Understanding Disease Trajectory in Amyotrophic Lateral Sclerosis

Al Khleifat, Ahmad

Awarding institution:
King's College London

Title: Understanding Disease Trajectory in Amyotrophic Lateral Sclerosis

Dr Ahmad Al Khleifat

Thesis submitted for the degree of Doctor of Philosophy

Department of Basic and Clinical Neuroscience

Institute of Psychiatry, Psychology and Neuroscience

King's College London

January 2019

Acknowledgements

This research was conducted in the Department of Clinical Neuroscience at King's College London under the supervision of Professor Ammar Al-Chalabi and Professor John Powell, to whom I am thankful for their continuous guidance. I have been so blessed to have mentors and academics who have looked out for me in a generous way, to whom I am forever grateful.

I want to express my deepest gratitude for all the help that I received from Professor Ammar. Thank you for every minute you spent teaching me, thank you for believing in me. I hope that one day I will be a Clinician Scientist as good as you. Thank you, Ammar.

I am grateful for all the guidance I received from Professor John Powell - your door was always open, and your words and wisdom shaped most of this thesis. Thank you, John.

I would like to thank the members of Ammar's team;
Alfredo, Aleksey, Ton, Anna, Ashley, Sarah and Andrea. None of this work would be possible without you.

I would like to thank all the members of King's MND clinic and MIROCALS clinical trial team.

I want to thank Dr Rubika Balendra, whom I worked with on several projects.

I would also like to acknowledge the Motor Neurone Disease Association and Project MinE. Their expansional support for ALS research and people with ALS is phenomenal.

I want to thank Professor Jan Veldink for giving me the opportunity to extend my work to Project MinE and for his trust in me as a Co-Chair in the Genomic Structural Variation working group.

A big thank you to Dr Abdul Hye, Dr Jemeen Sreedharan and Dr Nick Ashton for their encouragement and friendship.

I want to thank my father-in-law Jeremy Heading for his advice and encouragement. Your fatherly support is incomparable.

Last but not least, none of this would have been possible without my family's endless support and encouragement; my father and especially from my son Adam and my loving wife Hanna. Thank you, Hanna, for being my voice of reason, my heart of the matter, and my sounding board. Thank you for always helping me think clearly, for helping me find the answers to my questions, and for giving me the courage to try. I love you forever.

This research is dedicated to my mother. Your love, guidance, support and example continues to be the cornerstone of my personal and professional life. I hope this research will help, even if a little bit, people with Motor Neurone Disease.

“Happiness is the joy that we feel striving for our potential”

-Shawn Achor

Declaration

I hereby declare that, as the author of this thesis, the work reported here was performed by myself unless otherwise stated. Any collaborative effort and assistance provided by others have been acknowledged and all previous work referenced accordingly. This work has not been submitted for any other degree.

Abstract

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease predominantly of motor neurons, characterized by progressive weakness of voluntary muscles and death from respiratory failure due to diaphragmatic paralysis, typically within three years of onset. Despite the very poor prognosis, there is considerable variation in the survival rate, and up to 10% of people with ALS live more than eight years from first symptoms. There is a strong genetic contribution to ALS risk. In 5% of cases or more, a family history of ALS or frontotemporal dementia is obtained, and the Mendelian genes responsible for ALS in such families have now been identified in about 70% of cases. Even in apparently sporadic cases, twin and population studies show the heritability is about 60%. Although risk genes reveal information about the mechanism of causation of ALS, it is also important to identify genomic variants that modify survival. Such variation could potentially be targeted directly, as could any RNA or protein product, to improve ALS survival. People with ALS have a progressive symptomatic course, and an understanding of the genetic burden requires an understanding of the phenotype. Survival modelling is important, as is an understanding of the way the disease progresses. Clinical staging is a method of achieving this.

In this thesis I explore the relationship between ALS disease progression, survival and genetics, developing and using various tools for measurement and assay of these parameters. I show that clinical staging using a simple system is feasible, usable by clinicians and other healthcare professionals, and matches with their expectations. I explore the use of such a system to analyse clinical trials, illustrating this with data from a clinical trial of Riluzole. I develop a pipeline for the analysis and interpretation of next generation sequence data and apply this to ALS whole genome sequences. I show that factors influencing survival in ALS include structural genomic variation, telomere length and rare genomic variants.

Table of contents

Table of Contents

Table of Tables

Table of Figures

Table of equations

Chapter 1: An introduction to ALS

1.1 Introduction

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease predominantly of motor neurons, characterized by progressive weakness of voluntary muscles and death from respiratory failure due to diaphragmatic paralysis, typically within 3 years of onset. Despite the poor prognosis, there is considerable variation in the survival rate, and up to 10% of people with ALS live more than 8 years from first symptoms[1]. There is a strong genetic contribution to ALS risk [2,3]. In 5% of cases or more, a family history of ALS or frontotemporal dementia is obtained, and the Mendelian genes responsible for ALS in such families have now been identified in about 70% of cases[4]. Even in apparently sporadic cases, twin and population studies show that the heritability is about 60% [5], and as well as the Mendelian variants found in familial ALS, at least five susceptibility genes have been identified so far in which variants increase the risk of ALS: *UNC13A, MOBP, SCFD1, C21orf2* and *C9orf72* [6–9]. Although risk genes reveal information about the mechanism of causation of ALS, it is also important to identify gene variants that modify survival. Survival genes could potentially be targeted directly, or their product or interactors augmented to improve ALS survival. A number of common gene variants associated with ALS survival have been identified through genome-wide association studies or other genome-wide approaches[10].

1.2Understanding disease progression in ALS

ALS presents in many different ways (Table 1), and it has been recognised for many years that the different clinical presentations correspond with differences in survival[11,12]. Bulbar palsy, in which dysarthria followed by swallowing difficulty is the main presentation, is associated with the worst prognosis, and flail arm or flail leg syndrome, in which there is symmetrical, predominantly flaccid weakness of the limbs, is associated with the best prognosis[13]. Perhaps surprisingly, statistical methods such as latent class cluster analysis can analyse the same data and identify different clinical subtypes that predict prognosis with far more discrimination than can neurologist classifications[13]. Most cases of ALS are focal in onset and relentlessly progressive, often to contiguous regions, although there are some exceptions[14]. The spread could be the result of a "prion-like" spread of toxic proteins through phagocytosis (consumption of cells by other cells) or possibly through a time-to-failure model[15,16]. Lower motor neuron failure is the main cause of weakness in ALS and can be measured non-invasively to provide data to assess cellular patterns of spread[17]. Understanding the mechanisms of spread will aid the development of novel therapeutics and may aid models of prognosis.

The diagnosis of ALS is clinical, with the support of electrophysiological studies and the exclusion of mimics. In some cases, early diagnosis can be challenging, particularly if weakness is confined to one region for some time or is confined to a subset of motor neurons (upper motor neurons only or lower motor neurons only).

ALS is classified for research purposes by the El Escorial criteria and their revisions, which improve homogeneity in recruitment for clinical trials and other clinical studies[18–21]. ALS progression is measured functionally using the ALS Functional Rating Scale – Revised, which

uses 12 questions scored between zero (no function) and four (full function) to generate a summary score[22]. The scale is widely used but has some limitations, since the subscores correlate more accurately with progression in different clinical subtypes[23].

Classifying feature	**Name of phenotype**	**Description**
Motor neuron involvement	Amyotrophic lateral sclerosis (ALS)	Mixture of upper and lower motor neuron signs on clinical examination. Degree of certainty of diagnosis based on El Escorial criteria. May involve up to all regions.
	Primary lateral sclerosis or upper motor neuron predominant ALS	Clinical signs limited to upper motor neuron features. Generally slowly progressive but involving up to all regions. This phenotype is usually confirmed if there have been no lower motor neuron signs after 4 years.
	Progressive muscular atrophy or lower motor neuron predominant ALS	Clinical signs limited to lower motor neuron features. Slightly slower progression but can involve all regions. This phenotype is usually confirmed if there have been no upper motor neuron signs after 4 years.
Site of onset	Bulbar onset	Site of onset may be included in the description of ALS, as different disease onset patterns have different rates of progression. The two categories are bulbar and spinal.
	Spinal onset	
Disease focality	Progressive bulbar palsy	Condition involving the bulbar region and predominantly lower motor neurons. May progress to other regions.
	Pseudobulbar palsy	Condition involving the bulbar region and predominantly upper motor neurons. May progress to other regions.
	Flail arm	Predominantly lower motor neuron proximal symmetrical involvement in the upper limbs. Some upper motor neuron signs may be seen in the lower limbs.
	Flail leg	Lower motor neuron distal symmetrical involvement restricted to the lower limbs. May affect one side only.
Cognitive involvement	ALS with cognitive impairment	ALS with some cognitive involvement below the threshold criteria for frontotemporal dementia.
	ALS with frontotemporal dementia (ALS-FTD)	ALS with frank frontotemporal dementia.

Table 1 Clinical presentations of amyotrophic lateral sclerosis

Disease staging allows a simple description of the extent of physical or functional involvement in an affected person and can guide management. Such systems have been in widespread use in cancer for years. In ALS, two recent staging systems have been proposed: King's clinical staging and Milano-Torino staging (MiToS)[24,25]. The King's system is similar to cancer staging in that the clinical spread of disease is used to infer the extent of disease progression[26]. Spread is defined as involvement producing signs or symptoms in the El Escorial domains (one domain is stage 1, two domains is stage 2, and three domains is stage 3), with respiratory or nutritional failure characterising stage 4. The ALS functional rating scale can be used to estimate the King's stage with 92% correlation[27]. MiToS uses the ALS functional rating scale subscores to define functional stage[25]. Each system has benefits in describing ALS stage succinctly. The two disease staging systems are complementary[26]. King's staging summarises the clinical or anatomical spread of disease. Mapping disease progression to clinical stage rather than survival could be used as a secondary endpoint in clinical trials, which would shorten trial durations and provide meaningful information on which stage of the disease is prolonged by an effective therapy[26]. MiToS summarises the functional burden of disease. It would therefore be useful in showing a functional benefit in clinical trials. Comparison of the systems shows that functional stage lags behind clinical stage, reflecting the functional reserve available in an affected limb, and it has been proposed that a combined stage is used, as is standard in cancer, along the lines of K3M2, which would mean King's stage 3, MiToS stage 2[28].

It is now recognised that ALS involves non-motor systems[29,30]. Between 30 and 50% of people have cognitive impairment detectable on formal testing, resulting from involvement of the frontotemporal circuits[31,32]. Frank frontotemporal dementia occurs in about 5%, and in some families, people may have ALS, frontotemporal dementia, or both[31,33–35]. The clinical impact of frontotemporal impairment in ALS is now more easily recognised because of recent

advances in the tools available to detect it, such as the Edinburgh cognitive assessment score (ECAS)[36–38]. Other neurodegenerative diseases have also been linked to ALS, including spinocerebellar ataxia, in which case studies have reported the co-occurrence of ALS and cerebellar degeneration[39]. Schizophrenia may be more frequent in families with ALS, and there may also be an increased frequency of multiple sclerosis[40,41]. In many of these cases, genetic factors are responsible for some of the risk. For example, pathological expansion of a repeat sequence in the *C9orf72* gene is associated with ALS, frontotemporal dementia, or both, and the same mutation may increase the risk of schizophrenia, Parkinson's disease, and multiple sclerosis[42]. Expansion of a repeat sequence in the *ATXN2* gene is associated with ALS or, if more than 30 repeats are involved, with spinocerebellar ataxia[43]. Autonomic, skin, and eye movement changes are also seen. Thus, ALS is a neurodegenerative disease in which the brunt falls on the motor system, but, as for many other neurodegenerations, the clinical syndrome is also dispersed through other anatomical and physiological systems.

Respiratory impairment is usually an end-stage event in ALS. Despite this, because respiratory function is difficult to measure reliably with non-invasive methods, measurement of respiratory function is generally used as a guide to the use of respiratory support rather than prognostication[44,45]. There have been many attempts at prognostic modelling, using either clinical features alone or biological markers such as albumin, creatinine, or neurofilament levels[46,47]. Most studies find that longer survival is associated with younger age at symptom onset, presentation with limb dysfunction rather than swallowing or speech disturbance, and specific forms of ALS such as symmetrical patterns (e.g. flail arm syndrome) or upper motor neuron predominant forms. Conversely, cognitive impairment comprising executive dysfunction, rapid weight loss, and respiratory involvement at first examination, although not necessarily respiratory onset, predict a poor prognosis[48–54]. The best predictor of slow

progression, however, appears to be a long interval between symptom onset and diagnosis, probably because this reflects the rate of disease progression overall[55]. Genetic variations have been associated with survival duration, with the best studied being variation in the *UNC13A* gene, which also affects risk[56–58]. Variation in the *CAMTA1* gene has also been associated with survival[10]. Furthermore, some risk genes harbour specific variants that are themselves predictors of prognosis, while other variants in the same gene do not have that effect. For example, the p.Asp91Ala variation of the *SOD1* gene is associated with very slow progression, while the p.Ala5Val variant is associated with aggressive disease[59,60]. Statistical models can be used to provide clinically useful information for patients, the strongest message being that survival is extremely unreliably predicted in individuals, even though patterns can be seen in the data[61–63].

1.3 An introduction to genomic structural variation

Over the past two decades, a great deal of research has been focused on understanding the genetic architecture of ALS. As a result of advances in sequencing technology and the substantial decrease in cost, gene discovery in ALS has increased exponentially – doubling every four years (Figure 1)[3,64]. Consequently, more than 130 ALS genes have been putatively associated with ALS[65], although only 25 genes show replicable association[66]. Genome-wide association studies (GWAS) have been a considerable success (Figure 2), giving a valuable insight into underlying ALS pathology by examining common variants. With growing evidence of the role of genetics in this complex disease's pathology, low frequency variants (minor allele frequency 0.5-5%) and rare variants have become the next target for study[66,67]. Rare and low frequency variants usually have large effect sizes which can help in the search for causal genes but they are generally not tagged by GWAS microarrays (microarray technology profiles genes according to gene expression/functional activity)[66,68]. As a result, the application of next generation sequencing, which includes whole genome sequencing and whole exome sequencing, have become powerful tools to understand the genetic aetiology of complex traits. Choosing the most suitable next-generation sequencing platform can be challenging but the study design can guide this choice. For example, whole exome sequencing only covers the coding region of the genome, and thus, the cost can be cut and a bigger sample or a higher depth is achievable[69]. As the cost of whole-genome sequencing goes down, it might become more feasible for multiple reasons including that it has more variant/allele calling power than other sequencing techniques[70,71]. It also has consistent read depth (current sequencing technologies use an approach that breaks the genome into random ~200bp sections for sequencing and then assembles them into the presumed original genomic sequence using a high performance computer; thus by chance some regions are "read" more frequently than others, and the average number of reads (read depth) is important), which provides the analytical tools

with enough statistical certainty to detect or differentiate genetic variations and sequencing error. Moreover, whole genome sequencing provides better distribution of sequencing position and, most importantly, the coverage of the non-coding area, where structural variants like copy number variants and single nucleotide variants are thought to play an important role in disease mechanism[71]. There are multiple challenges associated with choosing whole genome sequencing, such as the considerable sequencing cost; which includes gathering a substantial amount of generated data before and after the analysis. Therefore, the main challenge lies in choosing the appropriate approach for the analysis when dealing with these large amounts of sequence data and how to filter this data according to priority and pathogenicity. Whole genome sequencing is a cutting-edge technique where the best methods for analysis, both in data generation and statistical analysis, are still being discovered. After the genome wide association studies (GWAS) era, researchers use stringent *P*-value thresholds for significance during quality control and analysis and report a new gene only when it is validated in a case-control study. The ideal scenario is to validate study findings by using a different data-set or by testing another method.

