SOD1* mutations have been estimated to be the cause of 12-23% of familial ALS cases and 2-3%

of sporadic ALS cases [45].

Since the discovery of *SOD1* as a cause of ALS, several other genes have been implicated in the disease, including *ANG* [46], *DCTN1* [47–49], *TARDBP* [50], *FUS* [51,52], *OPTN* [53–55], *ATXN2* [56], *ALS2* [57, 58], *SETX* [59–61], *SIGMAR1* [62], *VAPB* [63] and *UBQLN2* [64], and many more have been studied in the disease (figure 1.3). Some of the genes in figure 1.3, for example *APEX*, have been studied in ALS by virtue of LD with the causative gene, and many of the studies report negative findings, so this list of genes is somewhat longer than the list of genes currently known to cause ALS directly.

Despite many genes being implicated in ALS, a large proportion of its heritability remains unexplained. It is such observations that have motivated several GWAS attempts in ALS, which are reviewed in chapter 3. Of note, mutations in a further gene, *C9orf72*, have recently been successfully identifed as a cause of ALS with one of the lines of evidence that led to this discovery being a strong GWAS signal in the region [66,67] (although this was already a known locus identified through familial linkage studies [68]) and a few other speculative loci have been identified by GWAS (in particular *UNC13A* [67]). Additionally, NGS has had some success, with mutations in *VCP* being identified in ALS by NGS within an extended pedigree [69].

Although many research efforts have identified many genes to be involved in ALS aetiology, the pathophysiology of the disease is still incompletely understood, and this is in part due to the catalogue of genetic variants known to cause ALS being incomplete. Therefore, a better understanding of the underlying genetics of ALS is required, and the use of modern genomics technologies in addressing this issue is the focus of this thesis.

1.4 Scope and structure of thesis

This thesis presents research exploring the complex genetics of ALS, mainly within the Irish population (within this work, 'Ireland' and 'the Irish population' refer to the island

Figure 1.3: Genes that have been studied in ALS, according to a comprehensive list recorded on the ALS Online Genetics Database (ALSoD) [65]. ALSoD is an excellent resource listing extensive information derived from every study investgating the genetics of ALS.

of Ireland and the combined population of both The Republic of Ireland and Northern Ireland). The research is geared towards a better understanding of the underlying disease mechanisms behind ALS and is somewhat representative of the shifting trends in complex disease genetics research. Chapter 2 investigates genetic variation across a single locus and its effects on protein levels, building on previous observations describing *ANG* mutations in ALS. Chapter 3 scales from a single locus to the whole genome, augmenting an already existing whole-genome SNP dataset and assaying common variation across the genome in ALS. Chapter 4 uses candidate regions generated in chapter 3 to inform the selection of genes for rare variant discovery by NGS. Finally, chapter 5 analyses the data generated in chapters 3 and 4 to make inferences about optimal design in future research into the genetics of ALS in Ireland.

Chapter 2

Angiogenin in amyotrophic lateral sclerosis

2.1 Introduction

In 1999, Hayward *et al.* demonstrated an association of the D148E variant in *APEX*, on chromosome 14q11.2, with ALS susceptibility in 153 Scottish ALS patients [70]. Greenway *et al.* postulated that the causative variant may not lie within *APEX*, but instead within a nearby gene in LD with the associated variant [71]. Examining the local region for candidate ALS genes based on function, they speculatively identified *ANG* due to its functional similarity to *VEGF* [72], a gene which had previously been shown in animal studies to be linked to an ALS-like phenotype [73]. A subsequent sequencing study revealed an association of the rs11701 polymorphism in *ANG* with ALS and a novel mutation, K40I, in two patients. Two years later, the same authors published results showing that *ANG* mutations segregate with ALS in families, and are also observed in ALS patients with the 'sporadic' form of the disease [46]. They identified seven novel missense mutations in patients of European descent and showed common haplotypes for the K17I and K40I mutations in Irish and Scottish patients, suggesting common founders for the mutations.

The authors also demonstrated that the observed mutations predict loss of RNAse and angiogenic function.

Since the observations of Greenway *et al.*, however, the role of *ANG* in ALS has been the subject of some debate. Initial failure to replicate the findings in Italian cohorts [74,75] was countered by an Italian study [76] that demonstrated an *ANG* mutation in a patient that was absent in 332 controls, and a later study that showed several mutations in a large Italian ALS cohort [77]. *ANG* mutations have also been observed in French [78,79], German [80] and American [81] patients, but replication in further populations has not been demonstrated. Despite the uncertainty around the association of *ANG* mutations with ALS, the product of the *ANG* gene, angiogenin, has been shown to be an important neurodevelopmental protein with neuroprotective properties [82] and *ANG* mutations lead to loss of function in angiogenin [81,83], implicating the gene further in the pathogenesis of ALS.

Angiogenin is the 14.1 kD product of *ANG* [84] and was originally discovered as a result of its properties as a potent inducer of neovascularization [85]. It also functions as a ribonuclease [86] and it is upregulated in response to hypoxia [87,88]. It is functionally similar to vascular endothelial growth factor (VEGF) [72], for which ALS risk promoter haplotypes have been described in European ALS populations [89]. $Vegf^{\delta/\delta}$ mice show an adult-onset ALS phenotype [73] and when G93A *SOD1* ALS model rats are treated with intracerebroventricular Vegf, disease onset is prolonged [90]. Combined evidence from animal models suggests that VEGF isoforms have a neuromodulatory and neuroprotective role in the CNS [91].

Because altered regulation of *VEGF* is linked to ALS disease pathology, study of the possible altered regulation of *ANG* and angiogenin in ALS could be fruitful. Serum angiogenin levels have been shown to differ in ALS compared to controls in the Irish population [92]. However, the patterns of plasma and cerebrospinal fluid (CSF) angiogenin

expression have not previously been investigated, and there have been no studies to determine whether *ANG* haplotypes modulate protein expression, as is the case with VEGF. Further study is needed to delineate the role of angiogenin and genetic variation at the *ANG* locus in ALS.

2.1.1 Research aims

This chapter details work carried out to investigate a number of unknown factors related to the role of angiogenin in ALS. Using serum, plasma, DNA and CSF samples, the aim was to investigate how genetic variation at the *ANG* locus relates to expression of the angiogenin protein, both in ALS patients and in neurologically normal controls. This also afforded the opportunity to examine the relationship between plasma and CSF angiogenin levels, as well as re-evaluating the previously published finding of altered expression of angiogenin in ALS patients [92].

2.2 Methods

2.2.1 Sampling

In total, 859 ALS patients and 1,047 unrelated control subjects with no family history of ALS participated in the study. 467 participants were from Ireland, 763 were from Sweden and 676 were from Poland. DNA and serum samples were drawn from Irish and Polish participants; DNA, plasma and cerebrospinal fluid (CSF) samples were drawn from Swedish participants. The mean age of onset ($\pm$ SD) for ALS patients was 57.7 $\pm$ 12.9 years; the mean age of controls was 56.3 $\pm$ 14.7 years. The numbers of participants in each study group and their demographics are detailed in figure 2.1. All patients fulfilled the El Escorial criteria for clinially definite or probable ALS [28] and informed written consent was obtained from all participants. ALS patients with atypical phenotypes and Swedish ALS patients with mutations in *SOD1* were excluded from analysis.

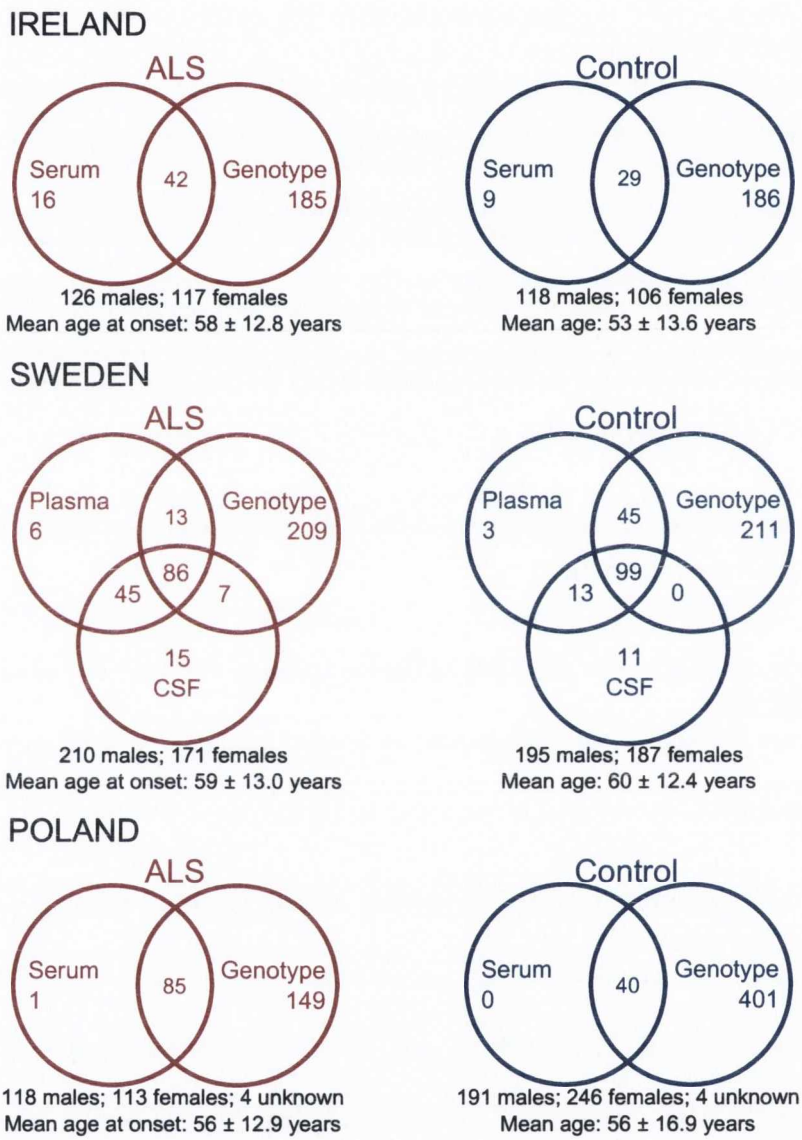


Figure 2.1: Numbers of individuals and demographics of the three study populations. Error values for mean ages represent standard deviation.

For Irish serum samples, approximately 4.5 ml of whole blood was drawn into a 4.9 ml Monovette® serum separation tube (Sarstedt, Nümbrecht, Germany) and immediately centrifuged at 3,000 RPM for 10 minutes. Approximately 1.5 ml of supernatant was then stored at -80°C until assay. Swedish plasma samples and Polish serum samples were drawn and extracted using similar methods by collaborators in Sweden and Poland. Swedish collaborators also provided CSF samples drawn by spinal tap. For Irish DNA samples, approximately 5 ml of whole blood was drawn into two 2.7 ml Monovette® tubes (Sarstedt) supplemented with ethylenediaminetetraacetic acid (EDTA) and DNA was extracted commercially by KBioscience (Herts, UK).

2.2.2 SNP genotyping

Using data from the CEPH panel of the International Hapmap Project data [5], 5 informative single nucleotide polymorphism (SNP)s with inter-marker $r^2 < 0.8$ and MAF $> 5\%$ were chosen for genotyping across the *ANG* locus, capturing the majority of genetic variation across the locus (assessed using the *Tagger* tool implemented within Haploview [93]). DNA samples from Ireland, Sweden and Poland were quantified using a Nanodrop ND-1000 (Thermo Fisher Scientific) and a minimum of 5 ng of DNA was provided for genotyping at KBioscience (Herts, UK) using competitive allele-specific PCR-based KASPar assays. Standard quality control checks were performed after genotyping (genotypes formed three distinct clusters, water controls were negative, call rate was greater than 90%).

2.2.3 Quantification of angiogenin in CSF, plasma and serum

Angiogenin concentrations in serum, plasma and CSF were measured by enzyme-linked immunosorbent assay (ELISA) according to manufacturer's guidelines (Quantikine Duoset, R&D Systems, Abingdon, UK) by collaborators in University of Limerick. All samples were assayed in duplicate and calibrated against serially diluted standards of known mass.

Pooled CSF and plasma quality control samples were both assayed in duplicate on each mitrotiter plate, allowing for estimation of the precision of the assay across all microtitre plates.

2.2.4 Statistical analysis

Unless otherwise stated, all statistical analyses were performed using the R statistical programming environment [94]. Assessment of allele frequencies and calculation of association statistics were conducted using the computer programmes Haploview [93] and PLINK [95]. Allelic association statistics were calculated using the χ^2 test, with the multiple testing issue being addressed by replication in the three populations and also using the permutation algorithm implemented within Haploview. Haplotype blocks were defined as a group of SNPs whose upper 95% confidence bound for D′ exceeded 98% with the lower bound above 70% [96] and a haplotype was examined if it occurred in more than 1% of individuals. Haplotypes were tested for association with ALS risk using the χ^2 test.

Because angiogenin has not been shown to have any binding partners in the blood, plasma and serum angiogenin levels were deemed comparable. This was confirmed by comparing the levels derived from the two blood components using the Mann-Whitney-Wilcoxon test.

A pipeline was developed using the R programming environment to analyse, quality control and visualise the ELISA and SNP data together (`LvGPlot.R`; see appendix B). Data for angiogenin levels were first assessed for the reported influence of age and sex [97]. Using data pooled from cases and controls in all three populations, angiogenin levels were regressed against age and sex and an outlier was identified and removed if its studentized residual exceeded the critical t statistic for the group's Bonferroni-corrected 5% significance threshold. This regression analysis was then re-iterated until no further outliers could be identified. Four Swedish plasma values and four Swedish CSF values were removed this

way.

The resulting linear models from the regression analyses were used to adjust the values in the respective groups based on age and sex. The influences of genotypes across the five SNPs were then assessed by analysis of variance (ANOVA) for each SNP and the differences between case and control angiogenin levels for each genotype were assessed for statistical significance using the Mann-Whitney-Wilcoxon test. Finally, using data from the Swedish population, corrected plasma angiogenin levels were assessed for correlation with corrected CSF angiogenin levels in ALS patients and in controls independently.

2.3 Results

2.3.1 *ANG* SNP and haplotype association

The mean genotyping call rate across all SNPs in the three populations was 98.4%. No SNP deviated significantly from Hardy-Weinberg equilibrium in any study population ($p > 0.01$ for all SNPs in all populations). Table 2.1 shows the results for the allelic association tests for the five SNPs, and linkage disequilibrium (LD) between SNPs is shown in figure 2.2. All five SNPs showed association with risk for ALS in the Irish study group, with one SNP, rs17114699, replicating in the Swedish population ($p_{Irish} = 0.03$; $p_{Swedish} = 0.001$). No SNP showed association in the Polish population. A haplotype block was identified in all three populations, incorporating SNPs rs9322855, rs8004382 and rs4470055 (figure 2.2). The AAG and CGA haplotypes at these three SNPs associated with ALS in the Irish data, while the AGG haplotype showed strong association with ALS in the Swedish data (table 2.2).

2.3.2 Plasma, serum and CSF angiogenin levels

Collaborators reported an inter-assay coefficient of variation of 6% and 8% for the high and low plasma quality control. An inter-assay coefficient of variation of 9% was obtained

Table 2.1: Allele frequencies and SNP association statistics in the *ANG* study populations

SNP	Alleles	IRELAND				SWEDEN				POLAND			
		RA	RA freq ALS;ctrl	Allelic association p	OR	RA	RA freq ALS;ctrl	Allelic association p	OR	RA	RA freq ALS;ctrl	Allelic association p	OR
rs9322855	A>C	C	0.50; 0.41	0.003*	1.57	A	0.56; 0.52	0.13	0.85	A	0.55; 0.55	0.92	0.99
rs8004382	G>A	G	0.57; 0.47	0.007*	1.50	G	0.55; 0.52	0.46	0.92	G	0.55; 0.52	0.46	0.92
rs4470055	G>A	A	0.29; 0.22	0.03*	1.47	A	0.25; 0.24	0.66	1.06	G	0.75; 0.72	0.36	0.88
rs17114699	G>T	T	0.16; 0.11	0.03*	1.53	T	0.14; 0.08	0.001*	1.78	G	0.89; 0.87	0.68	0.93
rs11701	T>G	G	0.18; 0.10	0.006*	1.88	G	0.13; 0.13	0.69	1.07	G	0.13; 0.10	0.14	1.3

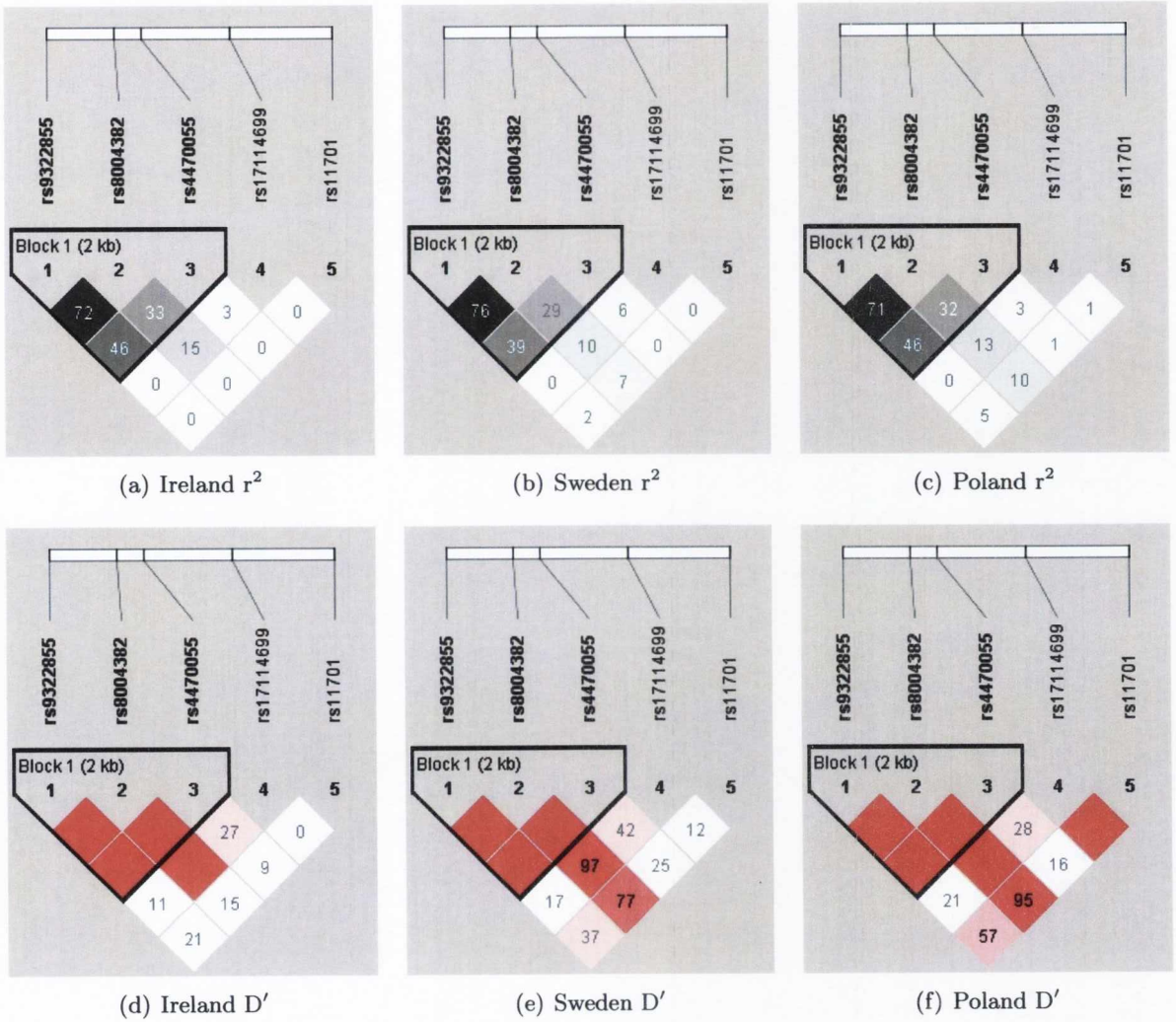
RA, risk allele; OR, odds ratio

* Significant p-value

Table 2.2: Haplotype frequencies and association statistics in the *ANG* study populations

Haplotype	IRELAND			SWEDEN			POLAND		
	Freq (ALS;ctrl)	p	Permuted p	Freq (ALS;ctrl)	p	Permuted p	Freq (ALS;ctrl)	p	Permuted p
AAG	0.45; 0.53	0.024*	0.13	0.46; 0.47	0.64	0.99	0.453; 0.474	0.4498	0.95
CGA	0.29; 0.22	0.023*	0.12	0.25; 0.25	0.68	1.00	0.256; 0.279	0.3627	0.90
CGG	0.18; 0.16	0.43	0.94	0.18; 0.24	0.027	0.16	0.193; 0.172	0.3424	0.88
AGG	0.07; 0.08	0.55	0.98	0.097; 0.045	<0.0001*	0.0006*	0.099; 0.075	0.1311	0.51

* Significant p-value



(a) Ireland r^2 (b) Sweden r^2 (c) Poland r^2

(d) Ireland D′ (e) Sweden D′ (f) Poland D′

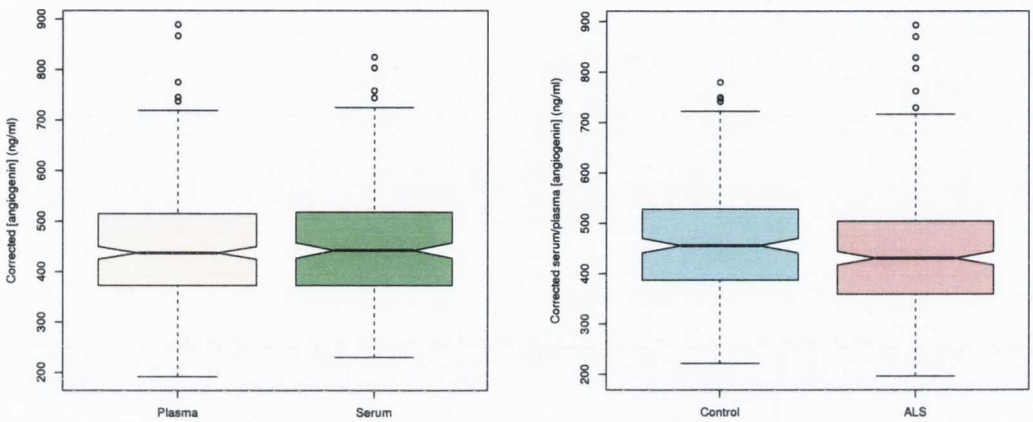
Figure 2.2: Linkage disequilibrium between the five *ANG* SNPs in the three populations, generated using Haploview [93]. A haplotype block was identified in all three populations, represented by the black line surrounding SNPs rs9322855, rs8004382 and rs4470055.

for CSF. In the initial regression analysis, age and sex both had a significant effect on angiogenin levels in plasma/serum and in CSF ($P(>|t|) < 0.0001$ for all covariates); the coefficients from the linear models were used to adjust the values for angiogenin levels. When corrected plasma and serum angiogenin levels were compared, they did not differ significantly (figure 2.3(a); $p = 0.93$, Mann-Whitney-Wilcoxon test), demonstrating that data could be pooled irrespective of the blood component from which the data were derived. Using data pooled from the three populations and after correcting for age and sex, angiogenin levels were significantly lower in ALS patients than in controls in plasma/serum (figure 2.3(b); mean $\pm$ SD = 438.2 $\pm$ 112.2 ng/ml for the ALS group and 467.6 $\pm$ 105.4 ng/ml for controls; $p = 0.001$, Mann-Whitney-Wilcoxon test) and in CSF (mean $\pm$ SD = 5.582 $\pm$ 1.754 ng/ml for the ALS group and 6.197 $\pm$ 1.987 ng/ml for controls; $p = 0.01$, Mann-Whitney-Wilcoxon test).

There was a significant positive correlation ($p < 0.0001$, Pearson product-moment correlation) between corrected CSF angiogenin levels and corrected plasma angiogenin levels in controls, whereas in ALS patients ($p = 0.21$) this was not statistically distinguishable from a correlation of 0 (figure 2.4; $r^2_{control} = 0.13$, $r^2_{ALS} = 0.011$)

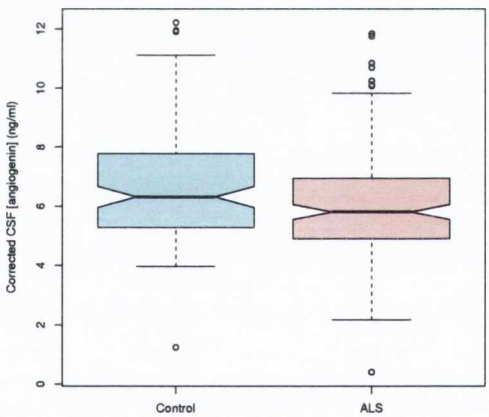
2.3.3 Contribution of SNP genotypes to angiogenin levels

In the initial regression analyses to correct for age and sex, angiogenin levels were found to vary considerably around the fitted models (multiple $r^2_{serum/plasma} = 0.074$; multiple $r^2_{CSF} = 0.16$). ANOVA was used to assess the contribution of genotype at each SNP to the overall variance in the data and the Mann-Whitney-Wilcoxon test was used to assess the differences between corrected plasma/serum levels in ALS patients and controls for each SNP, separated by genotype. Data were analysed both as independent populations and also as a pooled dataset. The results of these tests, along with the group means, are reported in figure 2.5.



(a)

(b)



(c)

Figure 2.3: Notched boxplots showing differences in angiogenin levels. (a) Corrected angiogenin levels in Swedish plasma samples compared with pooled Irish and Polish serum samples (b) Corrected serum/plasma angiogenin levels in controls compared with ALS patients. (c) Corrected CSF angiogenin levels in controls compared with ALS patients.

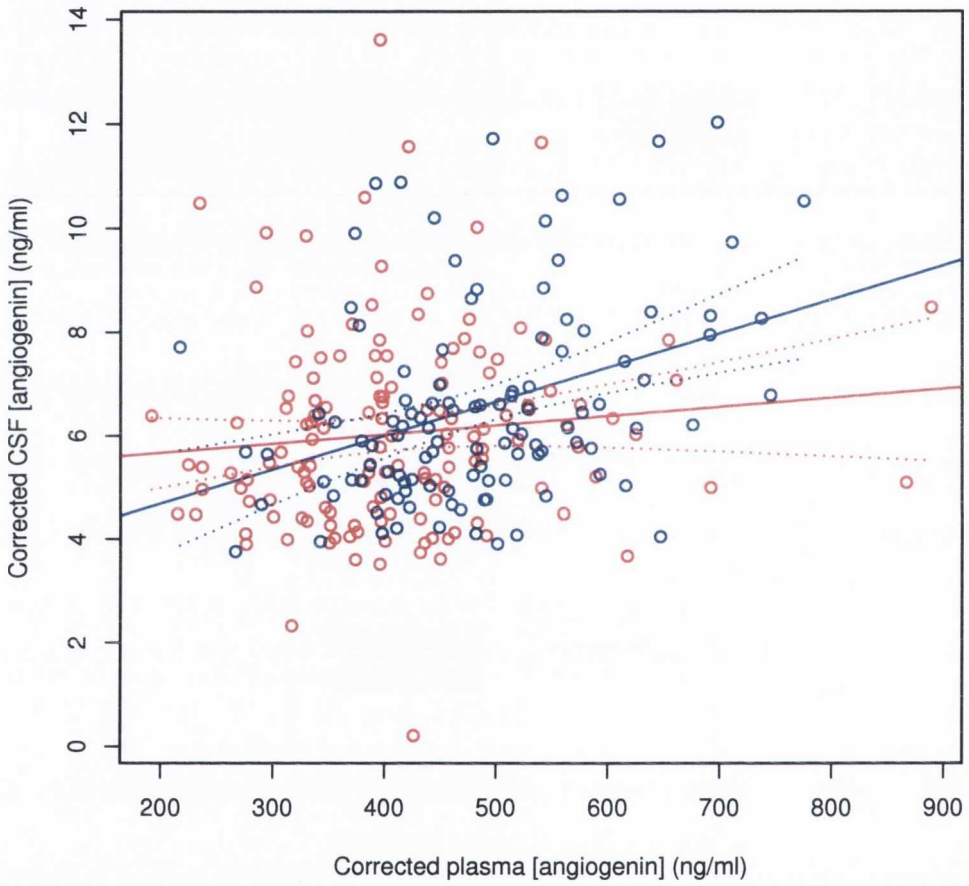


Figure 2.4: Correlation of CSF angiogenin levels with serum angiogenin levels in the Swedish population. ALS patients are shown in red and controls are shown in blue. Dashed lines indicate 95% confidence intervals of the regression lines. r^2 values are: controls, 0.13; ALS, 0.011.

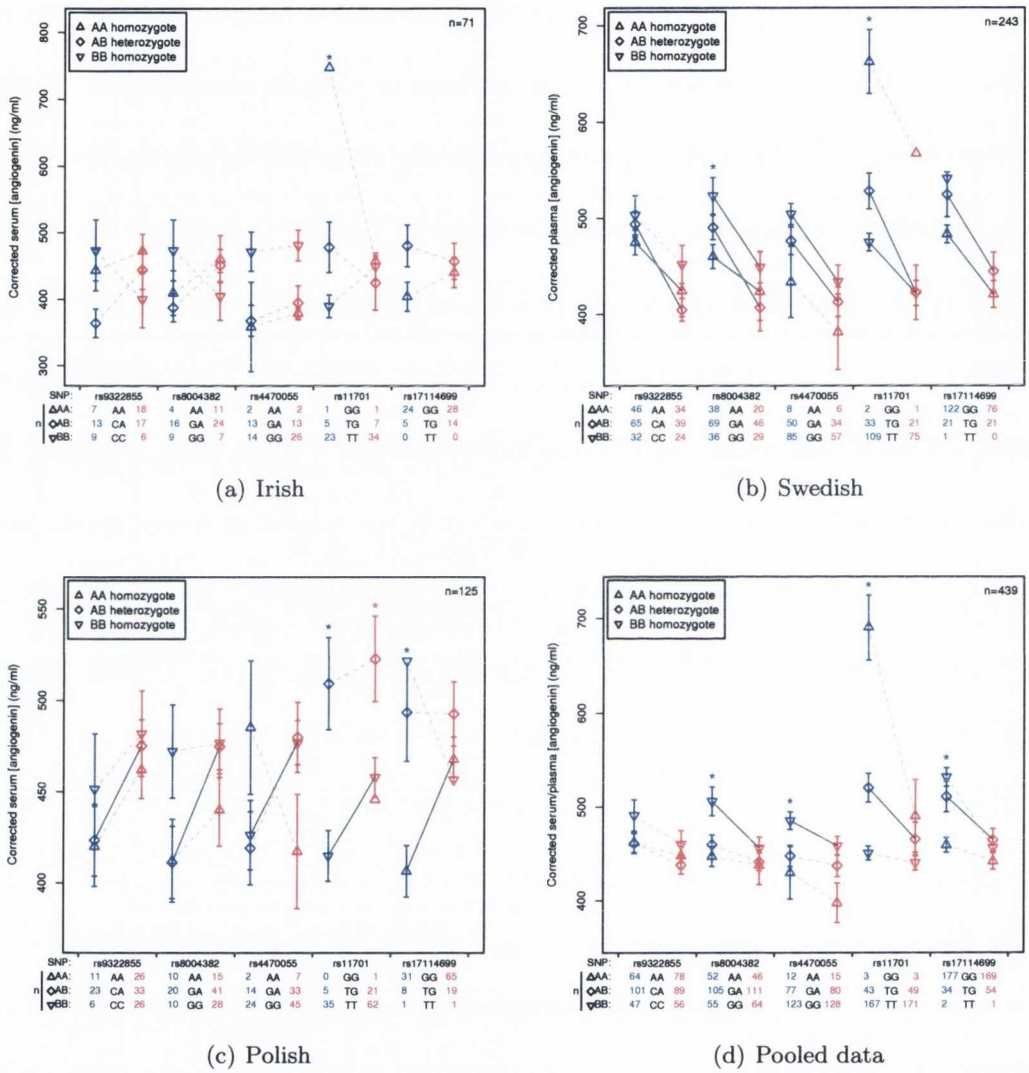
In the large Swedish dataset, an allele dose-dependent regulation of plasma angiogenin was readily observable for all SNPs in controls and perturbation of this pattern was seen in ALS patients at SNPs rs9322855, rs8004382 and rs11701. These findings are reflected in the pooled dataset. In the pooled data, allele dose-determined angiogenin levels for ALS patients were not consistent with controls for SNP rs17114699. Only at SNP rs11701 was a significant contribution of genotype to the variance in controls observable in all three populations; however, in the pooled dataset genotypes at every SNP except rs9322855 were shown to contribute significantly to variance in controls. No SNP contributed significantly to variance in ALS patients in any dataset, wih the exception of SNP rs11701 in the Polish dataset.

Figure 2.6 shows the same analysis applied to CSF levels in Swedish patients. The allele dose-dependent relationship between *ANG* SNP genotypes and angiogenin levels in CSF was not as readily observable as with plasma angiogenin. The same patterns as plasma were observed in CSF at SNPs rs8004382, rs11701 and rs17114699, however genotypes were not shown to contribute significantly to variance in levels for any of these SNPs. Dysregulation of CSF angiogenin levels was observed for the majority of the SNPs in ALS cases, with very little variation in the mean CSF angiogenin level for any SNP genotype.

2.4 Discussion

This study set out to identify the contribution of genetic variation across the *ANG* locus to levels of angiogenin detected in serum, plasma and CSF. It also assessed the relationship between serum/plasma angiogenin and CSF angiogenin, and has highlighted some discrepancies with the extant literature on the level of expression of angiogenin in ALS patients compared to controls.

Section 2.3.1 confirms the previously observed association between *ANG* polymor-



(a) Irish

(b) Swedish

(c) Polish

(d) Pooled data

Figure 2.5: Mean corrected serum or plasma angiogenin concentrations as a function of *ANG* SNP genotypes. ALS patients are shown in red and controls are shown in blue. Significant differences between ALS patients and controls are denoted by solid lines and significant F-statistics within groups are denoted by asterisks. Error bars are standard error of the mean. Numbers of observations for each genotype at each SNP are indicated in the table below each plot.

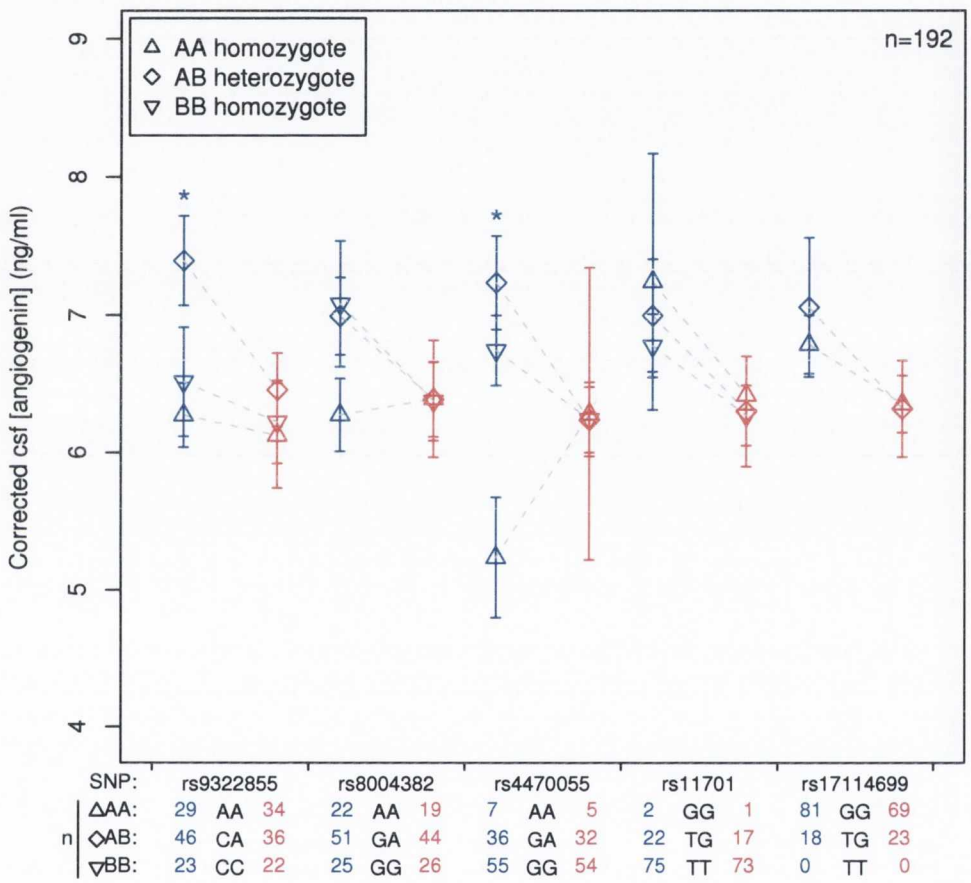


Figure 2.6: Mean corrected CSF angiogenin in the Swedish population as a function of *ANG* SNP genotypes. ALS patients are shown in red and controls are shown in blue. Per SNP genotype, differences between cases and controls were not statistically significant. Significant F-statistics within SNP genotype groups are denoted by asterisks. Error bars are standard error of the mean. Numbers of observations for each genotype at each SNP are indicated in the table below each plot.

phisms and ALS in the Irish population [46], with 5 SNPs across the *ANG* gene showing association with ALS. One SNP, rs1711699, replicated in the Swedish cohort, showing strong association with ALS risk ($p = 0.001$). It has also been shown that two *ANG* haplotypes in the Irish and one in the Swedish associate with ALS, with the AAG haplotype being protective in the Irish population, and the remainder of associated haplotypes being associated with increased risk for ALS. The combined SNP and haplotype association results add strength to the argument that *ANG* is implicated in the pathogenesis of sporadic ALS.

Although replication in the Swedish population improves the argument for confidence in the Irish findings, no SNP or haplotype associated with ALS in the Polish population. Similarly, in a recent screen for replication of findings from the Irish genome-wide association study for ALS risk [98] using a Polish dataset, the results were surprisingly uninformative [99]. The failure to replicate in the Polish population may reflect true population-based differences, with the aetiology of ALS in Poland being explained by a different set of genetic factors with different founders, or, at least, the absence of a founder for *ANG*-associated ALS. Thus, the complex genetics of ALS may be different between European populations, which has been suggested previously by van Es *et al.* [100], and the Swedish population may represent a better replication population for discoveries in Irish ALS patients than the Polish.

Figure 2.5 (notably parts a and c) demonstrates the need for large datasets when analysing data that vary so substantially by chance; this is also exemplified by the large spread of data in figure 2.4. However, the large Swedish dataset permits a reasonable estimate of the contribution of SNP genotypes to angiogenin levels, and it is evident that angiogenin expression in plasma is allele dose-dependent for SNPs across the *ANG* locus. This is also noted in data pooled from the three populations, although a caveat here is that if the arguments made about population-specific differences are correct, then the pooled

data would not be as representative for disease-driven phenomena as data that exclude the Polish levels. Nevertheless, the effect is indeed seen in the pooled data.

The relationship between angiogenin levels and allele dose that was observed in controls was not consistent in ALS patients. In the majority of cases, angiogenin levels were lower in ALS patients than in controls, regardless of SNP genotype. Moreover, SNP genotypes did not significantly contribute to the variance in the level of angiogenin in ALS patients, as determined by analysis of variance (with the exception of SNP genotypes for rs11701 in the Polish dataset), suggesting that the regulation of angiogenin levels by genetic variation across the *ANG* locus is not present in ALS cases.

The previous finding by Cronin *et al.* that serum angiogenin levels are elevated in ALS patients [92] was not replicated in this study. In fact, it was found that corrected serum/plasma angiogenin were significantly lower in ALS patients than in controls ($p < 0.001$, figure 2.3(b)). This discrepancy is most likely to be due to the statistical approach implemented when analyzing the data. In this study, a statistically robust estimate of the influence of age and sex on angiogenin levels was made using the complete dataset of values, totalling 532 individuals. Using the coefficients derived from the linear model applied to the data, the same correction factors were applied to both case and control data, under the assumption that age and sex determination of angiogenin levels would not be altered by disease status. Conversely, Cronin *et al.* made independent estimates in cases and controls, fitting separate models to the case cohort of just 79 patients and the control cohort of just 72 individuals, and correcting the levels thereafter. For data that vary so substantially around the fitted models, the better method is likely to be the one which used a greater number of values to estimate the regression coefficients. Indeed, applying the methods of Cronin *et al.* to this dataset, fitting different models to cases and controls, results in a significantly higher mean corrected angiogenin level in ALS patients than in controls ($p < 0.0001$), however, this is driven by the difference in the age and sex

coefficient estimates and the consequent different levels of correction applied to the two cohorts.

In neurologically normal controls, plasma angiogenin concentration weakly, but significantly, predicts CSF angiogenin concentration ($p < 0.0001$, $r^2 = 0.13$, figure 2.4). When corrected CSF angiogenin level is regressed against corrected plasma angiogenin level for ALS cases, however, the observed correlation ($r^2 = 0.011$) is not statistically distinguishable from a correlation of 0 ($p = 0.21$). This may suggest a tissue-specific dysregulation of angiogenin expression in ALS. This could be due to a number of factors, including perturbation of angiogenin transport in ALS, however an interesting possibility could be micro RNA (miRNA) regulation of angiogenin expression. Altered miRNA regulation of progranulin has been reported recently in frontotemporal dementia [101]. As progranulin is functionally similar to angiogenin, and there is significant clinical overlap between frontotemporal dementia and ALS [102], a similar form of altered regulation of angiogenin may apply in ALS. A search of the EBI's miRBase Sequence Database [103] using the online Microcosm web application reveals 19 potential miRNA binding sites in the *ANG* gene for 24 human miRNAs, some of which may be preferentially expressed in the central nervous system [104]. This suggests a possible mechanism for the observed tissue-specific differences indicating that further investigation of miRNA regulation of angiogenin could be fruitful.

In summary, this work has confirmed that *ANG* variants associate with ALS in the Irish and also in the Swedish. It also demonstrates that angiogenin expression is modulated by genetic variation across the *ANG* gene in an allele-dose dependent manner, and that this regulation is disrupted in ALS patients. The finding that plasma angiogenin level does not predict CSF angiogenin level in ALS patients suggests a tissue-specific regulation of angiogenin that may be perturbed by genetic or phenotypic variation in ALS. Cumulatively, the results suggest dysregulation of angiogenin in ALS. Given that it is un-

likely that the differences in levels are entirely driven by disease-associated mutations in *ANG*, the mechanisms behind the dysregulation probably lie within biochemical signalling pathways and regulatory networks to which angiogenin and *ANG* belong. Further study geared towards elucidating the complexities of regulation of angiogenin may yield further information about the pathogenesis of ALS and consequent potential for therapeutic intervention.

Chapter 3

Genome-wide SNP analysis in ALS

3.1 Introduction

There have been several ALS GWAS attempts carried out in multiple populations, with multifarious findings. An early GWAS was carried out by Schymick *et al.* in 2007 on 276 sporadic ALS patients and 271 controls in the USA [105]. The authors found that no SNP was associated with ALS at a level of significance that was strong enough to draw firm conclusions; however they reported an over-representation of genes associated with regulation of the actin cytoskeleton. As a commentary on the size of the study and the power to detect significant associations, the authors stressed the need for replication of their findings, a sentiment that would come to be reflected in several subsequent GWAS. A second GWAS from the USA followed shortly [106], which analyzed genome-wide SNPs in a discovery set of 386 ALS patients and 542 controls followed by genotyping of the top 384 SNPs in a replication panel of 766 patients and 750 controls. Top findings were then further assessed by re-analysis of the dataset of Schymick *et al.* [105]. Interestingly, the authors used a pooled genotyping approach, in which they performed only two genotyping

experiments for the full set of 386 patients and only one for the 542 controls, where each experiment was performed on a sample containing DNA mixed from many individuals. They subsequently used MAF estimates derived from the genotyping data to ascertain association statistics. A single SNP was associated at genome-wide significance (see section 3.1.1), mapping to the uncharacterized gene *FLJ10986*, which the authors demonstrated to be detectable in CSF.

A three-stage GWAS came from The Netherlands soon after the American studies [107], which identified associations with ALS in a panel of 461 patients and 450 controls, following up the 500 most associated SNPs in 291 Belgian and 272 further Dutch cases and 267 Belgian and 336 Dutch controls. The third stage was to genotype the 17 most associated SNPs from stage 2 in a Swedish population of 313 cases and 303 controls. The reported best result from this study was an association for a SNP lying within *ITPR2*, and when the authors assessed the expression levels of *ITPR2* mRNA in blood cells, it was found to be lower in ALS cases than in controls.

In 2008, Cronin *et al.* published the results of a GWAS of sporadic ALS in 221 Irish cases and 211 controls [98]. The authors used the first USA GWAS [105] and Dutch phase 1 GWAS [107] as replication datasets. In the study, no SNP associated with ALS at genome-wide significance in the Irish population alone or when the data from Ireland, the USA and the Netherlands were pooled. However, the top association of a SNP in *DPP6* was reported, a finding which was also published in almost the same dataset at roughly the same time by van Es *et al.* [108]. *DPP6* was later also identified as a variant in a Dutch study on copy number variation in ALS [109]. However, in a screen for replication of Cronin *et al.*'s GWAS findings in a Polish dataset [99], the *DPP6* result was not replicated; the authors suggested population-specific differences in disease allele frequencies. Replication has since also been troublesome in other populations [110–113].

Chio *et al.* cast further doubt on the realness of the *DPP6* associations, as well as the

ITPR2 findings of van Es *et al.* [113], by showing lack of replication of these findings in a two-stage Italian GWAS in 277 cases and 1,510 controls, supplemented with the 276 ALS cases from Schymick *et al.* [105] and 828 controls from the USA. The second stage of this GWAS focussed on 7,600 top SNPs from the first stage genotyped in a further 2,160 cases and 3,008 controls, all of European descent. The authors cited the possibility that the causative variants at these loci may not be in strong LD with the associated SNPs in the previous GWAS, therefore leading to false refutation of these loci in the Italian study, but concluded that the findings generally point towards greater genetic heterogeneity in the disease than previously anticipated.

Two strong association signals were observed in a GWAS by Laaksovirta *et al.* in the Finnish population, mapping to chromosome 21q22 (driven by the *SOD1* D90A allele) and the chromosome 9p locus previously identified in familial linkage studies [114–118]. Chromosome 9p21 was also identified in two large international GWAS [66,67], along with *UNC13A*, which has also been shown to correlate with a short disease duration [119]. These three studies [66,67,120] represent the most successful GWAS in ALS to date; this is probably mostly due to the large sample sizes involved, and in the case of the Finnish study, because of the very large proportion of the ALS in this population attributable to either *SOD1* or *C9orf72*, the gene later discovered to be responsible for the signal at chromosome 9p21 [121,122].

The findings of the many GWAS efforts in ALS have, however, demonstrated that ALS is likely to be a genetically heterogenous disease, and discovery of disease-associated SNPs by GWAS requires very large sample sizes. However, the difficulty in generating replicable associations is neither an indication that the disease does not have a large genetic component, nor is it a refutation of the common disease-common variant hypothesis. Indeed, the twin study-based heritability estimate of 0.61 for ALS [41] suggests that efforts to search the genome for loci involved in ALS aetiology should continue. This could be acheived

through further manipulation of GWAS datasets, but for the smaller studies it may require the focus of researchers to shift from the canopy of the strongest association signals towards the undergrowth of more moderately associated signals, so that true associations are not missed in, for example, two-stage designs that only consider top results in the first stage. Additionally, other patterns may exist in the SNP data that are uninterrogable by traditional GWAS design. Therefore, further generation and analysis of genome-wide SNP datasets is warranted. There are many considerations in the design of GWAS and the analysis of such data, the majority of which are based on statistical factors that become apparent when analyzing such large volumes of data.

3.1.1 Statistical considerations in GWAS design and analysis

With genome-wide SNP datasets, the contribution of genetic variation to disease aetiology can be assessed using a statistic as straightforward as the χ^2 test with one degree of freedom. However, given that there would be a non-trivial amount of chance variation within the data as well as systematic (disease-driven) variation, a high false positive rate would become a problem as a consequence of the very large number of independent statistical tests that are performed on the dataset (say, 500,000). Therefore, traditional p-value cut-offs determining significance are too high; with $\alpha = 0.05$ roughly 5% of the data (25,000 SNPs) would be associated by chance and with $\alpha = 0.01$ approximately 1% of the data (5,000 SNPs) would be associated by chance.

The most popular way to defend against such high numbers of type I errors is to define a Bonferroni-corrected p-value threshold, which is simply calculated by dividing α by the number of independent statistical tests that were performed. Therefore, for a dataset of 500,000 SNPs and an uncorrected p-value threshold of 0.05 a Bonferroni-corrected p-value threshold would be $\alpha = 1 \times 10^{-7}$. This, however, introduces a new problem: statistical power. In order to acheive power to detect such associations, very large sample sizes

are generally needed. To put this in context, the power of the 2008 Irish GWAS [98] to detect an association of a medium-frequency (25%) variant with ALS at genome-wide significance, assuming full penetrance under a multiplicative model and a modest genotype relative risk of 1.5, was 2% (assessed using the CaTS Power Calculator [123]).

Despite stringent significance thresholds being set to reduce the possibility of false positives, rigorous quality control of a genome-wide SNP dataset is required to eliminate the potential for spurious associations that can arise as a consequence of systematic biases present in the data. This is usually acheived through exclusion of SNPs that do not meet criteria such as Hardy-Weinberg equilibrium [124, 125], systematic missingness (for example, more missing genotypes in controls than in cases) and allele frequency. In a population-based GWAS, the presence of cryptically related individuals in the case or control cohort could influence association statistics also, as related individuals are likely to share more of their genotypes by descent from a common ancestor, and such individuals should be removed.

Another important factor to control with a GWAS is the ancestry of the genotyped individuals. In a case where more than one population is present in the case-control cohort, if the populations are not carefully balanced between cases and controls, this can lead to spurious associations through population stratification (figure 3.1). A good method for controlling against this is to select only individuals derived from the same population for genotyping, although the ancestry of individuals within a genotyped cohort can be checked using ancestry-informative markers compared against reference panels (for example, HapMap individuals) or by performing principal components analysis on genome-wide SNP markers [126], and outliers can subsequently be removed.