As non-synonymous rare variants identified to date can explain just 15% of ALS, new study designs are needed to maximize gene discovery. Various sample selection methods are useful, including studying population isolates, familial ALS, founder populations, and extreme phenotyping by selecting the tails of a phenotype distribution[67,72]. These sampling methods are valuable methods for reducing genetic complexity. As an example, higher homogeneity is observed in population isolates compared with cosmopolitan populations. These sample selection techniques are confounded by multiple limitations, including the sample-size, as usually these data-sets are small, particularly in extreme phenotyping. Furthermore, the uniqueness of these data-sets in population isolates is connected to challenging factors such as geographical or cultural elements[72,73].

With every gene discovery, the next discovery becomes more difficult because the remaining genes are more likely to be rare or of small effect and were missed by previously used techniques. To overcome these obstacles, collaboration between geneticists, bioinformaticians, and evolutionary biologists is an effective strategy. Furthermore, data sharing between research centres is increasing. This has led to the large scale project, ProjectMine[74], an international whole genome sequencing initiative, aiming to sequence 22,500 individuals, 15,000 with ALS, as well as the recent discovery of 4 new genes (*KIFA5*, *NEK1*, *C21ORF2*, *MOBP*, *SCFD1*) through a large collaborative GWAS[75]. This study also found evidence that more ALS genes remain to be found, The amount of heritability explained by genetic information captured on genome-wide microarrays is about 12%, implying that the remainder is in rare variants and other types of genetic variation such as copy number variation, microsatellite repeats, post-transcription RNA editing, and epigenetic changes[76]. These are likely to be the next targets of ALS genetics research and are reliant on international research consortia.

Moreover, there is a growing need to replicate projects like Project MinE in different parts of the world as most of the large-scale projects have been done in populations with homogeneous ancestry or in populations of low diversity, which might be useful in finding the genes most strongly involved, but is probably not enough to tackle an entire disease. The large and disproportionate genetic study of populations of European descent (96%), leaves very little room for understanding the diversity of ALS genetic architecture worldwide[77].

It has been 18 years since sequencing of the first human genome was completed, and most of the science gathered since then shows that we understand a great deal about a small number of important genes and very little about the larger number of remaining genes. This shows the importance of understanding the function of the vast majority of the genome, especially when genotype-phenotype correlation is not clear-cut. By examination of the pattern of gene study methods of ALS researchers (Figure 3), we can see that only 10% of ALS publications are

focussed on gene characterisation and prioritization, with the least study of copy number variation (a form of structural genomic variation comprising variations in segmental DNA nucleotide arrangements) and gene functional annotation. Both of these research areas are slowly coming under more intense focus as more is understood about the impact of structural variation on both gene function and patient phenotype[78,79]. Of over 6370 publications focused on copy number variations, only 7 are ALS publications (PubMed August 2018) (Figure 3). This can be explained by the lack of confidence in available analytic tools, and the difficulties of structural variant analysis, as it is computationally intensive and time consuming.

Structural variants (including copy number variations) comprise various forms of genomic imbalance, including insertion, deletion, inversion, duplication and inter-chromosomal translocation[80–82]. Structural variants also represent a major difference between individuals in health and disease[83–85]. Measuring the intensity of signals derived from a genotyping array is the most used method in detecting structural variants, but advances in next generation sequencing technologies offer an alternative. Combined with bioinformatics analytical tools, structural shifts and changes are measured then compared between cases and controls[86–88]. Other techniques can be applied for detection of structural variation, including optical mapping and NanoString[89], which are new technologies that use molecular florescent tagging and microscopic imaging to detect genomic variations. This might be limited in large scale whole genome projects but useful for validation of structural variations in a region of interest[90].

1.3.1 The role of next generation sequencing and bioinformatics in structural variation analysis

Producing sequencing data, whether it is whole genome sequencing (WGS), whole exome sequencing (WES) or from targeted gene panels, is common practice for the study of the genetic

basis of biological processes. In biomedical research, next generation sequencing data are widely used to investigate the genetic causes of disease, allowing for the study of genomic variants including single nucleotide variant (SNV), copy number variation, structural variants such as insertions or deletions, duplications and inversions.

There are several practical challenges when processing next generation sequencing data. For example, 40x WGS data for one sample produced on the Illumina Hiseq 2000, one of the most popular sequencers, is about 400 Gb in its raw format (fastq format)[91]. This size can be reduced to approximately one fourth when the data is compressed, using lossless formats such as fastq.gz (gzip-compressed version of fastq) and bam[92]. Such big files are not easy to handle for the average non-specialised scientist or lab, since they require sophisticated tools, bioinformatics skills and high-performance computing for their analysis. Indeed, as an example, consider mapping one of these files, typically about 1 billion 150-base-pair long reads, to the human genome, a key process in the analysis of WGS data. Assuming that a standard midrange desktop computer with 4 CPUs and 16 Gb of RAM is used with The Burrow Wheeler Aligner (BWA)[93], probably the most widely used mapper, aligning this data to the human reference genome would take about 1 day, and this would only be the first step of an next generation sequencing data analysis pipeline. Faster mappers exist. For example SNAP[94] would only take about 4 hours to complete the same job, using the same number of CPUs, but it requires about 65Gb RAM, making it an unsuitable choice if large memory High-Performance-Computing (HPC) facilities are not available.

Many gene variants have been identified that drive the degeneration of motor neurons in ALS, increase susceptibility to the disease or influence the rate of progression[95]. The ALSoD webserver[65] lists more than 120 such genes and loci which have been associated with ALS, although only a subset of these have been convincingly shown to be ALS-associated[3],

demonstrating one of the challenges of dealing with genetic data – interpretation of findings. Next-generation sequencing provides the ability to sequence extended genomic regions or a whole-genome relatively cheaply and rapidly, making it a powerful technique to uncover the genetic architecture of ALS[66]. However, there remain significant challenges, including interpreting and prioritizing the found variants and setting up the appropriate analysis pipeline to cover the necessary spectrum of genetic factors, which includes expansions, repeats, insertions/deletions (indels), structural variants and point mutations. For those outside the immediate field of ALS genetics, a group that includes researchers, hospital staff, general practitioners, and increasingly, patients who have paid to have their genome sequenced privately, the interpretation of findings is particularly challenging. Although various tools are available to predict the pathogenicity of a protein-changing variant, they do not always agree, further compounding the problem. Furthermore, with the increasing availability of next-generation sequencing data, non-specialists, including health care professionals and patients, are obtaining their genomic information without a corresponding ability to analyse and interpret it as the relevance of novel or existing variants in ALS genes is not always apparent. The same would be true of structural variant analysis, and its interpretation too requires care related to sample and platform selection, quality control, statistical analysis, results prioritisation, and replication strategy. Choosing appropriate bioinformatics tools is crucial for structural variation analysis, and there is a need for fast and computationally powerful tools to address these demands.

In conclusion, the genetic aetiology of ALS is multifactorial, and a significant proportion of genetic variance remains unexplained. This hidden heritability may be harboured in structural genomic variation as well as rare variants that may be unique to an affected individual or family. Structural variants comprise different forms of genomic imbalance including copy

number variants, insertions, deletions, inversions, duplications and inter-chromosomal translocations, as well as repeat sequences and repeat expansions.

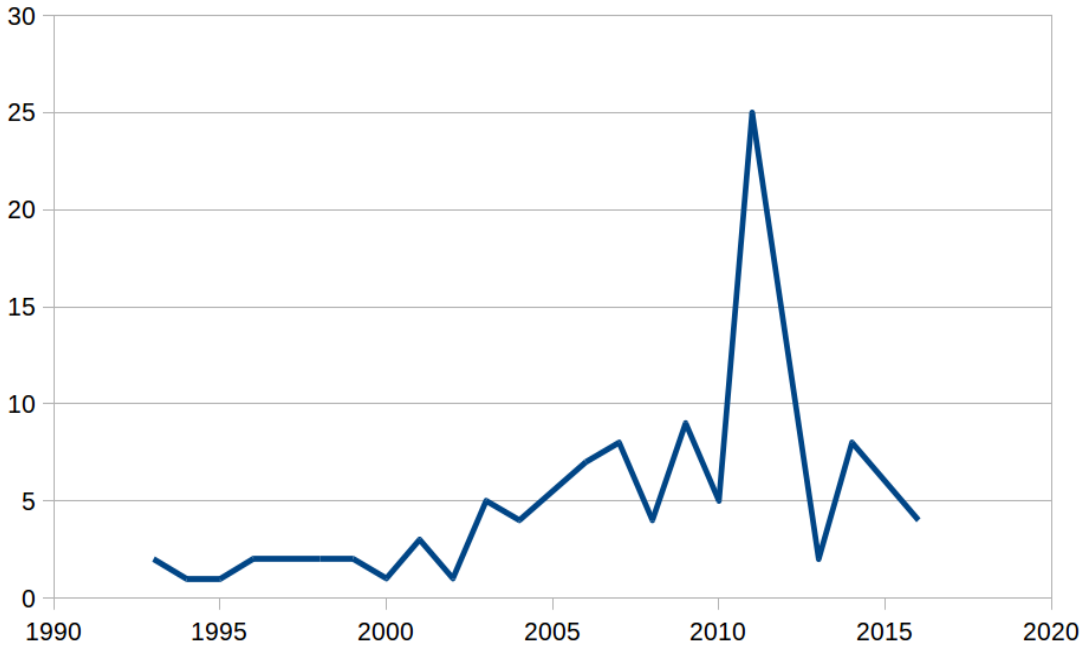


Figure 1 ALS gene discovery rate per year.

The data analysed in this chart is obtained from the ALS Online Genetics Database (alsod.iop.kcl.ac.uk August 2018).

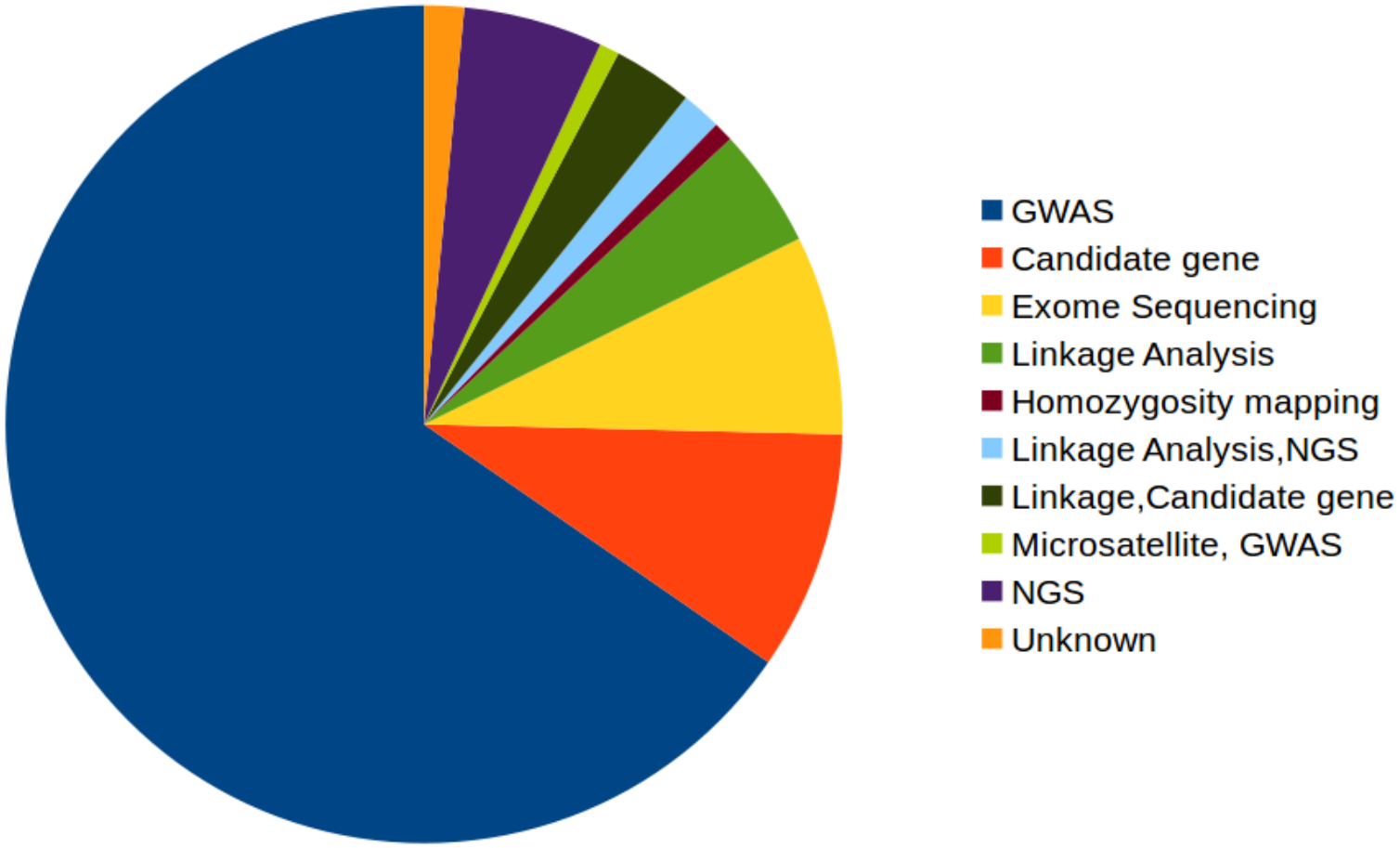


Figure 2 ALS gene studying methods.