Although the power to detect associations with disease is low with a small genome-wide SNP dataset, a GWAS is not the only application of such data for identification of regions that may be linked to disease aetiology. Alternative methods, such as analysis of

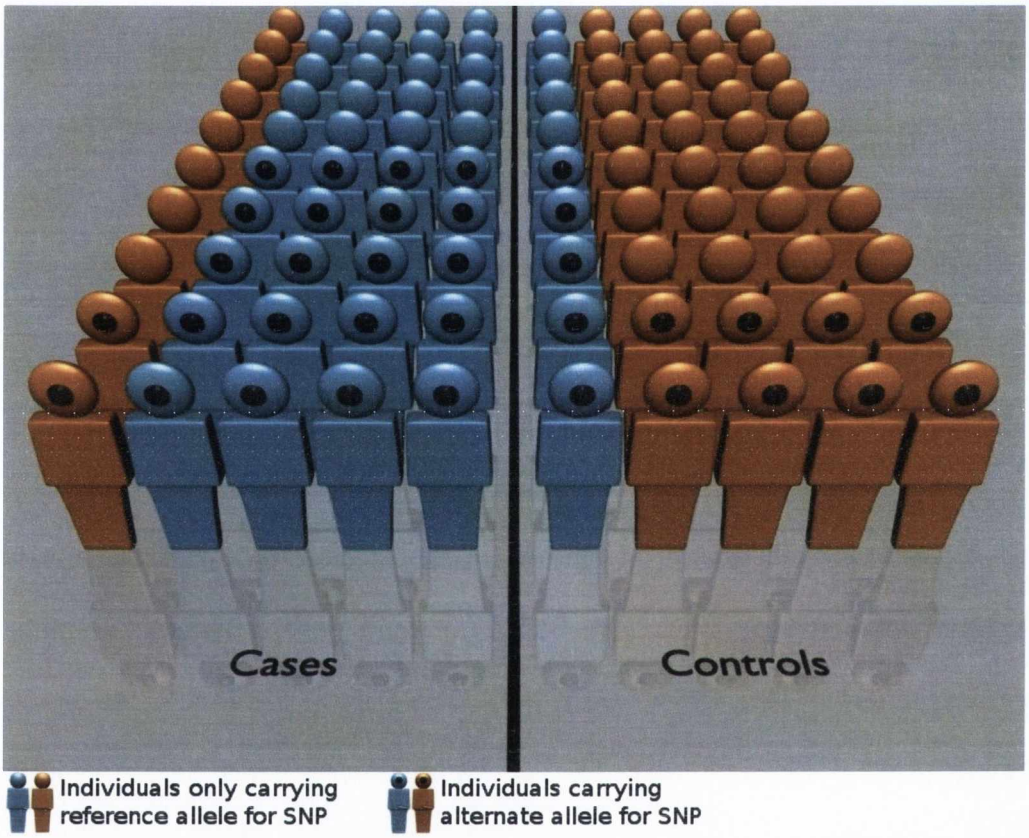


Figure 3.1: An illustration of an extreme example of population stratification in a case-control association study. In the orange population, 20% of individuals carry the alternate allele for the SNP of interest, whereas in the blue population it is present in 50%. In this situation, where there is an imbalanced assignment of the populations to the case and control cohorts, the odds ratio for having disease if the individual is a carrier of the minor allele of the SNP is $\frac{22/28}{13/37} = 2.236$, with a χ^2 statistic of $12.24, p = 4.68 \times 10^{-4}$. However, this strong association is being driven entirely by the population-differentiated nature of the SNP. Within either population, there is no association of the SNP with the disease (for both populations, $p = 1$.)

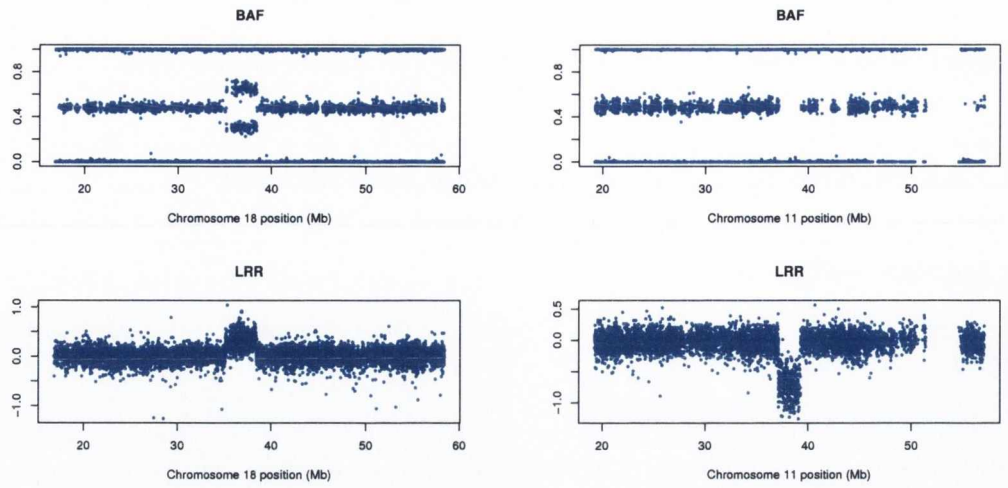
copy number variants (CNVs) or mapping of homozygous segments of the genome, have been shown to be successful in mapping disease loci in a number of studies.

3.1.2 Alternative uses of a genome-wide SNP dataset

Copy number variation

Copy number variation describes the situation where individuals differ in their number of copies of certain chromosomal segments. Whereas most parts of the human genome should be diploid, any individual could posess genomic segments that have either been duplicated or deleted, resulting in that individual posessing more or less of the genetic material corresponding to the segments. It is suggested that microhomology-mediated break-induced replication of genomic segments provides a mechanistic explanation for the phenomenon [127], which is an important source of genetic variation, accounting for up to 12% of possible human genetic variation [128] and even existing within individuals [129] and between monozygotic twins [130, 131]. Copy number variation is therefore an important possibility to consider when investigating the genetics of complex diseases.

Gold standard technologies used to map CNVs in the genome include array comparative genomic hybridization (aCGH) and quantitative polymerase chain reaction (qPCR). However, CNVs can also be inferred using data generated during SNP genotyping on a genome-wide SNP array. Such techniques typically make use of two values that can be generated from such data: the log R ratio (LRR), which is a normalized measure of the overall intensity of the signal measured from the SNP array (per SNP), and the B allele frequency (BAF), a normalized measure of the relative intensity of the 'B' allele's signal (per SNP) when the beadchip is scanned. CNVs can be identified in such data when long stretches of consecutive SNPs differ significantly from expected values (figure 3.2). Specifically, under copy-neutral conditions, assuming no variation in genome-wide intensity values, $LRR = 0$ and $BAF \in \{0, 0.5, 1\}$. In practice, there is a large amount of



(a) Copy number gain

(b) Copy number loss

Figure 3.2: Copy number variation in SNP intensity data. (a) A 2.8 Mb copy number gain in an ALS patient on chromsome 18. Each point represents a SNP, and the CNV lies in the middle of the two plots, marked by an overall increase in LRR and a separation of BAF from the expected {0,0.5,1} to {0,0.33,0.67,1}, representing the genotypes AAA,AAB,ABB and BBB. (b) A 2.1 Mb copy number deletion in a different ALS patient on chromosome 11. This time, the CNV is indicated by an overall decrease in LRR and a loss of the heterozygous state in BAF ($BAF \in \{0, 1\}$, representing hemizygous genotypes A and B only). Interestingly, two copy-neutral runs of homozygosity are also visible in this plot at around 41-42 and 43-44 Mb. The centromere is visible at around 52-55 Mb, demarcated by an absence of genotyped SNPs.

random variation in LRR and BAF across the genome (including systematic bias caused by local GC content) so algorithms designed to detect CNVs must be able to account for this in identifying true CNVs.

A paper which compared seven different methods for mapping CNVs using SNP intensity data [132] concluded that the best-performing algorithm assessed was the method implemented in QuantiSNP [133]. This uses an objective-Bayes (OB) hidden Markov model (HMM) to infer copy number variation and implements an expectation maximization (EM) algorithm to generate an associated statistic, the Bayes factor (reported by the algorithm as $\log(Bayes factor)$), which can be used to set a cutoff false positive rate for CNVs called by the algorithm. A similar method is PennCNV [134], which also uses a HMM to infer copy number variation but does not report quality statistics such as Bayes factor; instead the documentation recommends identifying outlying data *post hoc*.

Copy number variation has been assessed in ALS in the Irish population in two separate studies [109, 135]. In the first, Cronin *et al.* identified putative regions of copy number variation using QuantiSNP in 408 Irish and 868 Dutch individuals and found that although no CNVs associated significantly with ALS, a number of ALS-specific CNVs were observed in both the Irish and the Dutch datasets [135]. In a much larger multi-population study, Blauw *et al.* performed a genome-wide CNV association study using PennCNV with 1,875 cases and 8,731 controls, replicating top results in a cohort of 2,559 cases and 5,887 controls [109]. The main finding was an association of CNVs in *DPP6* with ALS; the authors also observed a number of rare CNVs specific to ALS.

In a large work that built up a comprehensive map of copy number variation across the genome [136], Conrad *et al.* concluded that common CNVs are unlikely to account for the missing heritability of complex genetic disorders that has remained following GWAS. This, however, does not speak for the potential contribution of rare or very rare CNVs to ALS susceptibility and so the assessment of copy number variation in ALS is still possibly a worthwile pursuit (although a conclusion of Blauw *et al.* was that 'rare CNVs with high effect size do not play a major role in ALS pathogenesis'). If neither common CNVs [136] nor rare CNVs [109] are going to be associated with ALS, there is little point in attempting per-CNV association tests like those performed by Cronin *et al.* [135] and Blauw *et al.* [109], and in most cases the statistical power to discover associations would suffer the same shortcomings as GWAS. It therefore may be more prudent in the design of any CNV experiment investigating the contribution of rare CNVs to ALS aetiology to search, for example, for ALS-specific copy number gains and deletions.

Runs of homozygosity

The labelling of non-familial ALS cases as sporadic is somewhat incompatible with the frequently stated fact that a large proportion of ALS cases should be explained by genetic

factors. One possible explanation for the seeming sporadic nature of many cases of ALS is that they are inherited recessively, meaning that individuals are required to inherit two copies of the disease allele in order to present with the disease. There is much less famiality observed with recessive traits, as affected members of a pedigree tend to segregate within a single generation; the disease seems to 'spread horizontally'. Even then, however, the disease may still seem sporadic in cases where only one sibling in a family is affected by virtue of the fact that each sibling only has a one-in-four chance of inheriting two copies of a disease allele from carrier parents. The argument that ALS may be inherited recessively in many cases is supported by evidence forwarded by Hemminki *et al.* [137], who observed higher risk for ALS between siblings with unaffected parents.

One method that can be used to search for recessive loci is homozygosity mapping [138]. This technique searches for portions of the genome that contain an unlikely number of consecutive markers that are homozygous; these are termed runs of homozygosity (ROHs). Homozygosity mapping successfully identified *OPTN* as an ALS gene from a screen for ROHs in six consanguineous families [53], and similar techniques have enjoyed recent success in a number of other conditions (six articles detailing homozygosity mapping in consanguineous families were published in November 2011 alone [139–144]). However, this technique requires the serendipitous finding of a pedigree (or many thereof) that has several affected members and a history of some kind of consanguinity, which is problematic. An alternative approach is to apply a homozygosity mapping technique on a population-based sample.

The idea of mapping homozygous stretches to identify recessive disease loci was suggested as early as 1987 [138], but the finding in 2006 that ROHs are common in the HapMap populations [145] popularized the technique of ROH mapping in genome-wide SNP datasets. In a larger study on non-HapMap European populations, McQuillan *et al.* [146] showed that ROHs are common in outbred populations, and the proportion of

the genome that lies within ROHs (F_{roh}) performs well at distinguishing populations, with a high correlation between F_{roh} and inbreeding estimates. Furthermore, Nalls *et al.* [147] demonstrated that younger generations have, on average, a lower proportion of their genome in ROHs than older generations, presumably as a consequence of increased mobilization within the human race causing a trend towards more cosmopolitan societies in modern times, resulting in increased panmixia and decreased consanguinity.

Population-based ROH mapping on unrelated individuals was applied by Lencz *et al.* [148] to identify several potential recessive loci for schizophrenia. Similar approaches were used by Nalls *et al.* [149] to map candidate genes associated with late-onset Alzheimer's disease and Hildebrandt *et al.* [150] described a ROH mapping technique that, when applied to a dataset of 72 mostly-unrelated individuals from outbred populations, correctly identified the loctions of known homozygous mutations.

Population-based ROH mapping has been shown by these studies to be a potentially powerful and fruitful technique. However, such studies rarely harness the rich abundance of extra information available to them: the actual genotypes being mapped as homozygous. This could help to determine whether ROHs identified as being linked to pathogenicity in multiple affected individuals are indeed identical-by-descent (IBD) segments of genome potentially harbouring recessive disease-causing mutations. Additionally, it could help to reduce false negatives at loci that have a high degree of homozygosity in the control population as well as in the case population. For example, figure 3.3 shows a scenario in which a genomic region would not be identified as associated with disease by simply counting the region's ROH status for cases and controls, despite the fact that the region does have one haplotype that, when homozygously inherited, results in recessive segregation of a disease-causing allele. Factoring in the genotypes within discovered ROHs could yield novel disease-linked loci.

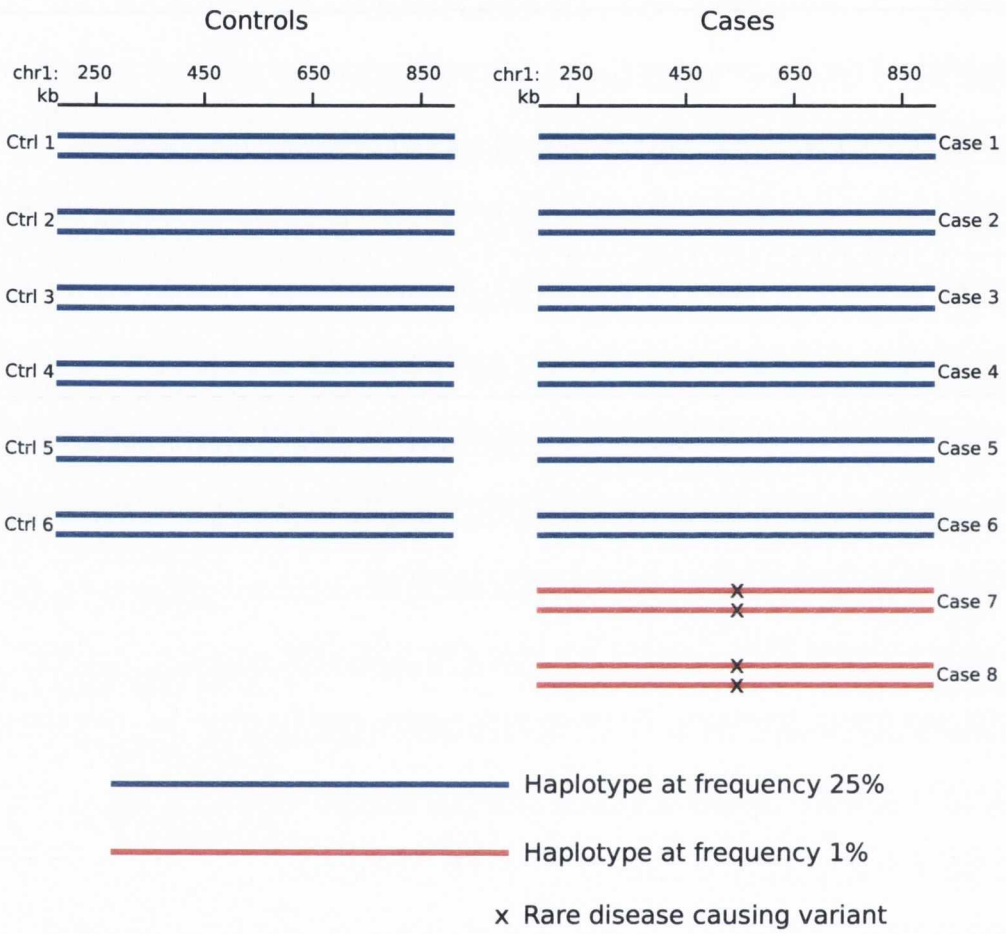


Figure 3.3: ROH mapping. A hypothetical scenario is depicted where ROHs are mapped in 100 cases and 100 controls. A haplotype that has a population frequency of 25% is detected as a ROH in 6 cases and 6 controls (blue), which would be the expectation, while a haplotype at population frequency of 1% is detected as a ROH in zero controls (power to detect this would be around 1%). However, because the case population is enriched for putative disease-causing variants, under assumptions of recessive inheritance the haplotype on which the disease-causing variant arose would be expected to appear homozygously more often than in the control population. A naïve ROH mapping/association strategy would not identify this region as associated with disease (8/100 cases, 6/100 controls; $p = 0.58$). However, if the genotypes were considered within the region, the red ROH would be identified as recurrently disease-specific.

3.1.3 Research aims

In this chapter, the Irish ALS GWAS dataset of 2008 [98] is augmented by further genotyping within the Irish ALS population and this larger dataset is used to generate candidate intervals for candidate exonic sequencing in chapter 4. Although straightforward association testing is one of the methods employed, an attempt is made to address power issues enforced by the small dataset by, where the analysis type permits, searching for features that are specific to ALS (that is, seen in exactly zero controls).

3.2 Methods

3.2.1 The 2008 dataset

561,466 SNP genotypes for 221 sporadic ALS cases and 211 controls from the 2008 Irish GWAS [98] were made available for use in this study. Additionally, for CNV analysis (section 3.2.4), the unprocessed intensity data for the same individuals were used for generation of the necessary datasets using Illumina® BeadStudio v2.0.

3.2.2 Genotyping of 308 further samples

Using the Irish ALS DNA bank, samples that had not been included in the original 2008 GWAS [98] were selected based on availability and quality of DNA, totalling 142 cases and 152 controls. For assessment of genotyping quality, eight samples that had been genotyped in the 2008 study were also included, and six samples in the new cohort were replicated, resulting in a total of 308 DNA samples being genotped. All patients had clinically definite or probable ALS [28] diagnosed by a neurologist with expertise in ALS, and of the 142 patients, 23 had positive family history of ALS.

Samples were genotyped for 620,901 polymorphic markers commercially by deCODE Genetics (Reykjavik, Iceland) using the Illumina® Human610-Quad BeadChip. Twenty

cases and eighteen controls did not pass initial call score-based quality checks at deCODE Genetics and were therefore not included in further processing of the data. Genotype calls for the remaining 270 samples were subjected to a number of quality control steps, first using Illumina® BeadStudio v2.0, then using PLINK v1.07 [95].

Using BeadStudio, all SNPs were reclustered following exclusion of individuals with low call rates. Any marker with a call rate lower than 98% or a cluster separation less than 0.3 was set to have missing genotypes across all samples. Call rates were then recalculated and checked for all samples, and final reports for all samples and markers were generated. Using an in-house script, the Illumina® final reports were parsed into a single PLINK-format .ped file for further quality control and analysis.

Using PLINK, 620,901 markers output from BeadStudio were merged with the unprocessed data from the 2008 study (561,466 markers) using the `--bmerge` option, resulting in a total of 630,738 markers in 352 cases and 355 controls. Markers with greater than 1% missing data were then removed using the PLINK `--geno` option, which also removed markers that had been set to missing across all samples in either of the two datasets, or markers that were not common to both datasets. With the 527,364 SNPs that remained, the PLINK `--update-map` option was used to update positions to NCBI build 36 coordinates (at the time of analysis the GRCh37 build was just being released). Using this approach, 598 SNPs were not remapped, so these SNPs were removed using the `--extract` command to retain only remapped SNPs.

The PLINK `--check-sex` option was used to identify individuals whose gender, as inferred by X chromosome heterozygosity, did not match the gender recorded for the individual in the DNA bank database. This identified three individuals whose sex did not match the database, as well as one individual whose X chromosome inbreeding estimate led to an ambiguous call about inferred gender ($F = 0.5878$). These individuals were excluded from further analysis.

The merged, remapped dataset was then checked for systematic missingness, which can arise from batch effects in sample treatment, using the `--test-missing` option with a threshold p-value of 0.05, removing 862 SNPs. Using both the `--hardy` and `--hardy2` options, 17,736 SNPs that deviated significantly ($p < 0.01$) from Hardy-Weinberg Equilibrium were removed. Following this extensive treatment, a further check for missingness within individuals was performed, although no individuals failed this check. Finally, 23,266 SNPs with a minor allele frequency of less than 0.01 were excluded (using `--maf`), along with 120 genotypes from non-autosomes, leaving a clean dataset of 484,882 markers.

Using the set of high-quality genotypes, samples were screened for cryptic relatedness by generating a matrix of identity by state (IBS) values with PLINK's `--cluster` and `--matrix` options. This method also identified replicate arrays, providing an indication of their accuracy. Sixteen pairs and one trio of individuals were identified as cryptically related (IBS score greater than 75%) using this approach; from each of the pairs, one individual was excluded and from the trio of individuals two were excluded.

Finally, to check for population stratification, which can cause mistaken assignment of population-differentiated SNPs as disease-associated SNPs, STRUCTURE v2.3.3 [151] and the SMARTPCA algorithm [126] implemented in EIGENSOFT were used. The STRUCTURE analysis compared the newly-generated genotypes against genotype data derived from The International HapMap Project's CEPH, YRI, CHB and JPT panels [5], using a subset of 1,969 unlinked ancestry-informative SNPs and an assumed population count (K) of 3. The same approach identified one individual with approximately 50% probability of being derived from the same population as the CHB/JPT panel in the 2008 dataset [98]; all newly-genotyped individuals had a high probability of being derived from the same population as the CEPH panel. Using SMARTPCA, the principal components of variation within the full genome-wide SNP datasets were assessed between cases and controls, showing no discernible population stratification.

The final dataset consisted of 484,882 high-quality genotypes in a panel of 344 unrelated Irish ALS cases and 331 age- and population-matched controls. The mean age of disease onset (± SD) in the ALS cohort was 61.7 ± 12.1 years and the mean age at sampling (± SD) in the control cohort was 58.4 ± 13.6 years. 73% of the ALS samples had spinal onset ALS; 27% had bulbar onset. The male:female ratio was 43%:57% for cases and 41%:59% for controls.

3.2.3 Allelic association

To test genome-wide SNPs for association with ALS, the PLINK `--model` option was implemented, which provides statistics for association with disease following a number of tests (basic allelic χ^2 test, Cochran-Armitage trend test, genotypic test, and tests under dominant and recessive models). The output from this analysis was then used as input for the PLINK `--clump` command, which can be used for generating associated genomic intervals based on very low p-values accompanied by neighbouring low p-values in regions of LD. For this, the threshold for index SNPs (the lower p-value threshold for primary identification of associated regions) was set at 0.0001, and the threshold for clumped SNPs (the higher p-value threshold for determination of associated intervals) was set at 0.01, with a modest r^2 value of 0.5 to indicate LD surrounding an associated SNP.

3.2.4 Analysis of putative copy number variation

To generate datasets for identification of regions of putative CNV, log-R ratio (LRR) and B-allele frequency (BAF) values for all SNPs that passed BeadStudio quality control steps were generated for all samples that passed the quality control steps described in section 3.2.2 using Beadstudio. Genomic regions of putative CNV were then identified using QuantiSNP v2.3 [133] and PennCNV (August 2009 release) [134]. QuantiSNP was run using default parameters; for PennCNV default parameters were also used but the

provided `hhall.*` files were used to set some of the parameters correctly for the 610k data. Inbuilt genomic GC content-based correction algorithms were implemented in both cases. The R statistical programming package [94] was then used to interpret the output, in particular the influence of variance in LRR and BAF, and suitable cutoffs and standards were determined based on the findings of these investigations.

The script `replicate_overlap.pl` was used to assess the concordance of the output of the two algorithms with datasets derived from replicate samples. Based on assessment of algorithm accuracy, the optimal strategy was to take the intersection of the results of both algorithms as the best estimate of true CNVs (figure 3.6); this was accomplished with `overlap.pl`, which also separated results into copy number gains or losses. Scripts `append.pl` and `count_status_per_SNP.pl` were used to combine all results into a single file and list case and control numbers for each SNP if a copy number loss or gain was detected. This facilitated the identification of recurrent ALS-specific CNVs.

3.2.5 Mapping runs of homozygosity

Runs of homozygosity were determined in the SNP dataset using the PLINK `--homozyg` argument. This algorithm takes a sliding window whose size is user-defined and scans across the genome, determining whether SNPs within the window look like they are in a ROH (based on user-provided definitions). Then, for each SNP, the proportion of windows that traversed the SNP that were homozygous are counted and, if this is above a user-defined threshold, the ROH is called.

The `--homozyg` algorithm takes many user-defined parameters, summarised in table 3.1; settings for these parameters were largely derived from analysis of the SNP dataset. Interpreting the logic forwarded by Lencz *et al.* [148], the minimum length l (in terms of number of consecutive SNPs) to call a ROH can be described in terms of the mean heterozygosity, $\overline{het}$, the number of SNPs in the dataset, n_s and the number of individuals

in the dataset, n_i, such that a proportion lower than α of the ROHs that are revealed could have occured by chance:

$$(1 - \overline{het})^l \cdot n_s \cdot n_i = \alpha. \tag{3.1}$$

This equation can be rearranged so that l can be calculated:

$$l = \frac{\log(\frac{\alpha}{n_s \cdot n_i})}{\log(1 - \overline{het})}. \tag{3.2}$$

However, the non-independence of SNP genotypes brought about by LD could result in the algorithm simply just describing extended LD, so it is useful to incorporate an estimation of the extent of LD in the dataset into the calculation. Specifically, Lencz *et al.* suggest that the minimum ROH length should be inflated proportionally to the ratio of the number of tag groups, n_t, identifiable within the dataset to the number of SNPs within the dataset. Thus, an LD-corrected statistic for l can be defined:

$$l = \frac{\log(\frac{\alpha}{n_s \cdot n_i}) \cdot n_s}{\log(1 - \overline{het}) \cdot n_t}. \tag{3.3}$$

Given 484,882 SNPs with a mean heterozygosity of 32%, 671 individuals, and 297,330 separable tag groups (determined using the PLINK `--indep-pairwise` argument to prune out SNPs with r^2 greater than 0.65), equation 3.3 evaluated to 104 SNPs. This number was assigned to the `--homozyg-snp` parameter to define the minimum run of consecutive homozygous SNPs required to call a ROH. ROHs were called using the default sliding window length (`--homozyg-window-snp`) of 50 SNPs.

With the definition of a ROH and the sliding window size both set in terms of number of SNPs, the parameters `--homozyg-window-kb` and `--homozyg-kb` were not required and so they were set to be very large (10,000 kb) and very small (1 kb) respectively, so that

Table 3.1: Parameters passed to the PLINK `--homozyg` algorithm

Parameter	Description	Setting
`--homozyg-window-kb`	Sliding window size in kb	10000
`--homozyg-kb`	Minimum ROH size in kb	1
`--homozyg-window-het`	Number of permitted heterozygotes	0
`--homozyg-window-missing`	Number of permitted missing genotypes	1
`--homozyg-window-threshold`	Proportion of overlapping 'homozygous' windows necessary to call a SNP 'in an ROH'	0.001
`--homozyg-window-snp`	Sliding window size in SNPs	50
`--homozyg-snp`	Minimum ROH size in SNPs	104
`--homozyg-density`	Required minimum density	50
`--homozyg-gap`	Pairwise maximum distance for two SNPs in one ROH	100
`--homozyg-group`	Invokes grouping algorithm	NA
`--homozyg-match`	Threshold for matching groups	1

they did not contribute to the definition of a ROH.

Within a ROH, no heterozygotes were permitted: `--homozyg-window-het` can be used to tolerate miscalled genotypes in a dataset (homozygotes miscalled as heterozygotes), but given that the genotype concordance rate (± SD) between replicate arrays was 99.99839 ± 1×10^{-5}%, the average ROH length required to incorporate one miscall was 62,112 SNPs, and this was therefore deemed a very unlikely event (for reference, there were 37,468 chromosome 1 SNPs in the dataset). However, when considering missing genotypes, the average missingness per individual was 0.07%, therefore the typical ROH length required to incorporate one missing genotype was 1428.57 SNPs. For this reason, one missing genotype was tolerated (using the `--homozyg-window-missing` parameter).

The final two parameters in table 3.1 are for grouping individuals based on mutual overlap of ROHs and allelic matching of overlapping regions. `--homozyg-group` invokes an algorithm that pools individuals together based on mutual overlap of homozygous regions and `--homozyg-match` subdivides these pools into groups of individuals whose alleles within the ROH match each other. The latter parameter was set to 1, meaning that alleles, where called, must match at 100% of the overlapping sites to be considered part of the same group.

Software was developed to parse, visualise and interpret the resulting output. Firstly,

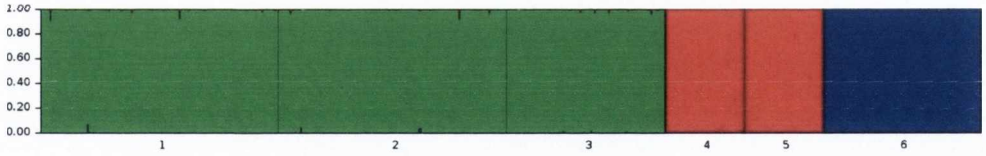
ALS-specific groups were extracted from the `.hom.overlap` output file using the script `parsePLINKROH.pl`, which extracts groups from pools when the phenotype for the group is exclusively ALS (with exactly zero controls). The PLINK algorithm had a tendency to output duplicate groups that would only be discovered following the treatment of `parsePLINKROH.pl`, so these were then removed using `removedups.pl`. Groups could be visually inspected using `showSNPs.php` to confirm *bona fide* homozygosity and allelic matching.

In order that only properly overlapping homozygous regions were carried forward for further analysis, `recipOverlap.pl` was used to calculate a score, S_i for each individual i's segment within a group of n individuals, based on the extent of its overlap with the rest of the group, as described by:

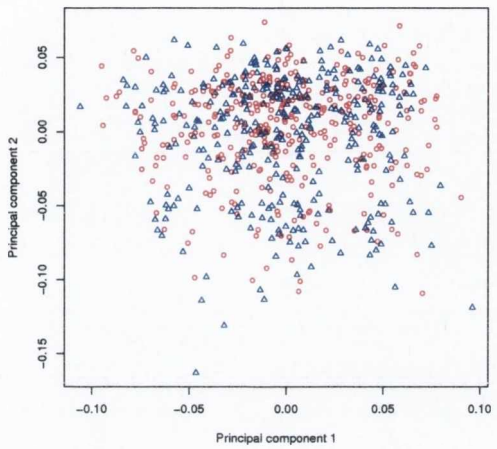
$$S_i = \sum_{j=1}^{n} \frac{\min(b_i, b_j) - \max(a_i, a_j)}{(n-1)(b_j - a_j)}, \tag{3.4}$$

where a is the start of a segment, b is the end of a segment and $i \notin j$. A segment was deleted if its S_i score was less than 0.5. Derivation of this formula can be found in appendix A.1.

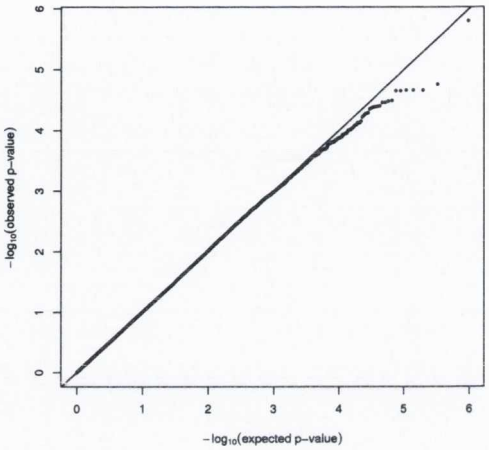
Finally, groups were only considered for further analysis if they survived the various treatment steps described with more than three individuals still present; sub-threshold groups were removed with `remove_small_tables.pl`. The resulting groups of high-quality, recurrent, overlapping, ALS-specific ROHs were collapsed down into single intervals using `create_intervals.pl`. In doing so, the interval was defined by the minimum and maximum genomic coordinates within the group, however particularly long ROHs (defined as intervals whose length exceeded the median length + 2 $\times$ the interquartile range of all the lengths across all groups) were not considered in this definition.



(a) STRUCTURE plot



(b) Principal components analysis



(c) Quantile-quantile plot

Figure 3.4: Checks for population stratification in the genome-wide SNP dataset. (a) STRUCTURE plot of the SNP dataset along with HapMap individuals under the assumption of 3 populations ($k = 3$). Population flags: 1, controls in the SNP dataset; 2, ALS cases; 3, CEPH HapMap individuals; 4, CHB HapMap individuals; 5, JPT HapMap individuals; 6, YRI HapMap individuals. (b) Scatterplot of the first two principal components of variation as determined by the SmartPCA algorithm implemented in EIGENSOFT. There was no discernable difference between the case (red circles) and control (blue triangles) populations for all principal components characterised (principal components 1 to 10). (c) Quantile-quantile plot of allelic association statistics showing observed p-values plotted against expected p-values, demonstrating that it is unlikely that there is any stratification within the sample causing spurious results.

3.3 Results

3.3.1 Genotyping

The majority of the results obtained during genotyping and quality control are detailed within section 3.2.2. The results from the STRUCTURE analysis to identify population outliers are shown in figure 3.4(a), demonstrating that all newly-genotyped individuals had a high probability of being European. Principal components analysis revealed no discernible population stratification between cases and controls (figure 3.4(b)).

3.3.2 Allelic association

Figure 3.5 summarises the results of the allelic association tests and figure 3.4(c) shows the quantile-quantile plot for expected distribution of the allelic association statistics. The critical Bonferroni-corrected p-value threshold for genome-wide significance, given 484,882 SNPs was 1.03×10^{-7}. No SNP was associated with ALS susceptability at genome-wide sigificance, however some came close. Table 3.2 shows all association statistics at $p < 1 \times 10^{-4}$.

3.3.3 CNV analysis

PennCNV called a total of 164,184 CNVs; QuantiSNP called 25,692. The median length of CNVs called by PennCNV was 6,735 bp; for QuantiSNP the median length was 93,558 bp.

Ten pairs of samples that had been assayed in duplicate (four 610k/610k pairs and six 610k/550k pairs) were assessed for consistency of CNVs called by QuantiSNP and PennCNV. A broad measure of accuracy was defined by the total length of concordant calls between replicates as a percentage of the total length of CNVs called within a replicate. Figure 3.6(a) shows the accuracy obtained for each replicate pair. The mean accuracy was calculated for PennCNV, QuantiSNP, QuantiSNP ($\log(Bayes factor) > 10$) and the overlap between PennCNV and QuantiSNP ($\log(Bayes factor) > 10$) and is shown in figure 3.6(b).

On inspection of the data, it was noted that variance within the LRR and BAF datasets had a profound effect on the number of CNVs called by PennCNV (figure 3.7 parts (a) and (b)). At the time of analysis, the documentation for PennCNV recommended that a cutoff of 0.24 for LRR_SD, the within-individual LRR standard deviation, be used to identify outliers, and that any individual showing greater than 50 CNVs is an outlier. While it is unlikely that these are arbitrarily chosen numbers, it is not obvious how these

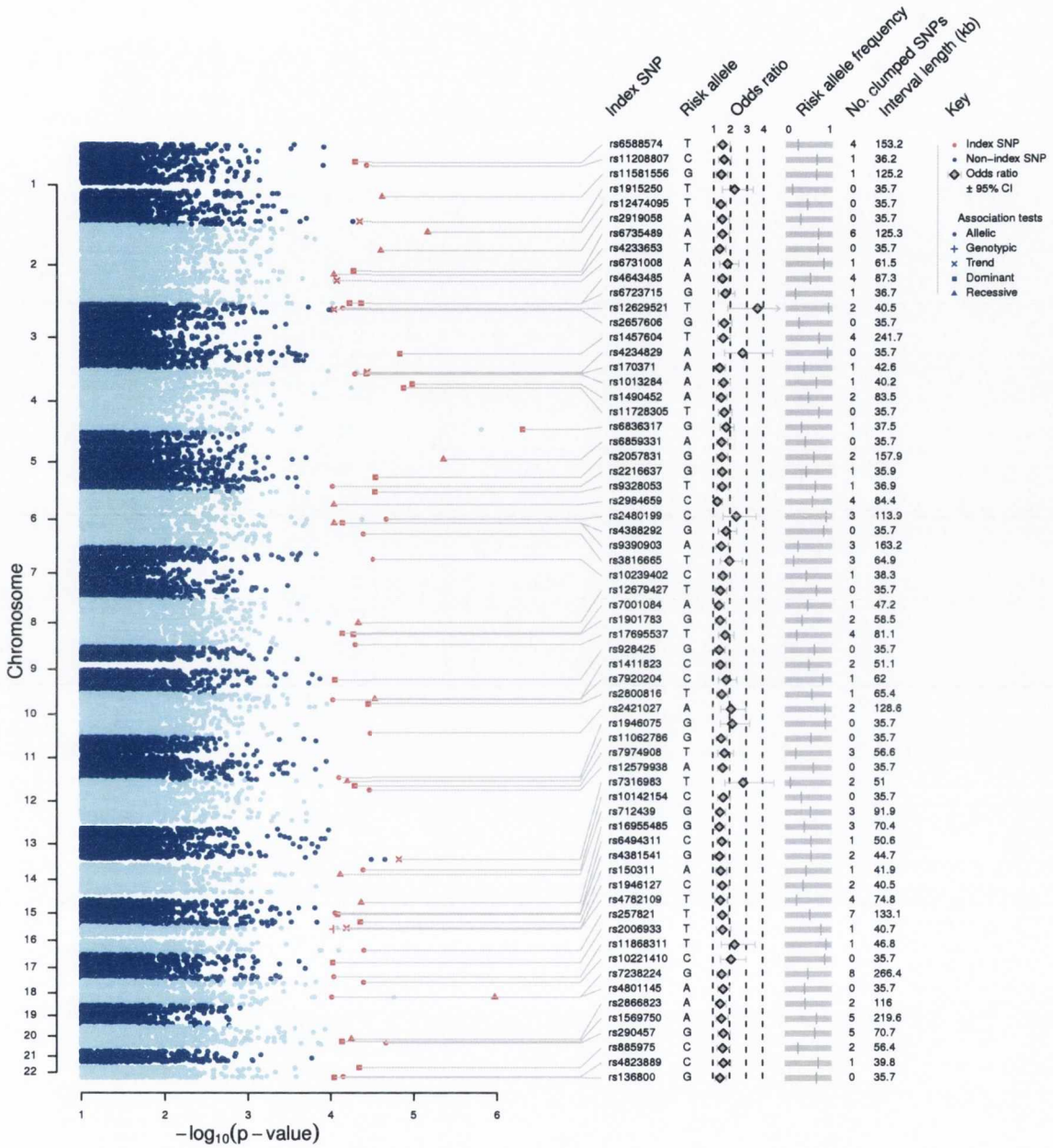
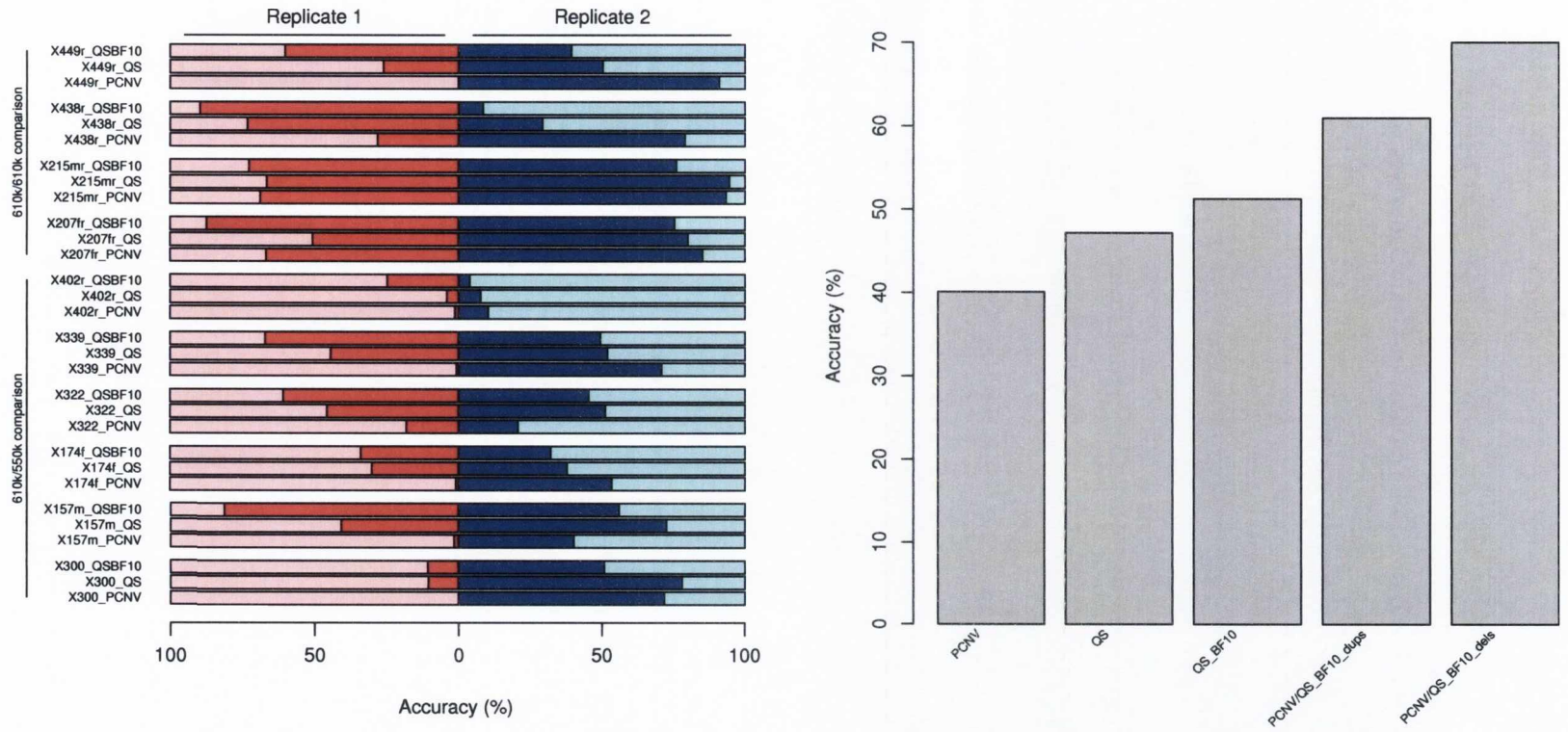


Figure 3.5: Statistics from tests of allelic association of 484,882 SNPs with ALS. Manhattan plot shows $-\log_{10}(p\text{-}values)$ for the basic allelic association test, with above-threshold results shown in red along with above-threshold results for the other four tests (in this case, 'threshold' means $p = 1 \times 10^{-4}$, which was the threshold for defining index SNPs with the PLINK `--clump` method). For index SNPs, only the best p-value from the five tests is shown. If a SNP's upper 95 % confidence bound for odds ratio exceeded 5, the error bar is truncated. Where no unique SNPs were clumped with an index SNP, the interval length represents a haplotype block length surrounding the index SNP (35.7 kb). Plot generated using `plotNatureStyle.R`; style based on Sawcer *et al.* (2011) [152].

Table 3.2: Association statistics at $p < 1 \times 10^{-4}$

Chromosome	Position (bp, NCBI36)	SNP rsID	RA	RAF	OR	$-\log_{10}(p)$	Test
4	182374181	rs6836317	G	0.3141	1.77 (1.4-2.23)	6.30	Dominant
18	50980784	rs4801145	A	0.4028	1.62 (1.3-2.01)	5.97	Recessive
5	82196556	rs6859331	A	0.4119	1.49 (1.2-1.86)	5.34	Recessive
2	20574316	rs12474095	T	0.4504	1.42 (1.14-1.76)	5.15	Recessive
4	44003303	rs1490452	A	0.4813	1.46 (1.18-1.82)	4.97	Dominant
4	55883043	rs11728305	T	0.7362	1.65 (1.29-2.11)	4.86	Dominant
3	150240824	rs1457604	T	0.7348	1.57 (1.23-2)	4.82	Dominant
13	114092979	rs7316983	T	0.06166	2.83 (1.72-4.67)	4.82	Trend
13	114108120	rs3813131	A	0.06148	2.83 (1.72-4.67)	4.82	Trend
20	52081236	rs290457	G	0.6535	1.63 (1.3-2.05)	4.66	Allelic
6	84856391	rs2497129	A	0.9163	2.38 (1.58-3.6)	4.66	Allelic
6	84859041	rs2480199	C	0.9163	2.38 (1.58-3.6)	4.66	Allelic
13	114108294	rs3813133	A	0.06074	2.78 (1.69-4.59)	4.64	Trend
1	160060235	rs11581556	G	0.6748	1.46 (1.16-1.83)	4.60	Recessive
2	77240715	rs2919058	A	0.2896	1.52 (1.2-1.92)	4.59	Recessive
5	137551328	rs2864	A	0.6133	1.49 (1.2-1.86)	4.53	Dominant
5	137563031	rs2057831	G	0.6126	1.5 (1.21-1.87)	4.53	Dominant
10	20117446	rs1411823	C	0.5007	1.44 (1.16-1.78)	4.52	Recessive
6	1389293	rs9328053	T	0.6533	1.53 (1.22-1.92)	4.52	Dominant
7	36754881	rs10239402	C	0.4348	1.58 (1.27-1.97)	4.49	Allelic
12	31689061	rs12579938	A	0.6148	1.59 (1.28-1.99)	4.46	Allelic
10	124193637	rs2421027	A	0.8895	2.09 (1.47-2.99)	4.46	Allelic
10	34300269	rs2800816	T	0.5815	1.53 (1.23-1.91)	4.43	Dominant
4	7823074	rs4234829	A	0.94296	2.74 (1.65-4.55)	4.41	Trend
4	12513904	rs223931	T	0.3733	1.38 (1.1-1.72)	4.40	Dominant
4	12520825	rs170371	A	0.3726	1.37 (1.1-1.71)	4.40	Dominant
20	52077174	rs158549	A	0.6719	1.54 (1.22-1.93)	4.40	Dominant
20	52079435	rs290453	C	0.6719	1.54 (1.22-1.93)	4.40	Dominant
1	66407508	rs11208807	C	0.683	1.62 (1.29-2.04)	4.40	Allelic
16	74701278	rs257821	T	0.5267	1.57 (1.26-1.94)	4.39	Allelic
20	52072283	rs158541	C	0.6748	1.56 (1.24-1.96)	4.39	Dominant
20	52097189	rs290421	C	0.6748	1.56 (1.24-1.96)	4.39	Dominant
18	5595988	rs10221410	C	0.8919	2.09 (1.46-3)	4.39	Allelic
14	33092583	rs10142154	C	0.3252	1.62 (1.28-2.04)	4.38	Allelic
6	129613022	rs3816665	T	0.1304	1.98 (1.42-2.75)	4.37	Allelic
15	27547807	rs16955485	G	0.3919	1.39 (1.12-1.73)	4.36	Recessive
6	84934534	rs6905922	C	0.92296	2.38 (1.55-3.66)	4.35	Allelic
15	87756365	rs150311	A	0.4393	1.47 (1.18-1.82)	4.35	Dominant
2	239336344	rs4643485	A	0.5252	1.53 (1.23-1.9)	4.34	Dominant
22	17810471	rs885975	C	0.2459	1.59 (1.24-2.05)	4.34	Dominant
1	237250396	rs1915250	T	0.08741	2.25 (1.5-3.38)	4.32	Trend
8	72405227	rs12679427	T	0.6874	1.43 (1.14-1.81)	4.31	Recessive
15	27559563	rs733612	C	0.3333	1.39 (1.11-1.75)	4.31	Recessive
12	18758875	rs7974908	T	0.1978	1.7 (1.29-2.23)	4.28	Dominant
8	138983253	rs17695537	T	0.2104	1.73 (1.32-2.26)	4.28	Allelic
4	13998452	rs1013284	A	0.6748	1.6 (1.27-2.02)	4.26	Allelic
1	56060311	rs6588574	T	0.2244	1.53 (1.18-1.98)	4.26	Dominant
8	107608916	rs1901783	G	0.3388	1.42 (1.13-1.78)	4.25	Dominant
2	141267819	rs6735489	A	0.72	1.53 (1.2-1.94)	4.25	Dominant
20	39912916	rs2866823	A	0.4148	1.44 (1.16-1.79)	4.25	Recessive
2	239580116	rs6723715	G	0.1689	1.7 (1.27-2.28)	4.20	Dominant
2	239581071	rs12328525	G	0.1689	1.7 (1.27-2.28)	4.20	Dominant
12	3729334	rs11062786	G	0.5578	1.49 (1.2-1.85)	4.19	Recessive
16	6181183	rs1946127	C	0.3652	1.56 (1.25-1.96)	4.19	Trend
22	45675237	rs4823889	C	0.7337	1.64 (1.28-2.09)	4.15	Allelic
20	47036647	rs1569750	A	0.6956	1.51 (1.2-1.91)	4.13	Dominant
6	96189807	rs4388292	G	0.857	1.77 (1.3-2.42)	4.12	Dominant
8	104296121	rs7001084	A	0.4763	1.35 (1.09-1.68)	4.12	Dominant
3	11981891	rs17035544	T	0.96444	3.64 (1.84-7.2)	4.12	Allelic
3	11982207	rs17035545	A	0.96444	3.64 (1.84-7.2)	4.12	Allelic
3	11986717	rs12629521	T	0.96444	3.64 (1.84-7.2)	4.12	Allelic
14	47524431	rs712439	G	0.5416	1.42 (1.15-1.77)	4.11	Recessive
16	74673563	rs977045	T	0.5274	1.54 (1.24-1.91)	4.09	Allelic
11	126147508	rs1946075	G	0.91037	2.17 (1.46-3.21)	4.09	Allelic
15	64880684	rs4381541	G	0.5556	1.4 (1.13-1.73)	4.08	Dominant
2	170679239	rs6731008	A	0.8526	1.84 (1.35-2.5)	4.05	Trend
15	60214655	rs6494311	C	0.56	1.54 (1.24-1.91)	4.05	Allelic
22	48224667	rs136800	G	0.6956	1.43 (1.13-1.8)	4.04	Dominant
17	67580692	rs11868311	T	0.92444	2.31 (1.5-3.55)	4.04	Allelic
9	100373273	rs928425	G	0.6341	1.37 (1.1-1.71)	4.03	Dominant
16	9965210	rs4782109	T	0.2111	1.45 (1.11-1.89)	4.03	Genotypic
17	23625259	rs2006933	T	0.8027	1.58 (1.21-2.08)	4.02	Dominant
6	96684599	rs9390903	A	0.2407	1.49 (1.16-1.92)	4.02	Recessive
6	39998244	rs2984659	C	0.5815	1.24 (1-1.55)	4.02	Recessive
3	16151301	rs2657606	G	0.2504	1.63 (1.27-2.1)	4.02	Trend
2	152195819	rs4233653	T	0.7074	1.36 (1.07-1.72)	4.01	Recessive
18	50592345	rs7238224	G	0.48	1.53 (1.24-1.9)	4.01	Allelic
10	21665980	rs7920204	C	0.843	1.81 (1.34-2.44)	4.01	Allelic
5	164912172	rs2216637	G	0.4311	1.54 (1.24-1.91)	4.00	Allelic

RA, risk allele; RAF, risk allele frequency; OR, odds ratio



(a) Per-replicate pair accuracy

(b) Accuracy of various strategies

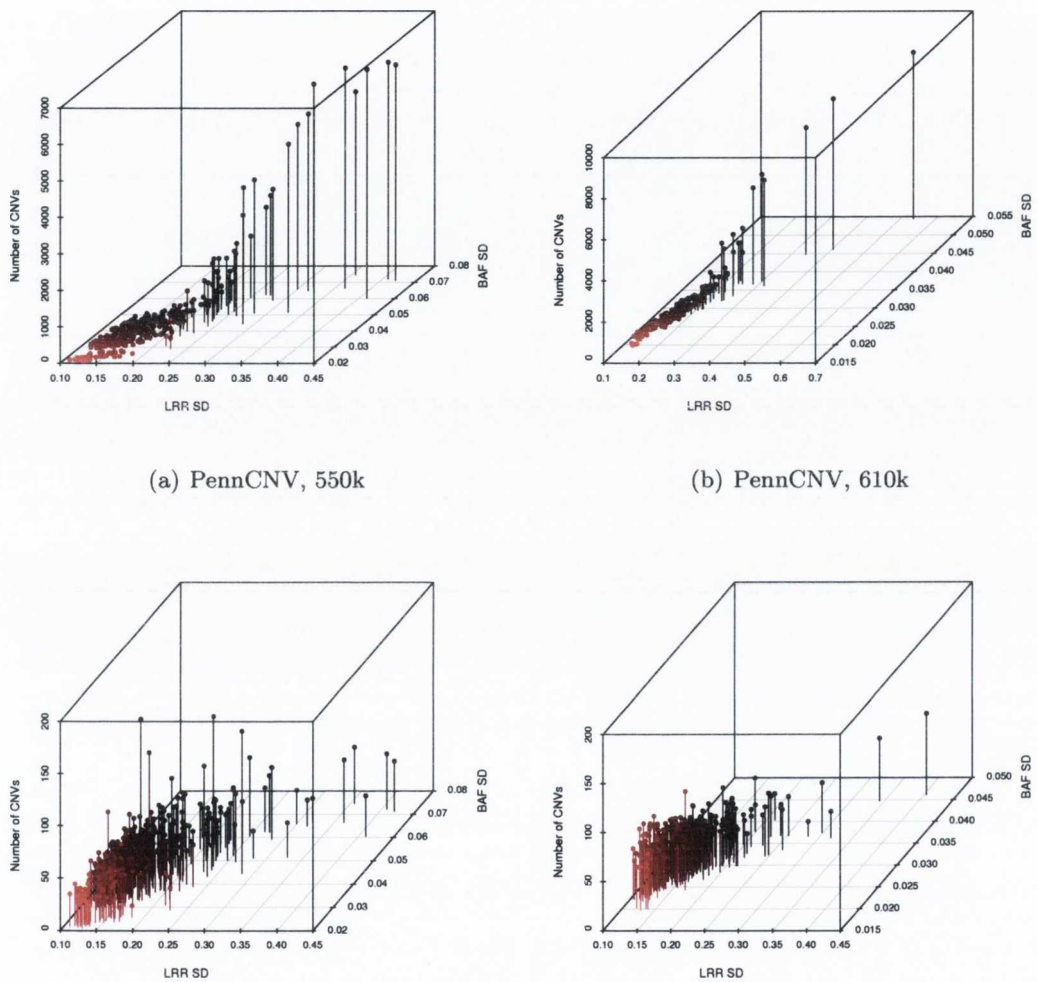
Figure 3.6: Accuracy of CNV mapping approaches. (a) Consensus between replicate arrays as a percentage of the total CNVs called by each algorithm. Darker bars represent the percentage overlap of the complementary replicate. PCNV, PennCNV; QS, QuantiSNP; QSBF10, QuantiSNP taking only CNVs with a $\log(Bayesfactor)$ of greater than 10. (b) Mean accuracy of five strategies for mapping valid CNVs. The last two bars represent datasets where only the consensus between the PennCNV output and the QuantiSNP ($\log(Bayesfactor) > 10$) output was considered, and this dataset has been divided into copy number duplications (dups) and copy number deletions (dels).

values were derived and the validity of their use in these datasets was unclear.