Pie chart shows each ALS gene studying methods through analysis of the number of publications (PubMed August 2018). The search was done using the key words which generated 1630 hits out of total ALS 16559 publications.

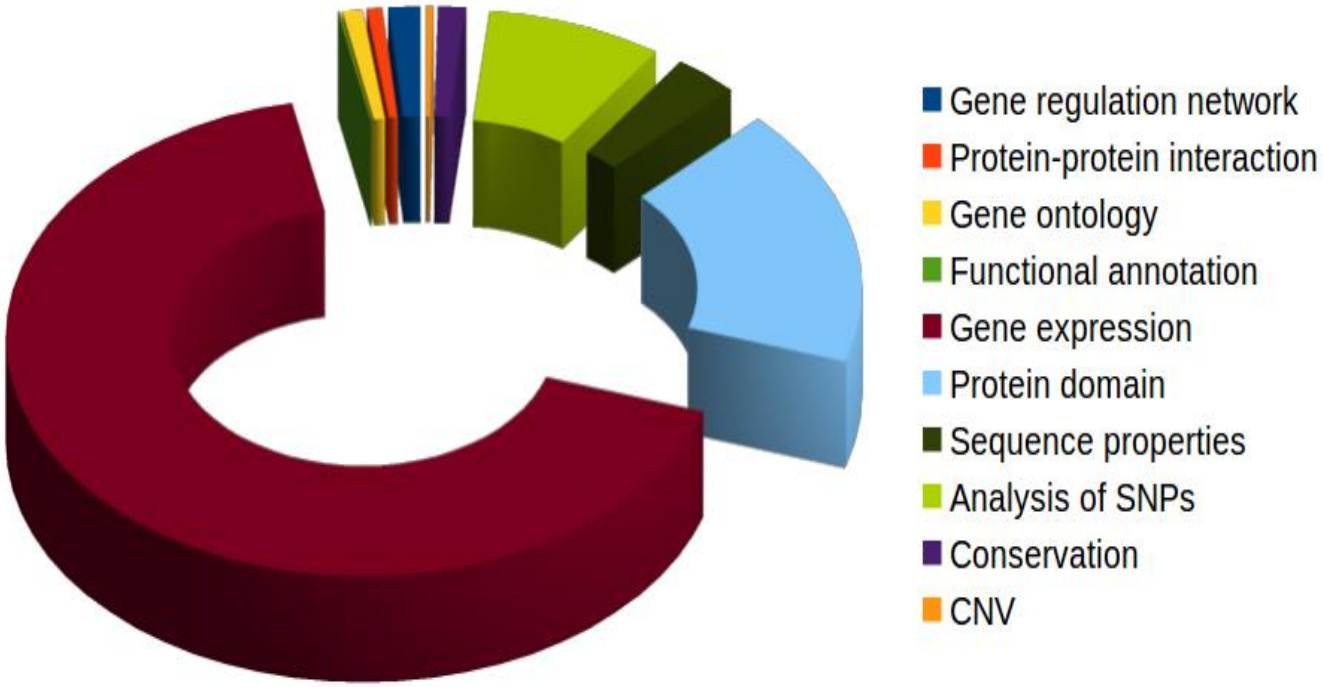


Figure 3 ALS gene study prioritizing methods.
Pie chart shows each ALS gene study prioritizing methods through analysis of the number of publications (PubMed August 2018). The search was done using the key words which generated 1630 hits out of total ALS 16559 publications

1.4 What causes ALS?

The causes of ALS are largely unknown. Significant advances have been made in understanding the genetic and environmental components of the disease. Here, I will explore what is known about why some people develop ALS and how the risk factors work together to cause disease.

1.4.1 Epidemiological studies of ALS

The incidence of ALS is about 2 per 100,000 person-years, and the prevalence is about 5 per 100,000 persons[96]. Because of the low prevalence, the average primary care physician will see one person in their lifetime, a typical UK neurologist will diagnose about 2 people a year, while a tertiary referral centre will see more than 100 people. Despite the low incidence, however, ALS is not particularly infrequent. The lifetime risk is about 1 in 300 by the age of 85, with the risk increasing steadily, at least until about the eighth decade of life. This is very similar to the risk for multiple sclerosis in the UK[97].

Tertiary referral centres see sufficient numbers of people that research studies may have adequate power for statistical analysis. However, there is a significant diagnostic delay in ALS, typically about a year, which seems to be independent of healthcare system and is probably related to low recognition by primary care physicians[98]. As a result, those attending specialist centres tend to be those with a better prognosis, who are younger, and who are more motivated[99,100]. In contrast, population-based registers capture all cases in a defined catchment population, regardless of attendance at a specialist clinic. Such registers have provided valuable insights into the epidemiology of ALS and offer an unbiased view of the condition[101].

ALS can affect people at any age. The mean age of onset is 56 in clinic registers but 70 in population-based registers[102,103]. In clinic registers, ALS is more frequent in men[104], with a male:female ratio of about 3:2, and the ratio becomes more equal with increasing age. In

population registers, although the male preponderance is still seen, the ratio may be closer to 1:1, an effect that can be attributed to the greater capture of older people with ALS[96].

1.4.2 Genetics and ALS

There are now more than 25 genes in which an association with ALS has been replicated, with the rate of gene discovery doubling every 4 years [3]. Although up to 10% of people may report a family history of ALS in a first-degree relative, detailed genealogical studies extending to more distant relatives and including related diagnoses suggest that more than 20% have a relevant family history. The genes responsible for familial ALS have now been identified for about 70% of all cases, but there is a significant genetic component, even in those without a family history[5,65]. Furthermore, statistical analysis shows that the distinction between familial and sporadic ALS is not clear-cut[105,106].

The most recent GWAS of ALS identified four new associations, three of which were successfully replicated[75]. An interesting feature of the study was that even though this was a study of people with apparently sporadic ALS, there were associations in genes previously identified from family-based studies – *C9orf72*, *TBK1*, and *NEK1* – further supporting the notion that familial and sporadic ALS are not mutually exclusive categories but rather a spectrum[7,75,107,108]. These three genes all harbour variants that are moderately penetrant. In other words, carrying a disease-associated variant does not mean ALS will inevitably follow. Current thinking is that common diseases are the consequence of the additive effects of small increases in risk from multiple common variations (polygenic), and rare diseases are the consequence of single gene variants that are themselves rare but have a large effect on the probability of disease (monogenic). For example, height and schizophrenia are polygenic traits, while Huntington's disease and Kennedy's disease are monogenic diseases. ALS sits

somewhere between these two extremes, with a lifetime prevalence that is far greater than is typical for a monogenic disease but far less than a common disease, and it is perhaps, therefore, to be expected that its genetic architecture also seems to sit somewhere between polygenic effects and monogenic high-penetrance disease.

There are three genes that have had a major impact on our understanding of ALS. ALS-linked dominant mutations in the superoxide dismutase gene *SOD1* were first identified in 1993, and since then mutations have been found in every exon of the gene[109]. The SOD1 protein is a free radical scavenger, and loss of function, increasing free radical damage in cells, is a logical hypothesis to consider. However, several well-characterised *SOD1* variants do not lead to a reduction in dismutase activity, and the evidence instead supports a toxic gain of function[110]. Transgenic *SOD1* mice develop a motor neuron degeneration and have been used to model the disease for treatment development[111]. The second important ALS gene is *TARDBP*, which codes for TDP-43, a protein regulating RNA expression and the major component of intracellular inclusions in ALS. The discovery of ALS-linked mutations in this gene was the first of many showing RNA processing defects to be important in ALS pathogenesis and, importantly, showed that the TDP-43 inclusions were not simply a passive marker of neuronal death but a crucial part of the disease pathway[112,113]. The third important genetic finding in ALS was of linkage[114,115] and then association[9,116,117] of a locus on chromosome 9, which led researchers to the identification of a massive expansion of a hexanucleotide repeat in intron 1 of the *C9orf72* gene[118,119]. This is the most frequent cause of ALS, being responsible for about 30% of familial and up to 10% of sporadic cases.

The focus of genetic research in ALS in the immediate future is therefore on rare variation. This is best discovered through high-throughput sequencing, and this technique has already identified several familial ALS genes. The major challenge facing researchers is how to

interpret the findings, since the identification of a rare variant in an ALS gene is not in itself strong evidence of relevance in that individual, and over-representation of rare variation in cases over controls in a particular gene does not provide sufficient information for genetic counselling on a specific variant[76].

1.4.3 Environmental risk factors

In contrast to genetics, environmental risk factors for ALS have been more difficult to identify. Such studies are expensive to perform, difficult to fund and are heavily reliant on recall[67]. As a result, they are susceptible to bias. Furthermore, unlike genome-wide analyses, in which a hypothesis-generating approach can be taken, it is not straightforward to assay all possible environmental factors, and so a selected subset of assumed risk factors is tested. Smoking has been associated with increased risk of ALS in some studies but may hold a higher risk in some subgroups[120]. Occupation, particularly military service with deployment, has been associated with risk of ALS, but the evidence mainly comes from the US, where there are large military datasets[4]. Physical activity is another widely studied risk factor, partly because of a number of high-profile sports players who have had ALS and because of people with ALS having a low BMI on presentation and having higher levels of leisure sports participation[121]. It is not clear whether having higher levels of physical activity raises the risk of ALS and, if it does, whether it is the activity itself or being genetically predisposed to high sporting prowess that is the mechanism[122–124]. Similarly, electric shock is not a risk factor in some analyses but is in others[125]. There is mixed evidence for the involvement of chemicals, such as heavy metals, ambient aromatic hydrocarbons, pesticides, and cyanotoxins[126–130]. Trauma, including head injury, also appears to be a risk factor in meta-analysis[131].

1.4.4 Inflammation and ALS

Evidence of an immuno-inflammatory component in ALS pathogenesis is compelling[132,133]. A pathological hallmark of the neuroinflammation is prominent microglia activation at involved sites. T-regulatory lymphocytes (Tregs) are important immunomodulatory cells that regulate the balance between activation and suppression of the immune response and control the microglia in the central nervous system: specifically, pushing them towards a state in which remodelling, and repair activities are activated. Defects in Treg levels or function have been found in ALS patients, becoming more frequent as the disease progresses. Treg levels are inversely correlated with disease severity, so that lower levels are seen in more severe disease, and survival is worse in those with Treg defects[134–136]. Studies are now underway to explore immune therapies that might improve Treg function and therefore improve survival[137].

1.4.5 Retroviruses and ALS

Poliovirus and other enteroviruses can cause a post-infectious myelitis with subsequent paralysis, and HIV infection can result in an ALS-like syndrome. Studies of serum and cerebrospinal fluid from ALS patients suggested that an activated endogenous retrovirus was associated with ALS[138]. Recently, the sequence has been identified as possibly HERV-K, an endogenous retrovirus that exists as an open reading frame in the human genome[139]. In mice, the *env* protein component of HERV-K is toxic to motor neurons. There is no evidence, yet that HERV-K is causative of the disease in humans, but studies are now underway to explore if antiretrovirals might slow progression and improve survival in ALS.

1.4.6 Conclusions

The apparently homogeneous phenotype of predominantly motor degeneration that is ALS can result from many different causes: genetic, epigenetic, environmental, and internal. Thus, many

different pathways converge on the final outcome of upper and lower motor neuron death. Careful analysis of incidence data in European population registers shows that, on average, each pathway comprises six molecular steps (Figure 4)[102,140]. The model explains many otherwise enigmatic features of ALS, such as the increasing risk with age, genetic pleiotropy (the same gene variation can result in different diseases), age-dependent penetrance of disease genes, the difficulty in identifying a single environmental cause, the observation that ALS appears to start in one region and spread, and that it is specific to motor neurons but can affect other cell types. The next challenge is to understand the extent to which the pathways overlap and therefore might be amenable to a common treatment strategy. Although ALS remains a uniformly fatal diagnosis, accelerating advances in our understanding bring the hope that an effective treatment can be found for this devastating disease.

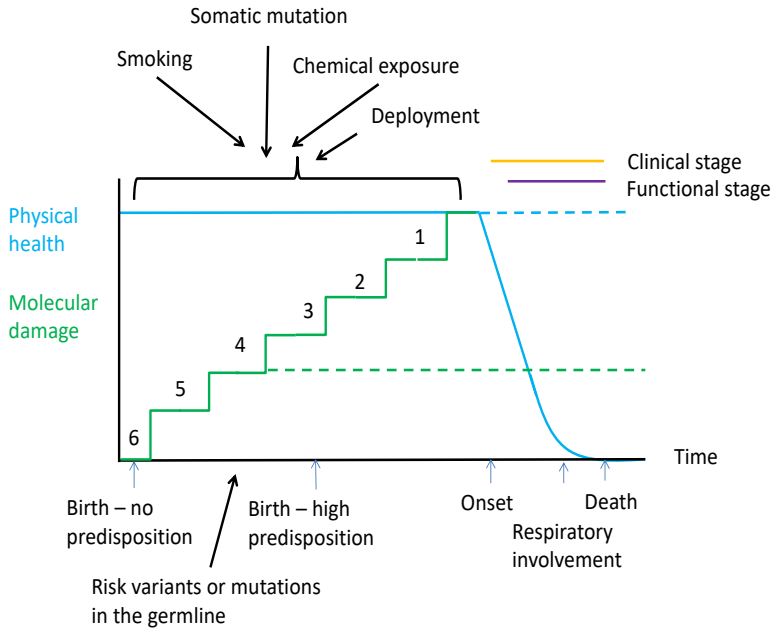


Figure 4. The time course of amyotrophic lateral sclerosis.

Time is represented along the x-axis; physical health and molecular damage are represented along the y-axis. With time, molecular damage increases in a step-wise way until it reaches a threshold, at which point physical health declines, representing disease onset. People with a family history of ALS may have a large genetic predisposition to ALS and so need fewer steps to reach the level of molecular damage that causes disease, corresponding to a younger age of onset. Lack of exposure to sufficient risk factors means that the disease does not manifest, even if a genetic cause is present, explaining reduced penetrance. There is not a 1:1 mapping of risk factors and steps, as the steps represent molecular hits that lead to cellular damage rather than actual exposures. Once physical symptoms have started, progression shows a log-linear decline until the onset of respiratory symptoms, where decline is exponential. Clinical and functional involvement can be measured by the King's clinical staging and Milano-Torino staging (MiToS) systems. A dotted line represents the hypothetical trajectory in an unaffected individual. Black arrows represent genetic and environmental risk factors. Numbers indicate remaining molecular hits until disease onset.

Chapter 2. Clinical staging in ALS

2.1 Introduction

In order to understand the role of genetic variation in disease progression, it is important to be able to measure disease progression; this can be done using staging[24,25]. The King's staging system consists of five disease stages, with Stage 5 being death, and Stages 1 to 4 based on the number of El Escorial central nervous system (CNS) regions involved in the disease, measured by weakness, wasting, spasticity, dysphagia or dysarthria, and the requirement for gastrostomy or non-invasive ventilation (NIV). The Milano-Torino (MiToS) Staging system comprises six stages, based on functional impairment as assessed by the revised ALS Functional Rating Scale (ALSFRS-R)[141].

2.1.1Aims:

To answer the following questions: How effective are existing clinical staging systems in determining ALS disease progression, and do systems for clinical staging classify disease progression in a way that matches clinical expectations?

2.1.2 Objectives

1. Determine clinical stage using both King's and MiToS systems in a set of clinical trial data.

2. Generate 17 clinical vignettes based on actual cases seen in the MND clinic and set up a Survey Monkey questionnaire for distribution to ALS specialists

3. Perform statistical tests to compare the clinical stages generated by the King's and MiToS systems with the expectations of ALS specialists.

2.1.3 Clinical Vignettes

We wrote 17 case vignettes of patients with ALS, representing different stages of the disease, with Stage 1 (3 vignettes), Stage 2 (7 vignettes), Stage 3 (3 vignettes) and Stage 4 (4 vignettes) represented. In 2016 and 2017 we ran two staging workshops during the European Network for the Cure of ALS (ENCALS) meetings.

The course was split into two tasks. In the first task, participants were asked to intuitively stage the clinical vignettes from stage one (early stage disease) to stage four (late stage disease). In the second task, the participants were asked to analyse the same vignettes using a taught clinical staging system. In all 17 clinical vignettes, the patients described had a diagnosis of ALS, with clinical examination and investigation consistent with the diagnosis. Participants were asked to provide information about their role, how long they had worked in the ALS field and to provide comments where appropriate if they wished to.

Participants included doctors, nurses, allied health care professionals and researchers with varying lengths of experience in ALS. We collected data on each participant's occupation and length of time working in ALS. During each workshop, we provided training in how to apply the staging SOP, and asked participants to stage the vignettes using the SOP.

In all these cases the patients described had a diagnosis of amyotrophic lateral sclerosis, with clinical examination and investigations consistent with this diagnosis. All the clinical vignettes are based on actual cases seen in MND clinic. The vignettes are available in supplementary material in the appendix.

2.2 A Standard Operating Procedure for King's staging

2.2.1 Introduction

In order for the King's staging system to be applied by different health care professionals and varying levels of experience working in ALS, as for example may be the case in a multicentre clinical trial, we designed a Standard Operating Procedure (SOP) for the use of the King's system. We then investigated whether it could be used by a variety of health care professionals.

2.2.2 Methods

2.2.2.1 Clinical Vignettes.

We wrote a SOP for using the King's clinical staging system for ALS (appendix supplementary material). All the 17 vignettes were included to in this study.

2.2.2.2 Statistical analysis

To measure the reliability of staging using the SOP across the entire cohort, we calculated a Spearman's Rank correlation coefficient between the actual King's clinical stage and the King's stage assessed according to the SOP. We also calculated Spearman's Rank correlation coefficients for different health care professional groups included in the study (doctors, nurses and allied health care professionals), and for those with less than 10 years, or 10 years or greater experience working in ALS.

As a further step to investigate the reliability of using the SOP to calculate clinical stage, we used the Bland-Altman method to calculate the difference between actual King's clinical stage and King's stage calculated by participants using the SOP, and the mean of the correct clinical

stage and stage using the SOP, determining the limits of agreement between these. To test for any systematic bias in the SOP, leading to over-estimation or under-estimation of stage, we calculated the Spearman's Rank correlation coefficient between the difference between actual King's clinical stage and stage using the SOP, and also the mean of the actual King's clinical stage and stage using the SOP.

Analyses were performed in SPSS v20.0 and GraphPad Prism v6.07.

2.2.3 Results

I helped plan the study, collated the data, performed the analyses, and wrote the paper. I am the first author on a paper in The Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration journal describing the work in this chapter.

The study consisted of 61 participants in total, with doctors (65.5%), nurses (9.1%) and other allied health care professionals (25.5%) represented in the cohort (Fig. 5A). There was an even distribution between those with less than 10 years' experience working in ALS (50.8%) and those with 10 years' or greater experience (48.2%) (Fig. 5B).

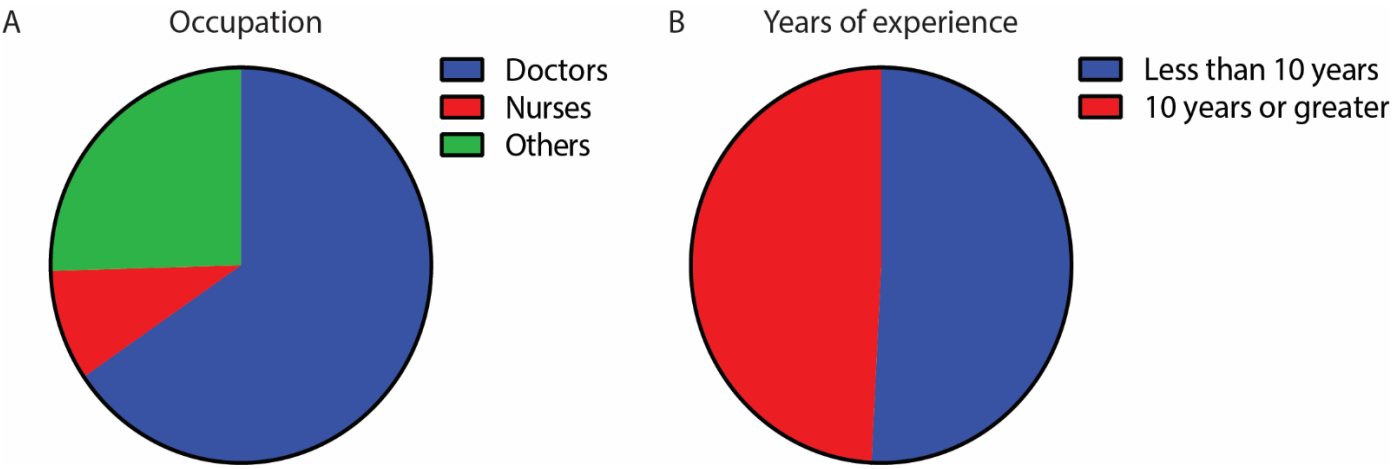


Figure 5. Demographics of participants included in the staging study

A) The study consisted of 61 participants in total, with doctors (65.5%), nurses (9.1%) and allied health care professionals (25.5%) represented in the cohort. B) There was an even

distribution between those with less than 10 years' experience working in ALS (50.8%) and those with 10 years' or greater experience (48.2%).

Use of a Standard Operating Procedure leads to reliable clinical staging

Across the cohort, we found that use of a SOP led to a high reliability of clinical staging of vignettes. The correlation between staging of the clinical vignettes using the SOP and the actual King's clinical stage was Spearman's Rho = 0.95, $p<0.001$ (Fig. 6). There was a very strong correlation between staging using the SOP and the actual King's clinical stage for every health care professional group, with similar correlations between each group: doctors (Spearman's Rho = 0.95, $p<0.001$), nurses (Spearman's Rho = 0.93, $p<0.001$) and allied health care professionals (Spearman's Rho = 0.94, $p<0.001$). The correlation between staging using the SOP and the actual King's clinical stage was the same for those with at least 10 years' experience working with patients with ALS (Spearman's Rho = 0.95, $p<0.001$) as for those with less than 10 years' experience (Spearman's Rho = 0.95, $p<0.001$). Overall, most participants correctly staged case vignettes in Stages 1 (95.9%), 2 (95.7%), 3 (84.6%) and 4 (98.9%) using the SOP (Fig. 7).

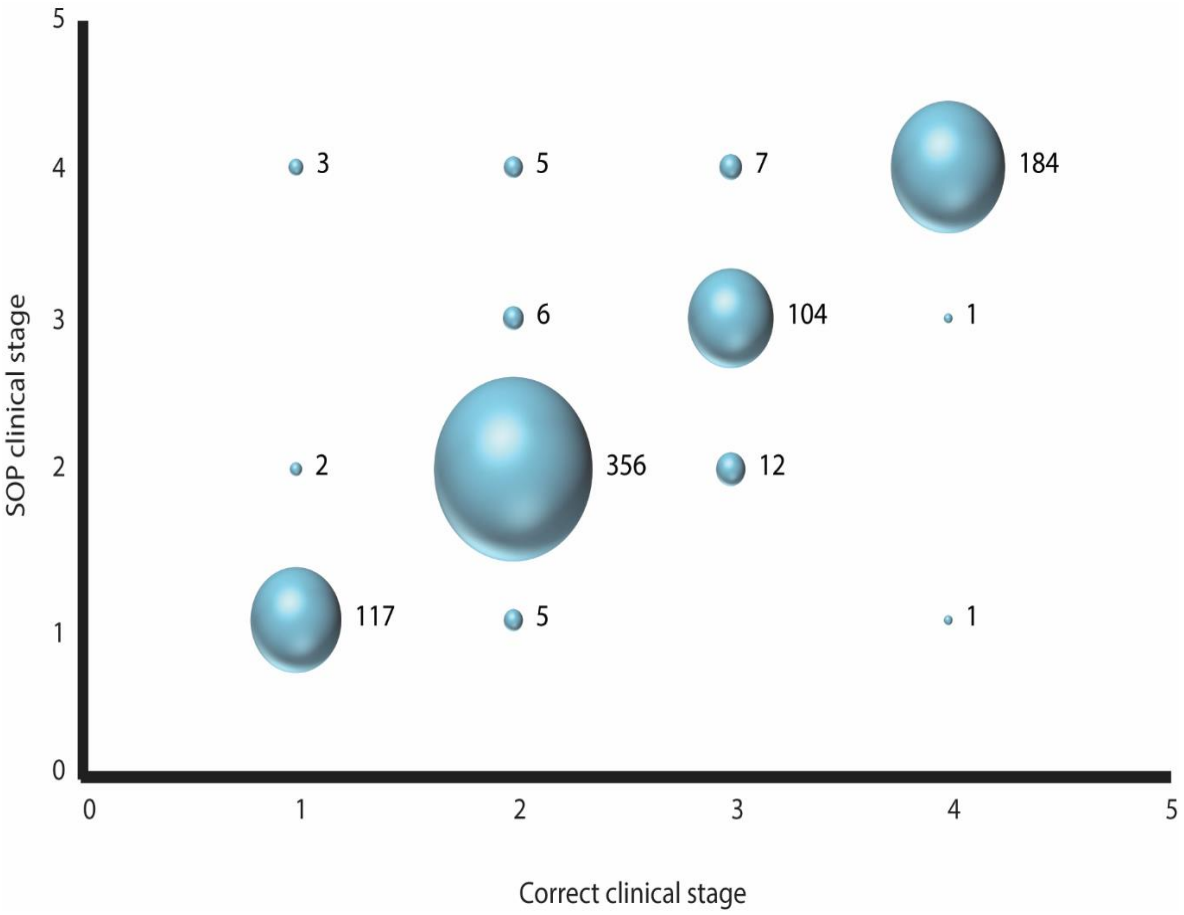


Figure 6 Use of a Standard Operating Procedure leads to reliable clinical staging

Participants were trained in how to use the staging Standard Operating Procedure (SOP, Supplementary Material). Most correctly staged case vignettes in Stages 1 (95.9%), 2 (95.7%), 3 (84.6%) and 4 (98.9%) when using the SOP. The numbers to the right of each bubble represent the number of answers within each group.

We used the Bland-Altman method to calculate the difference between correct clinical stage and stage using the SOP for all pairs and the mean of the stages for all pairs. The mean of the difference in stage between the two methods was 0.01 (95% CI of the mean -0.01 to 0.04, standard deviation 0.35) with the 95% confidence limits of agreement lying between -0.69 to 0.71. Therefore, the limits of agreement lie within a single stage, confirming that there is a clinically acceptable level of agreement between staging using the SOP and actual King's clinical stage. To test for systematic bias of the SOP over the range of stages, leading to over-staging or under-staging, we calculated a Spearman's Rank correlation coefficient between the

differences in stages and the means of stages, which showed a negligible relationship between the two (Spearman's Rho = 0.069), Therefore, there are unlikely to be any systematic biases in staging when using the SOP. As a further confirmation of this, there was a similar number of cases where staging using the SOP led to a higher stage than the actual clinical stage (23 cases) and cases where SOP staging led to a lower stage (20 cases), and these erroneously staged cases amounted to only 5.35% in total of the whole study.

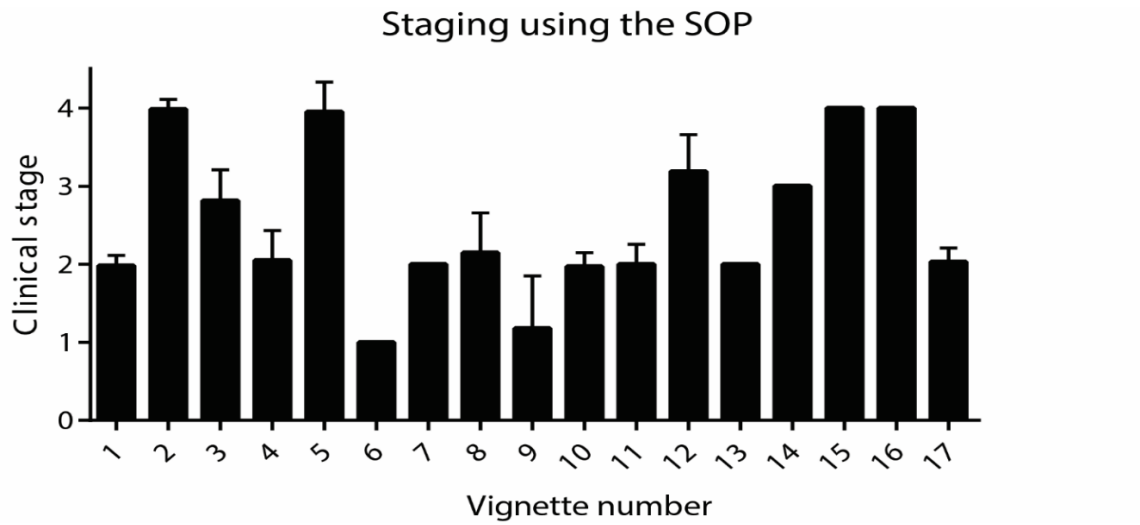


Figure 7 Variability of scores for staging by intuition or using the SOP for each case vignette.