With this in mind, a method was developed to use the empirical distribution of the datasets under analysis to decide suitable cutoff values (figure 3.8). This method makes the assumption that the data are derived from two superimposed distributions: one which is unaffected by any systematic bias in LRR_SD, and one which contains upwardly biased values. A second assumption of the method is that the unbiased distribution is lognormal, and that the intrusion of the biased dataset on the first and second quartiles of the log-transformed unbiased dataset is minimal. Working with these assumptions, a 'fitted' mean (representing the modal value) and SD for the log-transformed unbiased distribution can be estimated using the data that lie between $\min(LRR\_SD)$ and the maximum of the kernel density estimate (generated with the R `stats::density` function) for the dataset.

Using this method, 131 individuals from the 550k dataset and 39 from the 610k dataset were identified as having LRR_SD values that were too high to produce reliably high-quality inferences about copy number states when analysed with PennCNV. On the other hand, LRR_SD and BAF_SD had no discernible effect on the number of CNVs called by QuantiSNP (figure 3.7 (c and d)). For outlier identification in the QuantiSNP dataset, the certainty metric $\log(Bayes factor)$ was used to examine the data's tendency to converge to low CNV calls when high stringencies are applied (figure 3.8 (c and d). This resulted in far fewer individuals being flagged as outliers (just four in the 550k dataset).

As a consequence of all the quality control observations, the final strategy employed for generating the best dataset possible was to exclude the four individuals flagged in figure 3.8(c) (as well as all replicate individuals) and to take results that represented the intersection of PennCNV and QuantiSNP ($\log(Bayes factor) > 10$) results. Using this approach, a total of 2,492 CNVs were called in cases and 2,548 in controls. Candidate regions to be carried forward to analysis in chapter 4 were chosen if they were recurrent and ALS-specific. Using this criterion, there were 37 ALS-specific copy number losses and



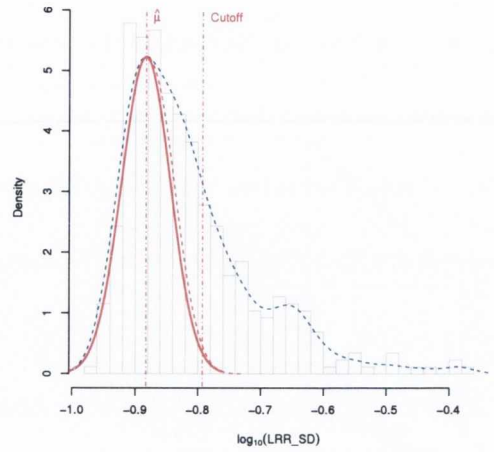
(a) PennCNV, 550k

(b) PennCNV, 610k

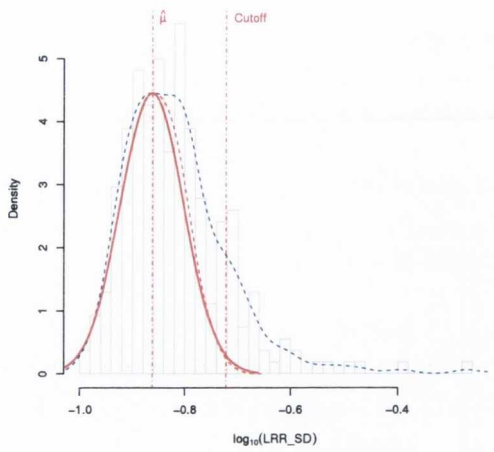
(c) QuantiSNP, 550k

(d) QuantiSNP, 610k

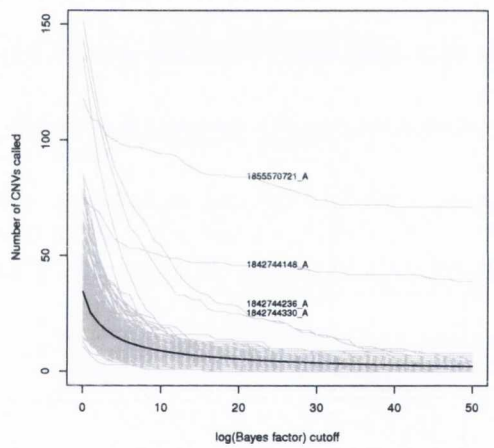
Figure 3.7: 3D scatterplots demonstrating the effect of *LRR_SD* and BAF standard deviation (*BAF_SD*) on the number of CNVs called by PennCNV and QuantiSNP (before processing). (a and b) *LRR_SD* and *BAF_SD* were reasonably correlated in both datasets from the PennCNV output ($r^2_{550k} = 0.4232$; $r^2_{610k} = 0.652$), suggesting collinearity between *LRR_SD* and *BAF_SD* and that the same system was causing an upward bias of both metrics. Therefore, for simplicity, only *LRR_SD* was used in further analyses of outliers. (c and d) *LRR_SD* and *BAF_SD* had no effect on the number of CNVs called by QuantiSNP.



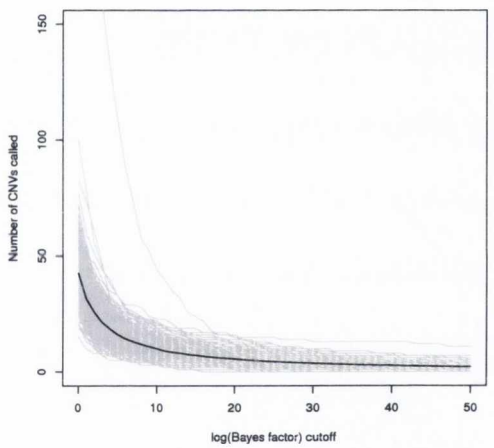
(a) PennCNV, 550k dataset



(b) PennCNV, 610k dataset



(c) QuantiSNP, 550k dataset



(d) QuantiSNP, 610k dataset

Figure 3.8: Outlier identification in the 550k and 610k CNV outputs. (a, b) PennCNV outlier identification using LRR_SD. The dashed blue line shows the kernel density estimation for the underlying distribution and $\hat{\mu}$ represents the 'fitted' mean of the unbiased distribution derived from the maximum of this estimate. Based on this, the dashed red line represents the assumed underlying (unbiased) distribution, generated by mirroring the kernel density estimate for $\{\min(LRR\_SD), \ldots, \hat{\mu}\}$ around $\hat{\mu}$. Taking data between $\min(LRR\_SD)$ and $\hat{\mu}$, the standard deviation $\hat{s}$ could be estimated, and a cutoff derived based on $2.326 \times \hat{s}$. The solid red line shows a normal distribution with mean $\hat{\mu}$ and standard deviation $\hat{s}$, demonstrating a close fit to normality of the theoretical distribution ($p_{550k} = 0.4196$; $p_{610k} = 0.4838$, Shapiro-Wilk test for normality). (c, d) The effect of varying the $\log(Bayesfactor)$ cutoff on number of valid CNVs called by QuantiSNP, and its use in identifying outliers. Each plot shows one grey line per individual, and the black line shows the mean of the individual lines. Outliers were identified from these plots arbitrarily based on whether the number of CNVs called dropped to below 20 when the $\log(Bayesfactor)$ exceeded 20; these individuals are marked.

25 ALS-specific copy number gains (table 3.3).

3.3.4 ROH analysis

Following mapping of ROHs using the parameters defined in section 3.2.5, there was no significant difference between cases and controls in terms of the number or average length of ROHs called (figure 3.9). The median number of ROHs called per individual was 43 and the median total length of ROHs, per individual, was 35.7 Mb. The median of all within-individual mean ROH lengths was 831.4 kb. A total of 7,989 recurrent regions of ROH in the overall dataset were called by PLINK. Table 3.4 shows the number of regions that remained after each subsequent processing step. In general, recurrent ALS-specific ROHs overlapped well within groups (figure 3.10). The final number of groups was 448, with the largest group containing 8 overlapping individuals; however, after consolidating overlapping groups into single intervals, only 270 intervals remained.

3.4 Discussion

This work has performed genome-wide SNP analysis on a previously-published GWAS dataset in the Irish population [98], made roughly 1.5× larger by further genotyping of cases and controls in the same population. Three methods were used to attempt to assess the contribution of genetic variation to ALS aetiology: SNP association, analysis of copy number variation and homozygosity mapping. All three methods yielded results that were carried forward to chapter 4 for greater depth of analysis by next-generation sequencing.

3.4.1 Summary of findings and their significance

Association

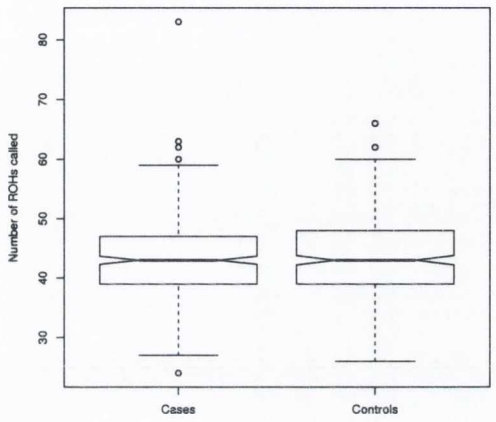
Using 484,882 high-quality genome-wide SNP genotypes, five different tests for association with ALS were performed: a basic allelic χ^2 test, the Cochran-Armitage trend test, a

Table 3.3: Recurrent, ALS-specific copy number gains and losses

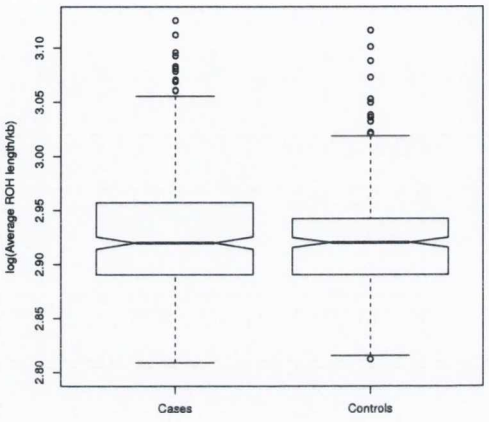
ALS-specific copy number gains				
Chromosome	Start position (bp, NCBI36)	End position (bp, NCBI36)	Length (bp)	No. cases
11	50950526	54533370	3582844	6
12	34496692	36718276	2221584	6
1	103904580	104063231	158651	5
8	2334306	2394805	60499	4
1	156783977	159944878	3160901	3
3	87410591	87993621	583030	3
8	144684227	144768796	84569	3
10	134641871	134682390	40519	3
3	83619218	84627172	1007954	2
4	57834727	58261378	426651	2
11	65013905	65193464	179559	2
4	129993825	130147254	153429	2
3	90088422	90173309	84887	2
1	2349841	2413982	64141	2
14	103624674	103683231	58557	2
1	9243828	9302280	58452	2
10	44550157	44594503	44346	2
9	138606913	138638786	31873	2
6	135009308	135039881	30573	2
5	104745638	104775166	29528	2
19	58206416	58232152	25736	2
17	53738052	53760857	22805	2
2	45188917	45201781	12864	2
14	105138663	105143361	4698	2
6	162798397	162800693	2296	2
ALS-specific copy number losses				
Chromosome	Start position (bp, NCBI36)	End position (bp, NCBI36)	Length (bp)	No. cases
15	85631534	92033269	6401735	5
8	2118532	2158362	39830	5
8	3987468	6156320	2168852	4
22	24017514	24240667	223153	4
6	162863051	162886421	23370	4
16	82466542	82484740	18198	4
5	109391074	109403379	12305	4
4	42400885	42404178	3293	4
20	28059305	28118678	59373	3
13	83000441	83055928	55487	3
17	63660583	63681642	21059	3
4	157183142	157186835	3693	3
3	190848118	190849457	1339	3
2	205824033	212891941	7067908	2
7	75981641	76348155	366514	2
1	166709802	166967709	257907	2
7	145088739	145332649	243910	2
17	30708148	30787791	79643	2
1	194097653	194138918	41265	2
19	48613901	48646755	32854	2
5	2094665	2119165	24500	2
5	98803229	98825827	22598	2
5	120714197	120731816	17619	2
5	109610790	109624892	14102	2
21	23820287	23834001	13714	2
16	3647748	3658849	11101	2
17	6237444	6247972	10528	2
9	11641144	11650080	8936	2
3	175382584	175391369	8785	2
7	1716791	1723762	6971	2
12	98526152	98532904	6752	2
3	56528363	56534273	5910	2
14	21816895	21822713	5818	2
10	135279590	135284293	4703	2
10	13096593	13100416	3823	2
5	103041190	103042663	1473	2
6	30889981	30890214	233	2

Table 3.4: Refinement of ROH results

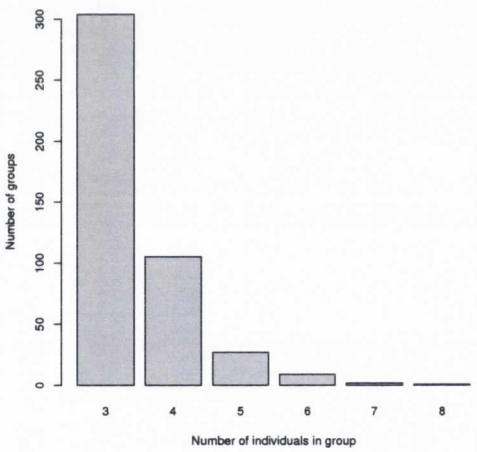
Step	Algorithm	Groups
Identification of recurrent ROHs	PLINK `--homozyg` argument	7,989
Extraction of ALS-specific groups	`parsePLINKROH.pl`	881
Removal of duplicate groups	`removedups.pl`	605
Extraction of adequately overlapping groups	`recipoverlap.pl` and `remove_small_tables.pl`	448
Consolidation of overlapping groups	`interval_overlap.pl`	270



(a) Number of ROHs



(b) Average length of ROHs



(c) Number of groups of each size

Figure 3.9: Statistics from ROH mapping. (a, b) Notched boxplots showing: (a) the number of ROHs called per individual; (b) the average length of ROHs per individual. There was no significant difference between cases and controls for either category ($p_{number} = 0.8453$; $p_{length} = 0.2192$, Mann-Whitney-Wilcoxon test). (c) The number of groups of ROHs with 3-8 individuals overlapping.

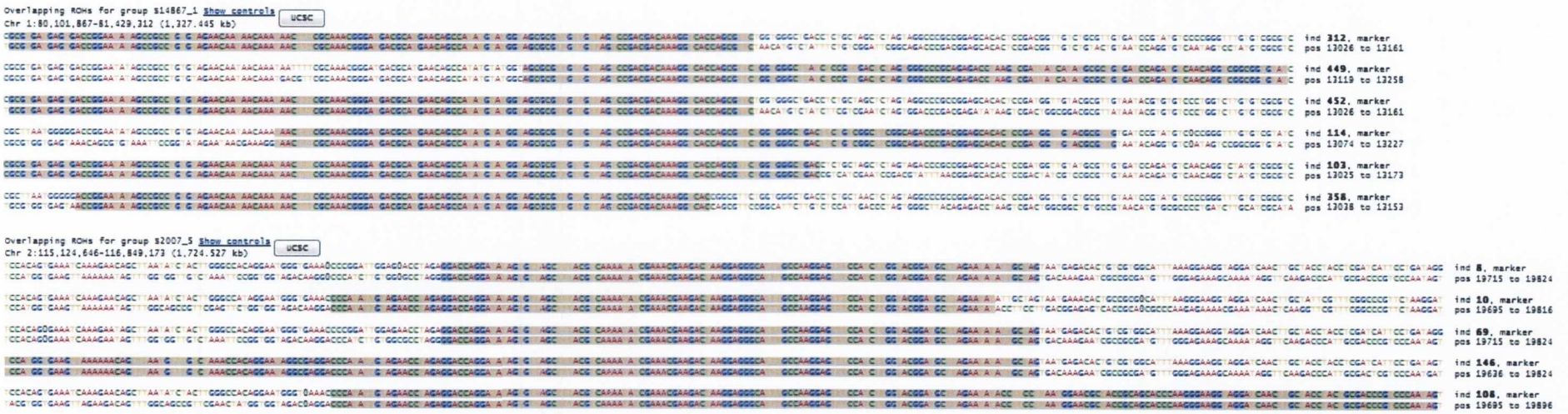
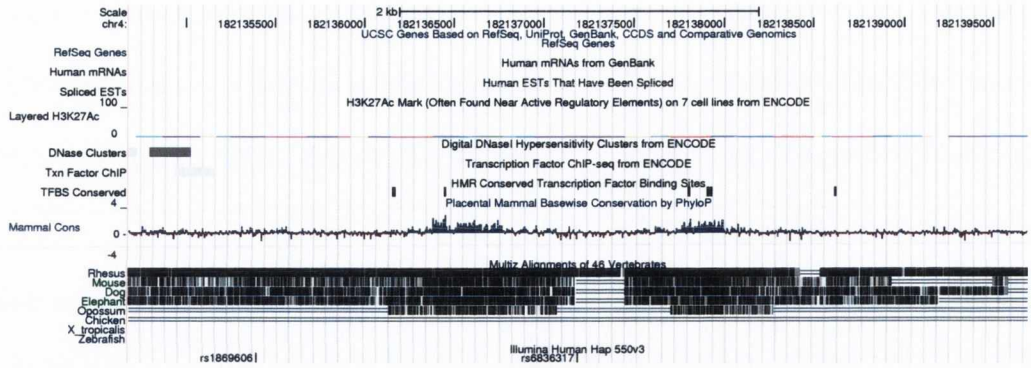


Figure 3.10: Screenshot of **showSNPs.php**, showing two recurrent ALS-specific ROH groups. Per-individual ROH regions are highlighted in grey. PLINK successfully sorted recurrent ROHs into allelically-matching pools which mutually overlapped well in most cases.

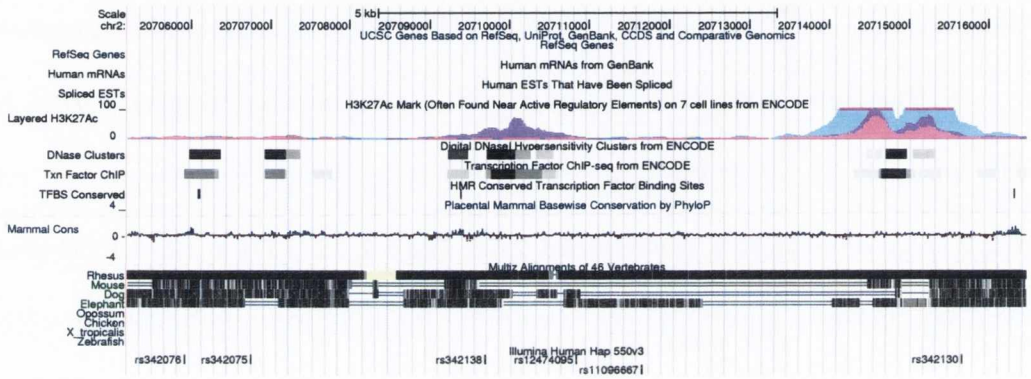
genotypic test, and tests under dominant and recessive models. No SNP showed association at genome-wide significance (1.03×10^{-7}), however several showed speculative association (table 3.2). Weak association signals in regions of LD around these 'peaks' were used to elucidate intervals that may contain causative alleles.

The best result for SNP association with ALS was for rs6836317 on chromosome 4q34.3, $p = 5.043 \times 10^{-7}$, OR = 1.767 (95% CI 1.399-2.233) under the dominant model (figure 3.5). This was narrowly higher than the critical Bonferroni-corrected p-value threshold of 1.03×10^{-7}. The SNP rs6836317 is intergenic, with the nearest RefSeq gene, *ODZ3*, lying 1.11 Mb telomeric to the SNP's GRCh37 genomic coordinate. However, it lies within a region of high conservation (determined from alignments [153] in the UCSC genome browser [154]) that overlaps with several conserved transcription factor binding sites (figure 3.11(a)), and it is 56.9 kb telomeric to the long noncoding RNA (ncRNA) *LINC00290*. Long ncRNAs have been shown to direct recruitment of *FUS* to the promoters *CCND1* in response to DNA damage signals [155]. The observation that ncRNAs are part of biological signalling pathways with the known ALS gene *FUS* make them an interesting candidate species for study into their implications in ALS. The result that an association peak lies in reasonably close proximity to this particular ncRNA corroborates such study further, although currently little is known about the biological importance of *LINC00290*.

The second-best result, rs4801145 on chromosome 18q21.2 ($p = 1.06 \times 10^{-6}$ under the recessive model, OR = 1.616 [95% CI 1.297-2.013]), was also intergenic, this time mapping 59.8 kb centromeric to *TCF4*, a gene for which deletions are implicated in Pitt-Hopkins syndrome [156]. This syndrome has several neurological components including severe mental retardation, microcephaly, respiratory pattern abnormalities, epilepsy and an excess of slow waves on electroencephalography [157]. This gene could therefore be implicated in neurological function and development, and it is possible that less severe variants in the gene could drive ALS pathology.



(a) rs6836317



(b) rs12474095

Figure 3.11: The genomic architecture of the region surrounding two of the top GWAS 'hits', visualized using the UCSC genome browser [154]. In both cases, the associated SNP is in the centre of the region depicted in the figure (SNPs are shown in the bottom track in both figures).

Two further SNPs were significant at the $p < 10^{-6}$ level, rs6859331 on chromosome 5q14.2 ($p = 4.521 \times 10^{-6}$, OR = 1.494 [95% CI 1.201-1.858]) and rs12474095 on chromosome 2p24.1 ($p = 7.074 \times 10^{-6}$ OR = 1.419 [95% CI 1.144-1.760]), both under recessive models. For rs6859331, the nearest gene is *TMEM167A*, which is 188 kb telomeric to the SNP, and for rs12474095 the closest gene, *RHOB*, is 61.6 kb telomeric. However, rs12474095 lies within a region for which there is strong evidence for the presence of regulatory elements, as visualized using the track in the UCSC genome browser derived from ENCODE data [158] (figure 3.11(b)).

The finding that no SNP showed association above the genome-wide p-value threshold of 1.03×10^{-7} was disappointing but not surprising given the low power to detect an association with just 344 cases and 331 controls. Nevertheless, Cronin *et al.* (2008) argue

that the slightly higher homogeneity of the genetic structure of the Irish population may mean that fewer genome-wide markers are truly independent and thus the stringency of Bonferroni correction in a dataset derived from this population may be higher than is necessary [98]. However, there were no single stand-out peaks of association as would be seen in larger studies such as those that correctly identified chromosome 9p21 as an associated locus [66, 67]. Instead, a handful of markers showed speculative association and a larger number showed weak association with ALS. In total, 79 independent SNPs showed association with ALS at $p < 1 \times 10^{-4}$ (table 3.2); these results were used to generate LD-based intervals that were carried forward to chapter 4 for further analysis by NGS.

Analysis of putative copy number variation

PennCNV [134] and QuantiSNP [133] were used to scan LRR and BAF values generated from the raw intensity data derived from SNP genotyping to identify regions of putative copy number variation. Quality assessment was made on the resulting data to control for false positives. The resulting dataset was used to search for recurrent copy number gains and losses that were specific to ALS cases.

PennCNV was shown to be less accurate than QuantiSNP (figure 3.6), although the raw output of QuantiSNP was not particularly accurate itself. The inaccuracy of PennCNV is discordant with a benchmark test of its accuracy [132] which praised its low false positive rate. This is likely to be due to differences in sample quality between data used in this study and the data used by Dellinger *et al.* [132]; indeed, figure 3.7 (a) and (b) shows that PennCNV is highly sensitive to upwardly-biased LRR_SD and BAF_SD values and data derived from samples with high variance in LRR and BAF have a much higher CNV call rate (and therefore, presumably, a higher false positive rate).

The use of the per-individual quality metric, $\log(Bayesfactor)$ helped to address the

inaccuracy of the QuantiSNP's results by only including CNV calls above a certainty threshold. However, PennCNV has no inbuilt method for quality-controlling the output on a per-individual basis, making the handling of false positives difficult. Taking only results that intersect with QuantiSNP helped to address this, although it was not perfect in that the 'inaccuracy' reported in figure 3.6 probably reflects a moderate false-negative rate as well as a moderate false-positive rate, and these false negatives would remain a problem if only intersecting results were taken forward for further analysis. Nevertheless, using the intersection between the two methods was deemed the most accurate strategy, which is in line with recommendations [159].

The best 'hit' for the CNV analysis was joint between two regions of putative copy number gain on chromosomes 11 and 12. However, both of these regions span the centromeres of the chromosomes, suggesting that they are likely to be artefacts of the calling algorithms. The best region of copy number loss identified was tied between a large 6.4 Mb region on chromosome 15 and a 39.8 kb region on chromosome 8. The chromosome 8 locus contained no genes but some evidence of regulatory regions. The chromosome 15 region, on the other hand, mapped to a gene-rich portion of the genome, with *SEMA4B* in the middle of the region. *SEMA4B* is an excellent candidate gene for ALS aetiology; it belongs to a family of genes that encode proteins involved in axon guidance and a related gene, *SEMA6A*, has been implicated in ALS by GWAS [160]. Additionally, another related gene, *SEMA6D*, lies within a familial linkage region for autosomal recessive juvenile onset ALS [161]. Evidence from the Database of Genomic Variants [162] only shows duplications for the genomic regions overlapping *SEMA4B*, adding weight to the argument that the discovery of deletions in this dataset could be an ALS-specific phenomenon driving some of the aetiology of ALS.

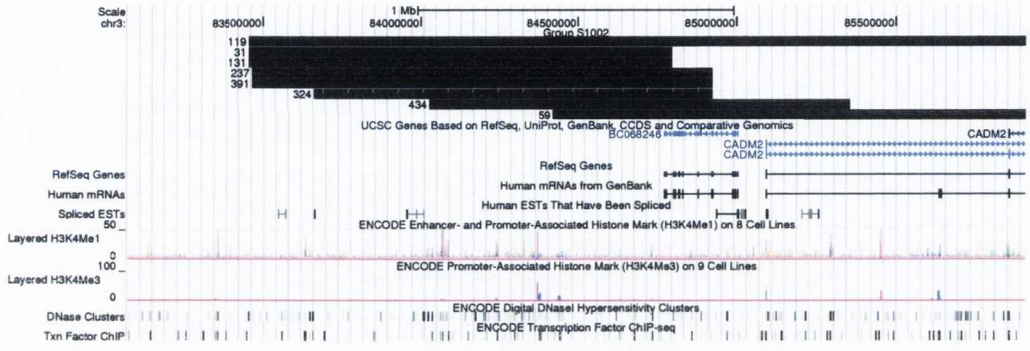
In total, 37 ALS-specific copy number losses and 25 ALS-specific copy number gains of various lengths were identified using the approaches described (table 3.3); these re-

gions were carried forward to chapter 4 for consideration in next generation sequencing experiments.

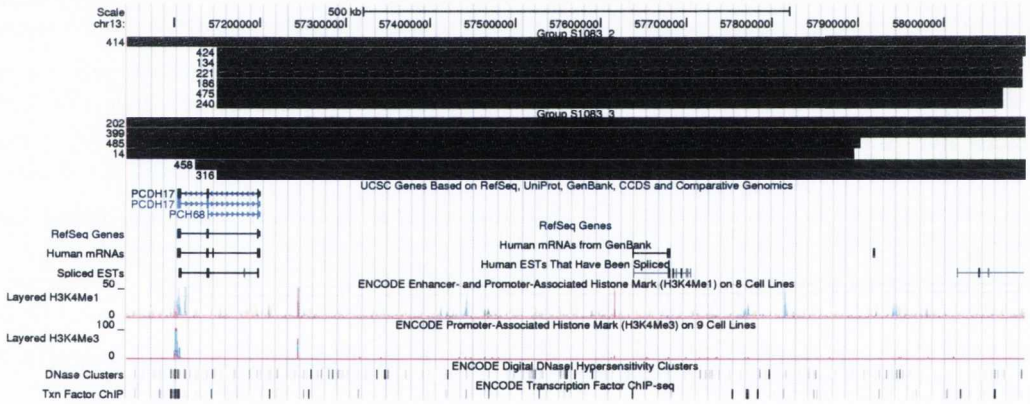
Homozygosity mapping

The ROH mapping technique successfully identified several overlapping allelically-matching runs of homozygous SNPs whose haplotypes were unique to ALS. Using these stringent parameters, the largest group identified was eight individuals that shared an allelically-matching ROH on chromosome 3p12.2–3p12.1 (figure 3.12(a)). This overlapped with an uncharacterized gene (identified as *BC068246* in figure 3.12(a)) as well as mapping to the 5′ end of *CADM2*. The bodies of seven of the eight ROHs mapped to a region of the genome showing strong evidence for the presence of regulatory elements, potentially the promoter region for *CADM2*. This gene is a reasonable candidate for ALS pathophysiology, as a member of the large immunoglobulin superfamily which contains cell adhesion molecules that act at synapses in the central nervous system [163]. Furthermore, a recent paper demonstrated a link between *CADM2* and autism spectrum disorder, using highly similar methods to those described in this study [164].

Possibly the best result, however, was for a region on chromosome 13q21.1 overlapping with the 3′ region of *PCDH17*. This region had two independent ROH groups map to similar locations, totalling 13 patients (figure 3.12(b)). *PCDH17* is a member of the cadherin superfamily, a group of genes whose protein products are responsible for cell-cell connections [165]. A recent study on cadherin expression in the various layers of the somatosensory cortex demonstrated a specific expression of mouse Pcdh17 in layer V of the somatosensory cortex, along with weak expression in layers VI and I [166]. Primary motor cortex layer V is the origin of upper motor neurones, so dysregulation of *PCDH17* expression could, in theory, lead to disruption of the maintenance of motor neurones and thus a syndrome like ALS. This ROH region also overlaps with the human mRNA



(a) ROHs on chr3p12.2–3p12.1



(b) ROHs on chr13q21.1

Figure 3.12: The best groups identified in the ROH analysis visualized using the UCSC genome browser [154]. (a) Eight individuals mapped to a region on chromosome 3p12.2–3p12.1; the end coordinates of ROHs for individuals 119 and 59 extend off the edge of the figure. (b) Two ROH groups mapped to the 5′ end of *PCDH17*, totalling 13 patients. The start and end coordinates for several of the individuals are truncated.

AK124674, which was one of 21,243 expressed RNA sequences discovered in a large-scale cDNA sequencing project [167], in this case, from amygdala tissue.

The technique employed to map ROHs has an important limitation. In order for a discovered ROH to survive control-based filtering, the haplotype identified in the case group has to be present homozygously in exactly zero controls. This necessitates a low haplotype frequency for ROHs to survive filtering, because the recessive disease allele would, in all likelihood, have arisen on an already existing haplotype, so future generations would have two variants of this same haplotype: one containing the disease allele and one not. Therefore, it is possible that a control could have a ROH that allelically matches a disease-causing ROH, while not carrying the disease allele homozygously. This limitation is addressed to some extent by the relatively low number of controls used to screen for non-pathogenic ROHs, thus rendering the power low to detect even medium-frequency haplotypes homozygously by chance. This, coupled with the reassuring fact that putatively disease-causing ROHs are observed in several cases, makes this limitation less of a problem.

A second limitation of this approach is that it could incorrectly identify copy number losses such as the scenario depicted in figure 3.2(b) as a ROH. However, given the relative rarity of CNVs in the genome compared to runs of homozygosity, this is not a large enough problem that the results of the ROH analysis would be affected.

Using this ROH mapping approach, 270 intervals were identified as loci potentially carrying recessive disease-causing mutations and these were brought forward to chapter 4 for further study by next-generation sequencing.

3.4.2 Conclusion

With a complex disease like ALS, which is likely to be caused by several genetic factors, many of which are rare and some of which are probably recessive and potentially incompletely penetrant, GWAS is an ambitious endeavour and its success is probably most

sensitive to sample size above all other factors. This is reflected in the body of literature that has been published on the subject [66, 67, 98, 113, 168]. However, GWAS is not the only analysis available with a genome-wide SNP dataset. In this chapter, an attempt has been made to draw on evidence from an Irish ALS genome-wide SNP dataset to identify ALS-specific regions of the genome that may harbour rare disease variants. Further interrogation by NGS of the intervals that have been identified is reported in chapter 4.

Chapter 4

Targeted high-throughput resequencing of ALS candidate genes

4.1 Introduction

Following many GWAS attempts in ALS, a substantial proportion of the heritability of the disease still remains unexplained, a finding that is also true for several other complex genetic conditions. It is likely that this is, in part, due to inadequate statistical power to detect associations by GWAS under the common disease-common variant hypothesis. It is also equally likely that multiple rare variants within various disease loci are contributing to disease pathogenesis. These may not be detectable by GWAS so the attention of many research scientists is now turning to methods for high-throughput assessment of rare variation in disease. The main method currently in use to perform such studies is high-throughput resequencing, also (at present) termed next-generation sequencing (NGS).

The principle behind current NGS methodologies is relatively straightforward. Instead of sequencing long stretches of DNA, which, at the time of writing, cannot be performed

in a high-throughput manner due to technological limitations, genomic DNA samples are highly fragmented and short sequencing reads (for example, 100 base pairs) are generated from these fragments by massively parallel sequencing. These reads are then aligned against a reference genome (allowing some differences in the sequence), from which sequence variants can be identified and examined for their potential pathogenicity. Figure 4.1 summarizes a typical pipeline for an NGS experiment from preparation of genomic DNA to generation of sequencing reads.

A number of different technological platforms exist to perform next-generation sequencing [169]. The common theme between all sequencing methods (including those available from Roche/454, Illumina and Applied Biosystems) is that millions of single-stranded template molecules are immobilized on a surface and are sequenced by synthesis of new, complementary strands from which fluorescence signals are detected representing the incorporation of labelled nucleotides. The order in which the fluorescence signals are read translate to the order of bases in the sequence reads. The differences between the technologies lie in the preparation of DNA, the way it is immobilized during sequencing, and the particular method of sequencing-by-synthesis. Roche/454 systems use an emulsion PCR-based template preparation method which results in highly amplified DNA molecules immobilized on beads which are then trapped within picotitre plate wells where sequencing by synthesis takes place. Similar DNA preparation methods are used in the Applied Biosystems platform, although sequencing is performed through ligation of labelled 8-mer probes [170].

The Illumina sequencing technology, which is implemented in the Genome Analyzer and the HiSeq 2000, uses alternative methodologies, summarized in figure 4.1. Template DNA is fragmented and adapter-ligated (usually with a PCR step), then following an optional target enrichment step, the DNA is added to a flow cell in a cluster generation station. Cluster generation is the process of repeated bridge PCR amplification/denatu-

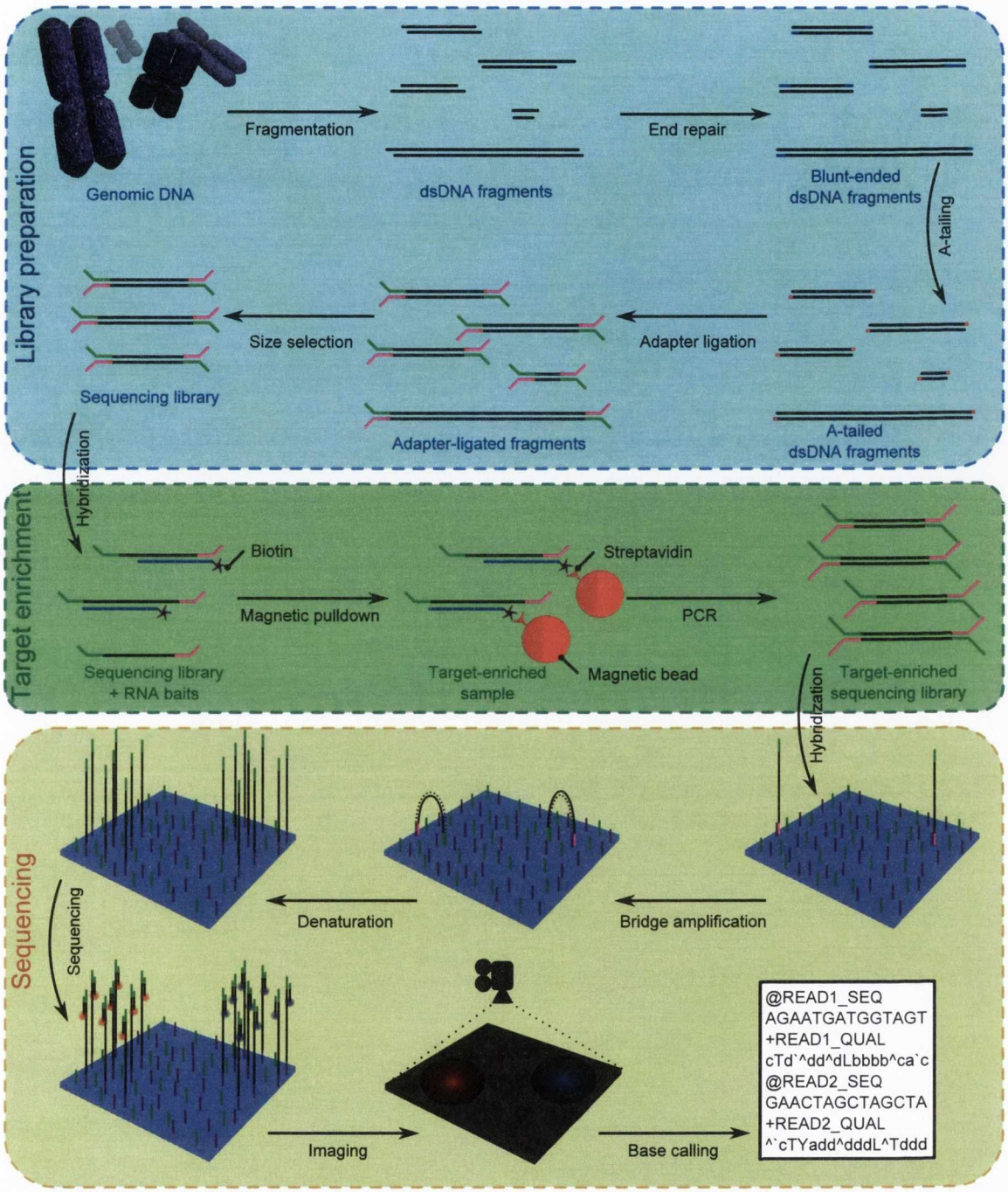


Figure 4.1: Overview of pipeline for DNA sequencing library preparation, target enrichment and next-generation sequencing using the Illumina sequencing method. dsDNA, double-stranded DNA; RNA, ribonucleic acid; PCR, polymerase chain reaction.

ration on the flowcell surface that results in single molecules being represented by many clonal copies in close proximity (a cluster). This way, when sequencing by synthesis is carried out, the fluorescence signals are detectable and consistent within a cluster, representing the sequence of bases being incorporated [170].

All NGS technologies generate an unavoidable amount of erroneous data, derived either during library preparation (for example, nucleotide misincorporation during PCR steps) or during the actual sequencing itself (for example, nucleotide misincorporation during sequencing by synthesis or base-calling error induced by uncertainty when reading the fluorescence signal due to overlapping emission spectra of fluorophores or run-time issues such as phasing and pre-phasing [171]). For this reason, quality scores are typically included with sequence reads representing the probabilities that base calls were made correctly. These scores can be used in downstream analyses to improve confidence in read mapping and variant calls.

Using NGS technologies, an enormous amount of data can be generated in a single experiment. However, for many sequencing experiments, the portion of the genome of interest is usually just a small fraction, and in such experiments the sequencing of the genome's entirety would be wasteful and potentially problematic in downstream analyses. For this reason, strategies for target enrichment have been devised that allow DNA samples to be refined down to just portions of the genome that are of interest. Such strategies have been developed to match the level of throughput that NGS affords and they permit the enrichment of many megabases of sequence. This is acheived through hybridization of probes that are complementary to the genomic intervals of interest, either through array-based technologies [172, 173] or in-solution technologies [174, 175].

Agilent's Sureselect technology [174] employs an in-solution capture method, using biotinylated 120-mer RNA probes designed to be complementary to the sequences of genomic target intervals. These probes can then easily be isolated using streptavidin-coated

magnetic beads; the strong biotin-streptavidin interaction permits specific enrichment of DNA that is bound to biotinylated RNA probes (figure 4.1). This in-solution method has been shown to be more effective than array-based methods in terms of sensitivity, specificity, uniformity and reproducibility [176].

4.1.1 Methodological and statistical considerations in the generation and analysis of NGS data

While NGS has multiple diverse applications (such as assessment of mRNA expression levels by RNA-seq [177] or investigation of protein-DNA interactions by ChIP-seq [178]), the focus of this work is on discovery of sequence variants in genomic DNA samples from patients with ALS. For this, NGS can be used to discover both single nucleotide variants (SNVs) and insertions/deletions (indels) through deep re-sequencing of candidate genomic intervals in several individuals. The design of such an experiment is dependent on many considerations derived from known capabilities and limitations of the technologies used, as well as expected findings and analysis methodologies.

An important consideration in the design of a project using NGS is target depth of coverage. In order that sequencing error or errors introduced in sample library preparation do not lead to false positive variant calls, NGS experiments are typically designed so that there is a large amount of redundancy in the coverage of aligned reads. This is also permissive to genotypes being called correctly, in particular for heterozygous variants and polymorphisms. Kenny *et al.* found that 8× coverage was sufficient to obatain 99% concordance with genotypes assayed using SNP arrays [179]. Conversely, Li *et al.* have demonstrated that a target depth of 4× coverage maximizes power to detect association of low-frequency variants with disease, given large sample sizes [180]. However, this may not be an optimal strategy if accurate inference of genotypes for individuals is of interest. In fact, for sequencing experiments where accurate genotype calls are required, the target

sequence depth is typically much higher. For a given mean sequencing depth, there will be substantial variation around this mean across the genome and for this reason, researchers typically aim for a target depth of 30× for genotype calling across many loci in many samples.

NGS sample preparation methodologies usually involve one or more PCR amplification steps so that a sufficient number of molecules are available to the sequencing machine. This results in a different kind of redundancy, where two sequenced molecules are representative of the same starting template molecule, thus providing no extra information and potentially being misleading in downstream analyses. Typically, such duplicate reads are removed in downstream processing of the data. However, these duplicate reads do have the potential to be informative of sequencing error, so the removal of duplicate reads in downstream sequence analysis steps would usually take this into account, keeping the read that is sequenced most accurately (determined by the mapping quality score of the read).

Given a set of mapped, unique sequence reads, variant calls can be made, for use in subsequent analyses. The benefit of NGS experiments is that, for an individual, the full complement of variants at a locus is identified (assuming adequate sequencing coverage), unlike GWAS which only assays genetic variation that is common within the population. This permits the assessment of rare variants, which are variants whose alternate allele has a frequency of less than 1%. However, if a particular rare variant is assumed to be the cause of a disease, sufficient sample numbers must be assayed in order to acheive adequate power to detect the variant at all. Conversely, if many individuals are sequenced, a large number of rare variants will be identified, and determining which variants are important from a disease perspective and which are simply background genetic variation becomes an issue. For this reason, a control dataset for comparison, representing background genetic variation is useful. The public data releases from the 1000 Genomes Project [20] are useful datasets for such purposes, although ideal comparative data would be derived from the

same population as the case cohort.

Nevertheless, the very large number of variants discovered in a large sequencing study, even with representative population-based controls, may prohibit the identification of obvious disease-causing variants. On the other hand, rare variants are in their nature less likely to be discovered in a restricted set of individuals, so it is possible that many pathogenic variants could easily be missed. For these reasons, given the hypothesis that multiple rare variants could be the cause of many complex genetic diseases, a prudent approach is to consider not just single variants that are discovered in the case cohort, but the sum of variants discovered within the bounds of a locus, for example, a gene. This approach is termed burden analysis, for which several techniques have been forwarded [181–184]. Such analyses benefit most from a well-matched control population.

4.1.2 Research aims

This chapter represents the work carried out in the initial case-only phase of a large project involving the sequencing of a set of candidate ALS genes in ALS cases and controls. The ultimate goal is to investigate rare genetic variation within ALS patients and identify disease-causing variants that lie within regions that have shown some evidence of potentially being involved in ALS aetiology. Candidate genes are chosen based on results from chapter 3 and are sequenced using the Illumina sequencing method in a cohort of Irish ALS patients. 1000 Genomes Project data are used for comparison.