Bars represent the mean and standard deviation of scores for staging using the SOP for each vignette. Variability in the answers was greatest for Vignettes 8, 9 and 12. Use of the Standard Operating Procedure for patients with Stage 4 disease.

Variability in the answers for staging using the SOP was greatest for Vignettes 8, 9 and 12 (Fig. 7). In vignette 9 a gastrostomy had been inserted for a reason other than ALS, due to oropharyngeal malignancy. Some participants had staged this case as Stage 4, despite the gastrostomy not being required as an intervention for ALS-related dysphagia. In vignette 9 a gastrostomy had been inserted for a reason other than ALS, due to oropharyngeal malignancy. Some participants had staged this case as Stage 4, despite the gastrostomy not being required as an intervention for ALS-related dysphagia. Although this is a rare scenario to see in the clinic, it is important to clearly differentiate between clinical intervention because of ALS

complications and other intervention related to other diseases such as the example used in vignette 9. The SOP clarifies that when the gastrostomy is required as an intervention for ALS-related dysphagia, Stage 4 is reached and has clear parameters for when these criteria are fulfilled. In vignettes 8 and 12 the patients did not yet meet the criteria for respiratory failure, however some participants had staged these cases as Stage 4. The SOP clarifies that Stage 4 is only reached when the UK National Institute of Health and Care Excellence guidelines for respiratory failure in ALS are reached, and the guidelines are stated in a summarised format within the SOP, although relaxed, country-specific guidelines are also acceptable.

2.2.4 Discussion

We have shown that using an SOP provides a highly reliable method of calculating clinical stages in patients with ALS. The 95% confidence limits of agreement of staging using the SOP were within less than a clinical stage and there were no clear systematic biases leading to over- or under-staging, further confirming that using the SOP for clinical staging is reliable.

We have demonstrated that using an SOP is effective across different groups of health care professionals and across different levels of experience working within ALS. ALS staging systems have utility in stratifying patients in clinical studies, as a marker of disease progression when assessing validity of biomarkers, and as outcome measures in clinical trials. For their use to be effective, it is critical the system is simple to understand and apply. This is of clear importance in ensuring staging can be applied reliably by different individuals, separated in space and time, as is the case for example in multicentre clinical trials. Participants were attendees at an ALS specialist meeting, therefore are likely to be representative of the individuals that would use the SOP and staging system in the future in clinical studies and

clinical trials. The impact of this study is that the SOP can now be implemented widely in future studies where it would be useful to prospectively collect staging data.

We found that the vignettes that were most variable were sometimes over-estimated as Stage 4. We have clarified in the SOP the exact definitions of when Stage 4 is reached, either when gastrostomy or NIV are required. It is likely that as this SOP is used more frequently alongside further training sessions, its repeated use will improve the reliability of clinical staging further. Some flexibility in the definitions of Stage 4 may however be necessary in implementations in different health systems.

Although the agreement between staging using the SOP and actual King's clinical stage was high, it was not perfect. Although 84.5 % of participants correctly staged stage 3, there appears to be a decrease in the ability to define stage 3, compared to the other stages. This might be explained that in order to identify the 3rd stage, three neurological signs have to be identified. This might be not straightforward as stage one or four. However, we found that the limits of agreement lay within a single stage, with systematic biases being negligible, and we detected a very low rate of errors (~5%) represented equally by cases of over- and under-staging, indicating that the extent of agreement is clinically acceptable. Overall, this suggests that the SOP is straightforward to understand as written.

A potential limitation of this study is that staging using the SOP was performed directly after a training session was delivered, and the accuracy of staging may have therefore been at its highest at this time. However, we would expect these reliability measures to continue to improve after repeated use of the SOP on subsequent occasions, and the SOP is now readily available for users to refer to in the future as required. A further limitation is that this cohort

represents a relatively small number of participants with an overrepresentation of doctors compared to nurses and other allied health care professionals. Further studies are therefore required to validate whether the SOP is reliable across a larger and more diverse group of health care professionals. Future validation could be achieved using an online survey platform.

2.2.5 Conclusion

We have demonstrated that the staging SOP provides a reliable method of calculating the clinical stage of a person with ALS and can be used prospectively by a range of health care professionals with different levels of experience.

2.3 Investigation of the relationship between King's clinical ALS stage and ALS stage as intuitively assigned by health care professionals.

2.3.1 Objective

As shown above, clinical stage in ALS can be assigned using King's staging with a simple protocol based on the number of CNS regions involved and the presence of significant nutritional or respiratory failure. It is important that the assigned clinical stage matches expectations, and generally corresponds with how a health care professional would intuitively stage the patient. We therefore investigated the relationship between King's clinical ALS stage and ALS stage as intuitively assigned by health care professionals.

2.3.2 Introduction

Disease duration in ALS varies greatly between individuals[54,62,142,143], even though median time to death is just 40 to 44 months from onset[24,29]. Disease progression rates therefore range from very aggressive, rapid disease to very slowly progressive disease [28,144–146]. Recently, we used King's staging to estimate the clinical stage at which Riluzole, a treatment for ALS, showed greatest effect[144]. The King's ALS stage corresponds well to anatomical clinical spread, and in this regard differs from the MiToS staging system, which corresponds more to the spread of functional impairment experienced by the patient[26].

It is important that clinical staging matches expectations and corresponds with how a health care professional would intuitively stage a patient based on their clinical experience with the disease, so that the meaning of any given stage can be easily interpreted. We therefore investigated the relationship between King's clinical ALS stage and ALS stage as intuitively assigned by health care professionals.

2.3.3 Methods

2.3.3.1 Clinical vignettes

We have used 10 case vignettes of patients with ALS, representing a spectrum of cases with different stages of disease, ranging from Stage 1 to 4, representing as a spectrum of cases with upper and lower limb involvement, upper and lower motor neuron signs and symptoms, bulbar and spinal site of onset, and cases requiring respiratory support and gastrostomy intervention. In all vignettes, the patients described had a diagnosis of ALS, with clinical examination and investigation consistent with the diagnosis.

2.3.3.2 Participants

Participants were asked to intuitively stage the clinical vignettes from Stage 1 (early stage disease) to Stage 4 (late stage disease). They were also asked to provide information about their role, how long they had worked in ALS and to provide comments where appropriate if they wished to. Sixty-one participants were classified into three groups in accordance with their role (doctors, nurses, other health care professionals/researchers). Participants were additionally classified based on their work experience into two groups, those with more than ten years' experience, and those with less than ten years' work experience. All the participants agreed to take part of this study. All the information was anonymized, kept confidential, and digitally stored behind a password.

2.3.3.3 Statistical analysis

To measure the reliability of intuitive staging across the entire cohort, we calculated a Spearman's Rank correlation coefficient between the actual King's clinical stage represented by each vignette and stage assessed intuitively. We also calculated Spearman's Rank

correlation coefficients for the different types of health care professional included in the study (doctors, nurses and allied health care professionals), and for those with less than 10 years, or 10 years or greater experience working in ALS. Analyses were performed in SPSS v20.0 and GraphPad Prism v6.07.

2.3.4 Results

I helped plan the study, collated the data, performed the analyses and wrote the paper. I am a co-first author on a paper to be submitted describing this work.

There was a good correlation between stages assigned using King's clinical staging and stages assigned intuitively by the course participants (Spearman's Rho = 0.64, $p<0.001$ and Fig. 8). This was true for every health care professional group: doctors (Spearman's Rho = 0.68, $p<0.001$), nurses (Spearman's Rho = 0.59, $p<0.001$) and allied health care professionals (Spearman's Rho = 0.58, $p<0.001$). Overall, the majority of participants declared the same stage intuitively as the actual King's stage for case vignettes representing King's Stages 1 and 2 (73.0% and 53.0% respectively), but a minority of participants did for case vignettes representing King's Stages 3 and 4 (41.1% and 48.9% respectively and figure 9). In all cases, the most frequent alternative answer for intuitive staging was only one stage different from the actual King's stage (Figure 8)

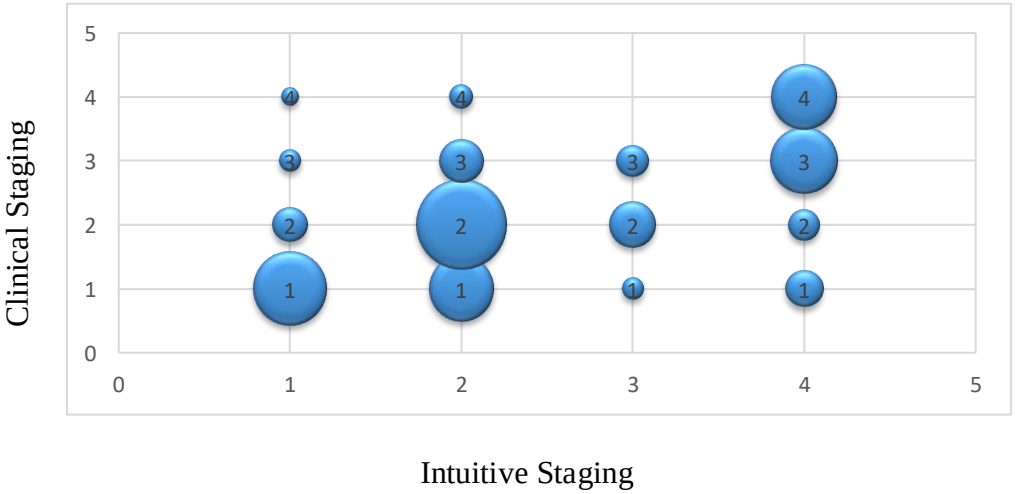


Figure 8. Health care professionals can intuitively stage patients with ALS. Participants were asked to score each vignette intuitively from Stage 1 (early stage disease) to Stage 4 (late stage disease). The

majority of participants correctly staged case vignettes in Stages 1 and 2 with intuition (73.0% and 53.0% respectively), however the minority of participants correctly staged case vignettes in Stages 3 and 4 (41.1% and 48.9% respectively). In all cases, the most frequent alternative selected was an adjacent stage. The numbers to the right of each bubble represent the number of answers within each group.

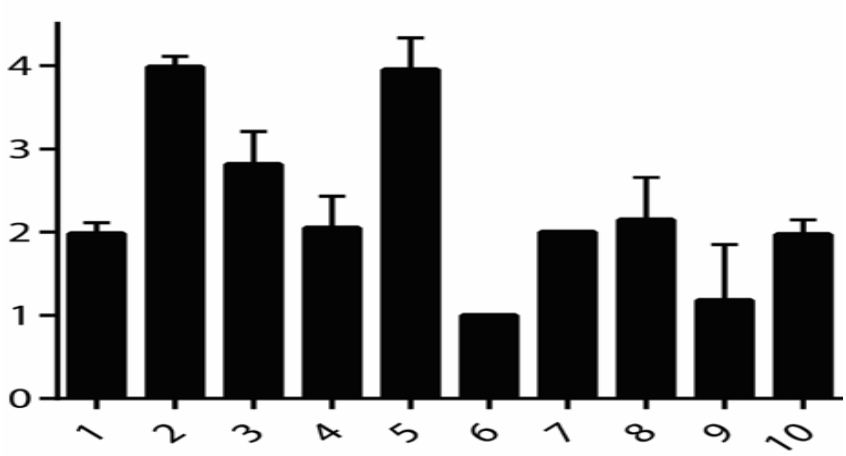


Figure 9.Variability of scores for staging by intuition for each case vignette.

Bars represent the mean and standard deviation of scores for staging using the clinical staging for each vignette. Variability in the answers was greatest for Vignettes 5, 8, and 9.

2.3.5 Discussion

Disease staging allows a simple description of the extent of disease progression. Staging systems have been used in cancer to guide patient management and treatment. In ALS, two recent staging systems have been proposed, King's clinical staging and Milano-Torino staging[24,25]. The King's system is similar to cancer staging in mapping clinical spread with disease progression. Both ALS staging systems are complimentary as we previously shown[26]. We have demonstrated that the King's clinical staging system correlates well with clinical stage assigned intuitively by health care professionals. It is important that clinical staging matches expectations and corresponds with how a health care professional would intuitively stage a patient based on their clinical experience with the disease, so that the meaning of any given stage can be easily interpreted.

We have shown that staging using the King's system was straightforward with minimal confusion between clinical staging and intuitive staging, as all the participants behaved similarly. King's system provides a clear system that all members of the multidisciplinary team can understand by supporting evidence-based actions that is prepared for a multidisciplinary team intervention if there is a significant change.

A weakness of this study is the limited number of participants in each subgroup. Although the agreement between actual King's clinical stage and intuitive staging was high, it was not perfect. However, we found that the limits of agreement lay within a single stage, with systematic biases being negligible, and we detected a very low rate of errors (~5%) represented equally by cases of over- and under-staging, indicating that the extent of agreement is clinically acceptable. Furthermore, stage two was overrepresented in the clinical vignettes but we think we managed to cover multiple clinical scenarios that might be seen in MND clinic.

We have shown that King's clinical ALS staging system corresponds with professional expectations, an important quality for a staging method.

2.3.6 Conclusion

In conclusion across a spectrum of ALS scenarios, King's clinical ALS stage corresponds to intuitive ALS stage as assigned by a range of health care professionals

2.4 Identification of the benefits and disadvantages of the King's and MiToS ALS staging systems and possible clinical uses

2.4.1 Introduction

Although the Milano-Torino (MiToS) functional staging and King's clinical staging systems both measure stage and show content validity, mapping correctly to disease progression, it is not clear to what extent they are collinear and therefore redundant. We therefore set out to compare the systems using data from a phase 3 randomised double-blind placebo-controlled trial of lithium carbonate in ALS (LiCALS) (EudraCT number 2008-006891-31)[147] , in which ALSFRS-R scores were recorded at 3 monthly intervals.

2.4.2 Methods

2.4.2.1 Patients

Anonymised data from the LiCALS clinical trial was reanalysed. Data consisted of ALSFRS-R scores, site of disease onset (bulbar or limb), gastrostomy timing, measures of respiratory function, and timing of non-invasive ventilation, recorded every three months during an eighteen-month trial enrolment. For all patients, date of death or last follow-up were also recorded.

2.4.2.2 Clinical staging systems

ALS clinical stage comparisons were undertaken using two staging systems; King's clinical staging and MiToS functional staging. As stages were not previously recorded during the LiCALS clinical trial, stages for both systems were determined retrospectively and derived from historical data, as previously described[24,25] (Figure 10). For simplicity, we encoded King's stages with prefix K, and MiToS stages with prefix M, so for example, K2M3 would represent King's stage 2 and MiToS stage 3.