4.2 Methods

4.2.1 Design of RNA sequence capture library

In order to enrich genomic DNA samples for exonic intervals of interest, an RNA bait library for in-solution target enrichment (Agilent Sureselect) was designed. Included in the design were exons of 395 genes that overlapped with associated intervals identified

Table 4.1: Genes included in the target gene set due to prior evidence linking to ALS

Symbol	Name	Evidence
ALS2	Alsin	[57,58]
ANG	Angiogenin	[46]
C9orf72	Chromosome 9 open reading frame 72	[66–68,121,122]
CHMP2B	Charged multivesicular body protein 2b	[185,186]
DCTN1	Dynactin	[47–49]
DPP6	Dipeptidyl aminopeptidase-like protein 6	[98,109,187]
ELP3	Elongator complex protein 3	[188]
FGGY	FGGY carbohydrate kinase domain containing	[106]
FIG4	Polyphosphoinositide phosphatase	[189]
FUS	Fused in sarcoma	[51,52]
GRN	Progranulin	[190]
HFE	Human hemochromatosis protein	[191,192]
IFNK	Interferon kappa	[66–68]
ITPR2	Inositol 1,4,5-trisphosphate receptor, type 2	[107]
MAPT	Microtubule-associated protein tau	[193]
MOBKL2B	Mps One Binder kinase activator-like 2B	[66–68]
NEFH	Neurofilament, heavy polypeptide	[194,195]
NEFL	Neurofilament, light polypeptide	[194,195]
NEFM	Neurofilament, medium polypeptide	[194,195]
NIPA1	Non-imprinted in Prader-Willi/Angelman syndrome	[109]
OPTN	Optineurin	[53–55]
PARK7	Parkinson disease protein 7	[196]
PON1	Paraoxonase 1	[197–200]
PON2	Paraoxonase 2	[197–200]
PON3	Paraoxonase 3	[197–200]
PRPH	Peripherin	[201–203]
SETX	Senataxin	[59–61]
SIGMAR1	Sigma non-opioid intracellular receptor 1	[62]
SMN1	Survival motor neuron protein 1	[204]
SMN2	Survival motor neuron protein 2	[204]
SOD1	Superoxide dismutase 1	[42,43]
SPG11	Spatacsin	[205]
TARDBP	TAR DNA-binding protein 43	[33,50]
UNC13A	Unc-13 homolog A	[67,119]
VAPB	VAMP-associated protein B	[63]
VCP	Valosin-containing protein	[69]

in section 3.3.2, 127 genes that overlapped with reccurrent ALS-specific deletions identified in section 3.3.3 and 1,411 genes that overlapped with recurrent ALS-specific runs of homozygosity identified in section 3.3.4. Extensive overlap between association, CNV and ROH datasets reduced the final number of candidate genes to 1,577. The sequencing experiment was also used as an opportunity to conduct a comprehensive Irish population-based screen of genes that have previously been implicated in ALS. Table 4.1 shows the known or suspected ALS genes that were included in the target set.

For all genes in the target set, genomic coordinates of exons for every known transcript

of each RefSeq gene (GRCh37 build) were downloaded from the UCSC genome browser [154]. The script `parse_exons.pl` was used to convert the output from the UCSC browser to lists of exon start and end positions, one list for genes on the positive strand and one list for the negative strand. Overlapping exonic coordinates representing alternate transcripts were then consolidated into single intervals using `interval_overlap.pl`. The resulting files containing genomic intervals for 2.77 Mb of target exons were uploaded to Agilent's online eArray tool (erray.chem.agilent.com/earray) so that 120-mer RNA probes could be designed against the human genome reference sequence. To avoid capture of repetitive sequence across the genome, the recommended maximum of 20 base pairs of overlap with repeat intervals identified by RepeatMasker was observed. This often resulted in the exclusion of large regions that could be rescued by more careful bait placement than the automated solution implemented by eArray. Such gaps were tolerated in the large candidate ALS gene set, but for screening the known or suspected ALS genes in table 4.1 this was suboptimal, so the script `rescue_intervals.pl` was used to identify gaps and attempt improved baiting for known or suspected ALS genes. Baits were tiled across target regions at 2× frequency. Occasionally, however, only one bait was generated for an interval so to avoid potential bias introduced by this the script `double_up_singletons.pl` identified these intervals and designed a second bait overlapping the singleton.

4.2.2 Subjects

A total of 106 individuals were chosen for sequencing based on availability and quality of DNA. Samples were also chosen to maximise the potential to detect variants; an attempt was made to have a representative sample from as many of the homozygous and CNV groups as possible as well as choosing individuals that were driving the association signals. All individuals had previously been genotyped in the GWAS projects (chapter 3 and [98]) and all patients had clinically definite or probable ALS [28], as determined by a neurologist

with expertise in ALS. The mean age of onset of the 106 individuals was 64.1 years and the ratio of spinal onset:bulbar onset ALS was 74%:26%. Four of the cases had familial ALS, as determined by patient-reported family history. The study was carried out as part of a larger project approved by the Beaumont Hospital ethics committee.

4.2.3 Sequencing library preparation and target enrichment

Sequencing libraries were constructed using genomic DNA from patients in preparation for multiplexed high-throughput resequencing of targeted exonic regions, using the multiplexing method of Craig *et al.* [206]. This method introduces, for each sample within a multiplexed pool, a unique 6-mer sequence to the 5′ end of the template DNA molecule in the adapter ligation step, which is consequently read by the sequencing machine and included at the beginning of each sequence read for that sample. A ten-sample pilot was first run on a single lane of an Illumina Genome Analyzer to confirm target depth of coverage (30×) was acheivable with the methods of multiplexing and target enrichment. Based on this, subsequent samples were prepared as 24-plexed libraries.

Sequencing libraries were prepared according to established protocols (figure 4.1, [207]). Briefly, for each sample, 1 µg of DNA was fragmented using either the Covaris™ adaptive focussed acoustics system or NEBNext™ dsDNA Fragmentase™. Fragmented DNA molecules were end-repaired using a cocktail of T4 polymerase, Klenow polymerase I, T4 polynucleotide kinase and, in the case of Fragmentase™, *E. coli* DNA ligase for Fragmentase™. A 5′ adenine overhang was then added using Klenow polymerase I fragment (3′ ⟶ 5′ exo-) and barcoded sequencing adapters (10 picomoles, table 4.2) were ligated to the A-tailed DNA fragments using Quick T4 DNA Ligase. All enzymes were purchased from New England Biolabs (Massachusetts, USA). Between steps, DNA was purified from the enzyme reactions using Agencourt® Ampure® XP beads (Beckman Coulter, California, USA).

Table 4.2: Adapter and primer sequences used in sequencing library preparation

Description	Sequence
Adapter 1	5′-p-XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
Adapter 2	5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXX*T-3′
Primer 1	5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC-3′
Primer 2	5′-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT-3′

Adapter1 and 2 were duplexed by gradual cooling from 99 °C
p 5′ phosphate group
* Phosphorothioate bond
XXXXX unique index for adapter

Table 4.3: PCR cycle conditions used after library preparation and target enrichment

Post-library preparation

°C	m:s	
98	0:30	
98	0:10	×5–7
65	0:30	
72	0:30	
72	10:00	
4	hold	

Post-target enrichment

°C	m:s	
98	0:30	
98	0:10	×11–13
57	0:30	
72	0:30	
72	7:00	
4	hold	

Adapter-ligated libraries were size-separated by electrophoresis on a 1.5% low melting point agarose gel supplemented with 1×SYBR green (Invitrogen, California, USA) and libraries sized between 300 and 400 bp were excised. DNA was subsequently purified from the gel using a Qiagen gel extraction kit (Qiagen GmbH, Hilden, Germany). Size-selected libraries were then amplified with between five and seven cycles of PCR using the primers detailed in table 4.2 and the PCR cycle conditions described in table 4.3.

Prepared sequencing libraries were quantified using the Quant-iT™ high-sensitivity assay performed with a Qubit® spectrophotometer (Invitrogen, California, USA) and pooled in equimolar quantities totalling 500 ng. Pooled samples were then subjected to target enrichment using the Agilent SureSelect RNA library described in section 4.2.1 according to manufacturer's protocol.

4.2.4 High-throughput resequencing of targeted exons

Target-enriched multiplexed sequencing libraries were quantified with a Qubit® spectrophotometer using the Quant-iT™ high-sensitivity assay, as well as the DNA-1000 assay

on an Agilent Bioanalyzer. Libraries were standardized to 5 µM concentration and provided for 80 bp paired-end sequencing on an Illumina Genome Analyzer at the TrinSeq facility in the Trinity Centre for Health Sciences, Saint James's Hospital, Dublin.

4.2.5 Sequence alignment and processing

Sequence data were provided as large text files in the FASTQ format [208]. This format lists sequence reads, four lines per read, where the first and third lines are headers (unique identifiers for the read), the second line contains the sequence itself and the fourth line contains an encoded quality string representing the Phred-scaled [209] probability that the base was called incorrectly by the sequencing machine. For each base position, the quality of the base call is represented by a character in standard American Standard Code for Information Interchange (ASCII) format.

Using these FASTQ files, data were split according to the unique barcode string present at the start of each read using the script `split_seq_data.pl`. This script creates an internal lookup table of identifiers for each barcode and matches the 6-mer barcode at the start of forward sequencing reads to individuals, outputting one new FASTQ file per individual. If a read's barcode does not match any barcodes in the lookup table, the script checks the reverse read's barcode for matches and outputs to the relevant file. If neither barcode has a match, the script outputs to a file that catches sequencing reads of unknown source.

Sequence reads for each individual were aligned to the hg19 build of the human genome using the Burrows-Wheeler Aligner (BWA) [210]. First, the `aln` method was implemented, finding the suffix array coordinates of the sequencing reads. In this step, the `-q` switch was set to 20, which results in the sequence reads being soft clipped according to the result of $argmax_x \sum_{i=x+1}^{l}(Q - q_i)$, where Q is the user-specified quality threshold, l is the read length and q_i is the i^{th} base position's quality score. Thus, only high-quality portions of

reads were considered in subsequent analyses. The `sampe` method of BWA was then used to generate paired-end alignments in the sequence alignment/map (SAM) format [211]. Individual SAM files were converted to compressed binary SAM files (BAM files) using SAMtools [211] and merged to form single BAM files, one per lane of sequencing, using Picard (http://picard.sourceforge.net).

To reduce false-positive SNPs resulting from poor alignment around insertions or deletions (indels) and to improve the potential to detect indels, local realignment was performed around clusters of SNPs as well as at known sites of indels using the Genome Analysis Toolkit (GATK) [212]. Firstly, the `IndelTargetCreator` algorithm was applied to the merged BAM files to generate target intervals for realignment, then the `IndelRealigner` method was implemented to find the optimal alternate alignment for problematic reads.

Following realignment, duplicate reads were removed using `MarkDuplicates`, a method within Picard which identifies and removes reads that have identical start and end coordinates, keeping only the read with the best mapping quality. This is necessary because the PCR steps involved in sequencing library preparation result in many copies being made of the same starting molecule; thus, if more than one of these copies is sequenced the two resulting reads are only truly representative of one sequence from the starting genome.

The base quality scores in the resulting unique, realigned BAM file were recalibrated using the GATK methods `CountCovariates` and `TableRecalibration` to remove any systematic bias in the assignment of quality scores and to generate a quality score distribution that is more representative of the underlying variance within quality scores. Four covariates were used for recalibration: dinucleotide, read group (groups of per-individual, per-sequencing run reads), quality score and machine cycle number. The distribution of quality scores were assessed before and after recalibration using GATK's `AnalyzeCovariates` tool.

4.2.6 Variant calling and annotation

The resulting high-quality sequence reads from section 4.2.5 were used to call variants in the sequence by implementing the `mpileup` method in SAMtools. This produced a variant call format (VCF) file [213], which was filtered for variant call score, keeping only variant calls with quality score greater than 20 (representing a 1% probability that the variant call is incorrect), which was then passed to ANNOVAR to annotate variants based on their positions within genes (http://www.openbioinformatics.org/annovar/). Variants were defined as non-silent if they were expected to alter protein structure based on gene annotation. This comprised all variant annotations except synonymous single nucleotide variants (SNVs). In the assessment of putative variant pathogenicity, only non-silent variants were considered.

As a comparison dataset, the 1000 Genomes Project's [20] October 2011 integrated variant call set (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/) was used extensively to assess the frequency of ALS variant calls in an alternative population of 1,092 ethnically diverse individuals. Variants discovered in the ALS dataset were defined as rare if, within the 1000 Genomes dataset, they were either not present or present at a frequency of less than 1%. This definition was used to refine the call set to variants that are rare globally using the script `filter_1kg.pl`. A comparison dataset of rare variants in the 1000 Genomes data was also produced this way and this was used with the script `count_variants_per_gene.pl` to compare the number of rare variants discovered in the ALS dataset to the number of rare variants discovered in the 1000 Genomes data. This permitted a speculative burden analysis on the data to reveal genes that have improbably high numbers of rare variants in the ALS dataset.

To identify possible recessive ALS-causing variants, intervals that were mapped homozygously in section 3.3.4 were passed to the script `find_homies.pl` to find homozygous genotypes for individuals in which the ROHs were originally mapped. Because minor al-

leles segregate homozygously only q^2 of the time, where q is the MAF, frequency-based variant filtration was not deemed necessary prior to this step. Variants were flagged as potential recessive ALS-causing mutations if they were present homozygously in the individuals in which the ROHs were mapped, and either not present in the 1000 Genomes dataset, or present only heterozygously.

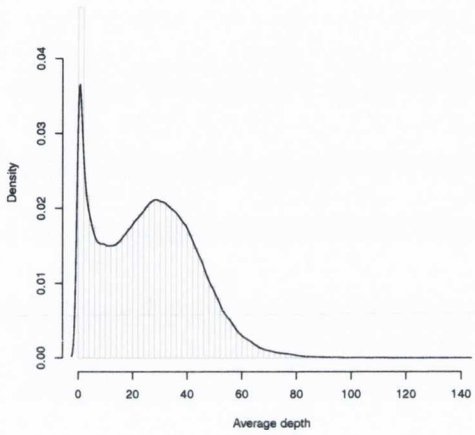
4.3 Results

4.3.1 Sequence alignment and processing

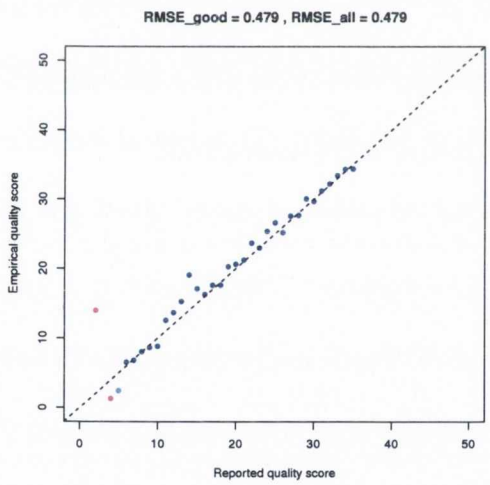
Following successful alignment of 529 million reads (95.46% of total) to the human genome, local realignment around indels and removal of reads marked as PCR duplicates (14% of alignable reads), the mean peak non-zero coverage for target genomic regions was 28.5× (figure 4.2(a)), representing an on-target rate of 25%. Base quality score recalibration successfully adjusted quality scores to be more representative of the bases' probabilities of mismatching the reference allele at variant sites (figures 4.2(b) and 4.2(c)).

4.3.2 Variant calling

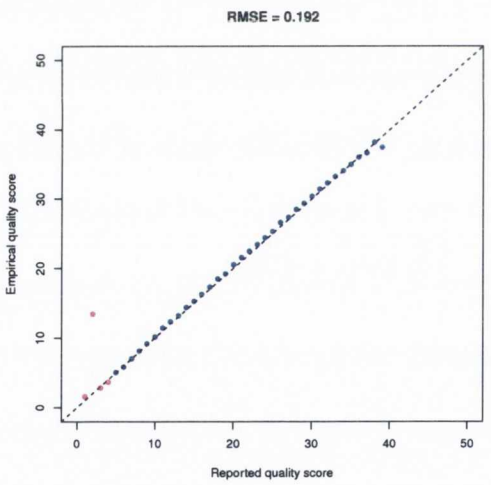
After initial calling of 6,903 non-reference sequence variants in the target gene set, 346 were discarded based on sub-threshold quality score, leaving a total of 6,557 high-quality variants within target intervals (figure 4.3(a)). Figure 4.3(b) shows the allele frequency spectrum of the remaining high-quality variants, demonstrating an excess of rare variants called in the target gene set. Of the high-quality variants, 103 were indels and 6,454 were single nucleotide variants. Figure 4.3(c) shows the relative quantities of variants based on annotation. 641 of the discovered variants were polymorphisms that are included in the Illumina genome-wide SNP dataset generated in chapter 3; genotypes for these SNPs with greater than eight supporting high-quality base calls [179] were used to check for concordance between the GWAS dataset and the current NGS dataset. A mean (± SD) of



(a) Coverage



(b) Before base quality score recalibration



(c) After base quality score recalibration

Figure 4.2: Statistics following sequence alignment and processing. (a) Mean coverage within target regions for all samples. The black line shows the kernel density estimate generated using R's `density` function (bandwidth=0.8043). (b) and (c) Example plots, for one representative individual, of observed versus expected base quality score before and after base quality score recalibration using GATK.

99.92 ± 0.15 % concordance for genotypes was observed, and no individual showed lower than 99.12% concordance.

4.3.3 Assessment of candidate disease variants

Rare variants

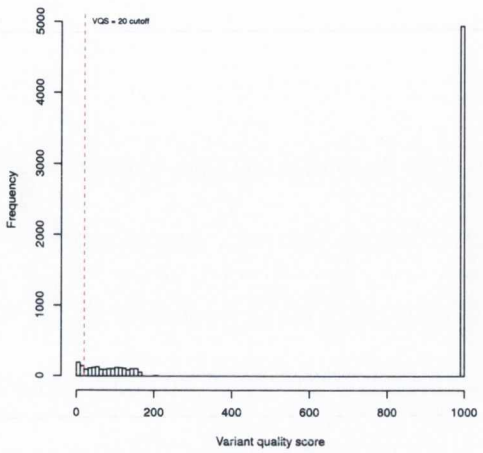
After only keeping variants at less than 1% in the 1000 Genomes dataset, there were 3001 remaining variants in 917 genes in the ALS dataset, of which 1036 were synonymous SNVs. The majority of these variants were also below 1% alternate allele frequency in the ALS dataset (figure 4.4(a)), and figure 4.4(b) shows a similar distribution of variant annotation to the unfiltered variant set shown in figure 4.3(c). The subset of rare variants discovered in the ALS dataset that are expected to affect protein structure at more than one amino acid are detailed in table 4.4.

Burden of rare variants in candidate genes

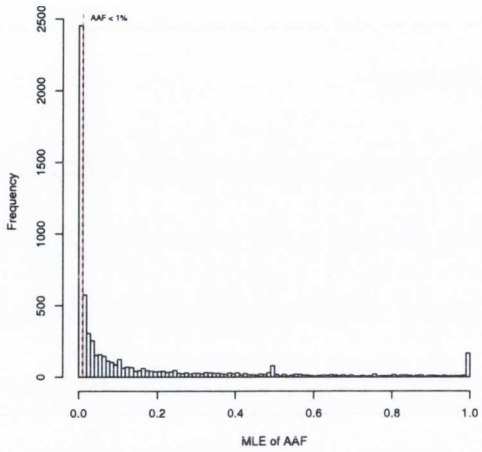
The relative quantity of rare variants discovered in the ALS dataset compared to the 1000 Genomes dataset was assessed for each gene. Figure 4.5 shows the ratio of rare variants in the ALS set to rare variants in the 1000 Genomes dataset, where both values are standardized to the length of the gene and the number of individuals sequenced in each experiment. Twelve genes were identified that had, in ALS, greater than 10× the standardized number of rare variants. These genes are detailed in table 4.5. One gene, *HYDIN*, stood out in particular; for this gene, every rare non-silent variant discovered in the ALS dataset is detailed in table 4.6.

Putative recessive disease-causing mutations

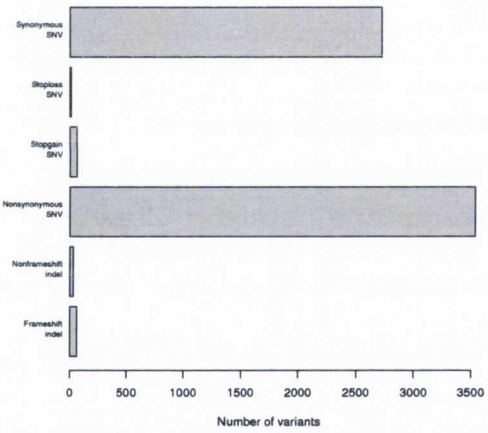
For homozygous intervals identified in section 3.3.4, genotypes for non-silent variants discovered in the ALS dataset were investigated for putative recessive disease-causing in-



(a) Variant quality score spectrum



(b) Alternate allele frequency spectrum



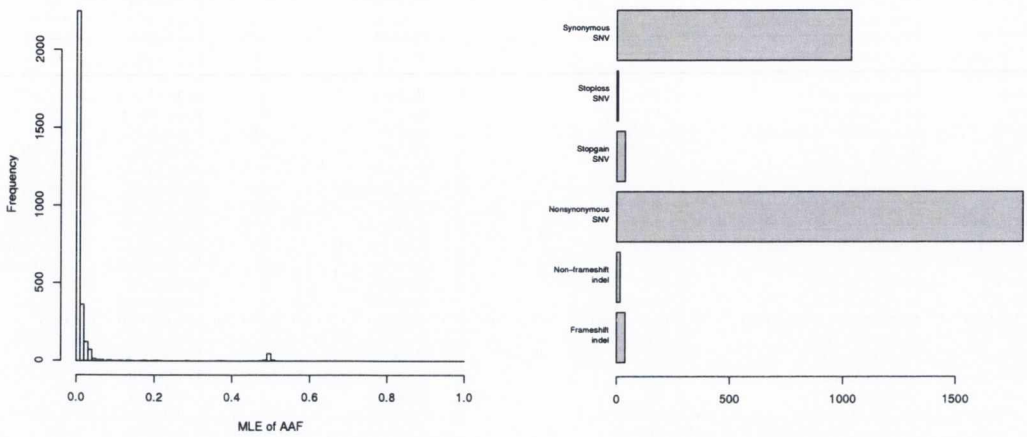
(c) Variant annotation

Figure 4.3: Characteristics of variants called in target intervals. (a) Distribution of variant quality scores determined by SAMtools `mpileup` method. The majority of variant calls were very high quality and for the remainder a cutoff for marginal quality calls was set at 20; 95.0% of variants were above this threshold. VQS, variant quality score. (b) Distribution of alternate allele frequencies in high-quality variant calls. A large proportion (37.4%) of discovered variants were rare (alternate allele frequency $< 1\%$) and there was a slight excess (2.6%) of variants with alternate allele frequency greater than 99%, which possibly represent variants for which the reference allele is actually the minor allele. MLE, maximum likelihood estimate; AAF, alternate allele frequency. (c) Relative quantities of six different types of variant. SNV, single nucleotide variant.

Table 4.4: Rare variants expected to alter protein structure at several amino acids

Gene	Type	Mutation	Coding change	Chr	Pos	AAF
ACADL	Frameshift deletion	1147_1148del	383_383del	2	211057579	0.004721
ACO2	Frameshift insertion	1926_1927insC	P642fs	22	41922430	0.1333
ADAM20	Stopgain SNV	G1814A	W605X	14	70989811	0.004717
ANKRD30A	Stopgain SNV	G328T	E110X	10	37419292	0.02367
ANKRD30A	Frameshift deletion	3771_3772del	1257_1258del	10	37508579	0.01415
AOAH	Stopgain SNV	G1453T	G485X	7	36561695	0.005455
ATF6	Stopgain SNV	C1294T	Q432X	1	161816345	0.004718
BEST3	Frameshift deletion	173delA	Q58fs	12	70088224	0.06137
BMP2K	Stopgain SNV	C3481T	Q1161X	4	79833182	0.00472
BTBD8	Stopgain SNV	T344G	L115X	1	92554449	0.004718
C10orf68	Frameshift deletion	1477_1478del	493_493del	10	33136820	0.08072
C10orf68	Stopgain SNV	C21G	Y7X	10	32873210	0.009704
C1orf168	Stopgain SNV	G1420T	E474X	1	57209907	0.004902
CALML4	Frameshift deletion	581delG	R194fs	15	68486363	0.004744
CCDC141	Frameshift deletion	282delG	E94fs	2	179843346	0.004735
CCDC66	Stopgain SNV	C1381T	R461X	3	56628033	0.009437
CENPE	Stopgain SNV	G6145T	E2049X	4	104061005	0.004721
CNTNAP5	Frameshift insertion	197_198insG	G66fs	2	124999786	0.032
CPA4	Stopgain SNV	G678A	W226X	7	129948221	0.004892
DEPDC5	Stopgain SNV	G2055A	W685X	22	32218727	0.02109
DEPDC5	Stopgain SNV	C1699T	R567X	22	32215040	0.004982
DNAH11	Stopgain SNV	C7301G	S2434X	7	21765441	0.00472
DNAH11	Stoploss SNV	A9957T	X3319C	7	21828869	0.004956
DNAH14	Frameshift deletion	1867delT	S623fs	1	225231640	0.00961
DNAH14	Stopgain SNV	G10171T	E3391X	1	225528175	0.009434
DNAH14	Stoploss SNV	A13548T	X4516Y	1	225586971	0.004736
DNAH14	Stopgain SNV	C409T	R137X	1	225152222	0.009627
DYSF	Frameshift insertion	3840_3841insC	G1280fs	2	71827918	0.1143
EFCAB4B	Frameshift deletion	2132_2133del	711_711del	12	3724584	0.00584
FAM111A	Frameshift insertion	791_792insGCAGATACTT	F264fs	11	58919932	0.009434
FMO3	Stopgain SNV	G913T	E305X	1	171083232	0.00472
GBP5	Frameshift insertion	500_501insCTGA	A167fs	1	89732765	0.009453
GBP5	Frameshift deletion	1314_1315del	438_439del	1	89729466	0.004736
HYDIN	Stopgain SNV	C1204T	R402X	16	71163647	0.004717
HYDIN	Stoploss SNV	T3052C	X1018Q	16	71061495	0.4985
HYDIN	Frameshift deletion	11708delT	I3903fs	16	70896017	0.4931
IFNK	Frameshift insertion	37_38insTTGT	W13fs	9	27524371	0.03678
KLLN	Frameshift deletion	339_340del	113_114del	10	89621905	0.0093
KRT76	Stopgain SNV	G826T	E276X	12	53167416	0.009758
LGSN	Frameshift insertion	1515_1516insA	L505fs	6	63989941	0.00481
LOC729020	Stopgain SNV	C178T	Q60X	10	105005931	0.004723
LRBA	Stoploss SNV	T3178G	X1060E	4	151773684	0.004717
LRP1B	Frameshift substitution	4170_4170TAA,	NA	2	141625832	0.03042
MAN2A2	Stopgain SNV	C1234T	Q412X	15	91452594	0.005143
MME	Frameshift deletion	467delC	P156fs	3	154834480	0.00488
MS4A14	Frameshift insertion	1905_1906insA	K635fs	11	60184346	0.004752
NDUFA6	Frameshift deletion	35_36del	12_12del	22	42486791	0.00844
OLFM4	Stopgain SNV	C640T	R214X	13	53617309	0.01457
OR4C45	Frameshift insertion	767_768insCT	P256fs	11	48367052	0.5
OR5M3	Stopgain SNV	C421T	R141X	11	56237553	0.004717
OR5T3	Frameshift deletion	999_1000del	333_334del	11	56020674	0.004719
OR8I2	Frameshift substitution	523_525CT,	NA	11	55861306	0.1038
OR8K3	Stopgain SNV	C778T	Q260X	11	56086560	0.02357
PAPSS2	Frameshift deletion	1759delG	E587fs	10	89505641	0.004792
PAPSS2	Frameshift deletion	381_381del	127_127del	10	89473067	0.03096
PDE11A	Frameshift deletion	907delT	F303fs	2	178681636	0.004723
PDE11A	Stopgain SNV	C169T	R57X	2	178879181	0.01417
PEG10	Stoploss SNV	T1717C	X573Q	7	94294357	0.004731
PKD1L3	Frameshift insertion	3689_3690insAACA	Q1230fs	16	71981420	0.7494
PKHD1L1	Stopgain SNV	C9124T	R3042X	8	110491814	0.004723
PON3	Stopgain SNV	C94T	R32X	7	95024007	0.004772
PPARD	Frameshift insertion	759_760insCA	T253fs	6	35392237	0.005751
PRKCH	Stopgain SNV	C811T	R271X	14	61917668	0.004719
PRSS48	Frameshift insertion	131_132insGTCAG	S44fs	4	152201026	0.5413
PTH2R	Stopgain SNV	C245A	S82X	2	209302328	0.03304
PTH2R	Frameshift deletion	594_597del	198_199del	2	209308157	0.004721
PTPMT1	Stoploss SNV	G605C	X202S	11	47593180	0.009652
RFX3	Stopgain SNV	G2000A	W667X	9	3248000	0.004723
SCIN	Stopgain SNV	G1212A	W404X	7	12689163	0.004733
SFI1	Stopgain SNV	C235T	R79X	22	31924818	0.004719
SLC13A1	Stopgain SNV	C34T	R12X	7	122839967	0.004719
SLC13A1	Stopgain SNV	G144A	W48X	7	122821111	0.004717
SLC17A2	Stopgain SNV	G265T	E89X	6	25921616	0.004782
SLC17A4	Stopgain SNV	C1297T	Q433X	6	25778182	0.009437
SLC9B1	Stopgain SNV	C913T	R305X	4	103832611	0.2263
SOAT2	Frameshift deletion	798delC	P266fs	12	53512153	0.004771
TTC9	Frameshift insertion	338_339insG	G113fs	14	71109184	0.2899
TTN	Stopgain SNV	C12190T	R4064X	2	179598098	0.004812
VWDE	Frameshift deletion	3644delG	C1215fs	7	12395838	0.004734
XIRP2	Frameshift deletion	3802_3803del	1268_1268del	2	168102370	0.004717
ZNF187	Frameshift insertion	236_237insG	C79fs	6	28239933	1
ZNF839	Frameshift deletion	524_525del	175_175del	14	102792557	0.004743

Chr, chromosome
Pos, base pair position in GRCh37 coordinates
AAF, alternate allele frequency
SNV, single nucleotide variant
NA, not applicable



(a) Alternate allele frequencies of rare variants

(b) Annotation of rare variants

Figure 4.4: Characteristics of rare ALS variants with alternate allele frequencies < 1% in 1000 Genomes data. MLE, maximum likelihood estimate; AAF, alternate allele frequency; SNV, single nucleotide variant.

Table 4.5: Genes demonstrating a burden of rare variants in ALS (10× more rare variants per base per person in ALS)

Gene	Length (bp)	n_{ALS}	n_{1kg}	$Score_{ALS}$	$Score_{1kg}$	Ratio
CREG2	3488	2	2	1.89×10^{-2}	1.83×10^{-3}	10.3
GABPB1	1635	1	1	9.43×10^{-3}	9.16×10^{-4}	10.3
GOLGA6B	3178	1	1	9.43×10^{-3}	9.16×10^{-4}	10.3
HYDIN	15685	49	20	4.62×10^{-1}	1.83×10^{-2}	25.2
OR14J1	966	3	3	2.83×10^{-2}	2.75×10^{-3}	10.3
OR1D5	939	2	1	1.89×10^{-2}	9.16×10^{-4}	20.6
OR4C3	990	16	8	1.51×10^{-1}	7.33×10^{-3}	20.6
OR4C45	919	11	3	1.04×10^{-1}	2.75×10^{-3}	37.8
OR8U1	930	8	8	7.55×10^{-2}	7.33×10^{-3}	10.3
OR8U8	906	1	1	9.43×10^{-3}	9.16×10^{-4}	10.3
RPS3	841	2	2	1.89×10^{-2}	1.83×10^{-3}	10.3
TTC9	5217	2	1	1.89×10^{-2}	9.16×10^{-4}	20.6
WDR26	6872	1	1	9.43×10^{-3}	9.16×10^{-4}	10.3

n_{ALS}, number of rare variants discovered in the ALS cohort
n_{1kg}, number of rare variants discovered in the 1000 Genomes cohort
$Score_{ALS}$, $Score_{1kg}$, Number of rare variants discovered ÷ (n×length of gene)
Ratio, $Score_{ALS} \div Score_{1kg}$

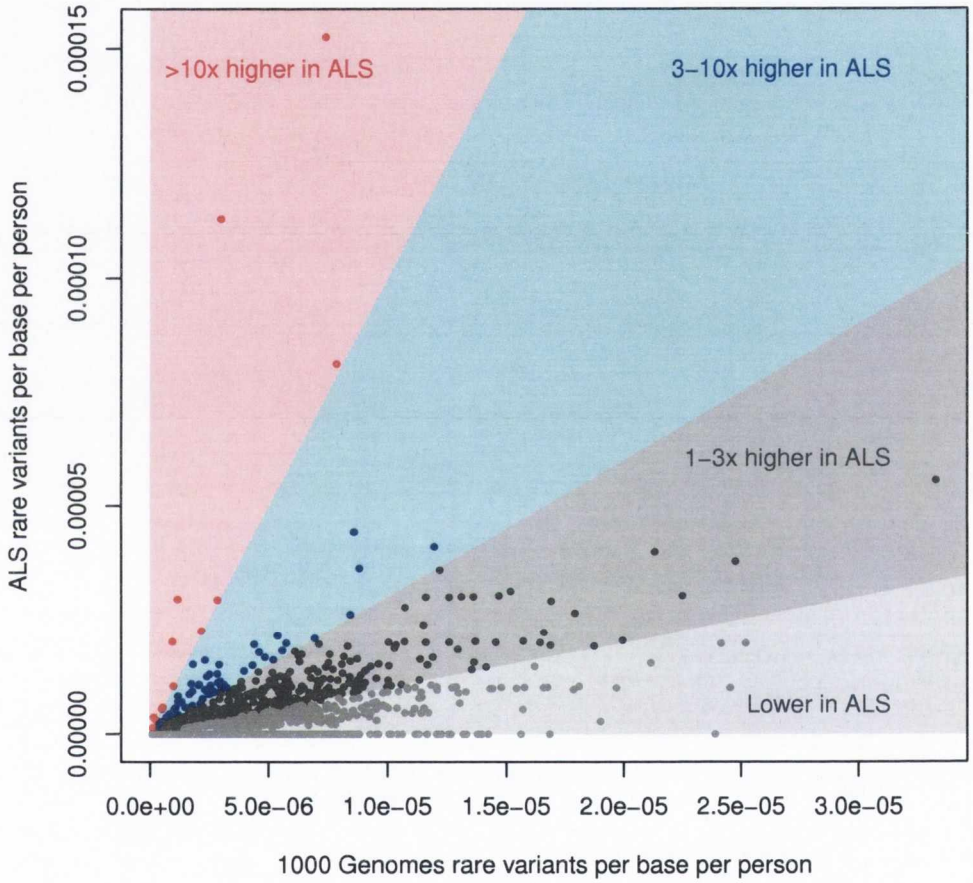


Figure 4.5: Burden of non-silent rare variants in ALS. Each point represents one gene, where the number of rare variants per gene has been divided by the total coding sequence length for the gene multiplied by the number of individuals that were sequenced in either the current ALS project or the 1000 Genomes Project.

Table 4.6: Variants discovered in *HYDIN* that are rare or not present in 1000 Genomes data

Mutation	Type	Pos	Coding change	Ref	Alt	AAF
G13535A	Nonsynonymous SNV	70867931	R4512H	C	T	0.005
A12869T	Nonsynonymous SNV	70883630	K4290M	T	A	0.005
C12805T	Nonsynonymous SNV	70883694	H4269Y	G	A	0.052
G12475C	Nonsynonymous SNV	70884524	E4159Q	C	G	0.500
A12260G	Nonsynonymous SNV	70891640	K4087R	T	C	0.176
G12121A	Nonsynonymous SNV	70893976	A4041T	C	T	0.023
G12073A	Nonsynonymous SNV	70894024	A4025T	C	T	0.208
A12010G	Nonsynonymous SNV	70894087	T4004A	T	C	0.467
11708delT	Frameshift deletion	70896017	I3903fs	GAA	GA	0.493
G11692A	Nonsynonymous SNV	70896033	V3898M	C	T	0.160
G11635A	Nonsynonymous SNV	70896090	D3879N	C	T	0.010
T11603G	Nonsynonymous SNV	70896122	M3868R	A	C	0.436
G11515C	Nonsynonymous SNV	70897039	V3839L	C	G	0.496
G11429A	Nonsynonymous SNV	70900111	R3810H	C	T	0.005
C11239T	Nonsynonymous SNV	70902541	R3747C	G	A	0.043
G11212A	Nonsynonymous SNV	70902568	A3738T	C	T	0.495
C10438T	Nonsynonymous SNV	70913316	R3480W	G	A	0.024
G10402A	Nonsynonymous SNV	70913352	A3468T	C	T	0.005
G10223A	Nonsynonymous SNV	70913649	R3408H	C	T	0.005
T9944C	Nonsynonymous SNV	70917855	L3315P	A	G	0.503
G9868C	Nonsynonymous SNV	70917931	A3290P	C	G	0.503
C9344G	Nonsynonymous SNV	70926334	T3115R	G	C	0.241
C8792T	Nonsynonymous SNV	70937582	P2931L	G	A	0.030
T8078G	Nonsynonymous SNV	70942688	I2693S	A	C	0.497
A7763G	Nonsynonymous SNV	70954513	K2588R	T	C	0.475
G7705A	Nonsynonymous SNV	70954571	D2569N	C	T	0.482
G7670A	Nonsynonymous SNV	70954606	G2557E	C	T	0.479
A7585G	Nonsynonymous SNV	70954691	K2529E	T	C	0.207
7545_7559del	Non-frameshift deletion	70954717	2515_2520del	GTGCGCTCCTTCTCC	NA	0.263
A7331T	Nonsynonymous SNV	70954945	N2444I	T	A	0.496
C6889G	Nonsynonymous SNV	70972620	R2297G	G	C	0.458
A6824G	Nonsynonymous SNV	70975565	Q2275R	T	C	0.213
A6427G	Nonsynonymous SNV	70986425	K2143E	T	C	0.034
C6256T	Nonsynonymous SNV	70989335	R2086C	G	A	0.507
T6136A	Nonsynonymous SNV	70993553	S2046T	A	T	0.116
G6050A	Nonsynonymous SNV	70993639	R2017H	C	T	0.068
G5852A	Nonsynonymous SNV	70995975	R1951Q	C	T	0.043
G5149A	Nonsynonymous SNV	71007809	V1717M	C	T	0.071
G3682C	Nonsynonymous SNV	71026076	V1228L	C	G	0.500
A3291G	Nonsynonymous SNV	71054116	I1097M	T	C	0.005
T3052C	Stoploss SNV	71061495	X1018Q	A	G	0.499
C2795T	Nonsynonymous SNV	71065636	T932I	G	A	0.005
A2149G	Nonsynonymous SNV	71101200	T717A	T	C	0.506
C1832A	Nonsynonymous SNV	71103393	T611N	G	T	0.019
T1763C	Nonsynonymous SNV	71113844	I588T	A	G	0.009
C1204T	Stopgain SNV	71163647	R402X	G	A	0.005
A1121G	Nonsynonymous SNV	71171057	Y374C	T	C	0.005
G856C	Nonsynonymous SNV	71186628	E286Q	C	G	0.005
G212A	Nonsynonymous SNV	71220668	R71Q	C	T	0.005

Pos, base pair position in GRCh37 coordinates
Ref, reference allele
Alt, alternate allele
AAF, alternate allele frequency
SNV, single nucleotide variant
NA, not applicable

heritance by searching for genotypes for the same variants in the 1000 Genomes dataset. Variants that were present in the ALS dataset in the correct individuals within the correct intervals, but that were not discovered in the 1000 Genomes dataset or that had no homozygous genotypes in the 1000 Genomes data, are detailed in table 4.7.

4.3.4 Assessment of rare variants in genes previously implicated in ALS

Rare, non-silent variants in known or suspected ALS genes (table 4.1) that were identified in the ALS dataset are detailed in table 4.8. For the majority of genes in table 4.1, no rare variants were discovered. Of the variants in table 4.8, only one has previously been shown to be involved in ALS (the G59S mutation in *DCTN1* [47]).

Table 4.7: Recessive variants in candidate regions not present homozygously in 1000 Genomes data

Gene	Type	Chr	Pos	Ref	Alt	AAF_{ALS}	AAF_{1kg}	Hom_A:Het:Hom_R	Coding change
BEST3	frameshift deletion	12	70088224	A	NA	0.06137	0	1:11:94	Q58fs
MAST4	nonsynonymous SNV	5	66441069	G	A	0.01951	0.0034	1:2:103	E797K
MRPL16	nonsynonymous SNV	11	59573980	G	A	0.02868	0.016	1:4:101	R199Q
OR5M3	nonsynonymous SNV	11	56237609	T	C	0.05847	0	2:0:104	M122T
OR5M3	nonsynonymous SNV	11	56237564	T	A	0.05847	0	3:6:97	V137D
PDE11A	nonframeshift insertion	2	178494179	NA	2008_2009insTCC	0.7871	0	65:37:4	S920insS*
PRSS48	frameshift insertion	4	152201026	NA	131_132insGTCAG	0.5413	0	36:45:25	S44fs
TUSC3	nonsynonymous SNV	8	15480643	A	G	0.01891	0.0029	1:2:103	I65V

Chr, chromosome
Pos, base pair position in GRCh37 coordinates
Ref, reference allele
Alt, alternate allele
AAF_{ALS}, alternate allele frequency in the ALS dataset
AAF_{1kg}, alternate allele frequency in the 1000 Genomes dataset
Hom_A:Het:Hom_R, number of individuals homozygous alternate, heterozygous and homozygous reference
SNV, single nucleotide variant
NA, not applicable
* *PDE11A* coding change denoted for transcript NM_016953

4.4 Discussion

The work detailed in this chapter set out to assess the incidence of rare variation in a set of ALS candidate genes using a cohort of 106 ALS patients. The work represents the initial case-only phase of a larger project which aims to look at further cases and population-matched controls in the same set of 1,577 candidate ALS genes. Initial results reported here are informative of the expected outcome and analysis strategy of the larger project. In particular, the need for a set of population matched controls is addressed.

4.4.1 Summary of findings and their significance

1,577 candidate genes, along with 36 genes previously implicated in ALS (table 4.1) were sequenced in a panel of 106 ALS patients. Genomic DNA samples were enriched for candidate genes using a custom Agilent SureSelect in-solution RNA bait library [174], yielding an on-target rate of just 25% after sequence alignment and processing. Furthermore, many target regions were covered at 0× depth for several individuals (figure 4.2(a)). Taken together, these two results represent a relatively underwhelming performance from the target enrichment method.

It is likely that the poor on-target rate is at least partially due to the indexing strategy that was employed for sample multiplexing. The protocol for SureSelect target enrichment involves the addition of several oligonucleotide blocks designed to hybridize to sequencing adapters during the RNA bait/sequencing library hybridization step, so that the only sequence that is visible to the RNA baits is target sequence in template DNA. The incorporation of a 6-mer oligonucleotide barcode between the sequencing adapter and the template molecule alters the sequence that is visible to the RNA baits and increases the likelihood that off-target sequence will be captured. Improved methodologies in the future might involve the addition of oligonucleotide blocks specific to the barcodes used. Alternatively, a different indexing method could be used which places the barcode in the

Table 4.8: Rare variants discovered in genes previously implicated in ALS

Gene	Type	Chr	Pos	Ref	Alt	Coding change	AAF
ALS2	nonsynonymous SNV	chr2	202609053	T	C	T700A	0.0047
ALS2	nonsynonymous SNV	chr2	202580536	G	A	P1288L	0.0049
ALS2	nonsynonymous SNV	chr2	202575717	T	C	I1373M	0.0048
ALS2	nonsynonymous SNV	chr2	202622481	G	C	P372R	0.0047
ALS2	nonsynonymous SNV	chr2	202622313	G	T	T428N	0.0047
DCTN1	nonsynonymous SNV	chr2	74605231	C	T	**G59S**	0.0048
DCTN1	nonsynonymous SNV	chr2	74605315	A	G	S31P	0.0053
DCTN1	nonsynonymous SNV	chr2	74594037	A	G	I780T	0.0048
ELP3	nonsynonymous SNV	chr8	27957364	G	A	A47T	0.0048
ELP3	nonsynonymous SNV	chr8	27957431	G	T	R69L	0.0048
HFE	nonsynonymous SNV	chr6	26091185	A	T	S65C	0.0143
IFNK	frameshift insertion	chr9	27524371	NA	TTGT	W13fs	0.0368
ITPR2	nonsynonymous SNV	chr12	26755563	C	T	R1180Q	0.0098
ITPR2	nonsynonymous SNV	chr12	26755617	A	C	V1162G	0.0050
ITPR2	nonsynonymous SNV	chr12	26755367	C	T	R1205Q	0.0047
ITPR2	nonsynonymous SNV	chr12	26493117	C	T	A2668T	0.0049
NEFM	nonsynonymous SNV	chr8	24774791	G	A	A99T	0.0047
OPTN	nonsynonymous SNV	chr10	13167989	C	G	Q398E	0.0047
PON1	nonsynonymous SNV	chr7	94937419	G	A	A201V	0.0047
PON2	nonsynonymous SNV	chr7	95039247	A	C	S209A	0.0047
PON2	nonsynonymous SNV	chr7	95034794	G	A	R293C	0.0094
PON3	stopgain SNV	chr7	95024007	G	A	R32X	0.0048
PON3	nonsynonymous SNV	chr7	95001590	T	C	M88V	0.0047
SETX	nonsynonymous SNV	chr9	135204010	T	C	K992R	0.0189
SETX	nonsynonymous SNV	chr9	135203756	C	T	D1077N	0.0047
SETX	nonsynonymous SNV	chr9	135202325	A	C	C1554G	0.0094
SETX	nonsynonymous SNV	chr9	135140020	A	G	I2547T	0.0142
SPG11	nonsynonymous SNV	chr15	44918690	C	T	A695T	0.0190
SPG11	nonsynonymous SNV	chr15	44907562	T	C	K1013E	0.0094
SPG11	nonsynonymous SNV	chr15	44890484	A	G	I1327T	0.0048
SPG11	nonsynonymous SNV	chr15	44859744	C	T	R2098H	0.0095
SPG11	nonsynonymous SNV	chr15	44855327	C	G	A2329P	0.0047
SPG11	nonsynonymous SNV	chr15	44944037	C	T	E370K	0.0283
SPG11	nonsynonymous SNV	chr15	44925740	A	C	D566E	0.0378
UNC13A	nonframeshift deletion	chr19	17766941	TCC	NA	344_345del	0.0324
UNC13A	nonsynonymous SNV	chr19	17746950	A	T	V1033D	0.0049

If a variant has previously been implicated in ALS, it is shown in bold
Chr, chromosome
Pos, base pair position in GRCh37 coordinates
Ref, reference allele
Alt, alternate allele
AAF, alternate allele frequency
SNV, single nucleotide variant
NA, not applicable

middle of the sequencing adapter, thus rendering it unable to interfere with template DNA sequence [214].

Despite the disappointing performance of the target enrichment strategy, a large volume of usable data was generated, from which many variant calls could be made. The first application of these variant calls was to confirm that genotypes called in individuals that had been sequenced were concordant with genotypes that had been generated in chapter 3. This confirmed concordance of well over 99%, which gave confidence in genotype calls and confidence that the correct individuals were sequenced.

The allele frequency spectra generated after variant calling demonstrated that the majority of variants that were discovered were less than 1% frequency (figures 4.3(b) and 4.4(a)). However, a small number of variants were present at much higher alternate allele frequencies. This could be due to a number of factors. Although unlikely, the 1000 Genomes Project could have acheived poor coverage for the region, thus missing the variant call, and the variant was therefore not flagged as common in the current analysis. Alternatively, these variants could represent errors in the reference sequence that were accounted for in the generation of the 1000 Genomes dataset. In many cases, however, these higher-frequency variants could represent alleles that have very low frequency worldwide, but that have drifted to higher frequency in the Irish population. Further sequencing of these genomic regions in a control cohort would be informative of population-specific variants, thus making the interpretation of pathogenicity of discovered variants easier.

Following variant filtration based on allele frequency in the 1000 Genomes Project data, 3,001 variants remained. Table 4.4 details the subset of these variants that are expected to alter protein structure at more than one amino acid. Filtering of variants based on this annotation reduced the set of candidate variants effectively from 3,001 to 82; however this list is not small enough to permit a straightforward follow-up in a larger population. Furthermore, this strategy rejects the potential pathogenicity of nonsynonymous SNVs,

for which there is little cause. In light of this, a further analysis was conducted that assessed the occurrence of multiple rare variants within individual genes, comparing the ALS dataset with the 1000 Genomes dataset. This revealed just 13 genes that showed greater than 10× the number of variants per base, per person, in ALS than in the 1000 Genomes dataset.

However, many of these genes showing an excess of rare variants are likely to be false positives. For example, just one rare variant was discovered in *GABPB1*, yet, by virtue of the difference in project size between the current study and the 1000 Genomes Project, this gene was flagged as having 10.3× the number of rare variants in the ALS dataset, per base of sequence per individual sequenced. This was the case for most of the variants in table 4.5. However, one gene, *HYDIN*, demonstrated a large number of rare variants in ALS compared to 1000 Genomes data (49 rare variants were discovered in 106 individuals sequenced in the ALS project versus 20 variants discovered in 1,092 individuals sequenced in the 1000 Genomes Project). Table 4.6 details the complete set of variants discovered in *HYDIN*. While it is expected that many of the variants described in table 4.6 are non-pathogenic polymorphisms, the excess of variants discovered in this gene compared to the expectation derived from 1000 Genomes data suggests that this gene could indeed be involved in ALS aetiology. In mice, Hydin mutations cause lethal communicating hydrocephalus with early onset [215], which is triggered by denudation of ependyma and neuroepithelium early in development [216], also leading to downstream neurological effects. Furthermore, structural variation in an ancestrally recent *HYDIN* paralogue [217] causes microcephaly, macrocephaly and behavioural abnormalities in humans [218]. Together, these studies implicate *HYDIN* in neurological function, contributing to the disease-related interpretation of the findings of multiple rare variants in this gene in ALS.

A separate analysis was conducted on discovered variants that made no assumptions

about frequency of alternate alleles. This analysis searched for recessive variants in genomic regions where they would be expected in certain individuals, as determined by the ROH mapping results described in section 3.3.4. This identifed 8 genes that were not present homozygously in the 1,092 individuals of the 1000 Genomes project and therefore may be involved in recessively-inherited ALS (table 4.7). One of these genes, *PDE11A*, was originally implicated in twelve individuals representing three different homozygous haplotype groups, and the recessive variant was discovered in all five of the individuals that were sequenced from these groups. However, the number of individuals that were shown to be homozygous for this variant in the ALS dataset (65 out of 106) raises suspicion; it is unlikely that a gene that accounts for over half of the cases of ALS has been undetected by previous GWAS and linkage studies. It is therefore more likely that the variants in table 4.7 with very low alternate allele frequencies are indicative of recessive disease-causing mutations. However, assessment of their prevalence in ALS will require further sequencing, as the occurrence of a homozygous genotype for an allele with a frequency q will only be q^2 if it is simply a population-based variant.

Assessment of rare variants in genes previously known or suspected in ALS was performed; table 4.8 shows the results. No variants were discovered in *SOD1*, *FUS* or *TARDBP*, which is surprising given that these genes carry the strongest evidence implicating them in ALS aetiology. However, the majority of samples sequenced in the current study were from individuals with sporadic ALS, whereas *SOD1*, *FUS* and *TARDBP* are mainly implicated in familial ALS. One variant that had previously been implicated in ALS aetiology, the G59S mutation in *DCTN1* [47], was identified in the current study at a frequency of 0.48% in the ALS cohort.

Many of the remainder of the variants listed in table 4.8 are interesting, in that they may suggest causative alleles for disease genes that had only previously been implicated in ALS by GWAS, for example *UNC13A* and *ITPR2*. In particular, the frameshift deletion in

UNC13A is interesting given its relatively high frequency of 3.24% in the sequenced cohort, which could represent a previously-undiscovered ALS variant that is high frequency in ALS patients in Ireland. However, this could also be a rare Irish population polymorphism, the identification of which would require further sequencing in unaffected Irish individuals. This is true also for the many other interesting findings reported in table 4.8.

4.4.2 Limitations

This study has attempted to identify disease-causing rare variants in sequencing data derived from a panel of ALS cases. Many of the doubts raised in the previous section about pathogenicity of discovered variants could be addressed through the sequencing of a representative panel of healthy controls to describe the background genetic structure of the Irish population. This way, it could be determined whether interesting variants are in fact simply low-frequency population polymorphisms within Ireland. This is a strong limitation of the current study and accurate inferences about the pathogenicity of variants identified cannot be assessed until a representative comparison dataset is available.

A second limitation is the number of individuals that have been sequenced in this study. Many of the genes in the target gene set were derived from the ROH analysis, but of the 329 individuals in which these ROHs were discovered, only 106 have been sequenced. The remaining 223 individuals may harbour recessive mutations that are rare enough that they would not be discovered unless representative individuals were sequenced. Indeed, several of the candidate gene regions mapped by ROH analysis did not have a representative individual in the sequenced panel.

A third limitation is that this study has only focussed on sequencing the protein coding portions of the candidate genes. It could be the case that many cases of ALS are explained by genetic factors that lie within either regulatory regions or non-coding portions of genes, such as the recently-discovered *C9orf72* hexanucleotide repeat expansion

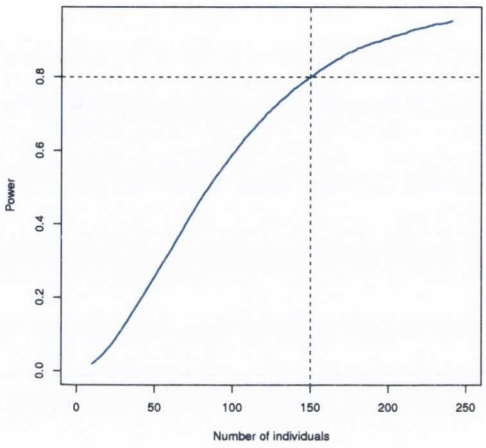
[121,122]. Additionally, the sequencing strategy is only capable of detecting SNVs or small indels; a variant such as the expansion in *C9orf72* would not be detected by the methods described in this chapter.

4.4.3 Future directions

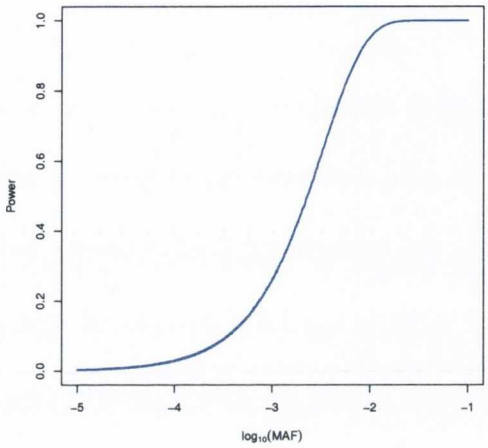
Future work should attempt to address the limitations mentioned in the previous section. Sequencing of a representative panel of neurologically normal Irish controls will help to ascertain the background level of genetic variation in the Irish population so that disease-causing variants can be distinguished from population polymorphisms. In order to have a reasonable chance of estimating the allele frequency of a population polymorphism, it would be preferable that it is observed more than once in the control cohort. Therefore, to identify genetic variation that is at the upper limit of the difinition of rare, a cohort of 150 control individuals would be required to detect the variant at least twice (figure 4.6(a)). The power to detect a range of MAFs with 150 individuals is shown in figure 4.6(b). 150 represents the minimum number of controls that should be sequenced to address the issue of identifying population polymorphisms.

However, a large proportion of this project has been based on the assumption of recessive inheritance of ALS. Because rare variants are very unlikely to be present in an individual homozygously, the number of individuals required to refute the role of recessive inheritance of the variant in disease is much greater. Figure 4.6(c) shows the power to detect a rare variant homozygously for a range of alternative allele frequencies and a range of sample sizes. This demonstrates that for even for a relatively common variant whose MAF is 5%, a large follow-up validation would be required to rule out the possibility that the disease allele segregated homozygously in an ALS case by chance (80% power with approximately 625 individuals).

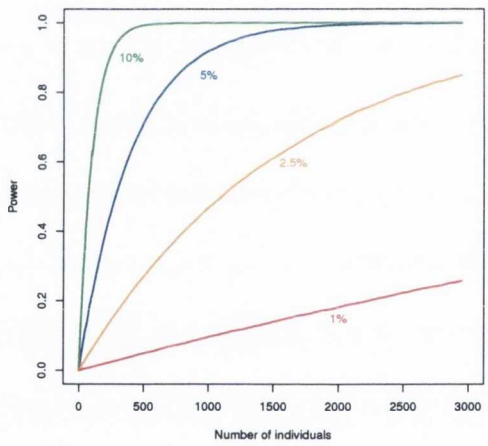
The relative lack of results for the recessive variant analysis has highlighted the need to



(a) Variants at MAF = 1% in two individuals



(b) Range of MAFs with 150 individuals



(c) Recessive variants

Figure 4.6: Power considerations when sequencing controls to assess population-based variants. (a) Power to detect a variant at MAF $\leq$ 1 % in at least two individuals for different study sizes. A population variant of MAF $\leq$ 1% will be detected at least twice 80% of the time when 150 individuals from that population are sequenced (dashed lines). (b) Power to detect a range of allele frequencies in at least one individual when 150 individuals are sequenced. (c) Power to detect variants at a range of MAFs homozygously in a range of sample sizes. These plots were produced by simulation, detail of which can be found in appendix A.2.

sequence the remainder of the representative portion of the case cohort in which the ROH regions were mapped. Nevertheless, it may be the case that fewer results are generated than expected, due to the limitation that only protein coding portions of the genome are being sequenced, and also that NGS methodologies are only capable of detecting certain types of variation. For example, although NGS is well-suited to detect indels, the recently-reported finding of polyglutamine repeats in *ATXN2* [56] would be relatively undetectable using NGS methodology, due to difficulty in alignment of larger indels, especially those representing repeat sequence. Hexanucleotide repeats recently reported in *C9orf72* [121, 122] would also be difficult to detect for similar reasons, in addition to the problem that this project has only assessed genetic variation within coding regions and *C9orf72* expansions are present in the non-coding portion of the gene.

For these reasons, it may be fruitful to re-visit the analyses of chapter 3 to identify genomic regions that may be interesting to follow-up in a study design that considers more than just coding regions. Furthermore, incorporation of techniques that are permissive to the detection of structural variation may represent an improved methodolgy above what is described in this chapter.

4.4.4 Conclusion

The work described in this chapter has made an attempt at addressing the contribution of rare genetic variation to ALS aetiology by NGS of the coding sequence of many candidate genes in a relatively small cohort of Irish ALS cases. The main conclusion from this study is that further sequencing is required, both to ensure that representative individuals are being sequenced for all the candidate genes, and to maximize the power to detect rare disease-causing variants. Furthermore, a representative panel of population-matched controls is necessary to identify polymorphisms that have no link to ALS pathogenicity. Nevertheless, the work described has hinted at potential ALS-causing disease genes, one of

the most notable of which is *HYDIN*. Furthermore, rare variants in several genes previously implicated in ALS aetiology have been identified, which may help to contribute to the understanding of the disease. Further sequencing will help to disentangle the findings of this work.

Chapter 5

Towards exome sequencing in ALS: an exploration of identity-by-descent in the Irish population

5.1 Introduction

Chapter 4 describes work that was carried out to identify rare putatively ALS-causing variants on the level of the population by sequencing the coding regions of a set of candidate genes in many individuals. Under the assumption that coding sequence alterations may be causing ALS, such an approach is excellent for maximizing the power to detect rare variants and for estimating allele frequencies of putative disease variants in cases and controls. However its success is heavily dependent on a well-chosen set of candidate genes that show good evidence for being involved in ALS. An alternative approach for rare variant discovery is to sequence the entire coding portion of the genome, which is known as the exome. Such an approach makes no assumptions about which genes may be causing the

disease, although, like the techniques described in chapter 4 it also makes the assumption that a variant that alters protein structure is driving disease aetiology. Nevertheless, exome sequencing is an effective way of reducing the sequencing burden through focussing on the protein coding portion (1.5%) of the genome [219]. Exome sequencing strategies use similar approaches to the target enrichment methods described in chapter 4 to prepare a genomic DNA sample so that it has a high copy number of exonic intervals compared to the rest of the genome. Exome capture methods are available from several manufacturers [220–222], using either solution-based or array-based target enrichment strategies.

With exome sequencing, the workflow is similar to the processes described in chapter 4 for generating sequencing reads from genomic DNA samples. However, depending on the scale of the project, the downstream data processing methods could differ substantially. With population-based rare variant discovery, it is possible to generate an estimate of the allele frequencies of discovered variants because a sufficient number of individuals have been genotyped. However, exome sequencing projects tend to be smaller, so approaches are often adopted that assess variant sets that are common to all individuals sequenced. In order for this approach to be successful, it would require a carefully chosen set of individuals in whom the same phenotype-causing variants are expected. This variant intersection strategy was successfully applied early in the development of exome sequencing by Ng *et al.* to identify rare variants that contribute to the aetiology of Miller syndrome [223] and since then it has been used extensively in multiple studies in a variety of diseases. Because of its strong genetic component [41], complex inheritance and large amount of unexplained heritability ALS is a good candidate disease for exome sequencing.

A successful exome sequencing study was recently performed on ALS by Johnson *et al.* [69], which identified a mutation in *VCP* as a cause of the disease in an Italian family. In this study, the authors used a number of filtration methods to reduce their set of discovered variants from over one hundred thousand to just four that were predicted to

be damaging. As well as ruling out any polymorphism that had previously been reported in dbSNP, they also only took variants that were common to three cases from the same family and excluded any that were present in the exomes of 200 controls. In addition, they had the privilege of prior evidence in the form of a linkage region which they could use to rule out 42% of the variants in one of the filtration steps. Several further filters were applied to reduce the variant set effectively.

In a population-based sample, the variant filtration methods employed would be somewhat different, but given that the sample does not contain individuals from the same family, the number of variants shared between cases would probably be lower, and therefore filtering based on overlap may be easier. In this scenario, an optimal strategy could be to sequence a few cases to identify variants that are common to everyone, and also sequence some controls to exclude putatively non-pathogenic variants. For example, if the exomes of two cases and one control were sequenced, the variant set of interest would be $\{(A \cap B) \setminus C\}$, where A and B are the variant sets independently discovered in the two cases a and b, and C is the variant set discovered in control c. Figure 5.1 shows the percentage of variants that remain as a result of various filtration strategies, given different numbers of cases and controls.

Naturally, such an approach assumes high penetrance, although this could potentially be addressed by the selection of hypernormal controls as has been argued for the design of GWAS [224]. In the case of ALS these could be individuals with little or no family history of neurodegeneration who lie within the extreme upper tail of the population distribution of age while demonstrating a healthy central nervous system. By selecting such an individual, the odds of the control sharing the pathogenic variant would hopefully be diminished, given the assumption that a carrier could not live to such an old age without manifesting some of the symptoms of ALS.

A second important assumption of this exome sequencing strategy is that the cases

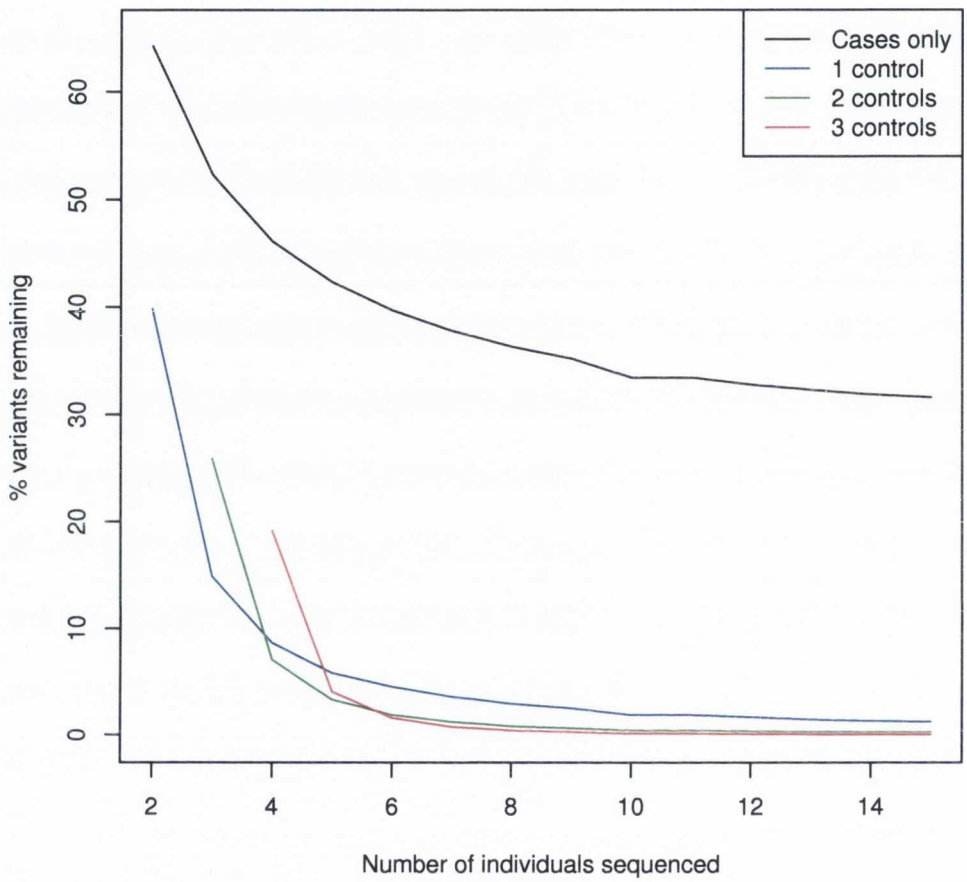


Figure 5.1: Strategies for exome sequencing of a small number of individuals, based on analysis of data generated in chapter 4. To generate this plot, groups of individuals ($n = 2 \ldots 15$) that had been sequenced in chapter 4 were selected and the variant sets discovered in each individual were used to count the percentage of variants that remained after filtering. The strategy for filtering was to intersect the complement of the union of all variants discovered in controls with the intersection of all variants discovered in cases. For example, for 5 individuals, two of which are controls, the variants that remain after filtering case variant sets A, B and C and control variant sets D and E would be $\{(A \cap B \cap C) \setminus (D \cup E)\}$.

sequenced all have a common disease phenotype due to segregation of the same disease-causing variant. Exclusion of samples that carry variants in genes known to cause ALS would be an important first step in meeting this assumption. Subsequently, the sequenced cases would have to be very carefully selected in order to maximise the probability that the same disease-causing variant is present in every case. One approach is to use a very specific and unique endophenotype. However, on its own this may not be sufficient evidence that the same variant is causing the disease due to allelic or genetic heterogeneity for the endophenotype, and, conversely, non-pathogenic phenotypic modifiers may reduce the effectiveness of this method.

Another approach to enrich the probability of detecting the same disease-causing variant is to select distantly-related individuals from the population, with the assumption that if they are descended from the same reasonably recent ancestor and they share the same disease phenotype, they probably share the same variant causing the disease. An obvious strategy in this case would be to select two individuals from the same extended pedigree (for example, third cousins). However, given that ALS is a late-onset disease, the collection of large families can be difficult because patients are often deceased before their relationships with other patients have been elucidated. Banking of DNA goes some way to address this, but the Irish DNA bank has historically focussed on the collection of representative population-based samples and there has been less emphasis on prioritizing the collection of multiple individuals from within families than in other countries. Although this is no longer the case, the identification of large pedigrees and subsequent collection of DNA samples from multiple affected individuals within pedigrees is a demanding process, and it would be considerably more difficult to perform retrospective investigations of pedigrees for individuals who are no longer alive.

One way to circumvent this problem is to infer the relationship between supposedly unrelated individuals using dense genotype data, for example the data generated and an-

alyzed in chapter 3. As discussed in section 3.2.2, one important quality-control step is the exclusion of cryptically-related individuals from the dataset to reduce the potential for spurious associations. However, methods used for identification of cryptically-related individuals (for example, identity-by-state [IBS] clustering or the $\hat{\pi}$ measure of proportion of SNPs that are IBD implemented in PLINK [95]) rely on relatively simple metrics based on cumulative statistics over many individual SNPs, and they do not take into account groups of genotypes (haplotypes). An alternative approach to using single-marker tests is to look for shared genomic segments that are identical-by-descent (IBD) between individuals in a population.

For such analyses to be effective, it is useful for the genotypes to be phased, meaning that the haplotypes on which the genotypes appear need to be ascertained. Although segmental IBD can be estimated in unphased genotypes [95, 225–227], accuracy is greatly improved if phased genotypes are used [228]. Given phased haplotypes, the extent of sharing of genomic segments between individuals can be estimated, and thus the degree of relatedness can be inferred. For example, while siblings would be expected to share 50% of their genomes, first cousins would share 12.5%, second cousins 3.13% and third cousins 0.78%. Several algorithms exist that can infer relatedness given phased haplotypes, including GERMLINE [226] and the `fastIBD` algorithm [228] implemented within BEAGLE [229].

Given its relatively small size of around 6.4 million individuals [230, 231], the population of Ireland may be a well-suited group for this kind of analysis, especially when a dataset derived from a late-onset disease such as ALS is under consideration. In the same sense that Nalls *et al.* argued that younger generations demonstrate fewer ROHs due to increased mobility and panmixia, there are probably pockets of strong inter-relatedness within Ireland as a consequence of a small gene pool and a history of lower mobility within the country. Given these assumptions, the Irish ALS population-based dataset may con-

tain hidden relatedness that can be used as a proxy for large pedigrees and exploited in the design of exome sequencing projects.

5.1.1 Research aims

The work detailed in this chapter describes an exploration of IBD in the Irish ALS population and in matched controls. For comparison, and to improve the accuracy of inferences, a second dataset is included in the analysis, representing British individuals from the 1958 British Birth Cohort genotyped as part of a study conducted by the Wellcome Trust Case-Control Constorium (WTCCC). The aim of this chapter is to use IBD inference to gain a better understanding of the inter-relatedness of a population of individuals assumed to be unrelated, in the hope that this will help to identify optimal ALS cases for the design of an exome sequencing project following strategies such as those depicted in figure 5.1. Specifically, the aims are:

i to determine whether IBD levels differ between the Irish and British populations;

ii to determine whether IBD levels differ between Irish ALS cases and Irish controls;

iii to identify genomic regions of higher IBD in Irish ALS cases than in Irish controls;

iv to assess geographical patterns of IBD within Ireland;

v to investigate any clustering of inter-relatedness within Ireland by IBD.

Results obtained from these explorations are used to make inferences about the optimal design of future exome sequencing projects.

5.2 Methods

5.2.1 Genotype data

Genome-wide SNP data for a total of 692 Irish and 2,708 British individuals were used in the study. Irish data included 620,901 genotypes generated in chapter 3 using the Illumina 610-Quad beadchip and 561,466 from the 2008 Irish ALS GWAS generated using the Illumina HumnHap550 beadchip [98]. Genotype data from the WTCCC's 1958 Birth Cohort panel, which includes 1,116,106 SNPs for 2,930 individuals genotyped on the Illumina Human1M-Duo platform, were retreived from the European Genome-Phenome Archive (EGA, http://www.ebi.ac.uk/ega) under accession EGAD00000000022, parsed using the script `parse_WTCCC.pl` (converting genotype likelihoods > 0.95 to genotypes, setting the rest to missing) and merged into a single PLINK-format .bed file. WTCCC-recommended exclusions for SNPs (based on similar criteria to those described in section 3.2.2) and individuals (based on heterozygous/missing proportion, PCA-based ancestry outliers, gender mismatches, A/B allele channel bias and identity of replicate SNP genotypes) were then removed from the dataset, leaving 900,374 genotypes for 2,718 individuals. For both the Irish and British datasets, individuals flagged as related based on the SNP genotypes were deliberately left in the dataset to improve phasing accuracy and for evaluation of the performance of the IBD method. Replicate datasets in the British cohort were identified by IBS clustering within PLINK (see section 3.2.2) and removed, leaving a dataset of 2,708 individuals.

To avoid allele encoding and strand issues when combining the Irish and British datasets, Irish genotypes were corrected to the Illumina 'top' allele encoding using a lookup table generated by Illumina GenomeStudio and the script `correct_strand.pl`. Irish and British data were merged together using the PLINK `--bmerge` option, and filtered to include only SNPs common to both datasets (using per-SNP missingness as a

proxy, determined using the PLINK `--geno` option). This resulted in a single large dataset containing genotypes for 477,356 SNPs in 3,400 individuals.

5.2.2 Haplotype phasing and IBD estimation

According to a benchmark test performed by Browning and Browning (2011) [232], for sample sizes of greater than 3,000 individuals, BEAGLE [229] achieves close to 98% accuracy with haplotype phase calls at a fraction of the computational cost when compared with two other algorithms, MACH [233] and IMPUTE2 [234]. In addition, for IBD estimation, BEAGLE version 3.3.2 incorporates the algorithm `fastIBD` [228], which is demonstrably more powerful than an alternative, GERMLINE [226], or the hidden Markov model method implemented within PLINK's `--read-genome` option, with far fewer false positives [228]. For these reasons, genotypes were phased using BEAGLE and IBD was estimated using the `fastIBD` algorithm.

First, the single large merged WTCCC/Irish genotypes file was converted to BEAGLE input format using `ped_to_bgl`, a convenient utility that ships with GERMLINE. The resulting file was used as input to BEAGLE, which was run using default parameters with the `fastibd` switch set to `true`. Total computation time was 121 hours and 3 seconds, after which a single .fibd.gz file was generated, containing the locations of every IBD segment identified by the `fastIBD` algorithm.

5.2.3 Assessment of IBD

Data were processed, interpreted and visualized using a combination of custom scripts and the R statistics package [94]. Firstly, following recommendations in the documentation for `fastIBD` [228], high quality IBD calls (fastIBD score $< 10^{-10}$) were extracted and only these were used in subsequent analyses. Data were then split into separate files for within-Ireland comparisons, within-Britain comparisons and between-population

comparisons. Irish data were also split by case-control status into separate files for case-case comparisons, case-control comparisons and control-control comparisons. These steps resulted in individual datasets representing individual IBD segments between pairs of individuals. Further similar datasets were generated representing total genomic IBD length between individual pairs and total number of IBD segments between pairs.

For comparison of genomic regions between cases and controls, a file was generated using `parse_for_cc_comparison.pl` representing, per SNP, the numbers of case and control pairs that were identified as IBD by `fastIBD`. Genome-wide IBD values for the two cohorts were plotted using R, as well as genome-wide values for the difference between control-control pairs and case-case pairs. For each plot, the total number of pairs of individuals IBD was divided by the total number of possible pairs for that dataset. For self-comparisons within a population of size n, there are $\frac{n^2-n}{2}$ possible combinations, whereas between populations of size n and m, there are nm possible combinations of individuals.

For the Irish case cohort, coordinates of the postal addresses of patients were available, representing the geographical location of individuals at time of diagnosis of ALS. To assess geographical clustering of IBD within Ireland, a map of Ireland was rasterized into blocks of approximately 125 km^2 and mean IBD values were calculated for all pairs of individuals that fell within a sliding window representing approximately 17,000 km^2 surrounding each block. These values were then visualized using R's `image` function.

Finally, lists of pairwise total IBD values for the Irish dataset were parsed into a large 692 $\times$692 matrix of pairwise total IBD length, which was then used to assess evidence of clustering of high IBD values. This was initially assessed using the heirarchical clustering methods available within R's `hclust` function; however, these methods were found to be insufficient for reordering the sparse matrix that the IBD values represented. To attempt to address this, a brute force method for reordering the matrix was written (`cluster_IBD.pl`) and applied to the dataset. The resulting reordered matrix was visualized as a heatmap

using the R `image` function.

5.3 Results

5.3.1 Comparison within and between populations

IBD extent was higher within the Irish population than within the British population in terms of overall length IBD (figure 5.2(a)), length of IBD segments (figure 5.2(b)) and number of IBD segments (figure 5.2(c), $p < 2.2 \times 10^{-16}$ for all metrics).

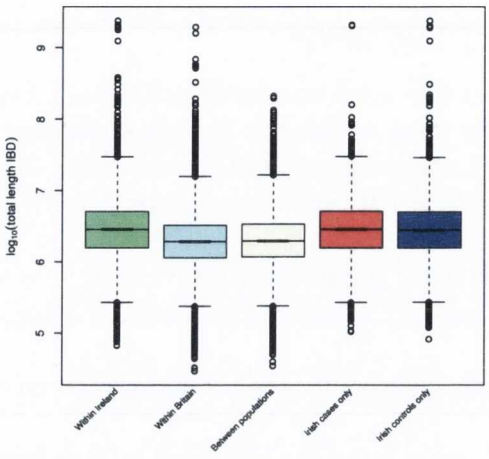
5.3.2 Geographical phenomena within Ireland

Using data derived from ALS patients, IBD showed some degree of geographical clustering within The Republic of Ireland, with some regions showing higher average IBD than the background average rate (figure 5.3). Unfortunately address data were not available for patients from Northern Ireland, so geographical phenomena could not be ascertained for these patients.

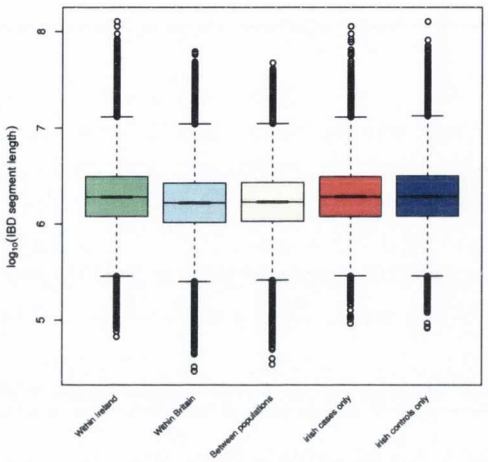
5.3.3 Case-control comparisons

Within the Irish dataset, IBD was not higher in case-case comparisons than in control-control comparisons for total IBD length ($p = 0.095$, figure 5.2(a)). However, control-control pairs showed a significantly higher number of IBD segments called compared to case-case pairs ($p = 0.010$, figure5.2(c)).

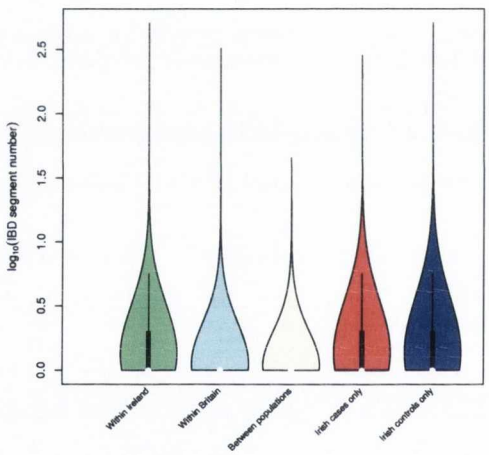
Genomic regional IBD estimates were calculated as the proportion of all possible case-case pairs or control-control pairs that were identified as IBD across the genome, and these values are plotted in figure 5.4 (a) and (b). Additionally, the difference between case-case pairs and control-control pairs was calculated per SNP and plotted in figure 5.4(c). There was no overall inflation of IBD in the case cohort, although some regions showed elevated IBD when compared to controls.



(a) Total length IBD

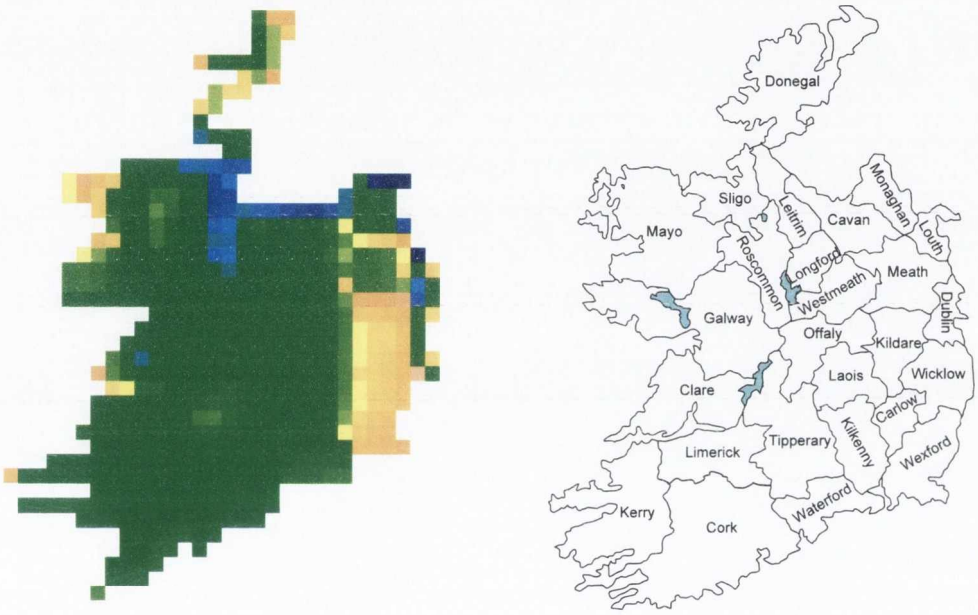


(b) IBD segment length



(c) Number of segments IBD

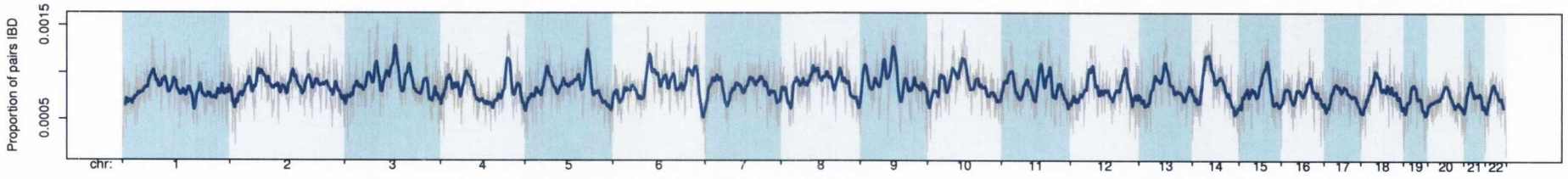
Figure 5.2: IBD within and between Irish and British populations. (a) Notched boxplots showing total IBD called between pairs of individuals within Irish and British datasets, and between datasets. Also shown is total pairwise IBD within cases and within controls in the Irish dataset. (b) As plot (a), but showing individual IBD segment lengths derived from pairwise comparisons within and between datasets. Average values in this boxplot are similar to (a) because the majority of pairs of individuals had very few IBD segments identified. (c) Violin plots of number of segments called between individual pairs within and between datasets. Plotted shapes correspond to kernel density estimates of the distributions of the datasets, and they demonstrate that within Ireland, there were typically many more IBD segments called in individual pairs than within Britain.



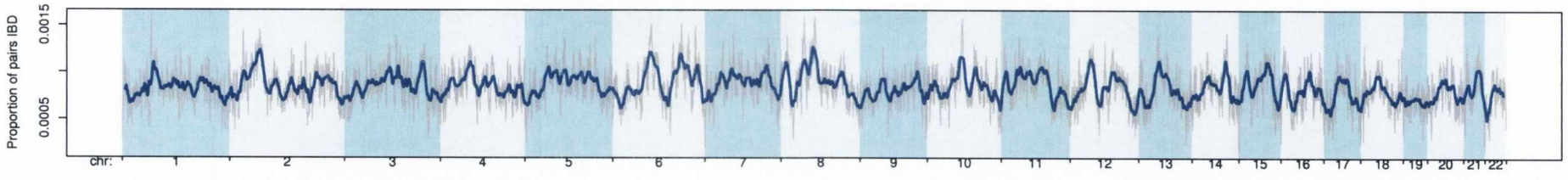
(a) IBD within regions in Ireland

(b) Map of Ireland for comparison

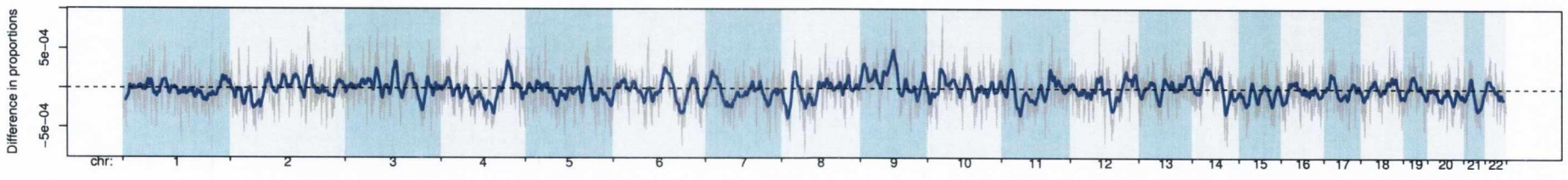
Figure 5.3: IBD by region within Ireland. High IBD values (yellow = high IBD; blue = low IBD) seem to cluster geographically, with extensive IBD seen in a block from Wexford to Meath, as well as pockets of high IBD in North Tipperary, Waterford, South Cork, Clare, Mayo, Sligo and Donegal. To generate this map, a sliding window was scanned across a rasterized map of Ireland and average IBD was measured within all individual pairs that were found within the sliding window. The map was then plotted using R's `image` function.



(a) Case-case pairs



(b) Control-control pairs



(c) Difference between case-case and control-control pairs

Figure 5.4: Genome-wide IBD plots for cases and controls. (a) and (b) show within-cohort IBD comparisons expressed as proportion of total possible comparisons ($\frac{n^2-n}{2}$) and (c) shows values for (b) minus (a). In each plot, the blue line represents a smoother applied to the data, calculated, for each SNP locus, by averaging values ± 1000 SNPs.

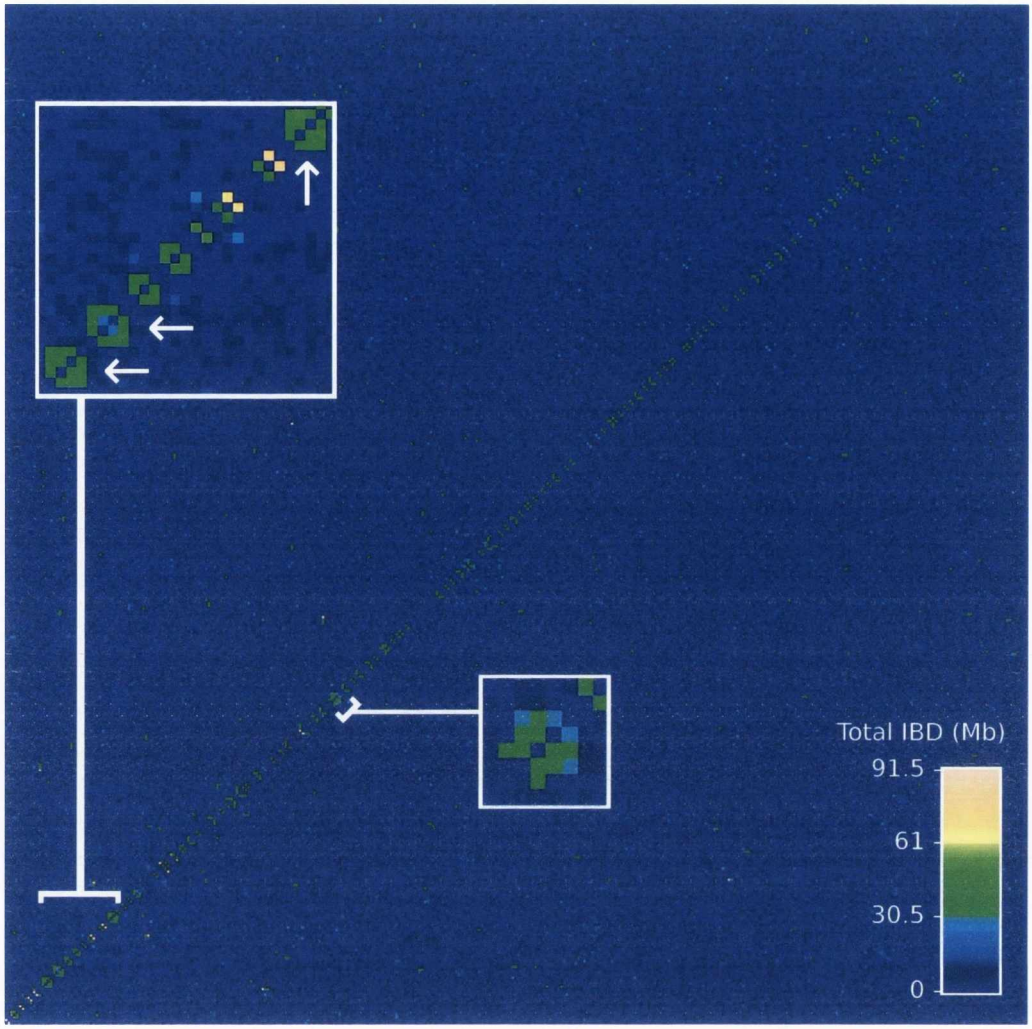


Figure 5.5: Clustering of high IBD values within the Irish dataset. The figure shows the top results when the matrix of pairwise total IBD is reordered by `cluster_IBD.pl`. In general, no very large, obvious clusters stood out in the dataset, although four clusters of high mutual IBD were identified containing four individuals each; these clusters are indicated. Additionally, more subtle clusters were identified in the data which are less obvious in the figure as the IBD values represented are very low (less than 15 Mb).

5.3.4 Evidence of clustering of IBD

R's heirarchical clustering methods performed poorly at identifying clusters of high IBD within the dataset. The brute force method `cluster_IBD.pl` also performed poorly, but did identify four clusters of high IBD containing four individuals each, and several clusters containing three individuals (figure 5.5). Nevertheless, high pairwise IBD values seemed to be relatively independent within the dataset.

5.4 Discussion

This study set out to explore the occurrence of IBD within Ireland, using British data as a comparison. The intention was to assess whether IBD inferences could be used to make optimal sampling decisions for inclusion in an exome sequencing project, however several other findings also resulted from the study.

IBD was found to be higher between Irish individuals than between British individuals. This could be driven by a number of factors affecting the genetic structure of the Irish population, including a smaller population size (approximately 6.4 million for Ireland [230, 231] versus approximately 60 million for Great Britain [231, 235]), historical emigrations and lower urbanization with less mobility within the country leading to less within-population panmixia. Within Ireland, within-case IBD was not higher than within-control IBD values, which is expected given that no single genetic factor of large effect size causes ALS. However, it was noted that the number of regions IBD were higher in control-control comparisons than in case-case comparisons, with no concomitant increase in overall IBD, suggesting that case-case pairs may be more related to one another than control-control pairs (an inter-related cohort should have fewer, longer IBD segments than a randomly-breeding cohort), but, on the level of the case cohort, the total length IBD is statistically indistinguishable from the background population level.

The similarity in total IBD between case-case comparisons and control-control comparisons is further supported by figure 5.4(c), which demonstrates that genomic regions are similar between cases and controls in terms of the proportion of pairs of individuals that are IBD. This differs from a study which applied the same technique using WTCCC bipolar disorder case-control data, finding that IBD levels were typically higher across the genome in the case cohort [228]. Given that ALS is a genetically heterogenous disease, it would not necessarily be expected that cohort ascertainment bias would cause an overall inflation of genome-wide IBD levels. However, regional differences in IBD levels may be in-

dicative of a common haplotype, or several common haplotypes, on which disease-causing mutations may have arisen in a subset of the case cohort.

Peaks in figure 5.4(c) may represent such disease allele-driven regional differences. The highest peak in this graph maps to chromosome 9q21, for which familial linkage with ALS has been demonstrated in the past [236]. Figure 5.6 shows the evidence from the current IBD study mapped to the location of the previous familial linkage region, suggesting that the genes identified by the linkage region mapped by Hosler *et al.* [236] could potentially be refined to just *RORB*, *TRPM6*, *CHAK2*, *C9orf40*, *BC043649*, *C9orf141*, *C9orf95* and *OSTF1*. The potential for an IBD mapping strategy such as this in identifying disease-causing genomic regions could be greater than, for example, haplotype association, as this method allows for the possibility that multiple disease variants have arisen on different haplotypes in the same region.

Albrechtsen *et al.* recently argued that genomic regions of high IBD could indicate the signatures of recent selection events [237] on standing genetic variation. To assess whether this is detectable in the differences between the Irish and British IBD values, the same plot as figure 5.4(c) was generated for the difference between the IBD proportion in the Irish population (cases and controls combined) and the IBD proportion in the British population (figure 5.7). This demonstrated, as expected, an overall excess of IBD in the Irish population compared to the British, as well as identifying several peaks representing speculative recent selection on the standing genetic variation within the Irish population.

Clustering of individuals in figure 5.5 demonstrated that it is difficult to assemble IBD estimates in multiple pairs of individuals into single large clusters of inter-relatedness. However, some clusters were identified representing up to four individuals with mutually high IBD levels. As well as these clusters, there were several more subtle groups distributed within this plot, where mutual IBD was low (in the region of 15 Mb, or 0.5 % of the genome), indicating that groups of related individuals do exist within Irish ALS dataset,

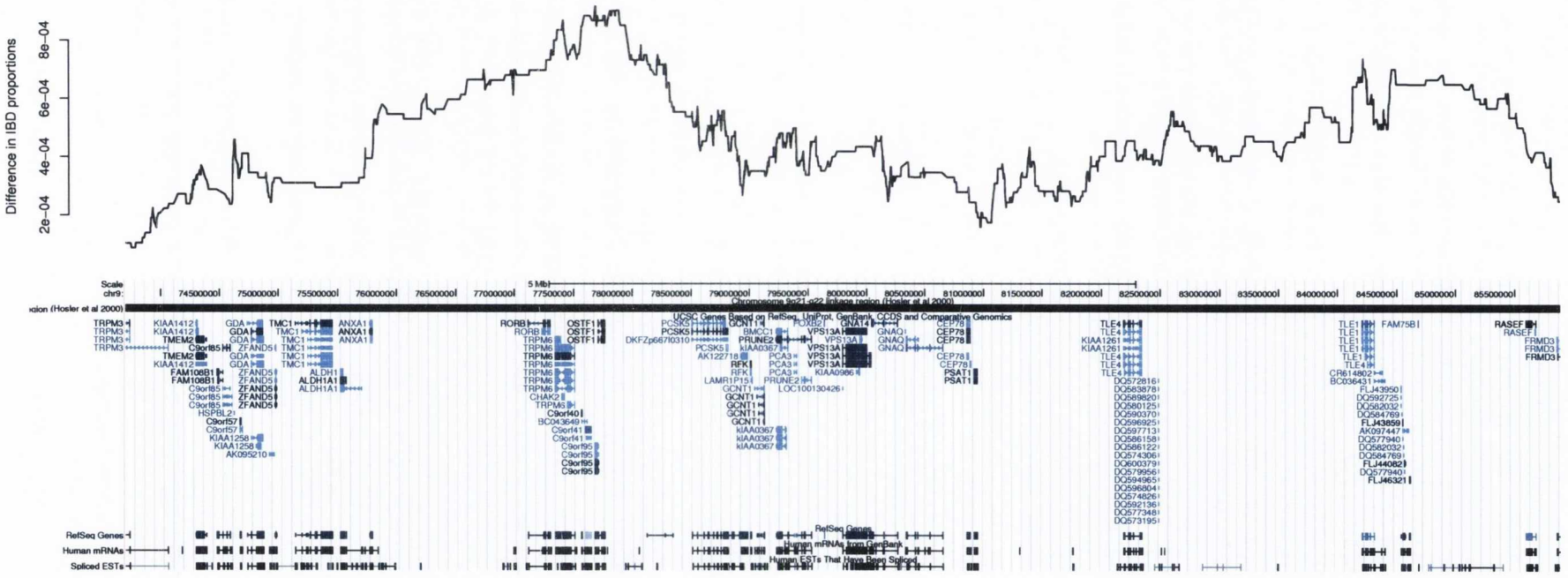


Figure 5.6: The chromosome 9q21 linkage region from Hosler *et al.* with current IBD evidence overlaid. The figure shows a screenshot from the UCSC genome browser [154] for the chromosome 9q21 familial ALS linkage region identified by Hosler *et al.* [236] with evidence from the current IBD analysis plotted above. The plot shows the same metric as in figure 5.4(c), that is, the difference between cases and controls for IBD identified within groups as a proportion of all possible combinations per group.

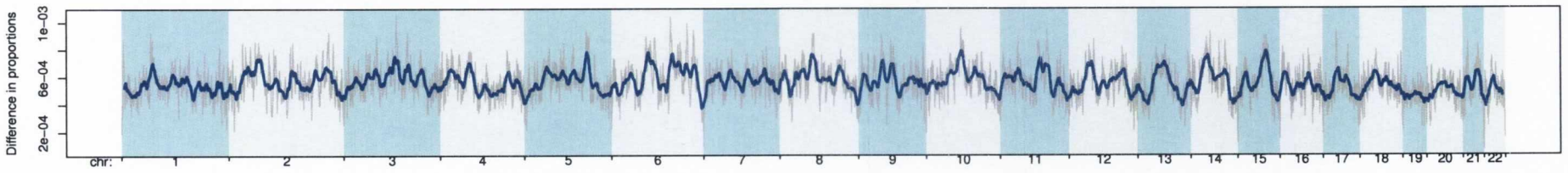


Figure 5.7: Genome-wide differences in IBD proportion between British and Irish populations. The plot shows the proportion of possible pairs identified as IBD in the Irish population minus the proportion of possible pairs identified as IBD in the British. Several peaks exist, which could represent recent selection in the Irish population [237]. The blue line represents a per-SNP $\pm$ 1000 SNPs smoother applied to the data.

representing individuals from very extended pedigrees, but detecting these groups is not trivial.

In spite of this, the best strategy for selecting individuals for exome sequencing could be to use evidence garnered from IBD studies, but also to incorporate further lines of evidence beyond these inferences. One potentially useful finding derived from this study was the finding that IBD tended to show some clustering geographically. Unfortunately addresses were not available for the control data, so it could not be determined whether this geographic clustering was a phenomenon of population genetics or one of case ascertainment. If it is actually driven by the latter, then this could reflect a within-population founder effect for a particular mutation which would be permissive to discovery by exome sequencing of related individuals derived from the geographic cluster.

A limitation of the geographical mapping approach, however, is that it has depended on the postal address of patients at the time of diagnosis. In many cases this would not be representative of the actual ancestral origin of the patient, so future patient sampling strategies should attempt to address this by collecting further information about the ancestry of the patient. Future work should also investigate any geographical clustering of control-control IBD, to attempt to resolve the issue of whether the patterns described are driven by within-population genetic stratification or by case ascertainment reflecting detection of actual signatures of familial ALS in extended pedigrees. Additionally, to exclude the possibility of algorithmic artefacts, the concordance of results with other IBD inference methods should be assessed.

5.4.1 Conclusion

This chapter set out to discover whether IBD in the Irish population can be used to inform the optimal design of future exome sequencing studies. The finding that IBD is higher in the Irish population than in the British suggests that it could be exploited to enhance

the probability of sharing a disease-causing variant in a case cohort. However, clustering of individuals based on high IBD revealed only small groups of inter-relatedness within the dataset. Nevertheless, only a small number of individuals is required for efficient variant discovery/filtration in the identification of disease-causing mutations (figure 5.1). For example, a group of four inter-related cases sequenced along with two hypernormal controls, filtered using the methods described in figure 5.1, would yield roughly the same end proportion of variants as sequencing/variant filtration with three inter-related cases and three hypernormal controls, at the same cost.

It would therefore not be unreasonable to conclude that the inferences made in this chapter about IBD could potentially be used to good effect in the design of an exome sequencing study. However, the results should be interpreted with care, ensuring that the chances of identifying the same disease-causing variant are maximized. A prudent first step would be the exclusion of any known disease-causing variant. Subsequent deep endophenotyping could add to IBD inferences, and assessment of potential geographical phenomena might help to validate the argument further. Apt use of these strategies, followed by exome sequencing, variant filtration, and follow-up validation of discovered variants in a population-based cohort, could potentially yield novel ALS genes and pathways that could subsequently lead to better intervention.

Chapter 6

Discussion

This thesis details work that studied the complex genetics of ALS using a number of approaches in datasets derived from Ireland and other European populations. The overall goal was to garner a more complete understanding of the contribution of genetic variation to the condition, by embracing the developing principles and technologies that have characterized the modern genomics era. In doing so, an appreciation was gained of the immensity of the task that is elucidating the genetics of ALS, not least because ALS is likely to have many contributory genetic factors. This is reflected in the multiple studies that have been published in the last two decades on the subject, identifying many genes that are certain to be in some way involved in the aetiology of the condition, yet with the majority of ALS cases remaining unexplained by genetics. However, in small increments the scientific community is edging closer to the complete answer, and this thesis exemplifies many of the efforts that are currently underway worldwide. The work described herein has not arrived at any single conclusion, but instead has pointed towards several avenues of research that would constitute warranted future work.