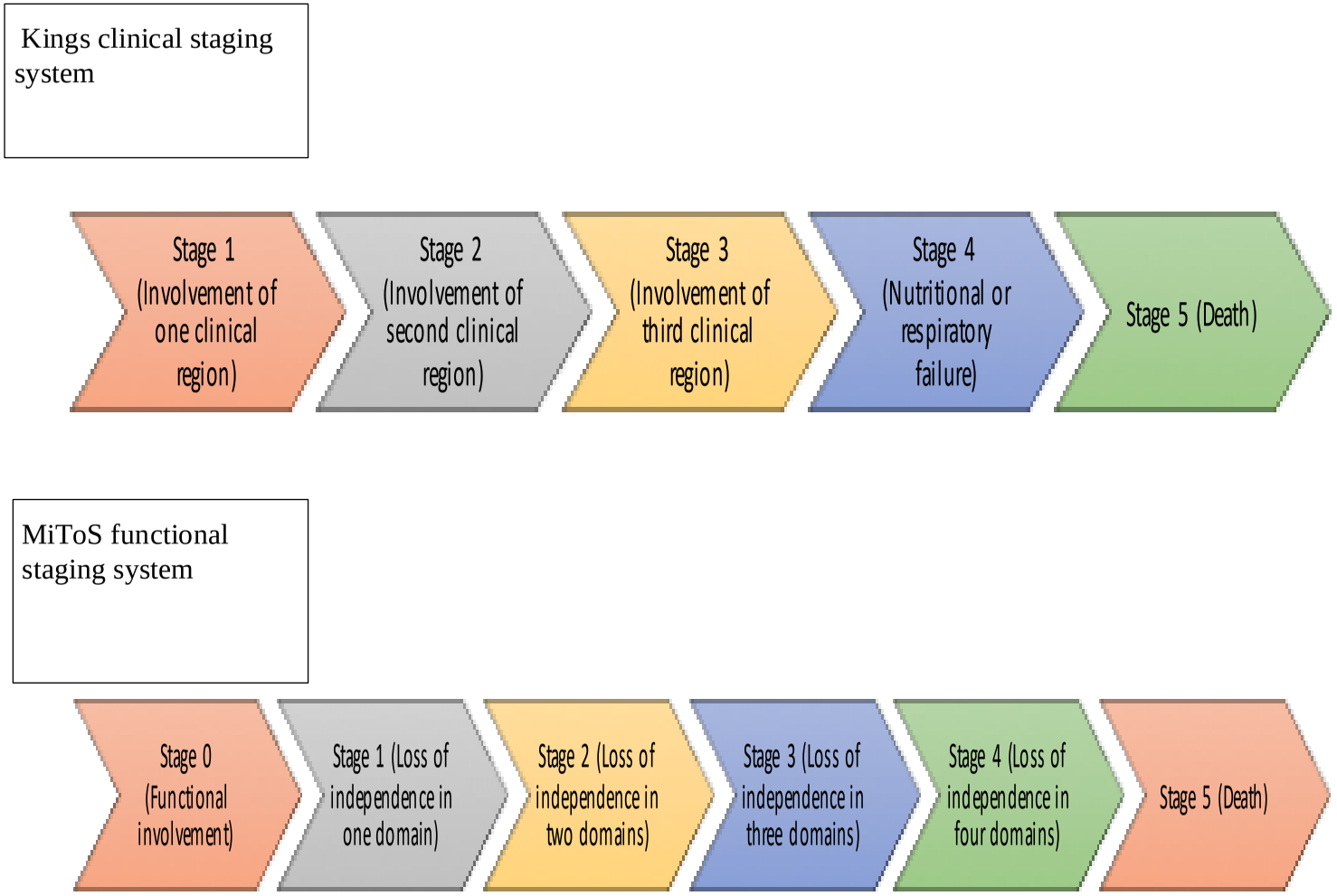


Figure 10**:** Flowchart of ALS staging systems (King's staging and MiToS staging

2.4.2.3 Statistical analysis

A Kaplan-Meier survival analysis and log-rank test were used to test differences in survival from disease onset for each categorical variable, site of onset (bulbar or spinal), family history (sporadic ALS or familial ALS), gender, and age of onset in 10-year categories. We also tested the proportion of patients dead or alive on 30 June 2011 (defined as the censor date) using a chi square test.

Standardised median times for reaching clinical stages were calculated as a proportion of time elapsed from onset to each disease stage across the duration of the disease for both King's and MiToS systems, with 0 representing symptom onset, and 1, death, using only data from deceased patients, as previously described[24].

Spearman's coefficient was used to test overall correlation between the two systems. Pairwise comparisons between the number of patients in each King's and MiToS stage were used to test the relationship between specific stages using a chi-square test. Standardised residuals were used to test which items were most responsible for any associations observed.

All statistical tests were carried out using SPSS v22.0 (SPSS Inc, Illinois, USA).

2.4.2.4 Ethics

The LiCALS clinical study was ethically approved by the South East London Research Ethics Committee reference 09/H1102/15. All participants involved, provided written consent. This current study does not require ethical approval due to analysis being conducted on fully anonymised pre-existing clinical trial data.

2.4.3 Results

2.4.3.1 Patient characteristics

I helped plan the study, collated the data, performed the analyses and wrote the paper. I am a co-first author on a paper to be submitted describing this work.
Data was available for 217 patients, of whom 95 had died by the end of the study. Patient characteristics are shown in Table 2. Median survival was 43.6 months, which is similar to that found in a previous study, 42.3 months[24]. There were no significant differences in survival by gender, family history or site of onset, and no differences were seen in the proportion still alive by the study end date. Older age at disease onset was associated with worse survival *p=0.01*, and consistent with this observation, the proportion of deaths compared with censored observations progressively increased as patients were classified into higher age groups: 56% of patients in the 75-84 year age group had died by the end of the trial, compared with only 14% for the 25-34 year age group.

	***n* (%)**		**Median time to death or last observation in months (95% CI)**		***p*-value**	**Death at last observation (%)**		***p*-value**
Gender					$p = 0.19$			$p = 0.13$
Male	151	(70)	47.8	(39.0-56.6)		61	(40)	
Female	66	(30)	37.9	(32.0-43.9)		34	(52)	
Site of Onset					$p = 0.24$			$p = 0.60$
Limb	170	(78)	40.1	(32.8-47.3)		76	(45)	
Bulbar	47	(22)	47.8	(-)		19	(40)	
Type					$p = 0.91$			$p = 0.60$
Sporadic	211	(97)	43.6	(36.6-50.5)		93	(44)	
Familial	6	(3)	-	(-)		2	(33)	
Age					***p* = 0.01**			***p* = 0.02**
25-34	7	(3)	-	(-)		1	(14)	
35-44	16	(7)	-	(-)		4	(25)	
45-54	58	(27)	-	(-)		18	(31)	
55-64	75	(35)	34.4	(29.0-39.7)		41	(55)	
65-74	52	(24)	37.9	(32.3-42.9)		26	(50)	
75-84	9	(4)	32.2	(20.0-44.3)		5	(56)	
Total	217	(100)	43.6	(36.6-50.5)		95	(44)	

Censor date was 30/06/2011

Table 2 Characteristics of LiCALS patients.

Median time to death or last observation and percentage of death at last observation compared using different categories (gender, site of onset, family history and age of onset).

2.4.3.2 Standardised median time

95 patients had died by the end of the study. The standardised median proportion of time elapsed from onset to each King's stage is shown in Table 3a and Figure 11a. Corresponding values for MiToS stages are shown in Table 3b and Figure 11b, showing a wider distribution of King's stages through the early and middle disease course, compared with a tendency for MiToS stages to be distributed later in the disease course.

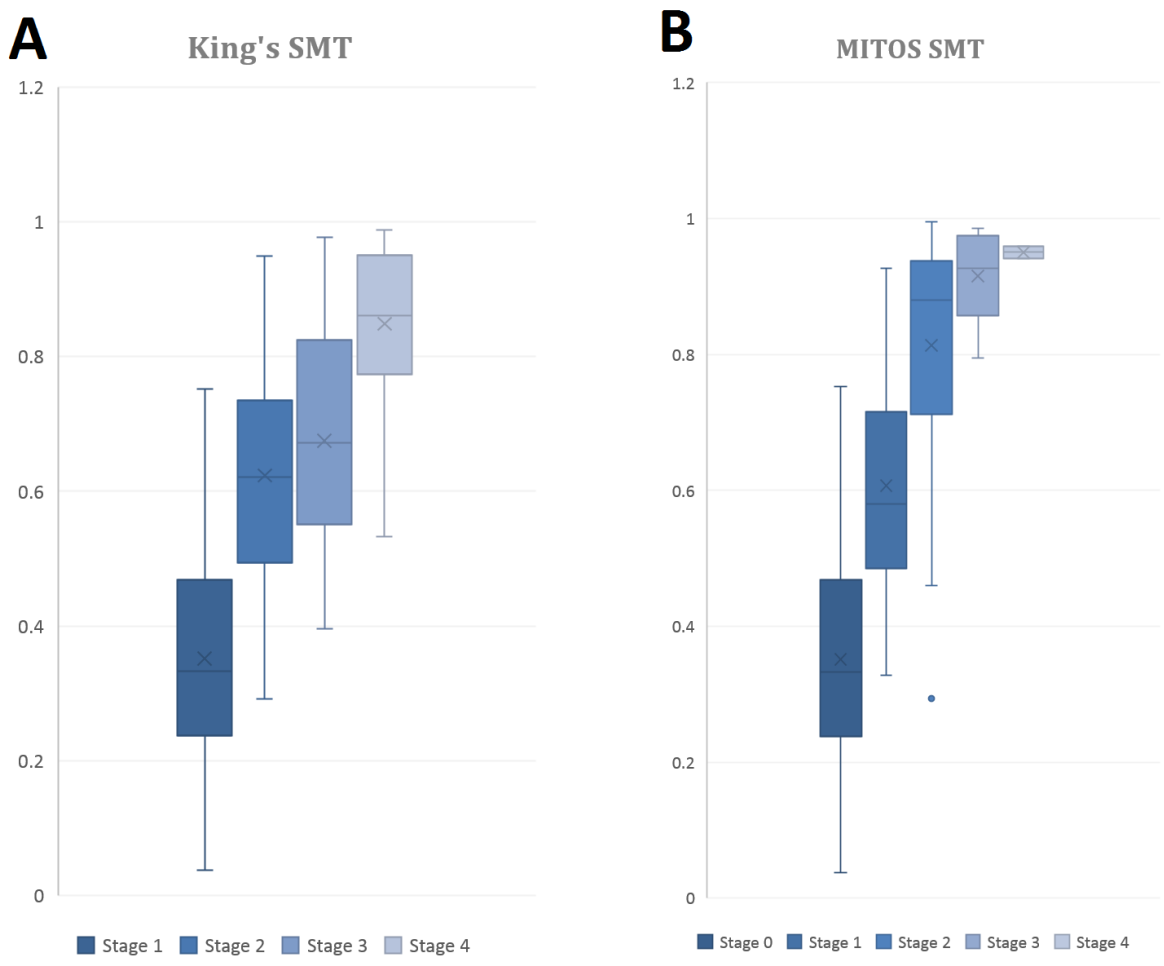


Figure 11 Box plot for Standardised Median Time (SMT) from onset to each disease stage.
(A) SMT for King's stages. (B) SMT for MiToS stages.

A)

King's staging system (n)	Median number of months from onset (IQR)		SMT (IQR)	
1 (95)	9.0	(5.4-13.0)	0.33	(0.24-0.46)
2 (49)	18.4	(12.8-22.6)	0.62	(0.51-0.73)
3 (67)	18.9	(12.6-24.6)	0.67	(0.55-0.82)
4 (32)	24.8	(17.4-30.9)	0.86	(0.79-0.95)
5 (95)	27.7	(22.0-34.0)	1.00	(1.00-1.00)

B)

Milano-Torino staging system (n)	Median number of months from onset (IQR)		SMT (IQR)	
0 (95)	9.0	(5.4-12.9)	0.33	(0.24-0.46)
1 (94)	16.5	(11.9-22.1)	0.58	(0.49-0.71)
2 (37)	25.0	(20.0-31.7)	0.88	(0.72-0.93)
3 (12)	25.1	(21.0-30.0)	0.93	(0.86-0.97)
4 (2)	27.0	(24.1-29.8)	0.95	(0.95-0.96)
5 (95)	27.7	(22.0-34.0)	1.00	(1.00-1.00)

Table 3Median number of months and Standardised Median Time (SMT) from onset to each stage. (A) King's staging system, (B) MiToS staging system.

2.4.3.3 Comparison of staging systems

To compare each staging system, King's and MiToS scores were plotted against frequency for all pairwise comparisons (Figure 12 & Table 3). King's stages 1 and 2 matched almost perfectly with MiToS stage 1 (K1M1 n=186, K2M1 n=310) with little overlap to MiToS stage 2 (K2M2 n=4) and none with MiToS stages 3 and 4. However for King's stage 3, more patients were defined as MiToS stages 2 (K3M2 n=37) and 3 (K3M3 n=9). In King's stage 4 all four MiToS stages were seen (K4M1 n=68, K4M2 n=77, K4M3 n=26, K4M4 n=6). A chi-square test confirmed association between some stages with the two staging systems ($p<0.001$) and standardised residuals showed the strongest association was of King's stage 4 with MiToS stage 2. A Spearman's rank correlation coefficient between the King's and MiToS systems showed a correlation of 0.48 ($p=0.01$).