6.1 Summary of findings and their significance

The work in this thesis was conducted as four separate corpora: an elucidation of the contribution of genetic variation at the *ANG* locus to angiogenin levels in ALS (chapter 2); an exploration of common variation through analysis of genome-wide SNP genotypes in ALS (chapter 3); an exploration of rare variation through analysis of NGS data in ALS (chapter 4) and an exploration of potential methods for designing future experiments through analysis of IBD in the Irish ALS population (chapter 5). Each chapter generated several findings on their own and indicated prudent future experimental design.

Chapter 2 described the contribution of genetic variation at the *ANG* locus to levels of angiogenin. The principal findings were that angiogenin levels are lower in ALS patients than in controls, that genotypes across the *ANG* locus determine the level of angiogenin in serum or plasma in an allele dose-dependent manner, and that this genetic regulation is somewhat disrupted in ALS. Furthermore, plasma angiogenin levels predict the level of angiogenin observed in CSF to some extent, but this is not observed in ALS cases.

While these findings are compatible with the previous discovery of *ANG* mutations in ALS [46], the majority of the patients that were assayed for this study would not have had mutations in *ANG*. Indeed, the work described in chapter 4 failed to identify any *ANG* mutations in any ALS patients, suggesting that this is a rare phenomenon. Therefore, the findings of dygsregulation of angiogenin in ALS suggest that, although *ANG* mutations are rare, it is possible that the biochemical signalling networks of which angiogenin is a member are perturbed in ALS. This is reinforced by the observation that even when the same genetic variation is present within the *ANG* locus, different patterns of angiogenin expression are observed when compared to controls, suggesting that an outside influence is modulating the effects of *ANG* genetic variation. Furthermore, there is indication that this is observed in a tissue-specific manner.

The interpretation of these observations could incorporate any one of several different

explanations. Future work should aim to demarcate the regulatory mechanisms involved in angiogenin expression and subsequently assess any disruption, or indeed mutations, that may be observed therein. This way, a better understanding of some subtypes of ALS may be gained, including (but not restricted to) ALS caused by mutations in *ANG*. A more complete understanding of the biochemical signalling pathways and networks in which all known ALS genes are involved would be a useful theme for future research, and it would generate a resource that would be invaluable for all ALS research worldwide.

The research detailed in chapter 3 was aimed at investigating the contribution of common genetic variation to the aetiology of ALS through analysis of genome-wide SNP data. This implemented three main approaches: standard GWAS, analysis of putative copy number variation and mapping of recurrent, allelically-matching ALS-specific ROHs. This led to the identification of several speculative intervals involved in ALS aetiology that were carried forward to chapter 4 for the design of the NGS experiments. No single SNP was associated with ALS at genome-wide significance levels, although one SNP (rs6836317) came close. Therefore, instead of assessing stand-out peaks, the undergrowth of more modestly-associated SNPs was considered in a design which attempted to identify intervals associated with ALS by considering neighbouring less-associated SNPs in LD with moderately associated SNPs. Although these were weak associations, when carried forward to chapter 4, of the 395 genes that overlapped with associated intervals, 350 were common to the gene set that was identified by ROH mapping and two were common to all three gene sets. These two genes (*CSMD1* and *ERBB4*) represent attractive candidates for future study.

Mapping of putative CNVs yielded relatively few candidate intervals (37 ALS-specific copy number losses and 25 ALS-specific copy number gains). So that the number of candidate genes for sequencing did not become prohibitively high in chapter 4, only ALS-specific deletions were considered for further analysis. This decision was made under the

assumption that the deletion of a gene could have similar effects to a stopgain or frameshift mutation, and therefore genes identified through mapping of copy number losses would be candidate genes for discovery of such mutations. Given the relative inaccuracy of the CNV mapping approach and the potential for false positives, the rejection of candidate genes identified through mapping of copy number gains was not considered detrimental to the design of the NGS experiments.

The parameters used to map ROHs were very stringent, allowing a ROH to be called only if it was above 104 SNPs in length. More relaxed parameters would yield more results, although this would potientially increase the false positive rate. For example, it was observed that decreasing the minimum ROH length by 20 SNPs resulted in many of the same ROH intervals being identified (still ALS-specific), but with many more individuals contributing to each interval. However, even with the stringent parameters applied in the ROH mapping approach, several genomic intervals were identified that showed evidence of potentially containing recessive disease-causing genetic variants. In many cases, loci were identified by more than one group of allelically-matching ROHs (for example, the case of *PCDH17* and *PDE11A*). These intervals represent very attractive candidates for further study, because they show evidence that different recessive variants within the same genes may be causing ALS.

Candidate intervals identified by all three mapping approaches used in chapter 3 were carried forward to chapter 4 for resequencing of exons of candidate genes within identified intervals. This ongoing project has identified several coding changes that could represent ALS-causing mutations. However, the strategy employed in this study was limited to the sequencing of only coding regions of the candidate genes. Following this restricted hypothesis may have resulted in several potentially ALS-associated mutations being overlooked. A number of examples are cited in chapter 3 in which candidate intervals overlap with either putative regulatory regions of the genome or with non-coding regions. For further

exploration of these interesting regions, the underlying hypotheses need to be extended to incorporate assumptions that ALS-causing mutations may not necessarily lie within coding regions, and, as has been shown recently for *ATXN2* and *C9orf72*, could potentially be genetic variation that is not detectable by naïve NGS approaches.

Despite its limitations, the investigation of genetic variation in coding regions of candidate genes in chapter 4 indicated many interesting findings that should be further pursued in future study. Overall, the findings of this chapter revealed that there are several rare variants that are observed in ALS patients that are not present in the comparison dataset of the 1000 Genomes Project. A small portion of these genes show an improbable excess of rare variants in the ALS cohort and some identified variants segregate homozygously in the expected intervals and within expected individuals. Furthermore, several rare variants have been identified in genes that have previously been implicated in ALS (of which one variant, *DCTN* G59S has previously been shown to cause ALS [47]).

One of the main conclusions drawn in this chapter was that a comparative dataset derived from the same population as the ALS cases is required in order to identify population polymorphisms so that incorrect inferences are not made about the pathogenicity of discovered variants. A cohort of at least 150 controls would be required to obtain 80% power to detect a polymorphism at 1% frequency at least twice, so that a reasonable estimate of the population allele frequency can be made. This would permit the reduction of the overall discovered variant set to a more manageable number. However, for the recessive hypothesis, a further, considerably larger, replication cohort would be required, whose size is dependent on the MAF of the variant in question.

It will require further sequencing of both cases and controls to demarcate properly the roles of discovered rare variants in ALS aetiology. However, speculative findings have been made, for example the role of multiple rare variants within *HYDIN* in the aetiology of ALS. Assessment of multiple rare variants within many loci has been difficult to carry out

in the past due to technological restrictions, however the advent of NGS has eliminated this limitation. As a demonstration of this, chapter 4 has harnessed the power of NGS to assay many rare variants in many loci *en masse* within a reasonably-sized population. The identification of multiple rare variants in *HYDIN* and its consequent potential as an ALS gene represents an exciting finding due to its previous implications in other neurological conditions [215, 216, 218].

Finally, an exploratory analysis was carried out in chapter 5 that aimed to describe the prevalence of IBD in the Irish population, in the hope that this could assist in the optimal design of future genetic studies, such as exome sequencing. A higher rate of IBD was observed within the Irish population when compared to within-Britain IBD, and this showed geographical clustering within particular regions. Furthermore, in a panel of individuals expected to be unrelated, it was observed that using pairwise inferences of IBD extent could identify small groups of individuals that were inter-related, representing members of single extended pedigrees that typically showed relatedness within the bounds of third cousins. Projection of expected outcome of exome sequencing studies based on analysis of data generated in chapter 4 suggested that as few as four ALS patients sequenced along with two hypernormal controls would be sufficient to optimize the capacity to refine discovered variants down to a subset of candidate variants involved in ALS aetiology. Therefore, the identification of these small clusters could be enough for selection of individuals for exome sequencing.

This study investigating IBD in the Irish population has implications for the design of future experiments. Exploiting inferences about relatedness between individuals that have been sequenced for the purposes of variant discovery is analagous to sequencing multiple distantly-related individuals from a familial pedigree. Such approaches have successfully revealed variants associated with disease in the past, including the identification of mutations in *VCP* in familial ALS [69]. Using the IBD mapping approaches described in this

chapter, an alternative method to selection from familial pedigrees has been proposed, which can be applied to population-based cohorts. This represents a valid future direction in the investigation of the contribution of genetic variation to the pathogenesis of ALS.

A final, unanticipated finding of the IBD profiling in chapter 5 was the identification of a within-cases IBD peak that maps to a previously-established familial linkage region on chromosome 9q21 [236]. This finding has speculatively refined the candidate genes within this linkage region to a subset of just eight genes. Furthermore, it demonstrated the potential for such IBD mapping approaches to identify regions of the genome linked to disease pathogenesis, which may represent a technique that is more powerful than haplotype association. Mapping regions of IBD allows for the potential that multiple (potentially oppositely-tagged) haplotyes could harbour different disease variants within the same locus. Taking individuals showing IBD within the 9q21 region, further exploration of the genes identified within the region could yield novel variants associated with ALS aetiology.

Taken together, the chapters 2-5 have yielded several interesting findings. The methods described have been heavily dependent on the development of projects such as the International HapMap Project [5,6] and the Human Genome Project [1] and the associated technological innovations leading to the possibility of GWAS and NGS.

6.2 Evaluation of genome-wide SNP analysis and NGS as methods in ALS research

GWAS has produced mixed results for ALS and it seems that the condition is too genetically heterogenous to permit study designs that are not statistically highly robust. The most successful results for GWAS in ALS have been the chromosome 9p21 locus and *UNC13A* [66, 67]; otherwise there has been difficulty in replicating other loci. Any future

GWAS effort in ALS should not proceed without careful consideration regarding power to detect variants given whatever study design is applied. In this thesis, an attempt has been made to address the small sample size by considering only results that were indicative of ALS-specific phenomena; this, however, is probably only advisable in a setting where a very large number of loci are going to be considered in the downstream investigations (in the case of the work in chapter 4, 1,577 genes).

Caution is also warranted in the design and interpretation of any NGS experiment in ALS. Individuals for exome sequencing should be carefully chosen to maximize the probability of detecting commonly shared disease-causing variants, and in population-based sequencing experiments, sufficient sample sizes should be chosen to obtain sufficient power to detect rare variants. The interpretation of population-based and exome-based sequencing studies should take into account the potential for recessive inheritance as a separate analysis pipeline to approaches that exclude any variant discovered in comparison control datasets.

It has been shown that exome sequencing can be used to good effect in detecting disease-causing variants in ALS [69] and it is likely that there will be future studies using similar methodologies that will identify other causative ALS genes. However, a limitation of current exome sequencing approaches is that the target enrichment kits used to capture the exome does not cover all consensus coding sequence annotated exons [221], which means that important information could be missed. Future revisions to target enrichment designs are likely to address these issues to some extent, but the presence of repeat sequence in the genome will always be a problem for complementary bait-based enrichment strategies.

In summary, GWAS and NGS are tools that hold great promise for research in the genetics of ALS, but they need to be applied with care. Given the proportion of ALS still unexplained by known genetic loci, yet the probable genetic contribution to the condition

[41], research using such methods should continue to be applied.

6.3 Future directions

From the findings of the work detailed within this thesis, a number of points of action have been identified that are worthy directions for future work. These relate both to ideas that have come about as a direct consequence of the work in this thesis, and ideas that are derived from developing trends and technologies in genomics research.

6.3.1 Future directions based on the findings of this thesis

Further sequencing

A primary goal of future work is to delineate the implications of the discovery of many rare sequence variants in chapter 4. Furthermore, a trove of variants are probably still to be found through sequencing of representative individuals from candidate locus mapping studies detailed in chapter 3. A sequencing experiment that matches that of chapter 4 in size would be a good starting point, and sequencing of at least 150 population-matched controls is essential. This way, population-based rare polymorphisms can be distinguished from rare disease-associated variants. Subsequent follow-up, through genotyping or more sequencing, in large replication populations will be necessary after these sequencing studies to validate arguments for recessive variants, as well as replicating any other variant discovererd.

Replication and functional follow-up with any genes implicated through sequencing

As mentioned, replication of variants will be necessary. A well-matched replication population in which similar genetic structure would be expected is the Scottish, due to recent ancestral links. However, replication of any discovered variant in further European pop-

ulations will help to validate any discovery further. Although it would be expected that mutations seen in Ireland should be seen in other countries in Europe, it is not necessarily a certainty in that founder effects for disease variants could be restricted to the geographical environs of the northwest of Europe. Nevertheless, assessment of discovered variants in other populations will be absolutely necessary to understand their contribution to ALS aetiology.

Further investigation of loci identified in chapter 3

The loci that were mapped in chapter 3 were only considered in chapter 4 under the hypothesis that protein coding changes are driving ALS pathogenesis. This may not necessarily be the situation in many cases of ALS, and most of the intervals identified in chapter 3 could just as well be interpreted from the viewpoint that noncoding and regulatory regions are perturbed in the disease. For this reason, it will be worthwhile to revisit the analyses and assess whether some of the observed patterns might be explained by non-coding changes such as the structural variation seen in *C9orf72* [121,122]. Assessment of putative changes could then possibly be as easy as a PCR-based check for expansions; however there may be several intervals identified in the analyses so a high-throughput solution may be more suited to the task.

Further investigation of the chromosome 9q21 locus

Identification of considerably higher IBD within cases than within controls for the locus overlapping chromosome 9q21 was an unexpected finding, but it represents a potentially very frutiful line of future work. Chapter 5 identifed just eight genes (*RORB*, *TRPM6*, *CHAK2*, *C9orf40*, *BC043649*, *C9orf141*, *C9orf95* and *OSTF1*) to which the original linkage region [236] could be refined; these genes should be the subject of further analysis. The relative absence of discussion around the chromosome 9q21 locus in the literature

following its mapping in 2000 is indicative that a potential opportunity to identify a novel ALS gene has been missed, possibly due, in part, to it not being identified in GWAS. However, the IBD evidence in chapter 4 is compatible with the possibility that multiple rare variants are present within the locus in a manner that may not be detectable by GWAS. Future work should investigate the haplotypes that were driving the IBD signals, and identification of novel disease-causing variants within the locus.

Further genotyping

Although the field of genomics has moved on, to some extent, from GWAS to NGS-centred methodologies, there may be room for a little more genome-wide SNP analysis in ALS, at least in the Irish population. Every ALS GWAS effort to date in Ireland (including the 2008 GWAS [98] and the work described in chapter 3) has been underpowered, yet population-specific ALS variants may be detectable, given more genotypes. ALS is a genetically heterogenous disease and particular variants may have drifted to higher frequencies within the Irish population (there is speculative suggestion in chapter 4 that the spectrum of ALS-causing variants in known genes within Ireland is quite different to that seen worldwide), the detection of which would require a well-powered Irish GWAS. Future work in this respect should carry out further genotyping in the ALS case cohort that has built up since the last batch of genotyping, as well as in controls (perhaps collaborating with research groups performing similar work where necessary), from which point GWAS and other SNP-related studies can be carried out. One such related study would be the inference of IBD described in chapter 5, where further relationships could be inferred and groups of individuals could be identifed (in combination with an endophenotyping approach) for studies assaying rare variation, such as exome sequencing.

6.3.2 Future directions for ALS genetics research

Whole-genome sequencing

The potential for exome sequencing in ALS has been discussed at length in this thesis. The principles of exome sequencing also apply to potential future work that will use whole-genome sequencing to identify risk variants in ALS. While this is still some way off in terms of affordability, technology is advancing at a rate such that it is worth considering the potential opportunities and pitfalls now. The issues surrounding variant filtration and interpretation in terms of pathogenicity will be present with whole-genome sequencing, but on a much greater scale (the human genome is over 1,000 times the length of the total sequence in the target enrichment kit described in chapter 4). For this reason, careful experimental design would be critical in whole-genome sequencing experiments, so that true disease-causing variants stand out from background genetic variation across the genome. However, whole-genome sequencing will afford the opportunity to assay genetic variation in regions not covered by exome capture strategies, which opens up a number of unexplored hypotheses. Nevertheless, the alignment, analysis and interpretation of whole-genome data will represent a huge challenge that will require, amongst other things, further development in the software tools currently used in order that this challenge can be effectively met.

Endophenotyping

To ensure success in sequencing projects geared towards the identification of pathogenic rare variants, a sensible first step would be the establishment of detailed endophenotypes within the case cohort. ALS is a phenotypically heterogenous disease, with a spectrum of associated cognitive and behavioural impairment [29, 30] and variability in site of onset, age of onset and disease duration. This phenotypic variation is likely to be driven, to some extent, by the genetic heterogeneity underlying the disease, and the isolation of a

particular endophenotype may help to enrich the probability that a common disease variant is shared amongst cases that share the endophenotype, thus facilitating its detection by techniques such as exome sequencing. For example, the recently-discovered hexanucleotide repeat expansion in *C9orf72* segregates with a phenotype of behavioural change and family history of FTD [238]. This approach of endophenotyping is also applicable to GWAS design. For example, in the AMD GWAS [16], the authors managed to detect a SNP association with AMD with a cohort of only 96 cases and 50 controls by carefully choosing a very specific disease endophenotype, as well as choosing hypernormal controls. Large-scale GWAS in ALS, taking advantage of detailed endophenotypes, may yield novel loci associated with particular variants of the disease.

Endogenotyping

The endophenotyping approach is intended to maximize the likelihood that a particular disease genotype is seen commonly to all cases with a particular form of the disease. This, however, only holds if the endophenotype is specific enough that it permits complete distinction of a sub-cohort of individuals from all other forms of the disease. In many cases, however, this will not be the case and in these instances, the best approach to maximize the probability of detecting a signal at a disease locus is to enrich the cohort as much as possible for as few putative disease-causing variants as possible by exclusion of all possible alternatives. Here, profiling of cases for variants known to be involved in ALS aetiology is a useful step in the experimental design. This way, cases can be identified for which the genetic cause is already explained, and the cohort of individuals under study can be enriched for any novel variants yet to be discovered. However, results of screening of known genes should be interpreted with caution. This is argued well by Felbecker *et al.* [44]; for genes known to be involved in ALS, especially *SOD1*, there has historically been a propensity to over-report findings without conclusive evidence for disease variant

pathogenicity (this is an argument for a comperehensive 'spring clean' of all known or suspected ALS mutations; such a study would serve as a valuable resource to the ALS genetics community). As with endophenotyping, the enriched cohort could then be used for a variety of disease locus discovery methods, such as GWAS or NGS.

'Isophenotyping'

The benefits of endophenotyping and endogenotyping apply to GWAS and NGS experiments within ALS. However, given the clinical overlap between ALS and FTD and the common broad theme of neurodegneration between ALS and a number of other diseases of the central nervous system, it may be beneficial to attempt studies such as GWAS where cases are combined between many related phenotypes under the hypothesis that genetic risk for general neurodegeneration may be conferred by common variation at a few loci, which is then modulated by other genetic or environmental factors. This way, power to detect associations could be dramatically increased (subject to apt selection of hypernormal controls), and if associations were discovered, the relationships between the similar phenotypes could be further explored to elucidate the common pathogenic mechanisms.

Expression studies

The work described within this thesis has focussed on the identification of genetic variation that confers susceptibility to ALS. This is dependent on either germline or *de novo* mutations being present in the individuals with ALS, and for these to be detectable using the methods of GWAS or NGS. An alternative approach to the problem of identifying genetic contribution to the disease could be to assay the expression of mRNA in patients affected with ALS, on a genome-wide scale, to assess whether there is perturbation of biomolecular signalling modules within the condition. This can be achieved through either microarray analysis or high-throughput RNA sequencing by NGS (RNA-seq) [177]. This top-down

approach has the benefit of being able to assay genetic variation on the level of the pathway, which may be more conducive to easy interpretation. Furthermore, it reveals answers that are closer to the biological end point than simply identifying mutations that may or may not be involved in the functional aetiology of the disease.

However, while this methodology has clear benefits, it raises issues in experimental design that are not a problem with GWAS or NGS using genomic DNA. One of the major problems is tissue choice. In order for accurate inferences to be made about perturbation of expression profiles in ALS, the preferred tissue from which mRNA is derived would be motor neurones. Sampling of such tissues would require autopsy *post-mortem*, which would have to be carried out within a narrow window of time due to the short half-life of RNA. This presents a problem in the majority of cases. A suboptimal alternative is to use RNA derived from leukocytes. A third option could be to culture induced pluripotent stem cells [239] derived from fibroblasts of patients and bring about their subsequent differentiation to motor neurones [240], from which RNA could then be derived.

Although ambitious, such studies represent a real opportunity in ALS research. Further to the information derived on expression levels, it has been suggested that RNA-seq could be used as an alternative to exome sequencing for assessing genetic variants and polymorphisms [241]. Although this would be affected by RNA editing and any other alteration in genetic sequence between genomic DNA and mRNA, it would be an efficient use of resources as it leads to two separate result sets from a single experiment.

Structural variation

With recently-reported findings of hexanucleotide repeat expansions in *C9orf72* [121,122] and polyglutamine repeats in *ATXN2* [56], the role of structural variation is becoming increasingly evident in the pathogenesis of ALS. This argument is further supported by previously-reported findings of putative CNVs associated with ALS [109, 135] and the

attempts described within this thesis to map CNVs specific to ALS. In light of these ideas, the discovery of structural variation associated with ALS pathogenesis is a justified future pursuit. Techniques such as aCGH would reveal the locations of large-scale structural variation with greater certainty than SNP array based methods, but with these techniques small scale variation could be missed. Very small structural changes can be discovered with NGS, as described in chapter 4, however indels above a certain size lead to problems with alignment and subsequent calling of the structural variant. Although its use has not yet been extensive, high-throughput sequencing of mate-pair libraries has been shown to be applicable to the detection of chromosomal rearrangements in disease genetics [242]. Such a technique could well be applied to ALS to perform a genome-wide screen of structural variation. However, it is likely to be through the combination of technologies that the best inferences are made.

Regulatory regions

The methods described in chapter 4 made the assumption that protein coding changes are responsible for some of the aetiology of ALS. This is likely to be true in many cases, but the possibility exists that sequence variants in non-coding regions of the genome contribute to ALS pathogenesis. Indeed, recurrent ROHs shown in figure 3.12(b) for *PCDH17* seemed to map better to upstream non-coding regions of the genome than to the gene itself. This was then reflected by a lack of discovery of recessive variants in *PCDH17* for the relevant individuals. The upstream region to which the *PCDH17* ROHs mapped contained several peaks that have been identified by the ENCODE project [158], representing putative enhancer or promoter regions (as well as containing a non-RefSeq mRNA). Future NGS experiments could be designed to take non-coding and regulatory regions into account by making use of data such as that generated by the ENCODE project.

To date, there has been little emphasis placed on the possibility that epigenetic varia-

tion may contribute to ALS aetiology. This is likely, in part, to be due to the difficulty in assaying this principle on a genome-wide scale. However, the Illumina Infinium assay has recently been modified to allow for genome-wide interrogation of DNA methylation [243] which allows for the design of epigenome-wide association studies (EWAS) [244], Such studies are likely to be vulnerable to the same statitstical issues in their design as GWAS, but through careful application, the principle could yield novel answers regarding the aetiology of ALS. ChIP-seq [178] is another method that holds promise for the future in determining the role of genetic regulation in disease.

Epistasis

A frequently-overlooked possibility in complex disease genetics is that genotypes for multiple SNPs may statistically interact, meaning that in order for a patient to manifest a disease they must have risk alleles for two (or more) separate SNPs. The principal reasons for this hypothesis not being popular in complex disease genetics are that associations with the SNPs would be very difficult to detect independently, and when considered together, the multiple testing burden for all pairwise comparisons with genome-wide SNPs becomes too extreme for any reasonable study to have statistical power under traditional association study designs. However, the potential for epistatic interaction on a genome-wide scale was investigated by Sha *et al.* [168] by adopting a two-stage approach where only top GWAS results were considered in the subsequent two-locus analysis. This approach reduced the computational and statistical burden of performing genome-wide multilocus association testing. Methods exist that can ameliorate the computational burden of full genome-wide epistasis modelling through machine learning approaches such as random forests [245] or multifactor dimensionality reduction [246]. Employment of such approaches on large genome-wide SNP datasets for ALS may reveal previously undetermined associations by considering multiple loci simultaneously.

Collaboration

Possibly the most important future direction for ALS genetics research is international collaboration between research groups. As was shown with GWAS, the only way a strong signal of association will be detected is by assessment of large numbers of samples [66,67], which is usually only obtainable through international collaboration. It will be similar with studies assaying rare variation also: verification of pathogenicity will require replication in further populations. This is especially true for population variants that cause recessive forms of ALS as the sample sizes required to discover rare population-based variants homozygously are enormous. A further benefit of international collaboration is the sharing of knowledge between different research groups with different areas of expertise and experience.

6.4 Conclusion

ALS is an unrelenting, incurable, fatal disease that strikes patients in the prime of their lives, often without warning. The inevitable death that comes as a result is rapid enough that patients and their families have little time to come to terms with the condition, but slow enough that the final months or years of patients' lives are spent in considerable discomfort with very low quality of life. In order that effective intervention can be designed, an understanding of the underlying disease mechanisms is an absolute requisite, and part of the work to this end is focussed on elucidation of the genetic causes for the condition. With a better understanding of the genetic mechanisms, a more complete picture of the molecular cell biology can be ascertained, and subsequent pharmacological intervention follows.

At the time of writing, it has been two decades since the first gene to be involved in the pathogenesis of ALS, *SOD1*, was discovered [42], and since then, a constellation of genes

have been proposed to be involved in the condition. Despite this, there is still no cure for the disease and its pathophysiology is only partially understood. Use of modern techniques derived from the revolution in genomics technologies holds promise to be useful in further elucidating the genetic cause of the condition, although a complete understanding of the mechanisms behind the pathogenesis of the disease will require extensive downstream molecular biology work. Nevertheless, through apt use of genome-wide SNP analysis and NGS, researchers are maximizing the prospect for completing the puzzle that the complex genetics of ALS presents.

In order for this to be realized, however, the experiments performed must be robust and replicable in order that the published findings serve to lead researchers in the correct directions. The many GWAS published in ALS have struggled to replicate one another, and the main successes have been derived from international collaboration permitting the necessary sample sizes to perform well-powered GWAS. Similar principles will be true for NGS studies involving exome sequencing, candidate gene sequencing and transcriptome sequencing: it will only be through careful experimental design and thorough analysis delivering replicable results that real progress will be made. Findings should draw on the contexts of other experiments, relating the observations of multiple studies in order that a more complete picture can be drawn.

Looking ahead from a viewpoint that is firmly rooted in the genomic era, surrounded by an arsenal of emerging tools, techniques, data and models, it is clear to see that there is a lot of work ahead for researchers in ALS genetics. It is hoped that ongoing work derived from the studies described in this thesis will contribute to the growing body of knowledge surrounding ALS genetics and that years to come will see such knowledge contribute to a cure for this devastating disease.

Appendix A

Selected scripts

A.1 `recipoverlap.pl` (section 3.2.5)

Section 3.2.5 mentions a formula that was used to assess the extent of overlap of a ROH segment with the ROH group to which it belonged:

$$S_i = \sum_{j=1}^{n} \frac{\min(b_i, b_j) - \max(a_i, a_j)}{(n-1)(b_j - a_j)}, \tag{A.1}$$

where a is the start of a segment, b is the end of a segment and $i \notin j$. The derivation of this equation is straightforward, but it may not seem obvious at first.

As an explanation of the origin of this formula, consider the example overlapping ROH group depicted in figure A.1. The objective is to ascertain how well each segment overlaps with the group as a whole, and this can be determined by assessing the overlap between each segment and every other segment individually, then averaging these for the group. As an example, the overlap of ROH1 will be described.

First, consider the overlap between ROH_1 (i) and ROH_2 (j). This is defined by the distance between the highest start position and the lowest end position, or:

$$\min(b_i, b_j) - \max(a_i, a_j). \tag{A.2}$$

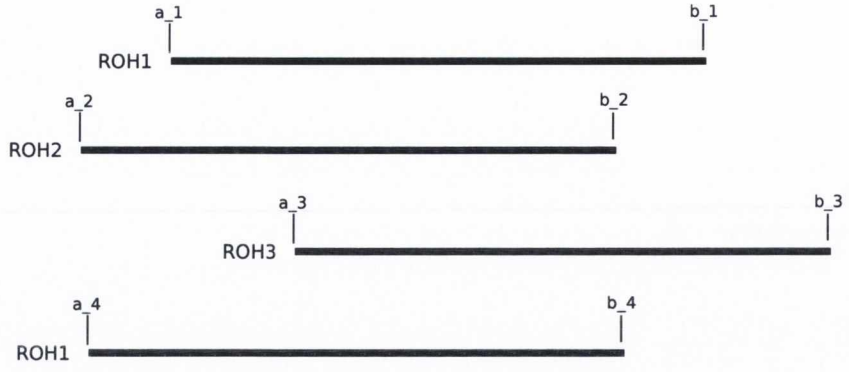


Figure A.1: An example ROH group.

To obtain the overlap between ROH1 (i) and all the segments in the group (all j), every i/j overlap is summed:

$$\sum_{j=1}^{n} \min(b_i, b_j) - \max(a_i, a_j), \tag{A.3}$$

and divided by the total number of pairwise comparisons made, which is the size of the group minus one:

$$\sum_{j=1}^{n} \frac{\min(b_i, b_j) - \max(a_i, a_j)}{(n-1)}. \tag{A.4}$$

This is standardized to the sizes of the segments j for which the comparisons were made by dividing each segment length by the total length of j's segment, $(b_j - a_j)$, such that

$$S_i = \sum_{j=1}^{n} \frac{\min(b_i, b_j) - \max(a_i, a_j)}{(n-1)(b_j - a_j)}. \tag{A.5}$$

In section 3.2.5, ROHs with overlap scores of 0.5 or greater were taken as high-quality. This method was written into the script `recip_overlap.pl` to automate the calculation of S_i scores in overlapping ROH groups.

A.2 Visualizing power (section 4.4)

Chapter 4 details the sequencing of 106 individuals geared towards detection of rare variants in ALS. One conclusion drawn from this work was that the sequencing of controls would be useful to identify population-based variants that are rare in the world but common in Ireland. An obvious question that arises from this notion is: how many controls would be necessary to detect Irish population-based variants?

In this instance, 'rare' is defined as a variant that is at a population MAF of 1% or less. So, to identify 'common' population-based variants we are interested in variants that are at a frequency of at least 1% in Ireland. Assuming a variant of MAF 1% is in Hardy-Weinberg equilibrium, the frequency of individuals carrying at least one copy of the minor allele is $0.01^2 + 2 \times 0.01 \times 0.99$ which evaluates to 0.0199, or 1.99%.

Detection of such variants can be modelled as sampling without replacement from a hypergeometric distribution. This is easily acheived using R's `rhyper` function, which is called using the syntax `rhyper(nn,n,m,k)`, where `nn` is the number of times sampling is repeated (the number of 'experiments' performed, `n` is the number of individuals not carrying the minor allele, `m` is the number of individuals carrying the minor allele and `k` is the number of individals sampled. This method returns a vector of values of length `nn`, such that the 'power' (the proportion of times the variant is detected) can simply be calculated using the R command `length(vector[vector>0])/length(vector)`, or in words, the number of observations greater than zero divided by the total number of observations.

This can be incorporated into a script such that the power to detect this variant given various different sample sizes can be plotted. To do this, a `for` loop is constructed and the `rhyper` function is called several times, for many sample sizes, and the results are stored in a vector.

```
power<-0
for( i in 1:250 )
{
  people<-rhyper(1000,1990,98010,i)
  power[i]<-length(people[people>0])/1000
}
```

This produces a vector of values that, when plotted, would show some variation, so the data can be smoothed by taking a moving average.

```
xpos<-0
smoother<-0
for( i in 10:length(power))
{
  xpos[i]<-i
  smoother[i] <- sum(power[(i-5):(i+4)])/10
}
plot(xpos,smoother,type="l",col="blue",lwd=2,xlab="Number of individuals",ylab="Power")
```

This generates a plot similar to figure 4.6(a).

Appendix B

Digital appendix contents

Some of the scripts written to facilitate data collection, parsing and interpretation within this thesis have been made available as a digital appendix, which can be downloaded at:

http://www.gen.tcd.ie/molpopgen/resources/rlmclda

This link will be maintained for at least five years after the submission date of the thesis. The majority of the scripts are written in Perl, with some scripts for use with the R statstical programming package, and some designed to be parsed by a web server such as Apache, for web browser-based visualisation of data. Most, if not all, of the scripts could easily be modified to meet the needs of other related projects, or could be run without any modification at all on datasets derived from such projects.

Table B.1 lists the contents of the directory, which is provided as a .tar.gz archive and can easily be extracted using the Linux `tar` tool by typing:

```
tar -xvf rlmclda.tar.gz
```

in the command line of a terminal (Linux and Mac) or in the command line of a Unix-like environment such as Cygwin (Windows).

Table B.1: Digital appendix contents

File	Section	Function
`LvGPlot.R`	2.2.4	Correction, visualization and analysis of levels vs genotypes for *ANG* data
`replicate_overlap.pl`	3.2.4	Assessment of concordance of QuantiSNP and PennCNV output
`overlap.pl`	3.2.4	Intersection of output of QuantiSNP and PennCNV
`append.pl`	3.2.4	Collation of CNV results into a single file
`count_status_per_SNP.pl`	3.2.4	Find case and control counts for CNV overlap per SNP
`parsePLINKROH.pl`	3.2.5	Parsing `.hom.overlap` file into case-specific overlapping ROHs only
`removedups.pl`	3.2.5	Removal of duplicate tables from `parsePLINKROH.pl` output
`showSNPs.php`	3.2.5	Visualization of genotypes in case-specific ROH regions
`showSNPs.css`	3.2.5	CSS style sheet to accompany `showSNPs.php` for colouring of SNP alleles
`recipOverlap.pl`	3.2.5	Calculation of S_i for within-group overlap of ROH
`plotNatureStyle.R`	3.3.2	Manhattan plot with various other data co-plotted, inspired by [152]
`parse_exons.pl`	4.2.1	Parsing of UCSC-format gene data into exonic intervals
`interval_overlap.pl`	4.2.1	Consolidation of overlapping intervals into single intervals
`rescue_intervals.pl`	4.2.1	Identification of poorly-baited regions due to repeat intervals
`double_up_singletons.pl`	4.2.1	Identification of singleton probes and generation of redundant probes
`split_seq_data.pl`	4.2.5	Splitting of FASTQ file based on barcodes in forward and reverse sequencing reads
`filter_1kg.pl`	4.2.6	Annotation of discovered variants if present above stated frequency in 1000 Genomes reference file
`count_variants_per_gene.pl`	4.2.6	Counting discovered variants per gene for burden analysis
`find_homies.pl`	4.2.6	Location of homozygous genotypes in expected intervals in expected individuals, conditional on absence of homozygosity for variant in comparison data
`power.R`	4.4.3	Power simulations for NGS experiments
`parse_WTCCC.pl`	5.2.1	Simple parsing script for WTCCC `.gen`→`.ped` conversion
`correct_strand.pl`	5.2.1	Correction of strand from Illumina FORWARD to TOP
`parse_for_cc_comparison.pl`	5.2.3	Counting extent of IBD in cases and controls per SNP for genome-wide plotting
`cluster_IBD.pl`	5.2.3	Brute-force IBD clustering method

Appendix C

Publications

OPEN ACCESS Freely available online

PLoS one

Angiogenin Levels and *ANG* Genotypes: Dysregulation in Amyotrophic Lateral Sclerosis

Russell Lewis McLaughlin[1,2]*, Julie Phukan[2,3], William McCormack[4], David S. Lynch[2,5], Matthew Greenway[5], Simon Cronin[2,5], Jean Saunders[6], Agnieska Slowik[7], Barbara Tomik[7], Peter M. Andersen[8], Daniel G. Bradley[1], Phil Jakeman[4], Orla Hardiman[2,3]

1 Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland, **2** Department of Neurology, Beaumont Hospital, Dublin, Ireland, **3** Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland, **4** Human Science Research Unit, Faculty of Education and Health Sciences, University of Limerick, Limerick, Ireland, **5** Royal College of Surgeons Ireland, Dublin, Ireland, **6** Statistical Consulting Unit, University of Limerick, Limerick, Ireland, **7** Department of Neurology, Jagiellonian University, Krakow, Poland, **8** Department of Pharmacology and Clinical Neurosciences, Umeå University Hospital, Umeå, Sweden

Abstract

Objective: To determine whether 5 single nucleotide polymorphisms (SNPs) associate with ALS in 3 different populations. We also assessed the contribution of genotype to angiogenin levels in plasma and CSF.

Methods: Allelic association statistics were calculated for polymorphisms in the *ANG* gene in 859 patients and 1047 controls from Sweden, Ireland and Poland. Plasma, serum and CSF angiogenin levels were quantified and stratified according to genotypes across the *ANG* gene. The contribution of SNP genotypes to variance in circulating angiogenin levels was estimated in patients and controls.

Results: All SNPs showed association with ALS in the Irish group. The SNP rs17114699 replicated in the Swedish cohort. No SNP associated in the Polish cohort. Age- and sex-corrected circulating angiogenin levels were significantly lower in patients than in controls ($p<0.001$). An allele dose-dependent regulation of angiogenin levels was observed in controls. This regulation was attenuated in the ALS cohort. A significant positive correlation between CSF plasma angiogenin levels was present in controls and abolished in ALS.

Conclusions: *ANG* variants associate with ALS in the Irish and Swedish populations, but not in the Polish. There is evidence of dysregulation of angiogenin expression in plasma and CSF in sporadic ALS. Angiogenin expression is likely to be important in the pathogenesis of ALS.

Citation: McLaughlin RL, Phukan J, McCormack W, Lynch DS, Greenway M, et al. (2010) Angiogenin Levels and *ANG* Genotypes: Dysregulation in Amyotrophic Lateral Sclerosis. PLoS ONE 5(11): e15402. doi:10.1371/journal.pone.0015402

Editor: Thomas Mailund, Aarhus University, Denmark

Received August 3, 2010; **Accepted** September 10, 2010; **Published** November 10, 2010

Funding: This work was supported by the Irish Health Research Board (grant number HRB CSA 2007/3); the Irish Motor Neurone Disease Research Foundation (grant number IMNDRF 07/01) and the Muscular Dystrophy Association (United States of America, grant number 2007/4252). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mclaugr@tcd.ie

Introduction

Angiogenin is the 14.1-kDa product of the hypoxia responsive gene *ANG* on chromosome 14. We have shown previously that mutations in *ANG* are associated with amyotrophic lateral sclerosis (ALS), and that *ANG* mutations predict loss of RNAse and angiogenic function [1]. Moreover, recent studies have suggested that angiogenin is an important neurodevelopmental protein with neuroprotective properties, and that mutant *ANG* impairs neurite outgrowth. [2–4].

Angiogenin is functionally similar to vascular endothelial growth factor (VEGF), altered regulation of which has also been associated with ALS [5,6]. 'At risk' promoter haplotypes in VEGF, which predict reduced expression of bioavailable isoforms, have been described in some European ALS populations [7] and combined with evidence from animal models, the data suggest that VEGF isoforms have a neuromodulatory and neuroprotective role in the CNS. Despite the functional similarity between angiogenin and VEGF, there have been few studies to date that have investigated angiogenin expression and regulation in ALS.

We have recently shown that serum angiogenin levels in ALS differ from controls [8]. The patterns of plasma and cerebrospinal fluid (CSF) angiogenin expression have not previously been investigated, and there have been no studies to determine whether *ANG* haplotypes modulate protein expression, as is the case with VEGF. We have sought to determine (i) whether angiogenin is detectable in CSF, (ii) whether there is a consistent relationship between plasma and CSF angiogenin levels, (iii) whether genetic variations in the *ANG* locus control angiogenin expression, and (iv) whether, as has been reported for VEGF [9–11], there is a dysregulation of angiogenin in sporadic ALS.

Methods

Participants

DNA and serum samples were drawn from Irish and Polish ALS patients; DNA, plasma and cerebrospinal fluid (CSF) samples were drawn from Swedish ALS patients. Unrelated control subjects with no family history of ALS were sampled from the same populations. The numbers of participants available in the three study populations and their demographics are detailed in figure 1. All patients fulfilled the El Escorial criteria for clinically definite or probable ALS [12]. Patients with atypical phenotypes and Swedish patients with mutations in the *SOD1* gene were excluded. Informed written consent was obtained from all participants and the study was approved by the ethics committees in Beaumont Hospital, Umeå University and the Jagelonian Institute.

SNP genotyping

Using data from the CEPH panel of the International HapMap Project [13], 5 informative haplotype-tagging single nucleotide polymorphisms (htSNPs) were selected covering the *ANG* gene with inter-marker r^2 below 0.8 and minor allele frequency above 5%. These htSNPs are detailed in table 1. Genotyping across these five htSNPs was performed commercially by KBiosciences (Herts, UK) using KASPar assays with standard quality-control criteria (genotypes formed three distinct clusters, water controls were negative and minor allele frequency was above 5%).

Quantification of angiogenin in CSF, plasma and serum

Serum and plasma were isolated from peripheral blood according to standard protocols. Since angiogenin has not been shown to have any interaction partners in the blood, plasma and serum angiogenin concentrations were considered to be comparable. Samples were stored at −80°C until assay. Angiogenin concentration was measured by enzyme-linked immunosorbent assay (ELISA) according to manufacturer's guidelines (Quantikine Duoset, R&D Systems, Abingdon, UK). All samples were assayed in duplicate and calibrated against serially diluted standards of known mass. Pooled CSF and plasma quality control (QC) samples were both assayed in duplicate on each mitrotitre plate, setting the precision of the assay across all microtitre plates. An

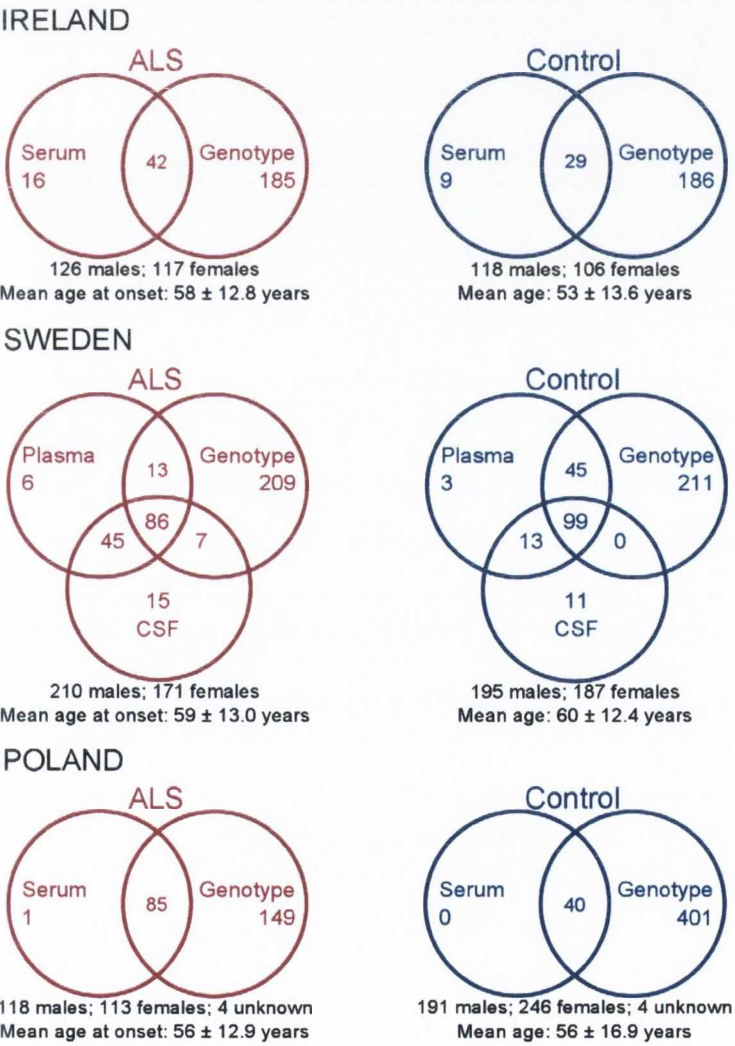


Figure 1. Numbers of individuals and demographics of the three study populations. Error values for mean ages represent standard deviation.

doi:10.1371/journal.pone.0015402.g001

Table 1. Allele frequencies and SNP association statistics in the three populations.

SNP	Alleles	IRELAND				SWEDEN				POLAND			
		RA	RA freq	Allelic association		RA	RA freq	Allelic association		RA	RA freq	Allelic association	
			ALS;ctrl	p	OR		ALS;ctrl	p	OR		ALS;ctrl	p	OR
rs9322855	A>C	C	0.50; 0.59	0.003*	1.57	A	0.56; 0.52	0.13	0.85	A	0.55; 0.55	0.92	0.99
rs8004382	G>A	G	0.57; 0.47	0.007*	1.50	G	0.55; 0.52	0.46	0.92	G	0.55; 0.52	0.46	0.92
rs4470055	G>A	A	0.29; 0.22	0.03*	1.47	A	0.25; 0.24	0.66	1.06	G	0.75; 0.72	0.36	0.88
rs17114699	G>T	T	0.16; 0.11	0.03*	1.53	T	0.14; 0.08	0.001*	1.78	G	0.89; 0.87	0.68	0.93
rs11701	T>G	G	0.18; 0.10	0.006*	1.88	G	0.13; 0.13	0.69	1.07	G	0.13; 0.10	0.14	1.3

RA, risk allele; OR, odds ratio.
*Significant p-value.
doi:10.1371/journal.pone.0015402.t001

inter-assay coefficient of variation (CV) of 6% and 8% was obtained for the high and low plasma QC respectively. An inter-assay CV of 9% was obtained for the CSF QC.

Statistical analysis

Unless otherwise stated, all statistical analyses were performed using the R statistical programming environment [14]. Assessment of allele frequencies were conducted using the computer programmes Haploview [15] and PLINK [16]. Allelic association statistics were calculated using the chi-squared test, with correction for multiple testing by replication in the three populations. Haplotype blocks were defined as a group of htSNPs whose upper 95% confidence bound for D' exceeded 98% with the lower bound above 70% [17] and a haplotype was examined if it occurred in more than 1% of individuals. Haplotypes were tested for association with ALS risk using the chi-squared test.

The data for angiogenin levels were assessed for the reported influence of age and sex [18]. Using data pooled from cases and controls in all three populations, angiogenin levels were regressed against age and sex and an outlier was identified and removed if its studentized residual exceeded the critical t statistic for the group's Bonferroni-corrected 5% significance threshold. The regression analysis was then re-iterated until no further outliers could be identified. Four Swedish plasma values and four Swedish CSF values were removed this way. The resulting linear models were used to adjust the values in the respective groups based on age and sex. The influences of genotypes across the five htSNPs were then assessed by analysis of variance (ANOVA) for each htSNP and the differences between case and control angiogenin levels for each genotype were assessed for statistical significance using the Mann-Whitney-Wilcoxon test. Finally, using data from the Swedish population, corrected plasma angiogenin levels were assessed for correlation with corrected CSF angiogenin levels in ALS patients and in controls independently.

Results

ANG SNP and haplotype association

The mean genotyping call rate across all htSNPs in the three populations was 98.4%. No htSNP deviated significantly from Hardy-Weinberg equilibrium in any study population. The results for the allelic association tests for the five htSNPs are shown in table 1. Linkage disequilibrium (LD) between htSNPs is shown in Figure S1. All five htSNPs showed association with risk for ALS in the Irish study group, with one htSNP, rs17114699, replicating in the Swedish population ($p_{Irish} = 0.03$; $p_{Swedish} = 0.001$). No htSNP showed association in the Polish population. A haplotype block was identified in all three populations, incorporating SNPs rs9322855, rs8004382 and rs4470055. The AAG and CGA haplotypes at these three SNPs associated with ALS in the Irish data, while the AGG haplotype showed strong association with ALS in the Swedish data (table 2).

Plasma, serum and CSF angiogenin levels

Age and sex both had a significant effect on angiogenin levels in plasma/serum and in CSF ($P(>|t|) < 0.0001$ for all covariates). Using data pooled from the three populations and after correcting for age and sex, angiogenin levels were significantly lower in ALS patients than in controls in plasma/serum (mean ± SD = 438.2±112.2 ng/ml for the ALS group and 467.6±105.4 ng/ml for controls; $p = 0.001$, Mann-Whitney-Wilcoxon test) and in CSF (mean ± SD = 5.582±1.754 ng/ml for the ALS group and 6.197±1.987 ng/ml for controls; $p = 0.01$, Mann-Whitney-Wilcoxon test). Angiogenin levels did not differ significantly depending on whether they were measured from serum or plasma ($p = 0.93$; Figure S2). There was a significant positive correlation ($p<0.0001$, Pearson product-moment correlation) between corrected CSF angiogenin levels and corrected plasma angiogenin levels in controls, whereas in ALS patients ($p = 0.21$) the observed correlation was attenuated (figure 3; $r^2_{control} = 0.13$, $r^2_{ALS} = 0.011$).

Contribution of SNP genotypes to angiogenin levels

Levels varied considerably around the fitted models (multiple $r^2_{serum/plasma} = 0.074$; multiple $r^2_{CSF} = 0.16$). ANOVA was used to assess the contribution of genotype at each htSNP to the overall variance in the data and the Mann-Whitney-Wilcoxon test was used to assess the differences between corrected plasma/serum levels in ALS patients and controls for each SNP, separated by genotype. Data were analysed both as independent populations and also as a pooled dataset. The results of these tests, along with the group means, are reported in figure 2.

In the large Swedish dataset, an allele dose-dependent regulation of plasma angiogenin was readily observable for all SNPs in controls and perturbation of this pattern was seen in ALS patients at SNPs rs8004382 and rs9322855. These findings are reflected in the pooled dataset. Only at SNP rs11701 was a significant contribution of genotype to the variance in controls observable in all three populations; however, in the pooled dataset genotypes at every SNP except rs9322855 were shown to contribute significantly to variance in controls. No SNP contributed significantly to variance in ALS patients in the pooled data, however this was observed at rs11701 in the Polish dataset.

Table 2. Haplotype frequencies and association statistics in the three populations.