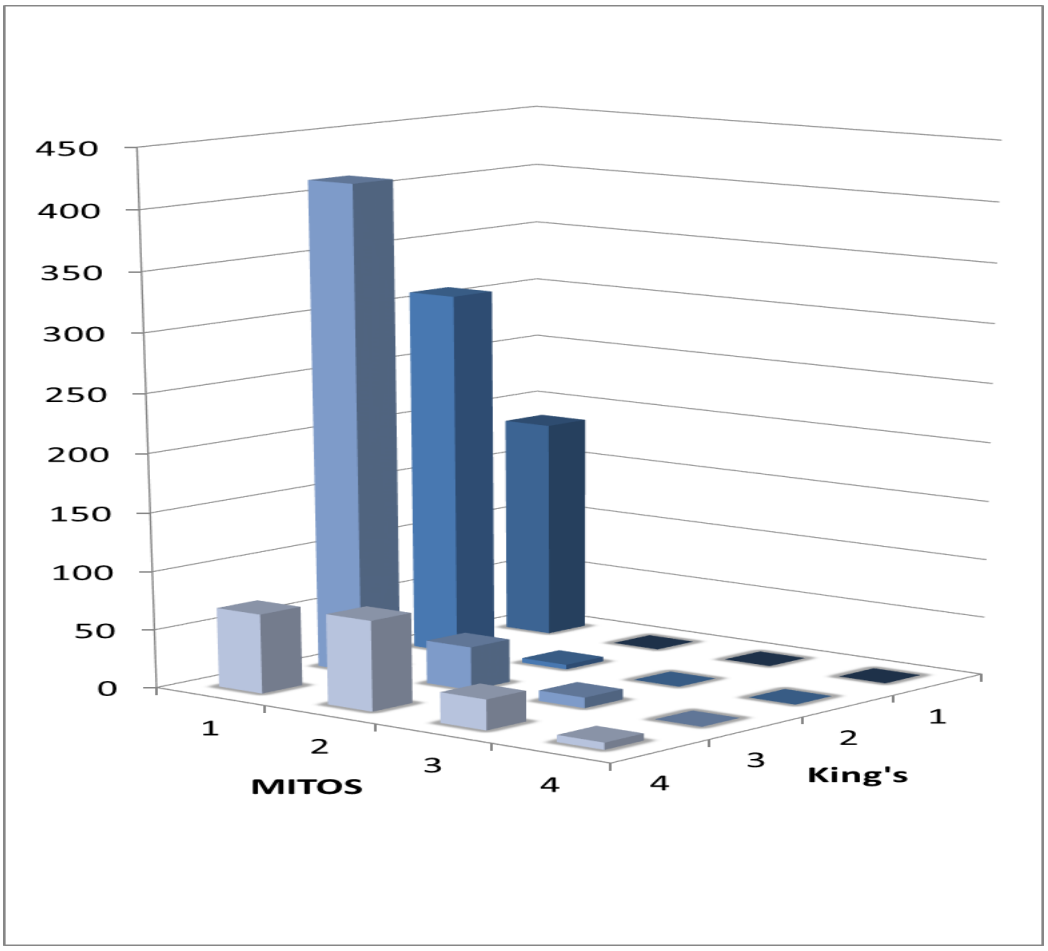


Figure 12 Bar chart showing the count of patients in each clinical stage by both systems.

2.4.4 Discussion

While the King's clinical staging system is able to differentiate early to mid-disease well, the MiToS staging is able to differentiate late stages in detail, which is in line with previous findings[29]. These results support the use of both systems when staging, as they summarise two different aspects of patient information. King's staging is mostly focussed on anatomical disease spread and significant involvement of respiratory muscles, whereas MiToS staging is aimed more towards the distinction of functional capabilities during the spread of the disease.

Because functional engagement necessarily follows anatomical involvement, MiToS stages inevitably tend to lag behind King's stages, manifesting as a higher resolution later on in the disease. This is most easily seen in Figure 12, where the relative distribution of individuals in each staging system is shown. The MiToS stages remains at a low resolution for the majority of King's stages up until K4, at which point there is significant differentiation of the MiToS stages.

These differences in disease description by the two systems are also shown by a Spearman's rank correlation of 0.48, showing some correspondence between the two systems. Association testing shows that King's stage 4 and MiToS stage 2 are the most strongly associated between all staging pairs.

Examination of the proportion of disease elapsed at each stage confirms that King's stages show more resolution through early to mid-disease and MiToS stages towards the end. Patients in King's stages 1, 2 and 3 are often in MiTos stages 0 or 1. We found that King's Stage 4 corresponds to MiToS stages 2, 3 and 4, and about 80% to 90% of the disease course. The benefit of MiToS staging in differentiating later stages of disease is in contrast to the ALSFRS-R scores from which it was derived, that exhibits a floor effect and lack of sensitivity in the later stages[24,25,29]. In other words, by combining information from different parts of the ALSFRS-R, MiToS staging is able to provide value over and above the ALSFRS-R score as a functional indicator for disease progression.

A limitation of this study is its use of clinical trial data rather than clinic or population data. However, this may be advantageous, as results are more likely to be relevant to daily clinical practice. We have previously shown that clinical trial data show a shift towards a greater

proportion of disease course passed for a given stage[28]. This occurrence is likely a result of left censoring due to the population being selected for trial participation and sourced from a biased clinic population.

The two disease staging systems described are complementary rather than redundant and provide different types of information. King's staging summarizes the clinical or anatomical spread of disease, while MiToS staging summarizes the functional burden of disease. A similar situation exists for cancers. The American Joint Committee on Cancer's TNM scale allocates a score for size, lymph node infiltration and metastasis as a functional indicator for disease progression[148], and this is combined with grouping of patients into one of four clinical stages that determine overall disease progression. King's stage prefixed K, and MiToS stage prefixed M, would allow a concise summary of disease spread and functional burden. We therefore propose using both to describe ALS stage.

2.4.5 Conclusions

In conclusion, these staging systems are complementary, with the King's clinical staging system having a higher resolution in early-mid diseases stages, and the MiToS system having a higher resolution in late disease stages. In the King's staging system there is more homogeneity between patients in the same stage, and a greater discrimination between patients in different disease stages.

Chapter 3. Understanding the relationship between the survival benefit of Riluzole and disease stage

3.1 Introduction

The course of ALS varies considerably between people, so that although the median survival is only about 2 or 3 years from symptom onset, the range of survival is large, and in some instances more than 20 years. This variability complicates the analysis of clinical trials, since time is used as a proxy for disease progression, but the wide variability means that in some people there will have been little change over the course of a trial, potentially reducing the power to detect an effect of treatment. To mitigate this, most clinical trials in ALS have strict entry criteria that aim to exclude the most slowly progressing patients. Disease staging is another approach, because it provides a framework for measuring disease progression regardless of whether the disease is aggressive or relatively slow in any one individual. Several methods have been previously proposed, but the most widespread systems are the Milano-Torino functional staging[11], and King's clinical staging[24] systems, both of which can be derived from standard clinical observation.

King's clinical stage ranges from 1, early disease, to 4, late disease, with stage 5 being death. The motor system is considered in the El Escorial domains of bulbar, upper limb and lower limb[149]. Stages 1, 2 and 3 correspond to symptomatic, functional or examination involvement of one, two or three domains respectively. Stage 4 corresponds to nutritional failure (10% of premorbid weight loss because of dysphagia), or significant respiratory failure (fulfilling guidelines for needing non-invasive ventilation). King's stage can be derived from the ALS Functional Rating Scale (ALSFRS-R) with a high correspondence to actual clinical stage, making it useful for retrospective analyses. It has been validated by confirmation in multiple

populations, and correlation with biomarkers. It has been used to assess the timing of cognitive change in ALS, and for health economics analysis[146].

Riluzole, a glutamatergic antagonist, is the only disease modifying treatment shown to extend life in a neurodegenerative disease in the UK, improving survival in ALS in an efficacy trial and dose-ranging study which completed recruitment in 1994[27,150]. There is no apparent effect on function, but any effect is difficult to assess in a clinical trial because those with the worst function drop out, either because travel becomes too onerous, or through death. The result is that only those with relatively preserved function remain, which reduces the power to detect an effect between treatment groups. A key question therefore is whether Riluzole extends life throughout the disease course, or only at an early or late stage of the disease. From a patient perspective, there is a major difference between a drug that prolongs the early disease stages and a drug that prolongs later disease stages. From a health economics standpoint, a survival benefit from extending early stages of disease might be preferable to prolongation at later, more expensive stages[149].

To understand the relationship between the survival benefit of Riluzole and disease stage, we retrospectively applied a clinical staging analysis to the original dose-ranging study of Riluzole[150].

3.2 Methods

3.2.1 Participants and data collection

Patients in Belgium, France, Germany, Spain, Canada, the USA, and the UK with probable or definite ALS as defined by the El Escorial criteria were eligible for the original Riluzole dose-ranging study, and enrolled between December, 1992, and November, 1993[151]. In this retrospective analysis, data were extracted from the electronic case record forms of the Riluzole dose ranging study[152]. There were 959 patients with case records in four treatment arms, 50mg Riluzole per day (n = 237), 100mg (n = 236), 200mg (n = 244) and placebo (n = 242). The censor date for the Riluzole survival data was set as the original study end date, 31 December 1994. To exclude an artefactual explanation for findings, a trial of Lithium carbonate in ALS was also tested for comparison, LiCALS; ISRCTN:83178718[147]. There were 217 patients with case records, 214 of whom were randomized, 107 to treatment and 107 to placebo.

3.2.2 Staging

Participants were staged retrospectively from the clinical trial data available at each study visit using an algorithm. For the LiCALS study, the algorithm, based on the ALSFRS-R, has been previously published and shown to correlate 92% with actual clinical stage[152]. The Riluzole dose-ranging study data were collected before the ALSFRS-R had been developed, and we therefore used the same principles to develop a corresponding algorithm based on functional scores from electronic case-records. Affected domains were established following the criteria of King's ALS staging system, using questions on the modified Norris scales, MRC muscle strength scores, El Escorial category, gastrostomy data and vital capacity. In both cases, tracheostomy and intubation were classified as equivalent to death for the purposes of analysis. Allocation to Stage 4 requires either nutritional failure sufficient to require gastrostomy, or respiratory failure sufficient to require non-invasive ventilation. We therefore used as proxies,

insertion of gastrostomy, or vital capacity ≤75% predicted. The vital capacity threshold was selected based on thresholds used in previous studies[53,151,153] and national guidelines[154]. Because clinical stage was being estimated, and previous studies have not shown transition to earlier stages, we maintained the highest stage recorded if a subsequent estimate of stage showed an apparent reversal. The lowest allocated clinical stage was stage 2 as the original trial consisted of only patients with El Escorial probable or definite ALS, and the analysis is therefore unable to answer questions about the effect of Riluzole at Stage 1.

3.2.3 Staging algorithm

A regional analysis was performed, using regions defined as Bulbar, Upper limb, Lower limb.

- Stages 1-3 were scored based on the number of affected regions. A region was considered affected if there was a drop in maximum score for that region in any of the relevant scales (Norris bulbar, Norris limb and muscle strength). One affected region = stage 1, two affected regions = stage 2, three affected regions = stage 3. The minimum stage at entry was stage 2 because the minimum El Escorial category at entry was Probable, which requires two regions to be affected.
- Stage 4 was scored if the person had a gastrostomy, defined by the date of insertion, or vital capacity ≤75% predicted value, whichever date fell earlier.
- Stage 5 was scored as the date of death, tracheostomy or intubation, whichever date fell earlier. The censor date was 31 December 1994. Any events occurring after this date were not recorded and treated as censored. No stage reversals were allowed.

3.2.4 Statistical analysis

Because stage at enrolment might differ between treatment groups, and might therefore influence our analysis, we performed a chi-square test of the independence of stage at

enrolment and treatment group. To test the hypothesis that the benefit of Riluzole treatment would be seen in all disease stages, the mean duration of each stage was estimated for each treatment arm. The Kaplan-Meier product limit distribution was used to compare treatment groups for the time taken to change stage for stages 2, 3 and 4. The test was repeated, stratifying for stage at enrolment. Cox regression was used to confirm any finding of an effect of treatment group on time in clinical stage, controlling for covariates. Regression models were built step-wise, adding in clinical stage at entry, an interaction term for treatment group and clinical stage at entry, age, and sex, with covariates discarded if the model fit was not significantly improved.

3.2.5 Sensitivity analysis

Several sensitivity analyses were performed to ensure findings were robust. Kaplan-Meier analyses were repeated after combining Riluzole treatment groups either using the doses found to significantly improve survival (100mg per day and 200mg per day), using current treatment recommendations (100mg per day), or using all doses (50mg per day, 100mg per day and 200mg per day), and with alternative vital capacity thresholds to define Stage 4 of ≤70% and ≤80% of predicted. Since clinical staging in ALS has not previously been used to estimate the timing of benefit in clinical trial data, we also performed the same analyses in the LiCALS data to exclude an artefact in trial data as a basis for findings. To confirm the results were not an artefact of the analysis method, we also used a second approach, Multi-state Outcome Analysis of Treatment (MOAT)[155]. MOAT does not tolerate missing data, and we therefore imputed missing or superseded disease stages by using the mean stage duration proportion by treatment groups across the study. Statistical tests were performed using IBM SPSS Statistics 24.0 (SPSS Inc., Illinois)[156], RStudio 1.0.143 (RStudio Inc., Boston)[157], R Foundation for Statistical Computing 3.4.1 (R Core Team, Vienna)[158] and SAS 9.4 (SAS Institute Inc., Cary)[159].

3.3 Results

I reformatted and inputted the dataset from the original Riluzole clinical trial and the original lithium carbonate in amyotrophic lateral sclerosis trial. I contributed to the design of the study, the analysis of the data and the writing. Statistical analysis including survival analysis was done by Ton Fang under my guidance. I am second author on a paper in The Lancet Neurology describing the work in this chapter.

I reformatted and inputted the dataset from the original Riluzole clinical trial and the original lithium carbonate in amyotrophic lateral sclerosis trial. I contributed to the design of the study, the analysis of the data and the writing. Statistical analysis including survival analysis was done by Ton Fang under my guidance. I am second author on a paper in The Lancet Neurology describing the work in this chapter.

Of the 959 participants who took part in the Riluzole trial, 355 enrolled at Stage 2, 451 at Stage 3 and 153 at Stage 4. Stage at enrolment did not differ between treatment arms (Supplementary Table 1) ($P = 0.22$). Counting the same patient at multiple stages where necessary, 355 patients reached Stage 2, 678 reached Stage 3 and 306 reached Stage 4. Although there was no difference in the proportions changing disease stage (Table 4A), in the treatment arms, there was a large increase in the proportion remaining in Stage 4 rather than other stages (Table 4B), suggesting that Riluzole increases survival by prolonging Stage 4. Kaplan-Meier analysis confirmed that Riluzole treated groups spent longer in Stage 4 than the placebo group ($P = 0.037$, Figure 13-A). Stratification by stage at enrolment confirmed the finding ($P = 0.027$). Time from Stage 2 or Stage 3 to subsequent stages or death was not significantly different between treatment arms and placebo, (Stage 2, $P = 0.83$, Figure 13-B, Stage 3, $P = 0.88$, Figure

13-C), and this remained true when combining treatment doses (Stage 2, $P = 0.99$, Figure 14-B, Stage 3, $P = 0.86$, Figure 14-C).