Haplotype	IRELAND			SWEDEN			POLAND		
	Freq (ALS;ctrl)	p	Permuted p	Freq (ALS;ctrl)	p	Permuted p	Freq (ALS;ctrl)	p	Permuted p
AAG	0.45; 0.53	0.024*	0.13	0.46; 0.47	0.64	0.99	0.453; 0.474	0.4498	0.95
CGA	0.29; 0.22	0.023*	0.12	0.25; 0.25	0.68	1.00	0.256; 0.279	0.3627	0.90
CGG	0.18; 0.16	0.43	0.94	0.18; 0.24	0.027	0.16	0.193; 0.172	0.3424	0.88
AGG	0.07; 0.08	0.55	0.98	0.097; 0.045	<0.0001*	0.0006*	0.099; 0.075	0.1311	0.51

*Significant p-value.

doi:10.1371/journal.pone.0015402.t002

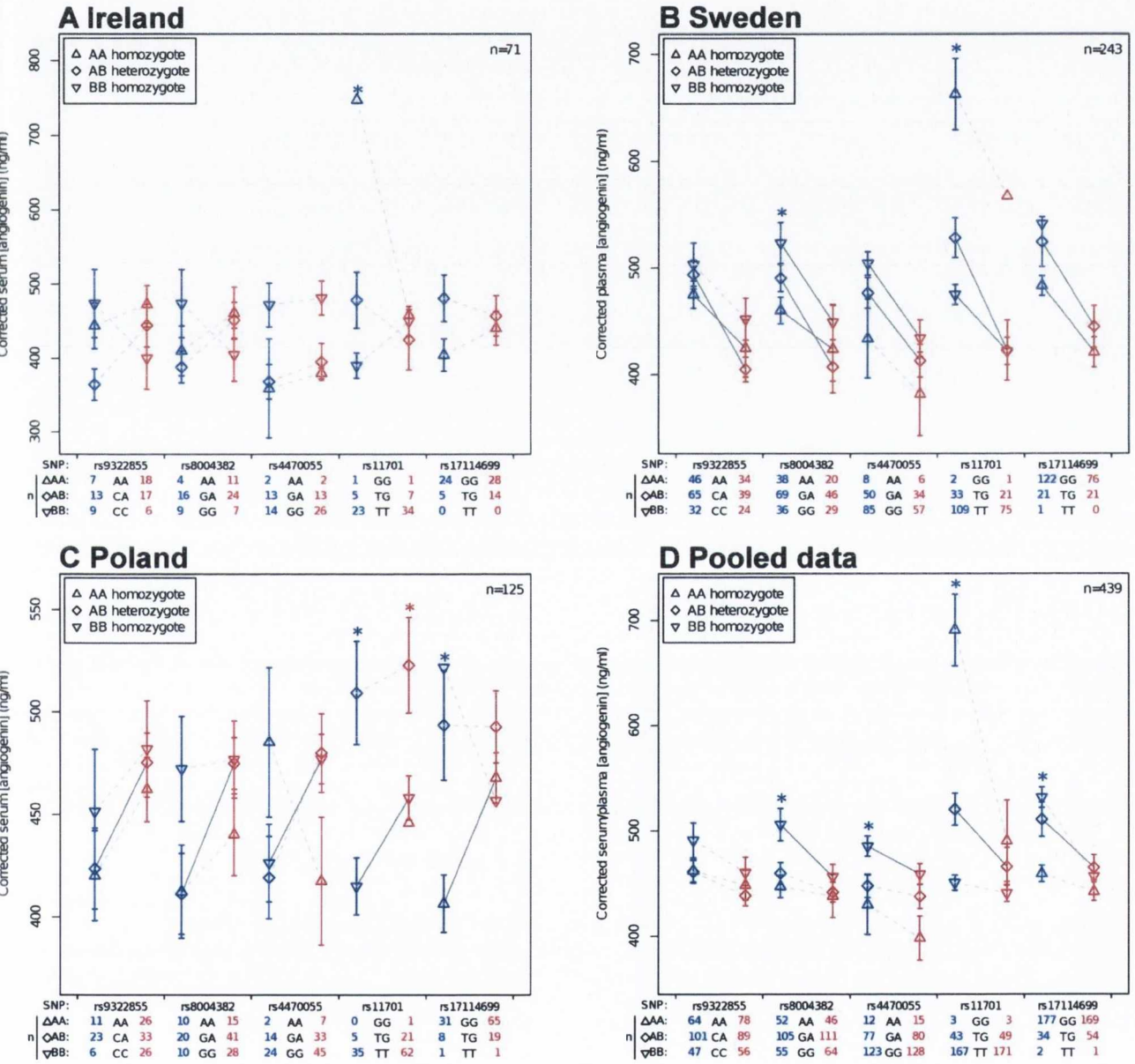


Figure 2. Mean corrected serum or plasma angiogenin concentrations as a function of *ANG* htSNP genotype. ALS patients are shown in red and controls are shown in blue. Significant differences between ALS patients and controls are denoted by solid lines and significant F-statistics within groups are denoted by asterisks. Error bars are standard error of the mean. Numbers of observations for each genotype at each SNP are indicated in the table below each plot.

doi:10.1371/journal.pone.0015402.g002

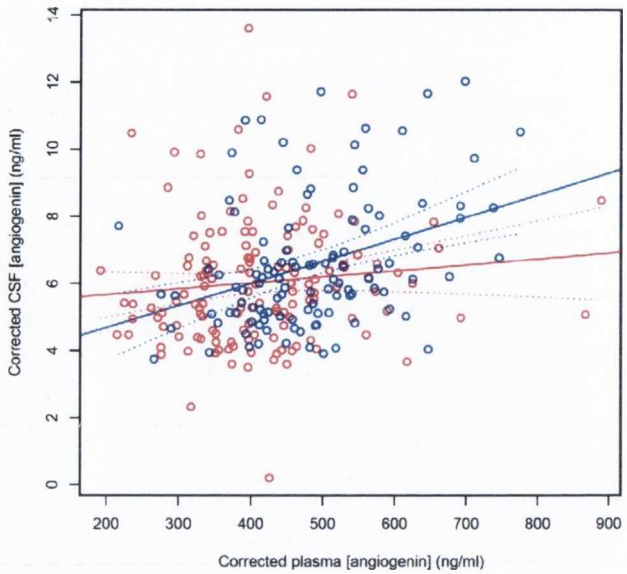


Figure 3. Correlation of CSF angiogenin levels with serum angiogenin levels in the Swedish population. ALS patients are shown in red and controls are shown in blue. Dashed lines indicate 95% confidence intervals of the regression lines. r^2 values are: controls, 0.13; ALS, 0.011.
doi:10.1371/journal.pone.0015402.g003

Discussion

This study confirms the previously observed association between *ANG* variants and ALS in the Irish population [1], with 5 htSNPs across the *ANG* gene showing association with ALS. One htSNP, rs1711699, replicated in the Swedish cohort, showing strong association with ALS risk ($p = 0.001$). We have also demonstrated that two *ANG* haplotypes in the Irish and one in the Swedish associate with ALS, adding strength to the argument that *ANG* is implicated in the pathogenesis of sporadic ALS. Although replication in the Swedish population increases our confidence in the Irish findings, no htSNP or haplotype associated with ALS in the Polish population. Similarly, in a recent screen for replication of findings from the Irish genome-wide association study for ALS risk [19] using a Polish dataset, the results were surprisingly uninformative [20]. The failure to replicate in the Polish population may reflect true population based differences, as has been recently demonstrated both in population genetics [21] and with respect to other risk genes in ALS [20]. Together, these findings suggest that the complex genetics of ALS differ between the Polish, Swedish and Irish populations.

Figure 2 (notably parts a and c) demonstrates the need for large datasets when analysing data that vary so substantially by chance. However, using pooled data we have shown that contribution of SNP genotypes to variance in angiogenin levels in serum is evident in neurologically normal individuals, and that this is abolished in ALS. In controls, this contribution of genotype to variance is allele dose-dependent. SNP genotypes at rs11701 were observed to contribute to variance in ALS patients in the Polish; this finding is consistent with the observation that no *ANG* SNP or haplotype associated with ALS in the Polish.

Using the current Irish dataset, we were unable to replicate our previous finding that serum angiogenin levels are higher in ALS patients compared to controls [8]. Using our current data pooled with Swedish and Polish populations, we have shown that angiogenin levels are in fact significantly lower in ALS patients than in neurologically normal controls ($p<0.001$). Moreover, sub-categorisation of ALS patients and controls by SNP genotypes maintains the significance of the case-control differences in angiogenin levels (figure 2).

The differences between the current data and our previous findings most likely relate to differences in our statistical management of the dataset. In the original study we considered the effects of covariates (age, sex) in ALS patients and controls independently. In the current analysis, we more correctly assumed that angiogenin levels in ALS patients would follow the same patterns based on age and sex as those observed in controls. Thus serum angiogenin levels were initially regressed against age and sex using combined data from cases and controls. This methodology permits a more robust estimate of the influence of age and sex on angiogenin levels, as it uses approximately twice as many values (541 values) as would be used if considering cases and controls separately. Indeed, re-analysis of the current dataset using our previous methodology yielded a significantly higher mean corrected angiogenin level in cases than in controls ($p<0.0001$); we now consider this to be a less accurate interpretation of the available data.

In neurologically normal controls, plasma angiogenin concentration predicts CSF angiogenin concentration ($p<0.0001$, figure 3). We have shown that this correlation is lost in ALS patients ($p = 0.21$ for patients), which may suggest a tissue-specific dysregulation of angiogenin expression in ALS. This could be due to a number of factors, including perturbation of angiogenin transport in ALS, however an interesting possibility could be micro RNA (miRNA) regulation of angiogenin expression. Altered miRNA regulation of progranulin has been reported recently in frontotemporal dementia [22]. As progranulin is functionally similar to angiogenin, and frontotemporal dementia is biologically related to ALS [23], a similar form of altered regulation of angiogenin may apply in ALS. A search the EBI's miRBase Sequence Database [24] using the online Microcosm web application reveals 19 potential miRNA binding sites in the *ANG* gene for 24 human miRNAs, some of which may be preferentially expressed in the central nervous system [25]. This suggests a possible mechanism for our observed tissue-specific differences indicating that further investigation of miRNA regulation of angiogenin is warranted.

In summary, we have confirmed that *ANG* variants associate with ALS in the Irish and also in the Swedish. We have also shown that angiogenin expression is modulated by genetic variation across the *ANG* gene in an allele-dose dependent manner, and that this regulation is disrupted in ALS patients. The finding that plasma angiogenin level does not predict CSF angiogenin level in ALS patients suggests a tissue-specific regulation of angiogenin levels that may be determined by genetic variation [18]. In light of these findings, further investigation of angiogenin regulation in ALS is justified.

Supporting Information

Figure S1 Linkage disequilibrium between the five *ANG* SNPs in the three populations. (PDF)

Figure S2 Boxplot comparing angiogenin levels measured in plasma from Swedish individuals (n = 320) and serum from Irish and Polish individuals (n = 220). The difference between the two datasets is not statistically significant ($p = 0.93$). (PDF)

Author Contributions

Conceived and designed the experiments: OH SC MG. Performed the experiments: JP SC RLM. Analyzed the data: RLM OH WM DSL PJ JS. Contributed reagents/materials/analysis tools: PMA BT AS DB PJ OH. Wrote the paper: RLM JP DSL SC OH.

References

1. Greenway MJ, Andersen PM, Russ C, Ennis S, Cashman S, et al. (2006) ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. Nat Genet 38: 411–413.
2. Subramanian V, Crabtree B, Acharya KR (2008) Human angiogenin is a neuroprotective factor and amyotrophic lateral sclerosis associated angiogenin variants affect neurite extension/pathfinding and survival of motor neurons. Hum Mol Genet 17: 130–149.
3. Wu D, Yu W, Kishikawa H, Folkerth RD, Iafrate AJ, et al. (2007) Angiogenin loss-of-function mutations in amyotrophic lateral sclerosis. Ann Neurol 62: 609–617.
4. Gellera C, Colombrita C, Ticozzi N, Castellotti B, Bragato C, et al. (2008) Identification of new ANG gene mutations in a large cohort of Italian patients with amyotrophic lateral sclerosis. Neurogenetics 9: 33–40.
5. Oosthuyse B, Moons L, Storkebaum E, Beck H, Nuyens D, et al. (2001) Deletion of the hypoxia-response element in the vascular endothelial growth factor promoter causes motor neuron degeneration. Nat Genet 28: 131–138.
6. Rosenstein JM, Krum JM (2004) New roles for VEGF in nervous tissue-beyond blood vessels. Exp Neurol 187: 246–253.
7. Lambrechts D, Storkebaum E, Morimoto M, Del-Favero J, Desmet F, et al. (2003) VEGFis a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. Nat Genet 34: 383–394.
8. Cronin S, Greenway MJ, Ennis S, Kieran D, Green A, et al. (2006) Elevated serum angiogenin levels in ALS. Neurology 67: 1833–1836.17.
9. Devos D, Moreau C, Lassalle P, Perez T, De Seze J, et al. (2004) Low levels of the vascular endothelial growth factor in CSF from early ALS patients. Neurology 62: 2127–2129.
10. Iłzecka J (2004) Cerebrospinal fluid vascular endothelial growth factor in patients with amyotrophic lateral sclerosis. Clin Neurol Neurosurg 106: 289–293.
11. Nygren I, Larsson A, Johansson A, Askmark H (2002) VEGF is increased in serum but not in spinal cord from patients with amyotrophic lateral sclerosis. Neuroreport 13: 2199–2201.
12. Brooks BR (1994) El Escorial World Federation of Neurology criteria for the diagnosis of amyotrophic lateral sclerosis Subcommittee on Motor Neuron Diseases/Amyotrophic Lateral Sclerosis of the World Federation of Neurology Research Group on Neuromuscular Diseases and the El Escorial "Clinical limits of amyotrophic lateral sclerosis" workshop contributors. J Neurol Sci 124(Suppl): 96–107.
13. Frazer K, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 31 million SNPs. Nature 449: 851–861.
14. R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3-900051-07-0.
15. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.
16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.18.
17. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The structure of haplotype blocks in the human genome. Science 296: 2225–2229.
18. Pantsulaia Ia, Trofimov S, Kobyliansky E, Livshits G (2006) Genetic and environmental determinants of circulating levels of angiogenin in community-based sample. Clin Endocrinol (Oxf) 64: 271–279.
19. Cronin S, Berger S, Ding J, Schymick JC, Washecka N, et al. (2008) A genome-wide association study of sporadic ALS in a homogenous Irish population. Hum Mol Genet 17: 768–774.
20. Cronin S, Tomik B, Bradley DG, Slowik A, Hardiman O (2009) Screening for replication of genome-wide SNP associations in sporadic ALS. Eur J Hum Genet 17: 213–218.
21. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98–101.
22. Rademakers R, Eriksen JL, Baker M, Robinson T, Ahmed Z, et al. (2008) Common variation in the miR-659 binding-site of GRN is a major risk factor for TDP43-positive frontotemporal dementia. Hum Mol Genet 17: 3631–3642.
23. Strong MJ (2008) The syndromes of frontotemporal dysfunction in amyotrophic lateral sclerosis. Amyotroph Lateral Scler 9: 323–338.
24. Griffiths-Jones S (2006) miRBase: the microRNA sequence database. Methods Mol Biol 342: 129–138.
25. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, et al. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. Cell 129: 1401–1414.

References

[1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, Feb 2001.

[2] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage,

F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291:1304–1351, Feb 2001.

[3] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis,

R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, and D. Altshuler. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409:928–933, Feb 2001.

[4] International HapMap Consortium. The International HapMap Project. *Nature*, 426:789–796, Dec 2003.

[5] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, Oct 2005.

[6] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861, Oct 2007.

[7] D. M. Altshuler, R. A. Gibbs, L. Peltonen, D. M. Altshuler, R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, L. Peltonen, E. Dermitzakis, P. E. Bonnen, D. M. Altshuler, R. A. Gibbs, P. I. de Bakker, P. Deloukas, S. B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, F. Yu, K. Chang, A. Hawes, L. R. Lewis, Y. Ren, D. Wheeler, R. A. Gibbs, D. M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J. M. Korn, K. Kristiansson, C. Lee, S. A. McCarrol, J. Nemesh, E. Dermitzakis, A. Keinan, S. B. Montgomery, S. Pollack, A. L. Price, N. Soranzo, P. E. Bonnen, R. A. Gibbs, C. Gonzaga-Jauregui, A. Keinan, A. L. Price, F. Yu, V. Anttila, W. Brodeur, M. J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, S. F. Schaffner, Q. Zhang, M. J. Ghori, R. McGinnis, W. McLaren, S. Pollack, A. L. Price, S. F. Schaffner, F. Takeuchi, S. R. Grossman, I. Shlyakhter, E. B. Hostetter, P. C. Sabeti, C. A. Adebamowo, M. W. Foster, D. R. Gordon, J. Licinio, M. C. Manca, P. A. Marshall, I. Matsuda, D. Ngare, V. O.

Wang, D. Reddy, C. N. Rotimi, C. D. Royal, R. R. Sharp, C. Zeng, L. D. Brooks, and J. E. McEwen. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–58, Sep 2010.

[8] I. Pe'er, P. I. de Bakker, J. Maller, R. Yelensky, D. Altshuler, and M. J. Daly. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.*, 38:663–667, Jun 2006.

[9] The chips are down. *Nature*, 444:256–257, Nov 2006.

[10] Jeffrey Perkel. SNP genotyping: six technologies that keyed a revolution. *Nature Methods*, 5(5):447–453, May 2008.

[11] M. Li, C. Li, and W. Guan. Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur. J. Hum. Genet.*, 16:635–643, May 2008.

[12] C. C. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.*, 5:e1000477, May 2009.

[13] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273:1516–1517, Sep 1996.

[14] E. S. Lander. The new genomics: global views of biology. *Science*, 274:536–539, Oct 1996.

[15] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, and E. S. Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.*, 22:231–238, Jul 1999.

[16] R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308:385–389, Apr 2005.

[17] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, 106:9362–9367, Jun 2009.

[18] S. P. Dickson, K. Wang, I. Krantz, H. Hakonarson, and D. B. Goldstein. Rare variants create synthetic genome-wide associations. *PLoS Biol.*, 8:e1000294, Jan 2010.

[19] N. R. Wray, S. M. Purcell, and P. M. Visscher. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol.*, 9:e1000579, 2011.

[20] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, Oct 2010.

[21] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74:5463–5467, Dec 1977.

[22] A. Chio, G. Mora, A. Calvo, L. Mazzini, E. Bottacchi, R. Mutani, A. Chio, A. Calvo, C. Moglia, P. Ghiglione, D. Cocito, R. Mutani, L. Durelli, B. Ferrero, A. Mauro, M. Leone, L. Mazzini, F. Monaco, N. Nasuelli, R. Comitangelo, L. Sosso, M. Gionco, M. Nobili, L. Appendino, C. Buffa, R. Cavallo, E. Oddenino, G. Ferrari, C. Geda, M. Favero, C. D. Bozzo, P. Santimaria, U. Massazza, A. Villani, R. Conti, G. Mora, F. Pisano, M. Palermo, F. Vergnano, M. Aguggia, M. T. Penza, F. Fassio, N. Di Vito, P. Meineri, D. Seliak, C. Cavestro, G. Astegiano, G. Corso, and E. Bottacchi. Epi-

demiology of ALS in Italy: a 10-year prospective population-based study. *Neurology*, 72:725–731, Feb 2009.

[23] A. Alonso, G. Logroscino, S. S. Jick, and M. A. Hernan. Incidence and lifetime risk of motor neuron disease in the United Kingdom: a population-based study. *Eur. J. Neurol.*, 16:745–751, Jun 2009.

[24] C. A. Johnston, B. R. Stanton, M. R. Turner, R. Gray, A. H. Blunt, D. Butt, M. A. Ampong, C. E. Shaw, P. N. Leigh, and A. Al-Chalabi. Amyotrophic lateral sclerosis in an urban setting: a population based study of inner city London. *J. Neurol.*, 253(12):1642–1643, Dec 2006.

[25] B. J. Traynor, M. B. Codd, B. Corr, C. Forde, E. Frost, and O. Hardiman. Incidence and prevalence of ALS in Ireland, 1995-1997: a population-based study. *Neurology*, 52:504–509, Feb 1999.

[26] J. C. Steele and P. L. McGeer. The ALS/PDC syndrome of Guam and the cycad hypothesis. *Neurology*, 70:1984–1990, May 2008.

[27] S. A. Banack and P. A. Cox. Biomagnification of cycad neurotoxins in flying foxes: implications for ALS-PDC in Guam. *Neurology*, 61:387–389, Aug 2003.

[28] B R Brooks. El Escorial World Federation of Neurology criteria for the diagnosis of amyotrophic lateral sclerosis Subcommittee on Motor Neuron Diseases/Amyotrophic Lateral Sclerosis of the World Federation of Neurology Research Group on Neuromuscular Diseases and the El Escorial "Clinical limits of amyotrophic lateral sclerosis" workshop contributors. *J Neurol Sci*, 124 Suppl:96–107, Jul 1994.

[29] J. Phukan, N. P. Pender, and O. Hardiman. Cognitive impairment in amyotrophic lateral sclerosis. *Lancet Neurol*, 6:994–1003, Nov 2007.

[30] J. Phukan, M. Elamin, P. Bede, N. Jordan, L. Gallagher, S. Byrne, C. Lynch, N. Pender, and O. Hardiman. The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study. *J Neurol Neurosurg Psychiatry*, Aug 2011.

[31] S. Sathasivam. Motor neurone disease: clinical features, diagnosis, diagnostic pitfalls and prognostic markers. *Singapore Med J*, 51:367–372, May 2010.

[32] H. Gray, T.P. Pick, and R. Howden. *Anatomy, descriptive and surgical.* Running Press, 1974.

[33] M. Neumann, D. M. Sampathu, L. K. Kwong, A. C. Truax, M. C. Micsenyi, T. T. Chou, J. Bruce, T. Schuck, M. Grossman, C. M. Clark, L. F. McCluskey, B. L. Miller, E. Masliah, I. R. Mackenzie, H. Feldman, W. Feiden, H. A. Kretzschmar, J. Q. Trojanowski, and V. M. Lee. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science*, 314:130–133, Oct 2006.

[34] H. X. Deng, H. Zhai, E. H. Bigio, J. Yan, F. Fecto, K. Ajroud, M. Mishra, S. Ajroud-Driss, S. Heller, R. Sufit, N. Siddique, E. Mugnaini, and T. Siddique. FUS-immunoreactive inclusions are a common feature in sporadic and non-SOD1 familial amyotrophic lateral sclerosis. *Ann. Neurol.*, 67:739–748, Jun 2010.

[35] K. M. Kurian, R. B. Forbes, S. Colville, and R. J. Swingler. Cause of death and clinical grading criteria in a cohort of amyotrophic lateral sclerosis cases undergoing autopsy from the Scottish Motor Neurone Disease Register. *J. Neurol. Neurosurg. Psychiatr.*, 80:84–87, Jan 2009.

[36] H. M. Bryson, B. Fulton, and P. Benfield. Riluzole. A review of its pharmacodynamic and pharmacokinetic properties and therapeutic potential in amyotrophic lateral sclerosis. *Drugs*, 52:549–563, Oct 1996.

[37] R. G. Miller, J. D. Mitchell, M. Lyon, and D. H. Moore. Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). *Cochrane Database Syst Rev*, page CD001447, 2007.

[38] P. Bede, D. Oliver, J. Stodart, L. van den Berg, Z. Simmons, D. O Brannagain, G. D. Borasio, and O. Hardiman. Palliative care in amyotrophic lateral sclerosis: a review of current international guidelines and initiatives. *J. Neurol. Neurosurg. Psychiatr.*, 82:413–418, Apr 2011.

[39] R. D. Azbill, X. Mu, and J. E. Springer. Riluzole increases high-affinity glutamate uptake in rat spinal cord synaptosomes. *Brain Res.*, 871:175–180, Jul 2000.

[40] S. Byrne, C. Walsh, C. Lynch, P. Bede, M. Elamin, K. Kenna, R. McLaughlin, and O. Hardiman. Rate of familial amyotrophic lateral sclerosis: a systematic review and meta-analysis. *J. Neurol. Neurosurg. Psychiatr.*, 82:623–627, Jun 2011.

[41] A. Al-Chalabi, F. Fang, M. F. Hanby, P. N. Leigh, C. E. Shaw, W. Ye, and F. Rijsdijk. An estimate of amyotrophic lateral sclerosis heritability using twin data. *J. Neurol. Neurosurg. Psychiatr.*, 81:1324–1326, Dec 2010.

[42] T. Siddique, D. A. Figlewicz, M. A. Pericak-Vance, J. L. Haines, G. Rouleau, A. J. Jeffers, P. Sapp, W. Y. Hung, J. Bebout, and D. McKenna-Yasek. Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity. *N. Engl. J. Med.*, 324:1381–1384, May 1991.

[43] D. R. Rosen, T. Siddique, D. Patterson, D. A. Figlewicz, P. Sapp, A. Hentati, D. Donaldson, J. Goto, J. P. O'Regan, and H. X. Deng. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature*, 362:59–62, Mar 1993.

[44] A. Felbecker, W. Camu, P. N. Valdmanis, A. D. Sperfeld, S. Waibel, P. Steinbach, G. A. Rouleau, A. C. Ludolph, and P. M. Andersen. Four familial ALS pedigrees discordant for two SOD1 mutations: are all SOD1 mutations pathogenic? *J. Neurol. Neurosurg. Psychiatr.*, 81:572–577, May 2010.

[45] P. M. Andersen. Amyotrophic lateral sclerosis associated with mutations in the CuZn superoxide dismutase gene. *Curr Neurol Neurosci Rep*, 6:37–46, Jan 2006.

[46] M. J. Greenway, P. M. Andersen, C. Russ, S. Ennis, S. Cashman, C. Donaghy, V. Patterson, R. Swingler, D. Kieran, J. Prehn, K. E. Morrison, A. Green, K. R. Acharya, R. H. Brown, and O. Hardiman. ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nat. Genet.*, 38:411–413, Apr 2006.

[47] I. Puls, C. Jonnakuty, B. H. LaMonte, E. L. Holzbaur, M. Tokito, E. Mann, M. K. Floeter, K. Bidus, D. Drayna, S. J. Oh, R. H. Brown, C. L. Ludlow, and K. H. Fischbeck. Mutant dynactin in motor neuron disease. *Nat. Genet.*, 33:455–456, Apr 2003.

[48] C. Munch, R. Sedlmeier, T. Meyer, V. Homberg, A. D. Sperfeld, A. Kurt, J. Prudlo, G. Peraus, C. O. Hanemann, G. Stumm, and A. C. Ludolph. Point mutations of the p150 subunit of dynactin (DCTN1) gene in ALS. *Neurology*, 63:724–726, Aug 2004.

[49] C. Munch, A. Rosenbohm, A. D. Sperfeld, I. Uttner, S. Reske, B. J. Krause, R. Sedlmeier, T. Meyer, C. O. Hanemann, G. Stumm, and A. C. Ludolph. Heterozygous R1101K mutation of the DCTN1 gene in a family with ALS and FTD. *Ann. Neurol.*, 58:777–780, Nov 2005.

[50] J. Sreedharan, I. P. Blair, V. B. Tripathi, X. Hu, C. Vance, B. Rogelj, S. Ackerley, J. C. Durnall, K. L. Williams, E. Buratti, F. Baralle, J. de Belleroche, J. D. Mitchell, P. N. Leigh, A. Al-Chalabi, C. C. Miller, G. Nicholson, and C. E. Shaw. TDP-43

mutations in familial and sporadic amyotrophic lateral sclerosis. *Science*, 319:1668–1672, Mar 2008.

[51] T. J. Kwiatkowski, D. A. Bosco, A. L. Leclerc, E. Tamrazian, C. R. Vanderburg, C. Russ, A. Davis, J. Gilchrist, E. J. Kasarskis, T. Munsat, P. Valdmanis, G. A. Rouleau, B. A. Hosler, P. Cortelli, P. J. de Jong, Y. Yoshinaga, J. L. Haines, M. A. Pericak-Vance, J. Yan, N. Ticozzi, T. Siddique, D. McKenna-Yasek, P. C. Sapp, H. R. Horvitz, J. E. Landers, and R. H. Brown. Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science*, 323:1205–1208, Feb 2009.

[52] C. Vance, B. Rogelj, T. Hortobagyi, K. J. De Vos, A. L. Nishimura, J. Sreedharan, X. Hu, B. Smith, D. Ruddy, P. Wright, J. Ganesalingam, K. L. Williams, V. Tripathi, S. Al-Saraj, A. Al-Chalabi, P. N. Leigh, I. P. Blair, G. Nicholson, J. de Belleroche, J. M. Gallo, C. C. Miller, and C. E. Shaw. Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science*, 323:1208–1211, Feb 2009.

[53] H. Maruyama, H. Morino, H. Ito, Y. Izumi, H. Kato, Y. Watanabe, Y. Kinoshita, M. Kamada, H. Nodera, H. Suzuki, O. Komure, S. Matsuura, K. Kobatake, N. Morimoto, K. Abe, N. Suzuki, M. Aoki, A. Kawata, T. Hirai, T. Kato, K. Ogasawara, A. Hirano, T. Takumi, H. Kusaka, K. Hagiwara, R. Kaji, and H. Kawakami. Mutations of optineurin in amyotrophic lateral sclerosis. *Nature*, 465:223–226, May 2010.

[54] R. Del Bo, C. Tiloca, V. Pensato, L. Corrado, A. Ratti, N. Ticozzi, S. Corti, B. Castellotti, L. Mazzini, G. Soraru, C. Cereda, S. D'Alfonso, C. Gellera, G. P. Comi, and V. Silani. Novel optineurin mutations in patients with familial and spo-

radic amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatr.*, 82:1239–1243, Nov 2011.

[55] M. van Blitterswijk, P. W. van Vught, M. A. van Es, H. J. Schelhaas, A. J. van der Kooi, M. de Visser, J. H. Veldink, and L. H. van den Berg. Novel optineurin mutations in sporadic amyotrophic lateral sclerosis patients. *Neurobiol Aging*, Jul 2011.

[56] A. C. Elden, H. J. Kim, M. P. Hart, A. S. Chen-Plotkin, B. S. Johnson, X. Fang, M. Armakola, F. Geser, R. Greene, M. M. Lu, A. Padmanabhan, D. Clay-Falcone, L. McCluskey, L. Elman, D. Juhr, P. J. Gruber, U. Rub, G. Auburger, J. Q. Trojanowski, V. M. Lee, V. M. Van Deerlin, N. M. Bonini, and A. D. Gitler. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature*, 466:1069–1075, Aug 2010.

[57] A. Hentati, K. Bejaoui, M. A. Pericak-Vance, F. Hentati, M. C. Speer, W. Y. Hung, D. A. Figlewicz, J. Haines, J. Rimmler, and C. Ben Hamida. Linkage of recessive familial amyotrophic lateral sclerosis to chromosome 2q33-q35. *Nat. Genet.*, 7:425–428, Jul 1994.

[58] S. Hadano, C. K. Hand, H. Osuga, Y. Yanagisawa, A. Otomo, R. S. Devon, N. Miyamoto, J. Showguchi-Miyata, Y. Okada, R. Singaraja, D. A. Figlewicz, T. Kwiatkowski, B. A. Hosler, T. Sagie, J. Skaug, J. Nasir, R. H. Brown, S. W. Scherer, G. A. Rouleau, M. R. Hayden, and J. E. Ikeda. A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nat. Genet.*, 29:166–173, Oct 2001.

[59] Y. Z. Chen, C. L. Bennett, H. M. Huynh, I. P. Blair, I. Puls, J. Irobi, I. Dierick, A. Abel, M. L. Kennerson, B. A. Rabin, G. A. Nicholson, M. Auer-Grumbach, K. Wagner, P. De Jonghe, J. W. Griffin, K. H. Fischbeck, V. Timmerman, D. R. Cornblath, and P. F. Chance. DNA/RNA helicase gene mutations in a form of

juvenile amyotrophic lateral sclerosis (ALS4). *Am. J. Hum. Genet.*, 74:1128–1135, Jun 2004.

[60] Z. H. Zhao, W. Z. Chen, Z. Y. Wu, N. Wang, G. X. Zhao, W. J. Chen, and S. X. Murong. A novel mutation in the senataxin gene identified in a Chinese patient with sporadic amyotrophic lateral sclerosis. *Amyotroph Lateral Scler*, 10:118–122, Apr 2009.

[61] F. Avemaria, C. Lunetta, C. Tarlarini, L. Mosca, E. Maestri, A. Marocchi, M. Melazzini, S. Penco, and M. Corbo. Mutation in the senataxin gene found in a patient affected by familial ALS with juvenile onset and slow progression. *Amyotroph Lateral Scler*, 12:228–230, May 2011.

[62] A. A. Luty, J. B. Kwok, C. Dobson-Stone, C. T. Loy, K. G. Coupland, H. Karlstrom, T. Sobow, J. Tchorzewska, A. Maruszak, M. Barcikowska, P. K. Panegyres, C. Zekanowski, W. S. Brooks, K. L. Williams, I. P. Blair, K. A. Mather, P. S. Sachdev, G. M. Halliday, and P. R. Schofield. Sigma nonopioid intracellular receptor 1 mutations cause frontotemporal lobar degeneration-motor neuron disease. *Ann. Neurol.*, 68:639–649, Nov 2010.

[63] A. L. Nishimura, M. Mitne-Neto, H. C. Silva, A. Richieri-Costa, S. Middleton, D. Cascio, F. Kok, J. R. Oliveira, T. Gillingwater, J. Webb, P. Skehel, and M. Zatz. A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. *Am. J. Hum. Genet.*, 75:822–831, Nov 2004.

[64] H. X. Deng, W. Chen, S. T. Hong, K. M. Boycott, G. H. Gorrie, N. Siddique, Y. Yang, F. Fecto, Y. Shi, H. Zhai, H. Jiang, M. Hirano, E. Rampersaud, G. H. Jansen, S. Donkervoort, E. H. Bigio, B. R. Brooks, K. Ajroud, R. L. Sufit, J. L. Haines, E. Mugnaini, M. A. Pericak-Vance, and T. Siddique. Mutations in UBQLN2

cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. *Nature*, 477:211–215, Sep 2011.

[65] R. Wroe, A. Wai-Ling Butler, P. M. Andersen, J. F. Powell, and A. Al-Chalabi. ALSOD: the Amyotrophic Lateral Sclerosis Online Database. *Amyotroph Lateral Scler*, 9:249–250, Aug 2008.

[66] A. Shatunov, K. Mok, S. Newhouse, M. E. Weale, B. Smith, C. Vance, L. Johnson, J. H. Veldink, M. A. van Es, L. H. van den Berg, W. Robberecht, P. Van Damme, O. Hardiman, A. E. Farmer, C. M. Lewis, A. W. Butler, O. Abel, P. M. Andersen, I. Fogh, V. Silani, A. Chio, B. J. Traynor, J. Melki, V. Meininger, J. E. Landers, P. McGuffin, J. D. Glass, H. Pall, P. N. Leigh, J. Hardy, R. H. Brown, J. F. Powell, R. W. Orrell, K. E. Morrison, P. J. Shaw, C. E. Shaw, and A. Al-Chalabi. Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. *Lancet Neurol*, 9:986–994, Oct 2010.

[67] M. A. van Es, J. H. Veldink, C. G. Saris, H. M. Blauw, P. W. van Vught, A. Birve, R. Lemmens, H. J. Schelhaas, E. J. Groen, M. H. Huisman, A. J. van der Kooi, M. de Visser, C. Dahlberg, K. Estrada, F. Rivadeneira, A. Hofman, M. J. Zwarts, P. T. van Doormaal, D. Rujescu, E. Strengman, I. Giegling, P. Muglia, B. Tomik, A. Slowik, A. G. Uitterlinden, C. Hendrich, S. Waibel, T. Meyer, A. C. Ludolph, J. D. Glass, S. Purcell, S. Cichon, M. M. Nothen, H. E. Wichmann, S. Schreiber, S. H. Vermeulen, L. A. Kiemeney, J. H. Wokke, S. Cronin, R. L. McLaughlin, O. Hardiman, K. Fumoto, R. J. Pasterkamp, V. Meininger, J. Melki, P. N. Leigh, C. E. Shaw, J. E. Landers, A. Al-Chalabi, R. H. Brown, W. Robberecht, P. M. Andersen, R. A. Ophoff, and L. H. van den Berg. Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat. Genet.*, 41:1083–1087, Oct 2009.

[68] M. Morita, A. Al-Chalabi, P. M. Andersen, B. Hosler, P. Sapp, E. Englund, J. E. Mitchell, J. J. Habgood, J. de Belleroche, J. Xi, W. Jongjaroenprasert, H. R. Horvitz, L. G. Gunnarsson, and R. H. Brown. A locus on chromosome 9p confers susceptibility to ALS and frontotemporal dementia. *Neurology*, 66:839–844, Mar 2006.

[69] J. O. Johnson, J. Mandrioli, M. Benatar, Y. Abramzon, V. M. Van Deerlin, J. Q. Trojanowski, J. R. Gibbs, M. Brunetti, S. Gronka, J. Wuu, J. Ding, L. McCluskey, M. Martinez-Lage, D. Falcone, D. G. Hernandez, S. Arepalli, S. Chong, J. C. Schymick, J. Rothstein, F. Landi, Y. D. Wang, A. Calvo, G. Mora, M. Sabatelli, M. R. Monsurro, S. Battistini, F. Salvi, R. Spataro, P. Sola, G. Borghero, G. Galassi, S. W. Scholz, J. P. Taylor, G. Restagno, A. Chio, B. J. Traynor, F. Giannini, C. Ricci, C. Moglia, I. Ossola, A. Canosa, S. Gallo, G. Tedeschi, P. Sola, I. Bartolomei, K. Marinou, L. Papetti, A. Conte, M. Luigetti, V. La Bella, P. Paladino, C. Caponnetto, P. Volanti, M. G. Marrosu, and M. R. Murru. Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron*, 68:857–864, Dec 2010.

[70] C. Hayward, S. Colville, R. J. Swingler, and D. J. Brock. Molecular genetic analysis of the APEX nuclease gene in amyotrophic lateral sclerosis. *Neurology*, 52:1899–1901, Jun 1999.

[71] M. J. Greenway, M. D. Alexander, S. Ennis, B. J. Traynor, B. Corr, E. Frost, A. Green, and O. Hardiman. A novel candidate region for ALS on chromosome 14q11.2. *Neurology*, 63:1936–1938, Nov 2004.

[72] J. H. Distler, A. Hirth, M. Kurowska-Stolarska, R. E. Gay, S. Gay, and O. Distler. Angiogenic and angiostatic factors in the molecular control of angiogenesis. *Q J Nucl Med*, 47:149–161, Sep 2003.

[73] B Oosthuyse, L Moons, E Storkebaum, H Beck, D Nuyens, K Brusselmans, J Van Dorpe, P Hellings, M Gorselink, S Heymans, G Theilmeier, M Dewerchin,

V Laudenbach, P Vermylen, H Raat, T Acker, V Vleminckx, L Van Den Bosch, N Cashman, H Fujisawa, M R Drost, R Sciot, F Bruyninckx, D J Hicklin, C Ince, P Gressens, F Lupu, K H Plate, W Robberecht, J M Herbert, D Collen, and P Carmeliet. Deletion of the hypoxia-response element in the vascular endothelial growth factor promoter causes motor neuron degeneration. *Nat Genet*, 28:131–138, Jun 2001.

[74] R. Del Bo, M. Scarlato, S. Ghezzi, F. Martinelli-Boneschi, S. Corti, F. Locatelli, D. Santoro, A. Prelle, C. Briani, M. Nardini, G. Siciliano, M. Mancuso, L. Murri, N. Bresolin, and G. P. Comi. Absence of angiogenic genes modification in Italian ALS patients. *Neurobiol. Aging*, 29:314–316, Feb 2008.

[75] L. Corrado, S. Battistini, S. Penco, L. Bergamaschi, L. Testa, C. Ricci, F. Giannini, G. Greco, M. C. Patrosso, S. Pileggi, R. Causarano, L. Mazzini, P. Momigliano-Richiardi, and S. D'Alfonso. Variations in the coding and regulatory sequences of the angiogenin (ANG) gene are not associated to ALS (amyotrophic lateral sclerosis) in the Italian population. *J. Neurol. Sci.*, 258:123–127, Jul 2007.

[76] F. L. Conforti, T. Sprovieri, R. Mazzei, C. Ungaro, V. La Bella, A. Tessitore, A. Patitucci, A. Magariello, A. L. Gabriele, G. Tedeschi, I. L. Simone, G. Majorana, P. Valentino, F. Condino, F. Bono, M. R. Monsurro, M. Muglia, and A. Quattrone. A novel Angiogenin gene mutation in a sporadic patient with amyotrophic lateral sclerosis from southern Italy. *Neuromuscul. Disord.*, 18:68–70, Jan 2008.

[77] C. Gellera, C. Colombrita, N. Ticozzi, B. Castellotti, C. Bragato, A. Ratti, F. Taroni, and V. Silani. Identification of new ANG gene mutations in a large cohort of Italian patients with amyotrophic lateral sclerosis. *Neurogenetics*, 9:33–40, Feb 2008.

[78] A. Paubel, J. Violette, M. Amy, J. Praline, V. Meininger, W. Camu, P. Corcia, C. R. Andres, P. Vourc'h, F. Dubas, G. Nicolas, G. Lemasson, E. Salort, F. Viader, L. Car-

luer, P. Clavelou, N. Guy, M. Giroud, C. Maugras, G. Besson, A. Destee, V. Danel, P. Couratier, M. Lacoste, E. Broussole, C. Vial, N. Vandenberghe, J. Pouget, D. Lardiller, A. Verschuren, W. Camu, G. Garrigues, N. Pageot, M. Debouverie, S. Pition, C. Desnuelle, M. H. Soriani, V. Meininger, F. Salachas, P. F. Pradat, M. Dib, G. Bruneteau, N. Leforestier, C. Tranchant, M. C. Fleury, J. C. Antoine, J. P. Camdessanche, M. C. Arne-Bes, P. Cintas, P. Corcia, and J. Praline. Mutations of the ANG gene in French patients with sporadic amyotrophic lateral sclerosis. *Arch. Neurol.*, 65:1333–1336, Oct 2008.

[79] S. Millecamps, F. Salachas, C. Cazeneuve, P. Gordon, B. Bricka, A. Camuzat, L. Guillot-Noel, O. Russaouen, G. Bruneteau, P. F. Pradat, N. Le Forestier, N. Vandenberghe, V. Danel-Brunaud, N. Guy, C. Thauvin-Robinet, L. Lacomblez, P. Couratier, D. Hannequin, D. Seilhean, I. Le Ber, P. Corcia, W. Camu, A. Brice, G. Rouleau, E. LeGuern, and V. Meininger. SOD1, ANG, VAPB, TARDBP, and FUS mutations in familial amyotrophic lateral sclerosis: genotype-phenotype correlations. *J. Med. Genet.*, 47:554–560, Aug 2010.

[80] R. Fernandez-Santiago, S. Hoenig, P. Lichtner, A. D. Sperfeld, M. Sharma, D. Berg, O. Weichenrieder, T. Illig, K. Eger, T. Meyer, J. Anneser, C. Munch, S. Zierz, T. Gasser, and A. Ludolph. Identification of novel Angiogenin (ANG) gene missense variants in German patients with amyotrophic lateral sclerosis. *J. Neurol.*, 256:1337–1342, Aug 2009.

[81] D Wu, W Yu, H Kishikawa, R D Folkerth, A J Iafrate, Y Shen, W Xin, K Sims, and G F Hu. Angiogenin loss-of-function mutations in amyotrophic lateral sclerosis. *Ann Neurol*, 62:609–617, Dec 2007.

[82] V. Subramanian, B. Crabtree, and K. R. Acharya. Human angiogenin is a neuroprotective factor and amyotrophic lateral sclerosis associated angiogenin variants affect

neurite extension/pathfinding and survival of motor neurons. *Hum. Mol. Genet.*, 17:130–149, Jan 2008.

[83] B. Crabtree, N. Thiyagarajan, S. H. Prior, P. Wilson, S. Iyer, T. Ferns, R. Shapiro, K. Brew, V. Subramanian, and K. R. Acharya. Characterization of human angiogenin variants implicated in amyotrophic lateral sclerosis. *Biochemistry*, 46:11810–11818, Oct 2007.

[84] S. Weremowicz, E. A. Fox, C. C. Morton, and B. L. Vallee. Localization of the human angiogenin gene to chromosome band 14q11, proximal to the T cell receptor alpha/delta locus. *Am. J. Hum. Genet.*, 47:973–981, Dec 1990.

[85] J. W. Fett, D. J. Strydom, R. R. Lobb, E. M. Alderman, J. L. Bethune, J. F. Riordan, and B. L. Vallee. Isolation and characterization of angiogenin, an angiogenic protein from human carcinoma cells. *Biochemistry*, 24:5480–5486, Sep 1985.

[86] R. Shapiro, J. F. Riordan, and B. L. Vallee. Characteristic ribonucleolytic activity of human angiogenin. *Biochemistry*, 25:3527–3532, Jun 1986.

[87] A. Hartmann, M. Kunz, S. Kostlin, R. Gillitzer, A. Toksoy, E. B. Brocker, and C. E. Klein. Hypoxia-induced up-regulation of angiogenin in human malignant melanoma. *Cancer Res.*, 59:1578–1583, Apr 1999.

[88] J. R. White, R. A. Harris, S. R. Lee, M. H. Craigon, K. Binley, T. Price, G. L. Beard, C. R. Mundy, and S. Naylor. Genetic amplification of the transcriptional response to hypoxia as a novel means of identifying regulators of angiogenesis. *Genomics*, 83:1–8, Jan 2004.

[89] D Lambrechts, E Storkebaum, M Morimoto, J Del-Favero, F Desmet, S L Marklund, S Wyns, V Thijs, J Andersson, I van Marion, A Al-Chalabi, S Bornes, R Musson, V Hansen, L Beckman, R Adolfsson, H S Pall, H Prats, S Vermeire, P Rutgeerts,

S Katayama, T Awata, N Leigh, L Lang-Lazdunski, M Dewerchin, C Shaw, L Moons, R Vlietinck, K E Morrison, W Robberecht, C Van Broeckhoven, D Collen, P M Andersen, and P Carmeliet. VEGF is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. *Nat Genet*, 34:383–394, Aug 2003.

[90] E. Storkebaum, D. Lambrechts, M. Dewerchin, M. P. Moreno-Murciano, S. Appelmans, H. Oh, P. Van Damme, B. Rutten, W. Y. Man, M. De Mol, S. Wyns, D. Manka, K. Vermeulen, L. Van Den Bosch, N. Mertens, C. Schmitz, W. Robberecht, E. M. Conway, D. Collen, L. Moons, and P. Carmeliet. Treatment of motoneuron degeneration by intracerebroventricular delivery of VEGF in a rat model of ALS. *Nat. Neurosci.*, 8:85–92, Jan 2005.

[91] E. Storkebaum, D. Lambrechts, and P. Carmeliet. VEGF: once regarded as a specific angiogenic factor, now implicated in neuroprotection. *Bioessays*, 26:943–954, Sep 2004.

[92] S. Cronin, M. J. Greenway, S. Ennis, D. Kieran, A. Green, J. H. Prehn, and O. Hardiman. Elevated serum angiogenin levels in ALS. *Neurology*, 67:1833–1836, Nov 2006.

[93] J C Barrett, B Fry, J Maller, and M J Daly. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21:263–265, Jan 2005.

[94] Development Core Team R. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[95] S Purcell, B Neale, K Todd-Brown, L Thomas, M A Ferreira, D Bender, J Maller, P Sklar, P I de Bakker, M J Daly, and P C Sham. PLINK: a tool set for whole-genome

association and population-based linkage analyses. *Am J Hum Genet*, 81:559–575, Sep 2007.

[96] S B Gabriel, S F Schaffner, H Nguyen, J M Moore, J Roy, B Blumenstiel, J Higgins, M DeFelice, A Lochner, M Faggart, S N Liu-Cordero, C Rotimi, A Adeyemo, R Cooper, R Ward, E S Lander, M J Daly, and D Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, Jun 2002.

[97] I a Pantsulaia, S Trofimov, E Kobyliansky, and G Livshits. Genetic and environmental determinants of circulating levels of angiogenin in community-based sample. *Clin Endocrinol*, 64:271–279, Mar 2006.

[98] S Cronin, S Berger, J Ding, J C Schymick, N Washecka, D G Hernandez, M J Greenway, D G Bradley, B J Traynor, and O Hardiman. A genome-wide association study of sporadic ALS in a homogenous Irish population. *Hum Mol Genet*, 17:768–774, Mar 2008.

[99] S Cronin, B Tomik, D G Bradley, A Slowik, and O Hardiman. Screening for replication of genome-wide SNP associations in sporadic ALS. *Eur J Hum Genet*, 17:213–218, Feb 2009.

[100] M. A. van Es, C. Dahlberg, A. Birve, J. H. Veldink, L. H. van den Berg, and P. M. Andersen. Large-scale SOD1 mutation screening provides evidence for genetic heterogeneity in amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatr.*, 81:562–566, May 2010.

[101] R Rademakers, J L Eriksen, M Baker, T Robinson, Z Ahmed, S J Lincoln, N Finch, N J Rutherford, R J Crook, K A Josephs, B F Boeve, D S Knopman, R C Petersen, J E Parisi, R J Caselli, Z K Wszolek, R J Uitti, H Feldman, M L Hutton, I R Mackenzie, N R Graff-Radford, and D W Dickson. Common variation in the miR-

659 binding-site of GRN is a major risk factor for TDP43-positive frontotemporal dementia. *Hum Mol Genet*, 17:3631–3642, Dec 2008.

[102] M J Strong. The syndromes of frontotemporal dysfunction in amyotrophic lateral sclerosis. *Amyotroph Lateral Scler*, 9:323–338, Dec 2008.

[103] S Griffiths-Jones. miRBase: the microRNA sequence database. *Methods Mol Biol*, 342:129–138, 2006.

[104] P Landgraf, M Rusu, R Sheridan, A Sewer, N Iovino, A Aravin, S Pfeffer, A Rice, A O Kamphorst, M Landthaler, C Lin, N D Socci, L Hermida, V Fulci, S Chiaretti, R Fo, J Schliwka, U Fuchs, A Novosel, R U Mller, B Schermer, U Bissels, J Inman, Q Phan, M Chien, D B Weir, R Choksi, G De Vita, D Frezzetti, H I Trompeter, V Hornung, G Teng, G Hartmann, M Palkovits, R Di Lauro, P Wernet, G Macino, C E Rogler, J W Nagle, J Ju, F N Papavasiliou, T Benzing, P Lichter, W Tam, M J Brownstein, A Bosio, A Borkhardt, J J Russo, C Sander, M Zavolan, and T Tuschl. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129:1401–1414, Jun 2007.

[105] J. C. Schymick, S. W. Scholz, H. C. Fung, A. Britton, S. Arepalli, J. R. Gibbs, F. Lombardo, M. Matarin, D. Kasperaviciute, D. G. Hernandez, C. Crews, L. Bruijn, J. Rothstein, G. Mora, G. Restagno, A. Chio, A. Singleton, J. Hardy, and B. J. Traynor. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol*, 6:322–328, Apr 2007.

[106] T. Dunckley, M. J. Huentelman, D. W. Craig, J. V. Pearson, S. Szelinger, K. Joshipura, R. F. Halperin, C. Stamper, K. R. Jensen, D. Letizia, S. E. Hesterlee, A. Pestronk, T. Levine, T. Bertorini, M. C. Graves, T. Mozaffar, C. E. Jackson, P. Bosch, A. McVey, A. Dick, R. Barohn, C. Lomen-Hoerth, J. Rosenfeld, D. T. O'connor,

K. Zhang, R. Crook, H. Ryberg, M. Hutton, J. Katz, E. P. Simpson, H. Mitsumoto, R. Bowser, R. G. Miller, S. H. Appel, and D. A. Stephan. Whole-genome analysis of sporadic amyotrophic lateral sclerosis. *N. Engl. J. Med.*, 357:775–788, Aug 2007.

[107] M. A. van Es, P. W. Van Vught, H. M. Blauw, L. Franke, C. G. Saris, P. M. Andersen, L. Van Den Bosch, S. W. de Jong, R. van 't Slot, A. Birve, R. Lemmens, V. de Jong, F. Baas, H. J. Schelhaas, K. Sleegers, C. Van Broeckhoven, J. H. Wokke, C. Wijmenga, W. Robberecht, J. H. Veldink, R. A. Ophoff, and L. H. van den Berg. ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study. *Lancet Neurol*, 6:869–877, Oct 2007.

[108] M. A. van Es, P. W. van Vught, H. M. Blauw, L. Franke, C. G. Saris, L. Van den Bosch, S. W. de Jong, V. de Jong, F. Baas, R. van't Slot, R. Lemmens, H. J. Schelhaas, A. Birve, K. Sleegers, C. Van Broeckhoven, J. C. Schymick, B. J. Traynor, J. H. Wokke, C. Wijmenga, W. Robberecht, P. M. Andersen, J. H. Veldink, R. A. Ophoff, and L. H. van den Berg. Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.*, 40:29–31, Jan 2008.

[109] H. M. Blauw, A. Al-Chalabi, P. M. Andersen, P. W. van Vught, F. P. Diekstra, M. A. van Es, C. G. Saris, E. J. Groen, W. van Rheenen, M. Koppers, R. Van't Slot, E. Strengman, K. Estrada, F. Rivadeneira, A. Hofman, A. G. Uitterlinden, L. A. Kiemeney, S. H. Vermeulen, A. Birve, S. Waibel, T. Meyer, S. Cronin, R. L. McLaughlin, O. Hardiman, P. C. Sapp, M. D. Tobin, L. V. Wain, B. Tomik, A. Slowik, R. Lemmens, D. Rujescu, C. Schulte, T. Gasser, R. H. Brown, J. E. Landers, W. Robberecht, A. C. Ludolph, R. A. Ophoff, J. H. Veldink, and L. H. van den Berg. A large genome scan for rare CNVs in amyotrophic lateral sclerosis. *Hum. Mol. Genet.*, 19:4091–4099, Oct 2010.

[110] I. Fogh, S. D'Alfonso, C. Gellera, A. Ratti, C. Cereda, S. Penco, L. Corrado, G. Soraru, B. Castellotti, C. Tiloca, S. Gagliardi, L. Cozzi, M. K. Lupton, N. Ticozzi, L. Mazzini, C. E. Shaw, A. Al-Chalabi, J. Powell, and V. Silani. No association of DPP6 with amyotrophic lateral sclerosis in an Italian population. *Neurobiol. Aging*, 32:966–967, May 2011.

[111] H. Daoud, P. N. Valdmanis, P. A. Dion, and G. A. Rouleau. Analysis of DPP6 and FGGY as candidate genes for amyotrophic lateral sclerosis. *Amyotroph Lateral Scler*, 11:389–391, Aug 2010.

[112] X. G. Li, J. H. Zhang, M. Q. Xie, M. S. Liu, B. H. Li, Y. H. Zhao, H. T. Ren, and L. Y. Cui. Association between DPP6 polymorphism and the risk of sporadic amyotrophic lateral sclerosis in Chinese patients. *Chin. Med. J.*, 122:2989–2992, Dec 2009.

[113] A. Chiò, J. C. Schymick, G. Restagno, S. W. Scholz, F. Lombardo, S. L. Lai, G. Mora, H. C. Fung, A. Britton, S. Arepalli, J. R. Gibbs, M. Nalls, S. Berger, L. C. Kwee, E. Z. Oddone, J. Ding, C. Crews, I. Rafferty, N. Washecka, D. Hernandez, L. Ferrucci, S. Bandinelli, J. Guralnik, F. Macciardi, F. Torri, S. Lupoli, S. J. Chanock, G. Thomas, D. J. Hunter, C. Gieger, H. E. Wichmann, A. Calvo, R. Mutani, S. Battistini, F. Giannini, C. Caponnetto, G. L. Mancardi, V. La Bella, F. Valentino, M. R. Monsurro, G. Tedeschi, K. Marinou, M. Sabatelli, A. Conte, J. Mandrioli, P. Sola, F. Salvi, I. Bartolomei, G. Siciliano, C. Carlesi, R. W. Orrell, K. Talbot, Z. Simmons, J. Connor, E. P. Pioro, T. Dunkley, D. A. Stephan, D. Kasperaviciute, E. M. Fisher, S. Jabonka, M. Sendtner, M. Beck, L. Bruijn, J. Rothstein, S. Schmidt, A. Singleton, J. Hardy, and B. J. Traynor. A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. *Hum. Mol. Genet.*, 18:1524–1532, Apr 2009.

[114] M. Morita, A. Al-Chalabi, P. M. Andersen, B. Hosler, P. Sapp, E. Englund, J. E. Mitchell, J. J. Habgood, J. de Belleroche, J. Xi, W. Jongjaroenprasert, H. R. Horvitz, L. G. Gunnarsson, and R. H. Brown. A locus on chromosome 9p confers susceptibility to ALS and frontotemporal dementia. *Neurology*, 66:839–844, Mar 2006.

[115] C. Vance, A. Al-Chalabi, D. Ruddy, B. N. Smith, X. Hu, J. Sreedharan, T. Siddique, H. J. Schelhaas, B. Kusters, D. Troost, F. Baas, V. de Jong, and C. E. Shaw. Familial amyotrophic lateral sclerosis with frontotemporal dementia is linked to a locus on chromosome 9p13.2-21.3. *Brain*, 129:868–876, Apr 2006.

[116] P. N. Valdmanis, N. Dupre, J. P. Bouchard, W. Camu, F. Salachas, V. Meininger, M. Strong, and G. A. Rouleau. Three families with amyotrophic lateral sclerosis and frontotemporal dementia with evidence of linkage to chromosome 9p. *Arch. Neurol.*, 64:240–245, Feb 2007.

[117] A. A. Luty, J. B. Kwok, E. M. Thompson, P. Blumbergs, W. S. Brooks, C. T. Loy, C. Dobson-Stone, P. K. Panegyres, J. Hecker, G. A. Nicholson, G. M. Halliday, and P. R. Schofield. 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, 2008.

[118] I. Le Ber, A. Camuzat, E. Berger, D. Hannequin, A. Laquerriere, V. Golfier, D. Seilhean, G. Viennet, P. Couratier, P. Verpillat, S. Heath, W. Camu, O. Martinaud, L. Lacomblez, M. Vercelletto, F. Salachas, F. Sellal, M. Didic, C. Thomas-Anterion, M. Puel, B. F. Michel, C. Besse, C. Duyckaerts, V. Meininger, D. Campion, B. Dubois, A. Brice, A. Brice, F. Blanc, W. Camu, F. Clerget-Darpoux, P. Corcia, M. Didic, V. de la Sayette, C. Desnuelle, B. Dubois, C. Duyckaerts, M. O. Habert, E. Guedj, D. Hannequin, L. Lacomblez, I. Le Ber, R. Levy, V. Meininger, B. F. Michel, F. Pasquier, C. Thomas-Anterion, M. Puel, F. Salachas, F. Sellal, M. Ver-

celletto, and P. Verpillat. Chromosome 9p-linked families with frontotemporal dementia associated with motor neuron disease. *Neurology*, 72:1669–1676, May 2009.

[119] F. P. Diekstra, P. W. van Vught, W. V. Rheenen, M. Koppers, R. J. Pasterkamp, M. A. van Es, H. J. Schelhaas, M. de Visser, W. Robberecht, P. Van Damme, P. M. Andersen, L. H. van den Berg, and J. H. Veldink. UNC13A is a modifier of survival in amyotrophic lateral sclerosis. *Neurobiol Aging*, Nov 2011.

[120] H. Laaksovirta, T. Peuralinna, J. C. Schymick, S. W. Scholz, S. L. Lai, L. Myllykangas, R. Sulkava, L. Jansson, D. G. Hernandez, J. R. Gibbs, M. A. Nalls, D. Heckerman, P. J. Tienari, and B. J. Traynor. Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. *Lancet Neurol*, 9:978–985, Oct 2010.

[121] M. Dejesus-Hernandez, I. R. Mackenzie, B. F. Boeve, A. L. Boxer, M. Baker, N. J. Rutherford, A. M. Nicholson, N. A. Finch, H. Flynn, J. Adamson, N. Kouri, A. Wojtas, P. Sengdy, G. Y. Hsiung, A. Karydas, W. W. Seeley, K. A. Josephs, G. Coppola, D. H. Geschwind, Z. K. Wszolek, H. Feldman, D. S. Knopman, R. C. Petersen, B. L. Miller, D. W. Dickson, K. B. Boylan, N. R. Graff-Radford, and R. Rademakers. Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron*, 72:245–256, Oct 2011.

[122] A. E. Renton, E. Majounie, A. Waite, J. Simon-Sanchez, S. Rollinson, J. R. Gibbs, J. C. Schymick, H. Laaksovirta, J. C. van Swieten, L. Myllykangas, H. Kalimo, A. Paetau, Y. Abramzon, A. M. Remes, A. Kaganovich, S. W. Scholz, J. Duckworth, J. Ding, D. W. Harmer, D. G. Hernandez, J. O. Johnson, K. Mok, M. Ryten, D. Trabzuni, R. J. Guerreiro, R. W. Orrell, J. Neal, A. Murray, J. Pearson, I. E. Jansen, D. Sondervan, H. Seelaar, D. Blake, K. Young, N. Halliwell, J. B. Callister, G. Toulson, A. Richardson, A. Gerhard, J. Snowden, D. Mann, D. Neary, M. A.

Nalls, T. Peuralinna, L. Jansson, V. M. Isoviita, A. L. Kaivorinne, M. Holtta-Vuori, E. Ikonen, R. Sulkava, M. Benatar, J. Wuu, A. Chio, G. Restagno, G. Borghero, M. Sabatelli, D. Heckerman, E. Rogaeva, L. Zinman, J. D. Rothstein, M. Sendtner, C. Drepper, E. E. Eichler, C. Alkan, Z. Abdullaev, S. D. Pack, A. Dutra, E. Pak, J. Hardy, A. Singleton, N. M. Williams, P. Heutink, S. Pickering-Brown, H. R. Morris, P. J. Tienari, and B. J. Traynor. A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD. *Neuron*, 72:257–268, Oct 2011.

[123] A. D. Skol, L. J. Scott, G. R. Abecasis, and M. Boehnke. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.*, 38:209–213, Feb 2006.

[124] G. H. Hardy. MENDELIAN PROPORTIONS IN A MIXED POPULATION. *Science*, 28:49–50, Jul 1908.

[125] W. Weinberg. ber den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins fr vaterlndische Naturkunde in Wrttemberg*, 64:368–382, Jan 1908.

[126] N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet.*, 2:e190, Dec 2006.

[127] P. J. Hastings, G. Ira, and J. R. Lupski. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.*, 5:e1000327, Jan 2009.

[128] P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira. Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, 10:551–564, Aug 2009.

[129] A. Piotrowski, C. E. Bruder, R. Andersson, T. Diaz de Stahl, U. Menzel, J. Sandgren, A. Poplawski, D. von Tell, C. Crasto, A. Bogdan, R. Bartoszewski, Z. Bebok,

M. Krzyzanowski, Z. Jankowski, E. C. Partridge, J. Komorowski, and J. P. Dumanski. Somatic mosaicism for copy number variation in differentiated human tissues. *Hum. Mutat.*, 29:1118–1124, Sep 2008.

[130] C. E. Bruder, A. Piotrowski, A. A. Gijsbers, R. Andersson, S. Erickson, T. Diaz de Stahl, U. Menzel, J. Sandgren, D. von Tell, A. Poplawski, M. Crowley, C. Crasto, E. C. Partridge, H. Tiwari, D. B. Allison, J. Komorowski, G. J. van Ommen, D. I. Boomsma, N. L. Pedersen, J. T. den Dunnen, K. Wirdefeldt, and J. P. Dumanski. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.*, 82:763–771, Mar 2008.

[131] R. Pamphlett and J. M. Morahan. Copy number imbalances in blood and hair in monozygotic twins discordant for amyotrophic lateral sclerosis. *J Clin Neurosci*, 18:1231–1234, Sep 2011.

[132] A. E. Dellinger, S. M. Saw, L. K. Goh, M. Seielstad, T. L. Young, and Y. J. Li. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res.*, 38:e105, May 2010.

[133] S. Colella, C. Yau, J. M. Taylor, G. Mirza, H. Butler, P. Clouston, A. S. Bassett, A. Seller, C. C. Holmes, and J. Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.*, 35:2013–2025, 2007.

[134] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, 17:1665–1674, Nov 2007.

[135] S. Cronin, H. M. Blauw, J. H. Veldink, M. A. van Es, R. A. Ophoff, D. G. Bradley, L. H. van den Berg, and O. Hardiman. Analysis of genome-wide copy number variation in Irish and Dutch ALS populations. *Hum. Mol. Genet.*, 17:3392–3398, Nov 2008.

[136] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles. Origins and functional impact of copy number variation in the human genome. *Nature*, 464:704–712, Apr 2010.

[137] K. Hemminki, X. Li, J. Sundquist, and K. Sundquist. Familial risks for amyotrophic lateral sclerosis and autoimmune diseases. *Neurogenetics*, 10:111–116, Apr 2009.

[138] E. S. Lander and D. Botstein. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, 236:1567–1570, Jun 1987.

[139] B. J. van den Bosch, M. Gerards, W. Sluiter, A. P. Stegmann, E. L. Jongen, D. M. Hellebrekers, R. Oegema, E. H. Lambrichs, H. Prokisch, K. Danhauser, K. Schoonderwoerd, I. F. de Coo, and H. J. Smeets. Defective NDUFA9 as a novel cause of neonatally fatal complex I disease. *J Med Genet*, Nov 2011.

[140] A. Dauber, T. T. Nguyen, E. Sochett, D. E. Cole, R. Horst, S. A. Abrams, T. O. Carpenter, and J. N. Hirschhorn. Genetic Defect in CYP24A1, the Vitamin D 24-Hydroxylase Gene, in a Patient with Severe Infantile Hypercalcemia. *J Clin Endocrinol Metab*, Nov 2011.

[141] M. A. Aldahmesh, A. O. Khan, J. Mohamed, and F. S. Alkuraya. Novel recessive BFSP2 and PITX3 mutations: Insights into mutational mechanisms from consan-

guineous populations. *Genet. Med.*, 13:978–981, Nov 2011.

[142] R. Abou Jamra, S. Wohlfart, M. Zweier, S. Uebe, L. Priebe, A. Ekici, S. Giesebrecht, A. Abboud, M. A. Al Khateeb, M. Fakher, S. Hamdan, A. Ismael, S. Muhammad, M. M. Nothen, J. Schumacher, and A. Reis. Homozygosity mapping in 64 Syrian consanguineous families with non-specific intellectual disability reveals 11 novel loci and high heterogeneity. *Eur. J. Hum. Genet.*, 19:1161–1166, Nov 2011.

[143] V. Martinez-Glez, M. Valencia, J. A. Caparros-Martin, M. Aglan, S. Temtamy, J. Tenorio, V. Pulido, U. Lindert, M. Rohrbach, D. Eyre, C. Giunta, P. Lapunzina, and V. L. Ruiz-Perez. Identification of a mutation causing deficient BMP1/mTLD proteolytic activity in autosomal recessive osteogenesis imperfecta. *Hum Mutat*, Nov 2011.

[144] S. B. Dana, L. Jana, S. Helena, K. Marcela, T. Marie, M. Petr, and S. Pavel. DFNB49 is an important cause of non-syndromic deafness in Czech Roma patients but not in the general Czech population. *Clin Genet*, Nov 2011.

[145] J. Gibson, N. E. Morton, and A. Collins. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.*, 15:789–795, Mar 2006.

[146] R. McQuillan, A. L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, L. Barac-Lauc, N. Smolej-Narancic, B. Janicijevic, O. Polasek, A. Tenesa, A. K. Macleod, S. M. Farrington, P. Rudan, C. Hayward, V. Vitart, I. Rudan, S. H. Wild, M. G. Dunlop, A. F. Wright, H. Campbell, and J. F. Wilson. Runs of homozygosity in European populations. *Am. J. Hum. Genet.*, 83:359–372, Sep 2008.

[147] M. A. Nalls, J. Simon-Sanchez, J. R. Gibbs, C. Paisan-Ruiz, J. T. Bras, T. Tanaka, M. Matarin, S. Scholz, C. Weitz, T. B. Harris, L. Ferrucci, J. Hardy, and A. B.

Singleton. Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. *PLoS Genet.*, 5:e1000415, Mar 2009.

[148] T. Lencz, C. Lambert, P. DeRosse, K. E. Burdick, T. V. Morgan, J. M. Kane, R. Kucherlapati, and A. K. Malhotra. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.*, 104:19942–19947, Dec 2007.

[149] M. A. Nalls, R. J. Guerreiro, J. Simon-Sanchez, J. T. Bras, B. J. Traynor, J. R. Gibbs, L. Launer, J. Hardy, and A. B. Singleton. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics*, 10:183–190, Jul 2009.

[150] F. Hildebrandt, S. F. Heeringa, F. Ruschendorf, M. Attanasio, G. Nurnberg, C. Becker, D. Seelow, N. Huebner, G. Chernin, C. N. Vlangos, W. Zhou, J. F. O'Toole, B. E. Hoskins, M. T. Wolf, B. G. Hinkes, H. Chaib, S. Ashraf, D. S. Schoeb, B. Ovunc, S. J. Allen, V. Vega-Warner, E. Wise, H. M. Harville, R. H. Lyons, J. Washburn, J. Macdonald, P. Nurnberg, and E. A. Otto. A systematic approach to mapping recessive disease genes in individuals from outbred populations. *PLoS Genet.*, 5:e1000353, Jan 2009.

[151] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, Jun 2000.

[152] S. Sawcer, G. Hellenthal, M. Pirinen, C. C. Spencer, N. A. Patsopoulos, L. Moutsianas, A. Dilthey, Z. Su, C. Freeman, S. E. Hunt, S. Edkins, E. Gray, D. R. Booth, S. C. Potter, A. Goris, G. Band, A. B. Oturai, A. Strange, J. Saarela, C. Bellenguez, B. Fontaine, M. Gillman, B. Hemmer, R. Gwilliam, F. Zipp, A. Jayakumar, R. Martin, S. Leslie, S. Hawkins, E. Giannoulatou, S. D'alfonso, H. Blackburn, F. M. Boneschi, J. Liddle, H. F. Harbo, M. L. Perez, A. Spurkland, M. J. Waller,

M. P. Mycko, M. Ricketts, M. Comabella, N. Hammond, I. Kockum, O. T. McCann, M. Ban, P. Whittaker, A. Kemppinen, P. Weston, C. Hawkins, S. Widaa, J. Zajicek, S. Dronov, N. Robertson, S. J. Bumpstead, L. F. Barcellos, R. Ravindrarajah, R. Abraham, L. Alfredsson, K. Ardlie, C. Aubin, A. Baker, K. Baker, S. E. Baranzini, L. Bergamaschi, R. Bergamaschi, A. Bernstein, A. Berthele, M. Boggild, J. P. Bradfield, D. Brassat, S. A. Broadley, D. Buck, H. Butzkueven, R. Capra, W. M. Carroll, P. Cavalla, E. G. Celius, S. Cepok, R. Chiavacci, F. Clerget-Darpoux, K. Clysters, G. Comi, M. Cossburn, I. Cournu-Rebeix, M. B. Cox, W. Cozen, B. A. Cree, A. H. Cross, D. Cusi, M. J. Daly, E. Davis, P. I. de Bakker, M. Debouverie, M. B. D'hooghe, K. Dixon, R. Dobosi, B. Dubois, D. Ellinghaus, I. Elovaara, F. Esposito, C. Fontenille, S. Foote, A. Franke, D. Galimberti, A. Ghezzi, J. Glessner, R. Gomez, O. Gout, C. Graham, S. F. Grant, F. R. Guerini, H. Hakonarson, P. Hall, A. Hamsten, H. P. Hartung, R. N. Heard, S. Heath, J. Hobart, M. Hoshi, C. Infante-Duarte, G. Ingram, W. Ingram, T. Islam, M. Jagodic, M. Kabesch, A. G. Kermode, T. J. Kilpatrick, C. Kim, N. Klopp, K. Koivisto, M. Larsson, M. Lathrop, J. S. Lechner-Scott, M. A. Leone, V. Leppa, U. Liljedahl, I. L. Bomfim, R. R. Lincoln, J. Link, J. Liu, A. R. Lorentzen, S. Lupoli, F. Macciardi, T. Mack, M. Marriott, V. Martinelli, D. Mason, J. L. McCauley, F. Mentch, I. L. Mero, T. Mihalova, X. Montalban, J. Mottershead, K. M. Myhr, P. Naldi, W. Ollier, A. Page, A. Palotie, J. Pelletier, L. Piccio, T. Pickersgill, F. Piehl, S. Pobywajlo, H. L. Quach, P. P. Ramsay, M. Reunanen, R. Reynolds, J. D. Rioux, M. Rodegher, S. Roesner, J. P. Rubio, I. M. Ruckert, M. Salvetti, E. Salvi, A. Santaniello, C. A. Schaefer, S. Schreiber, C. Schulze, R. J. Scott, F. Sellebjerg, K. W. Selmaj, D. Sexton, L. Shen, B. Simms-Acuna, S. Skidmore, P. M. Sleiman, C. Smestad, P. S. Sørensen, H. B. Søndergaard, J. Stankovich, R. C. Strange, A. M. Sulonen, E. Sundqvist, A. C. Syvanen, F. Taddeo, B. Taylor, J. M. Blackwell, P. Tienari, E. Bramon, A. Tourbah, M. A. Brown, E. Tronczyn-

ska, J. P. Casas, N. Tubridy, A. Corvin, J. Vickery, J. Jankowski, P. Villoslada, H. S. Markus, K. Wang, C. G. Mathew, J. Wason, C. N. Palmer, H. E. Wichmann, R. Plomin, E. Willoughby, A. Rautanen, J. Winkelmann, M. Wittig, R. C. Trembath, J. Yaouanq, A. C. Viswanathan, H. Zhang, N. W. Wood, R. Zuvich, P. Deloukas, C. Langford, A. Duncanson, J. R. Oksenberg, M. A. Pericak-Vance, J. L. Haines, T. Olsson, J. Hillert, A. J. Ivinson, P. L. De Jager, L. Peltonen, G. J. Stewart, D. A. Hafler, S. L. Hauser, G. McVean, P. Donnelly, and A. Compston. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476:214–219, Aug 2011.

[153] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15:1034–1050, Aug 2005.

[154] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res.*, 12:996–1006, Jun 2002.

[155] X. Wang, S. Arai, X. Song, D. Reichart, K. Du, G. Pascual, P. Tempst, M. G. Rosenfeld, C. K. Glass, and R. Kurokawa. Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature*, 454:126–130, Jul 2008.

[156] J. Amiel, M. Rio, L. de Pontual, R. Redon, V. Malan, N. Boddaert, P. Plouin, N. P. Carter, S. Lyonnet, A. Munnich, and L. Colleaux. Mutations in TCF4, encoding a class I basic helix-loop-helix transcription factor, are responsible for Pitt-Hopkins syndrome, a severe epileptic encephalopathy associated with autonomic dysfunction. *Am. J. Hum. Genet.*, 80:988–993, May 2007.

[157] D. Pitt and I. Hopkins. A syndrome of mental retardation, wide mouth and intermittent overbreathing. *Aust Paediatr J*, 14:182–184, Sep 1978.

[158] The ENCODE Project Consortium. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447:799–816, Jun 2007.

[159] L. Winchester, C. Yau, and J. Ragoussis. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic*, 8:353–366, Sep 2009.

[160] J. E. Landers, J. Melki, V. Meininger, J. D. Glass, L. H. van den Berg, M. A. van Es, P. C. Sapp, P. W. van Vught, D. M. McKenna-Yasek, H. M. Blauw, T. J. Cho, M. Polak, L. Shi, A. M. Wills, W. J. Broom, N. Ticozzi, V. Silani, A. Ozoguz, I. Rodriguez-Leyva, J. H. Veldink, A. J. Ivinson, C. G. Saris, B. A. Hosler, A. Barnes-Nessa, N. Couture, J. H. Wokke, T. J. Kwiatkowski, R. A. Ophoff, S. Cronin, O. Hardiman, F. P. Diekstra, P. N. Leigh, C. E. Shaw, C. L. Simpson, V. K. Hansen, J. F. Powell, P. Corcia, F. Salachas, S. Heath, P. Galan, F. Georges, H. R. Horvitz, M. Lathrop, S. Purcell, A. Al-Chalabi, and R. H. Brown. Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. U.S.A.*, 106:9004–9009, Jun 2009.

[161] A. Hentati, K. Ouahchi, M. A. Pericak-Vance, D. Nijhawan, A. Ahmad, Y. Yang, J. Rimmler, W. Hung, B. Schlotter, A. Ahmed, M. Ben Hamida, F. Hentati, and T. Siddique. Linkage of a commoner form of recessive amyotrophic lateral sclerosis to chromosome 15q15-q22 markers. *Neurogenetics*, 2:55–60, Dec 1998.

[162] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee. Detection of large-scale variation in the human genome. *Nat. Genet.*, 36:949–951, Sep 2004.

[163] T. Biederer. Bioinformatic characterization of the SynCAM family of immunoglobulin-like domain-containing adhesion molecules. *Genomics*, 87:139–150, Jan 2006.

[164] J. P. Casey, T. Magalhaes, J. M. Conroy, R. Regan, N. Shah, R. Anney, D. C. Shields, B. S. Abrahams, J. Almeida, E. Bacchelli, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, P. F. Bolton, T. Bourgeron, S. Brennan, P. Cali, C. Correia, C. Corsello, M. Coutanche, G. Dawson, M. de Jonge, R. Delorme, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, S. Foley, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Green, S. J. Guter, H. Hakonarson, R. Holt, G. Hughes, V. Hus, R. Igliozzi, C. Kim, S. M. Klauck, A. Kolevzon, J. A. Lamb, M. Leboyer, A. Le Couteur, B. L. Leventhal, C. Lord, S. C. Lund, E. Maestrini, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougle, J. McGrath, W. M. McMahon, A. Merikangas, J. Miller, F. Minopoli, G. K. Mirza, J. Munson, S. F. Nelson, G. Nygren, G. Oliveira, A. T. Pagnamenta, K. Papanikolaou, J. R. Parr, B. Parrini, A. Pickles, D. Pinto, J. Piven, D. J. Posey, A. Poustka, F. Poustka, J. Ragoussis, B. Roge, M. L. Rutter, A. F. Sequeira, L. Soorya, I. Sousa, N. Sykes, V. Stoppioni, R. Tancredi, M. Tauber, A. P. Thompson, S. Thomson, J. Tsiantis, H. Van Engeland, J. B. Vincent, F. Volkmar, J. A. Vorstman, S. Wallace, K. Wang, T. H. Wassink, K. White, K. Wing, K. Wittemeyer, B. L. Yaspan, L. Zwaigenbaum, C. Betancur, J. D. Buxbaum, R. M. Cantor, E. H. Cook, H. Coon, M. L. Cuccaro, D. H. Geschwind, J. L. Haines, J. Hallmayer, A. P. Monaco, J. I. Nurnberger, M. A. Pericak-Vance, G. D. Schellenberg, S. W. Scherer, J. S. Sutcliffe, P. Szatmari, V. J. Vieland, E. M. Wijsman, A. Green, M. Gill, L. Gallagher, A. Vicente, and S. Ennis. A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. *Hum Genet*, Oct 2011.

[165] B. D. Angst, C. Marcozzi, and A. I. Magee. The cadherin superfamily: diversity in form and function. *J. Cell. Sci.*, 114:629–641, Feb 2001.

[166] K. Krishna-K, N. Hertel, and C. Redies. Cadherin expression in the somatosensory cortex: evidence for a combinatorial molecular code at the single-cell level. *Neuroscience*, 175:37–48, Feb 2011.

[167] T. Ota, Y. Suzuki, T. Nishikawa, T. Otsuki, T. Sugiyama, R. Irie, A. Wakamatsu, K. Hayashi, H. Sato, K. Nagai, K. Kimura, H. Makita, M. Sekine, M. Obayashi, T. Nishi, T. Shibahara, T. Tanaka, S. Ishii, J. Yamamoto, K. Saito, Y. Kawai, Y. Isono, Y. Nakamura, K. Nagahari, K. Murakami, T. Yasuda, T. Iwayanagi, M. Wagatsuma, A. Shiratori, H. Sudo, T. Hosoiri, Y. Kaku, H. Kodaira, H. Kondo, M. Sugawara, M. Takahashi, K. Kanda, T. Yokoi, T. Furuya, E. Kikkawa, Y. Omura, K. Abe, K. Kamihara, N. Katsuta, K. Sato, M. Tanikawa, M. Yamazaki, K. Ninomiya, T. Ishibashi, H. Yamashita, K. Murakawa, K. Fujimori, H. Tanai, M. Kimata, M. Watanabe, S. Hiraoka, Y. Chiba, S. Ishida, Y. Ono, S. Takiguchi, S. Watanabe, M. Yosida, T. Hotuta, J. Kusano, K. Kanehori, A. Takahashi-Fujii, H. Hara, T. O. Tanase, Y. Nomura, S. Togiya, F. Komai, R. Hara, K. Takeuchi, M. Arita, N. Imose, K. Musashino, H. Yuuki, A. Oshima, N. Sasaki, S. Aotsuka, Y. Yoshikawa, H. Matsunawa, T. Ichihara, N. Shiohata, S. Sano, S. Moriya, H. Momiyama, N. Satoh, S. Takami, Y. Terashima, O. Suzuki, S. Nakagawa, A. Senoh, H. Mizoguchi, Y. Goto, F. Shimizu, H. Wakebe, H. Hishigaki, T. Watanabe, A. Sugiyama, M. Takemoto, B. Kawakami, M. Yamazaki, K. Watanabe, A. Kumagai, S. Itakura, Y. Fukuzumi, Y. Fujimori, M. Komiyama, H. Tashiro, A. Tanigami, T. Fujiwara, T. Ono, K. Yamada, Y. Fujii, K. Ozaki, M. Hirao, Y. Ohmori, A. Kawabata, T. Hikiji, N. Kobatake, H. Inagaki, Y. Ikema, S. Okamoto, R. Okitani, T. Kawakami, S. Noguchi, T. Itoh, K. Shigeta, T. Senba, K. Matsumura, Y. Nakajima, T. Mizuno,

M. Morinaga, M. Sasaki, T. Togashi, M. Oyama, H. Hata, M. Watanabe, T. Komatsu, J. Mizushima-Sugano, T. Satoh, Y. Shirai, Y. Takahashi, K. Nakagawa, K. Okumura, T. Nagase, N. Nomura, H. Kikuchi, Y. Masuho, R. Yamashita, K. Nakai, T. Yada, Y. Nakamura, O. Ohara, T. Isogai, and S. Sugano. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.*, 36:40–45, Jan 2004.

[168] Q. Sha, Z. Zhang, J. C. Schymick, B. J. Traynor, and S. Zhang. Genome-wide association reveals three SNPs associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis. *BMC Med. Genet.*, 10:86, 2009.

[169] M. L. Metzker. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11:31–46, Jan 2010.

[170] E. R. Mardis. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402, 2008.

[171] M. Kircher, U. Stenzel, and J. Kelso. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.*, 10:R83, 2009.

[172] T. J. Albert, M. N. Molla, D. M. Muzny, L. Nazareth, D. Wheeler, X. Song, T. A. Richmond, C. M. Middle, M. J. Rodesch, C. J. Packard, G. M. Weinstock, and R. A. Gibbs. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, 4:903–905, Nov 2007.

[173] D. T. Okou, K. M. Steinberg, C. Middle, D. J. Cutler, T. J. Albert, and M. E. Zwick. Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods*, 4:907–909, Nov 2007.

[174] A. Gnirke, A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander,

and C. Nusbaum. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, 27:182–189, Feb 2009.

[175] R. Tewhey, M. Nakano, X. Wang, C. Pabon-Pena, B. Novak, A. Giuffre, E. Lin, S. Happe, D. N. Roberts, E. M. LeProust, E. J. Topol, O. Harismendy, and K. A. Frazer. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol.*, 10:R116, 2009.

[176] L. Mamanova, A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure, and D. J. Turner. Target-enrichment strategies for next-generation sequencing. *Nat. Methods*, 7:111–118, Feb 2010.

[177] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10:57–63, Jan 2009.

[178] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316:1497–1502, Jun 2007.

[179] E. M. Kenny, P. Cormican, W. P. Gilks, A. S. Gates, C. T. O'Dushlaine, C. Pinto, A. P. Corvin, M. Gill, and D. W. Morris. Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Res.*, 18:31–38, Feb 2011.

[180] Y. Li, C. Sidore, H. M. Kang, M. Boehnke, and G. R. Abecasis. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.*, 21:940–951, Jun 2011.

[181] S. Morgenthaler and W. G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.*, 615:28–56, Feb 2007.

[182] B. Li and S. M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, 83:311–321, Sep 2008.

[183] B. E. Madsen and S. R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, 5:e1000384, Feb 2009.

[184] B. M. Neale, M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly. Testing for an unusual distribution of rare variants. *PLoS Genet.*, 7:e1001322, Mar 2011.

[185] N. Parkinson, P. G. Ince, M. O. Smith, R. Highley, G. Skibinski, P. M. Andersen, K. E. Morrison, H. S. Pall, O. Hardiman, J. Collinge, P. J. Shaw, and E. M. Fisher. ALS phenotypes with mutations in CHMP2B (charged multivesicular body protein 2B). *Neurology*, 67:1074–1077, Sep 2006.

[186] L. E. Cox, L. Ferraiuolo, E. F. Goodall, P. R. Heath, A. Higginbottom, H. Mortiboys, H. C. Hollinger, J. A. Hartley, A. Brockington, C. E. Burness, K. E. Morrison, S. B. Wharton, A. J. Grierson, P. G. Ince, J. Kirby, and P. J. Shaw. Mutations in CHMP2B in lower motor neuron predominant amyotrophic lateral sclerosis (ALS). *PLoS ONE*, 5:e9872, 2010.

[187] M. A. van Es, P. W. van Vught, H. M. Blauw, L. Franke, C. G. Saris, L. Van den Bosch, S. W. de Jong, V. de Jong, F. Baas, R. van't Slot, R. Lemmens, H. J. Schelhaas, A. Birve, K. Sleegers, C. Van Broeckhoven, J. C. Schymick, B. J. Traynor, J. H. Wokke, C. Wijmenga, W. Robberecht, P. M. Andersen, J. H. Veldink, R. A. Ophoff, and L. H. van den Berg. Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.*, 40:29–31, Jan 2008.

[188] C. L. Simpson, R. Lemmens, K. Miskiewicz, W. J. Broom, V. K. Hansen, P. W. van Vught, J. E. Landers, P. Sapp, L. Van Den Bosch, J. Knight, B. M. Neale, M. R. Turner, J. H. Veldink, R. A. Ophoff, V. B. Tripathi, A. Beleza, M. N. Shah, P. Proitsi, A. Van Hoecke, P. Carmeliet, H. R. Horvitz, P. N. Leigh, C. E. Shaw, L. H. van den Berg, P. C. Sham, J. F. Powell, P. Verstreken, R. H. Brown, W. Robberecht, and A. Al-Chalabi. Variants of the elongator protein 3 (ELP3) gene are associated with motor neuron degeneration. *Hum. Mol. Genet.*, 18:472–481, Feb 2009.

[189] C. Y. Chow, J. E. Landers, S. K. Bergren, P. C. Sapp, A. E. Grant, J. M. Jones, L. Everett, G. M. Lenk, D. M. McKenna-Yasek, L. S. Weisman, D. Figlewicz, R. H. Brown, and M. H. Meisler. Deleterious variants of FIG4, a phosphoinositide phosphatase, in patients with ALS. *Am. J. Hum. Genet.*, 84:85–88, Jan 2009.

[190] K. Sleegers, N. Brouwers, S. Maurer-Stroh, M. A. van Es, P. Van Damme, P. W. van Vught, J. van der Zee, S. Serneels, T. De Pooter, M. Van den Broeck, M. Cruts, J. Schymkowitz, P. De Jonghe, F. Rousseau, L. H. van den Berg, W. Robberecht, and C. Van Broeckhoven. Progranulin genetic variability contributes to amyotrophic lateral sclerosis. *Neurology*, 71:253–259, Jul 2008.

[191] X. S. Wang, S. Lee, Z. Simmons, P. Boyer, K. Scott, W. Liu, and J. Connor. Increased incidence of the Hfe mutation in amyotrophic lateral sclerosis and related cellular consequences. *J. Neurol. Sci.*, 227:27–33, Dec 2004.

[192] E. F. Goodall, M. J. Greenway, I. van Marion, C. B. Carroll, O. Hardiman, and K. E. Morrison. Association of the H63D polymorphism in the hemochromatosis gene with sporadic ALS. *Neurology*, 65:934–937, Sep 2005.

[193] P. D. Sundar, C. E. Yu, W. Sieh, E. Steinbart, R. M. Garruto, K. Oyanagi, U. K. Craig, T. D. Bird, E. M. Wijsman, D. R. Galasko, and G. D. Schellenberg. Two

sites in the MAPT region confer genetic risk for Guam ALS/PDC and dementia. *Hum. Mol. Genet.*, 16:295–306, Feb 2007.

[194] D. A. Figlewicz, A. Krizus, M. G. Martinoli, V. Meininger, M. Dib, G. A. Rouleau, and J. P. Julien. Variants of the heavy neurofilament subunit are associated with the development of amyotrophic lateral sclerosis. *Hum. Mol. Genet.*, 3:1757–1761, Oct 1994.

[195] V. Skvortsova, M. Shadrina, P. Slominsky, G. Levitsky, E. Kondratieva, A. Zherebtsova, N. Levitskaya, A. Alekhin, A. Serdyuk, and S. Limborska. Analysis of heavy neurofilament subunit gene polymorphism in Russian patients with sporadic motor neuron disease (MND). *Eur. J. Hum. Genet.*, 12:241–244, Mar 2004.

[196] G. Annesi, G. Savettieri, P. Pugliese, M. D'Amelio, P. Tarantino, P. Ragonese, V. La Bella, T. Piccoli, D. Civitelli, F. Annesi, B. Fierro, F. Piccoli, G. Arabia, M. Caracciolo, I. C. Ciro Candiano, and A. Quattrone. DJ-1 mutations and parkinsonism-dementia-amyotrophic lateral sclerosis complex. *Ann. Neurol.*, 58:803–807, Nov 2005.

[197] M. Saeed, N. Siddique, W. Y. Hung, E. Usacheva, E. Liu, R. L. Sufit, S. L. Heller, J. L. Haines, M. Pericak-Vance, and T. Siddique. Paraoxonase cluster polymorphisms are associated with sporadic ALS. *Neurology*, 67:771–776, Sep 2006.

[198] A. Slowik, B. Tomik, P. P. Wolkow, D. Partyka, W. Turaj, M. T. Malecki, J. Pera, T. Dziedzic, A. Szczudlik, and D. A. Figlewicz. Paraoxonase gene polymorphisms and sporadic ALS. *Neurology*, 67:766–770, Sep 2006.

[199] S. Cronin, M. J. Greenway, J. H. Prehn, and O. Hardiman. Paraoxonase promoter and intronic variants modify risk of sporadic amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatr.*, 78:984–986, Sep 2007.

[200] N. Ticozzi, A. L. LeClerc, P. J. Keagle, J. D. Glass, A. M. Wills, M. van Blitterswijk, D. A. Bosco, I. Rodriguez-Leyva, C. Gellera, A. Ratti, F. Taroni, D. McKenna-Yasek, P. C. Sapp, V. Silani, C. E. Furlong, R. H. Brown, and J. E. Landers. Paraoxonase gene mutations in amyotrophic lateral sclerosis. *Ann. Neurol.*, 68:102–107, Jul 2010.

[201] F. Gros-Louis, R. Lariviere, G. Gowing, S. Laurent, W. Camu, J. P. Bouchard, V. Meininger, G. A. Rouleau, and J. P. Julien. A frameshift deletion in peripherin gene associated with amyotrophic lateral sclerosis. *J. Biol. Chem.*, 279:45951–45956, Oct 2004.

[202] C. L. Leung, C. Z. He, P. Kaufmann, S. S. Chin, A. Naini, R. K. Liem, H. Mitsumoto, and A. P. Hays. A pathogenic peripherin gene mutation in a patient with amyotrophic lateral sclerosis. *Brain Pathol.*, 14:290–296, Jul 2004.

[203] L. Corrado, Y. Carlomagno, L. Falasco, S. Mellone, M. Godi, E. Cova, C. Cereda, L. Testa, L. Mazzini, and S. D'Alfonso. A novel peripherin gene (PRPH) mutation identified in one sporadic amyotrophic lateral sclerosis patient. *Neurobiol. Aging*, 32:1–6, Mar 2011.

[204] P. Corcia, W. Camu, J. Praline, P. H. Gordon, P. Vourch, and C. Andres. The importance of the SMN genes in the genetics of sporadic ALS. *Amyotroph Lateral Scler*, 10:436–440, 2009.

[205] A. Orlacchio, C. Babalini, A. Borreca, C. Patrono, R. Massa, S. Basaran, R. P. Munhoz, E. A. Rogaeva, P. H. St George-Hyslop, G. Bernardi, and T. Kawarai. SPATACSIN mutations cause autosomal recessive juvenile amyotrophic lateral sclerosis. *Brain*, 133:591–598, Feb 2010.

[206] D. W. Craig, J. V. Pearson, S. Szelinger, A. Sekar, M. Redman, J. J. Corneveaux, T. L. Pawlowski, T. Laub, G. Nunn, D. A. Stephan, N. Homer, and M. J. Huentel-

man. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods*, 5:887–893, Oct 2008.

[207] M. A. Quail, I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow, and D. J. Turner. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, 5:1005–1010, Dec 2008.

[208] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, 38:1767–1771, Apr 2010.

[209] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, 8:186–194, Mar 1998.

[210] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–1760, Jul 2009.

[211] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079, Aug 2009.

[212] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20:1297–1303, Sep 2010.

[213] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27:2156–2158, Aug 2011.

[214] M. Meyer and M. Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*, 2010:pdb.prot5448, Jun 2010.

[215] B. E. Davy and M. L. Robinson. Congenital hydrocephalus in hy3 mice is caused by a frameshift mutation in Hydin, a large novel gene. *Hum. Mol. Genet.*, 12:1163–1170, May 2003.

[216] P. Paez, L. F. Batiz, R. Roales-Bujan, L. M. Rodriguez-Perez, S. Rodriguez, A. J. Jimenez, E. M. Rodriguez, and J. M. Perez-Figares. Patterned neuropathologic events occurring in hyh congenital hydrocephalic mutant mice. *J. Neuropathol. Exp. Neurol.*, 66:1082–1092, Dec 2007.

[217] N. A. Doggett, G. Xie, L. J. Meincke, R. D. Sutherland, M. O. Mundt, N. S. Berbari, B. E. Davy, M. L. Robinson, M. K. Rudd, J. L. Weber, R. L. Stallings, and C. Han. A 360-kb interchromosomal duplication of the human HYDIN locus. *Genomics*, 88:762–771, Dec 2006.

[218] N. Brunetti-Pierri, J. S. Berg, F. Scaglia, J. Belmont, C. A. Bacino, T. Sahoo, S. R. Lalani, B. Graham, B. Lee, M. Shinawi, J. Shen, S. H. Kang, A. Pursley, T. Lotze, G. Kennedy, S. Lansky-Shafer, C. Weaver, E. R. Roeder, T. A. Grebe, G. L. Arnold, T. Hutchison, T. Reimschisel, S. Amato, M. T. Geragthy, J. W. Innis, E. Obersztyn, B. Nowakowska, S. S. Rosengren, P. I. Bader, D. K. Grange, S. Naqvi, A. D. Garnica, S. M. Bernes, C. T. Fong, A. Summers, W. D. Walters, J. R. Lupski, P. Stankiewicz, S. W. Cheung, and A. Patel. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat. Genet.*, 40:1466–1471, Dec 2008.

[219] P. C. Ng, S. Levy, J. Huang, T. B. Stockwell, B. P. Walenz, K. Li, N. Axelrod, D. A. Busam, R. L. Strausberg, and J. C. Venter. Genetic variation in an individual

human exome. *PLoS Genet.*, 4:e1000160, 2008.

[220] M. J. Clark, R. Chen, H. Y. Lam, K. J. Karczewski, R. Chen, G. Euskirchen, A. J. Butte, and M. Snyder. Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.*, 29:908–914, 2011.

[221] A. M. Sulonen, P. Ellonen, H. Almusa, M. Lepisto, S. Eldfors, S. Hannula, T. Miettinen, H. Tyynismaa, P. Salo, C. Heckman, H. Joensuu, T. Raivio, A. Suomalainen, and J. Saarela. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol*, 12:R94, Sep 2011.

[222] Asan, Y. Xu, H. Jiang, C. Tyler-Smith, Y. Xue, T. Jiang, J. Wang, M. Wu, X. Liu, G. Tian, J. Wang, J. Wang, H. Yang, and X. Zhang. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol*, 12:R95, Sep 2011.

[223] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad. Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, 42:30–35, Jan 2010.

[224] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, 9:356–369, May 2008.

[225] E. A. Thompson. The IBD process along four chromosomes. *Theor Popul Biol*, 73:369–373, May 2008.

[226] A. Gusev, J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler, J. L. Breslow, J. M. Friedman, and I. Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Res.*, 19:318–326, Feb 2009.

[227] E. L. Stevens, G. Heckenberg, E. D. Roberson, J. D. Baugher, T. J. Downey, and J. Pevsner. Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genet.*, 7:e1002287, Sep 2011.

[228] B. L. Browning and S. R. Browning. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.*, 88:173–182, Feb 2011.

[229] S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, 81:1084–1097, Nov 2007.

[230] Central Statistics Office. Census of population 2011 preliminary results. Technical report, Central Statistics Office of Ireland, 2011.

[231] Northern Ireland Statistics and Research Agency. Population and migration estimates Northern Ireland (2009) – statistical report. Technical report, Northern Ireland Statistics and Research Agency, 2009.

[232] S. R. Browning and B. L. Browning. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.*, 12:703–714, Oct 2011.

[233] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, 34:816–834, Dec 2010.

[234] B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, 5:e1000529, Jun 2009.

[235] UK Office for National Statistics. Statistical bulletin: annual mid-year population estimates, 2010. Technical report, Office for National Statistics, UK, 2010.

[236] B. A. Hosler, T. Siddique, P. C. Sapp, W. Sailor, M. C. Huang, A. Hossain, J. R. Daube, M. Nance, C. Fan, J. Kaplan, W. Y. Hung, D. McKenna-Yasek, J. L. Haines, M. A. Pericak-Vance, H. R. Horvitz, and R. H. Brown. Linkage of familial amyotrophic lateral sclerosis with frontotemporal dementia to chromosome 9q21-q22. *JAMA*, 284:1664–1669, Oct 2000.

[237] A. Albrechtsen, I. Moltke, and R. Nielsen. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186:295–308, Sep 2010.

[238] S. Byrne, M. Elamin, P. Bede, A. Shatunov, R. L. McLaughlin, K. Kenna, N. Jordan, B. Wynne, C. OBrien, M. Heverin, C. Lynch, B. Corr, C. Walsh, A. Bodke, D. G. Bradley, N. Pender, A Al-Chalabi, and O. Hardiman. Phenotype, genotype and population--based frequency of C9ORF72 repeat expansion in ALS. *(Under review)*, Dec 2011.

[239] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126:663–676, Aug 2006.

[240] M. Nizzardo, C. Simone, M. Falcone, F. Locatelli, G. Riboldi, G. P. Comi, and S. Corti. Human motor neuron generation from embryonic stem cells and induced pluripotent stem cells. *Cell. Mol. Life Sci.*, 67:3837–3847, Nov 2010.

[241] E. T. Cirulli, A. Singh, K. V. Shianna, D. Ge, J. P. Smith, J. M. Maia, E. L. Heinzen, J. J. Goedert, and D. B. Goldstein. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol.*, 11:R57, 2010.

[242] O. A. Hampton, C. A. Miller, M. Koriabine, J. Li, P. Den Hollander, L. Carbone, M. Nefedov, B. F. Ten Hallers, A. V. Lee, P. J. De Jong, and A. Milosavljevic. Long-range massively parallel mate pair sequencing detects distinct mutations and

similar patterns of structural mutability in two breast cancer cell lines. *Cancer Genet*, 204:447–457, Aug 2011.

[243] P. W. Laird. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.*, 11:191–203, Mar 2010.

[244] V. K. Rakyan, T. A. Down, D. J. Balding, and S. Beck. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, 12:529–541, Aug 2011.

[245] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh. Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.*, 28:171–182, Feb 2005.

[246] L. W. Hahn, M. D. Ritchie, and J. H. Moore. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19:376–382, Feb 2003.