Table 4A

Time transitioning to a later stage or death

Treatment	*Stage 2*	*Stage 3*	*Stage 4*
50 mg/day Riluzole	256 (152)	241 (138)	224 (158)
100 mg/day Riluzole	243 (161)	265 (150)	248 (180)
200 mg/day Riluzole	243 (156)	230 (143)	233 (154)
Placebo	223 (157)	248 (143)	233 (154)

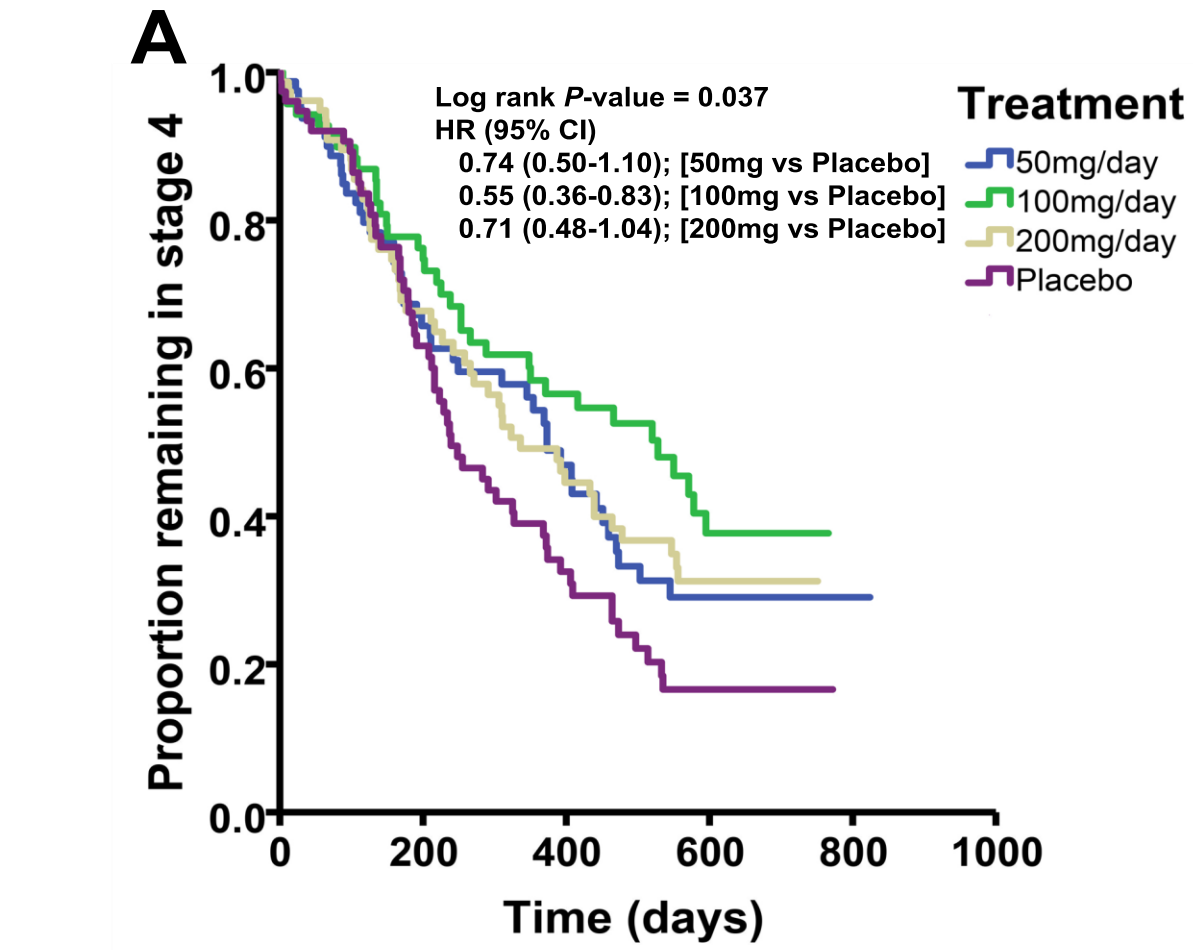
Table 4B

Time maintaining the same stage over the trial

Treatment	*Stage 2*	*Stage 3*	*Stage 4*
50 mg/day Riluzole	570 (92)	492 (158)	404 (404)
100 mg/day Riluzole	560	491	490
200 mg/day Riluzole	576	450	507
Placebo	568	485	391

Table 4 Stage transition times

The mean time patients spent transitioning to a later stage (Table 4A) or maintaining the same stage (Table 4B) is shown. Mean time spent calculated on a per patient basis and averaged over all patients in that group. There is a dose-dependent increase in time spent in stage 4, not seen for the other stages.



Number at risk				
50mg/day	81	43	24	11
100mg/day	70	48	30	14
200mg/day	78	48	29	15
Placebo	77	42	20	7

Figure 13. Patients progressing from each stage of amyotrophic lateral sclerosis with Riluzole or placebo.

Figure 13-A. Treatment with 100 mg Riluzole per day significantly prolonged time in stage 4 compared with placebo (p=0·037).

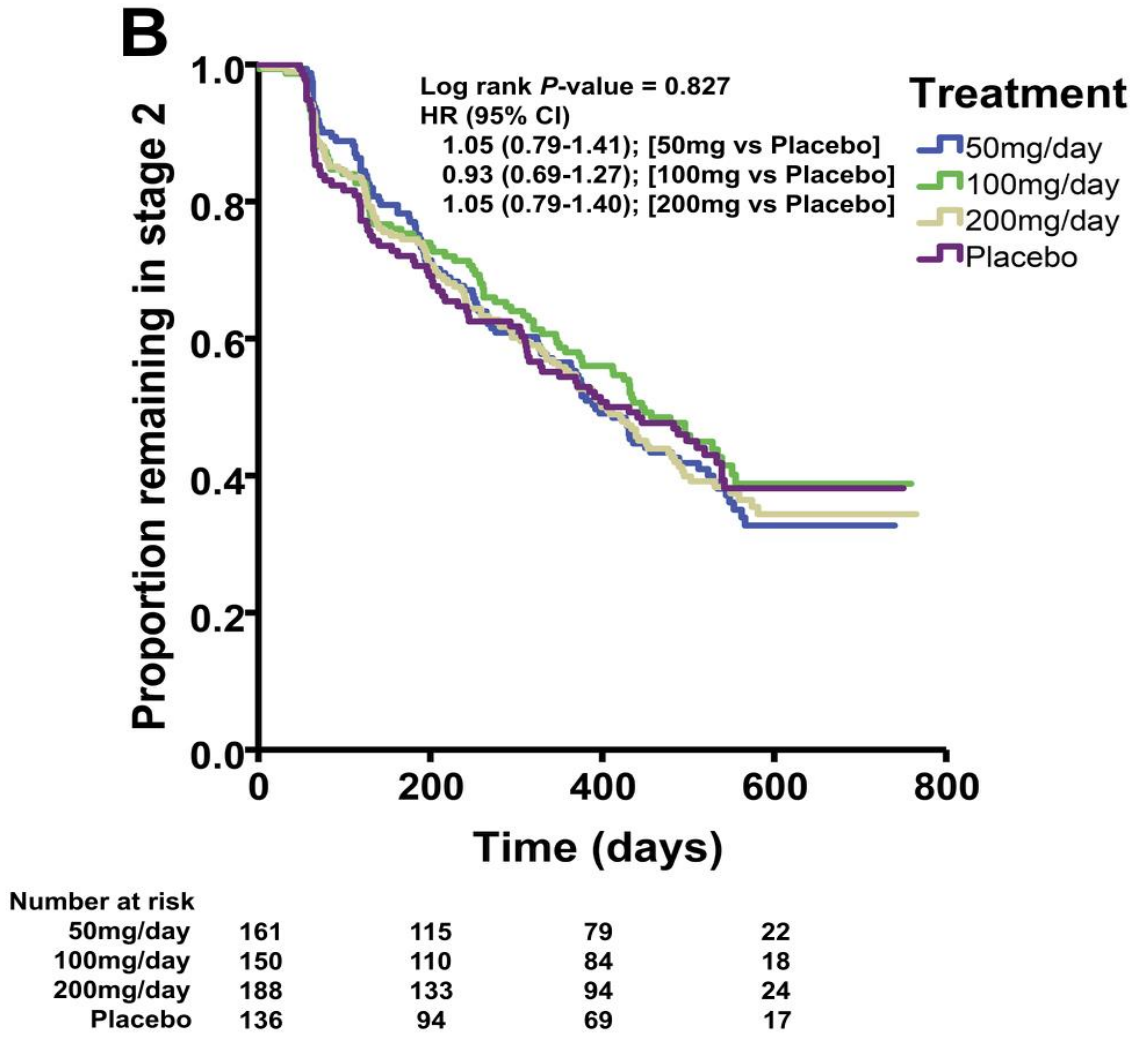


Figure 13-B. Patients progressing from each stage of amyotrophic lateral sclerosis with Riluzole or placebo. Treatment with all doses did not prolong time in stage 2 compared with placebo (p=0·827).

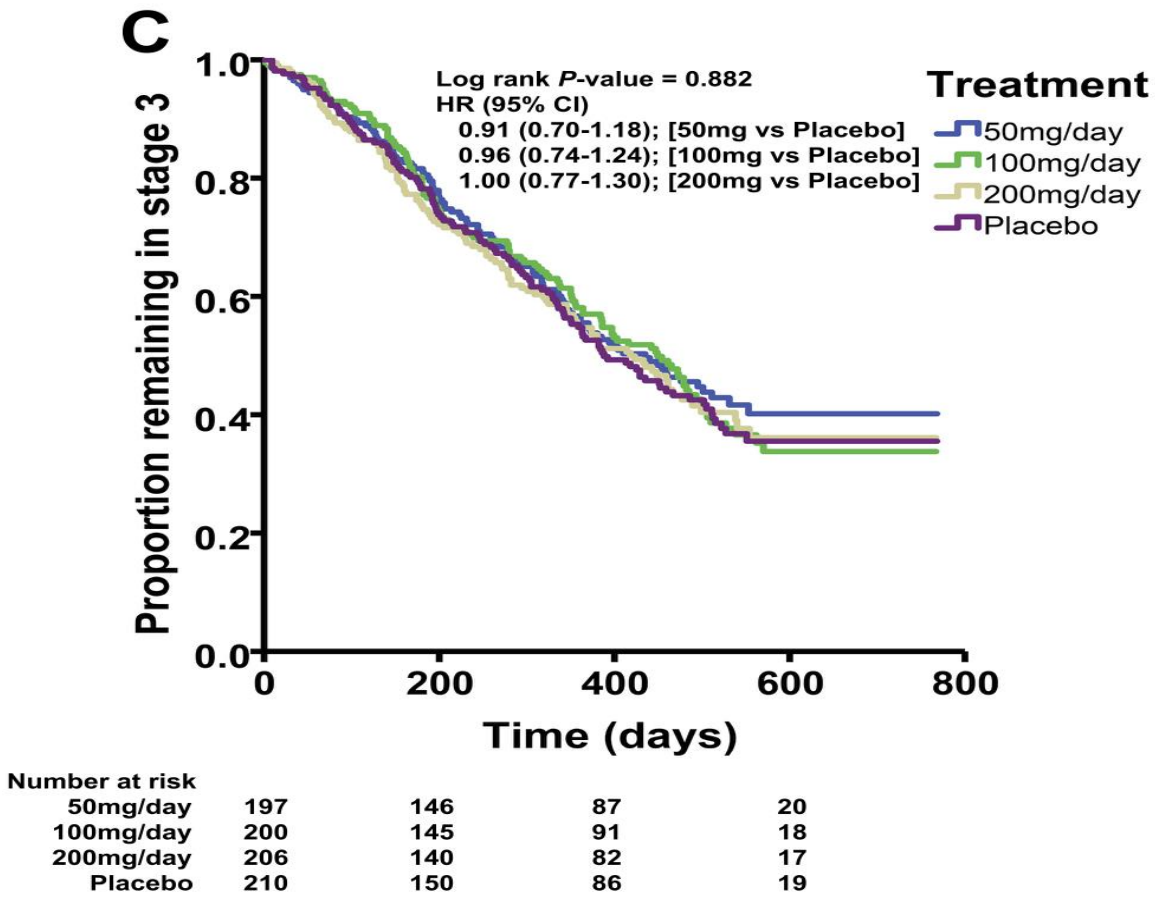


Figure 13-C. Patients progressing from each stage of amyotrophic lateral sclerosis with Riluzole or placebo. Treatment with all doses did not prolong time in stage 3 compared with placebo (p=0·882). HR=hazard ratio.

Restricting the analysis to the recommended treatment dose of 100mg still showed an extension of Stage 4 (P = 0.003), which remained the case after stratifying by stage at enrolment (P = 0.003). Cox regression confirmed an effect of treatment group on time in Stage 4 (P = 0.009), independent of the effect of clinical stage at entry ($P = 2.0 \times 10^{-9}$) with no evidence for an effect of the interaction of treatment group and stage at entry, age or sex. The findings were the same if treatment groups were tested in combination (treatment group P = 0.006, stage at entry $P = 7.0 \times 10^{-9}$), or restricted to the recommended treatment dose of 100mg per day (treatment group $P = 3.9 \times 10^{-4}$, stage at entry $P = 1.3 \times 10^{-5}$) (Table 5).

Variable	***P*-value**	**Hazard ratio**	**95% CI**
A. Individual treatment groups			
Treatment effect overall	0.009		
50mg/day Riluzole compared with Placebo	0.138	0.742	(0.501 - 1.101)
100mg/day Riluzole compared with Placebo	0.001	0.480	(0.313 - 0.735)
200mg/day Riluzole compared with Placebo	0.085	0.710	(0.481 - 1.048)
Stage at entry effect overall	< 0.001		
Entry at stage 2 compared with 4	0.027	1.697	(1.063 - 2.709)
Entry at stage 3 compared with 4	< 0.001	2.966	(2.118 - 4.152)
B. Combined treatment groups compared with placebo			
Treatment at any dose compared with placebo	0.006	0.638	(0.464 - 0.878)
Stage at entry effect overall	< 0.001		
Entry at stage 2 compared with 4	0.033	1.664	(1.043 - 2.655)
Entry at stage 3 compared with 4	< 0.001	2.825	(2.026 - 3.939)
C. Recommended treatment dose compared with placebo			
100mg/day treatment compared with placebo	< 0.001	0.456	(0.295 - 0.704)
Stage at entry effect overall	< 0.001		
Entry at stage 2 compared with 4	0.041	2.034	(1.031 - 4.016)
Entry at stage 3 compared with 4	< 0.001	3.154	(1.957 - 5.080)
D. Combined higher treatment doses compared with placebo			
100mg or 200mg/day treatment compared with placebo	0.002	0.578	(0.409 - 0.816)
Stage at entry effect overall	< 0.001		
Entry at stage 2 compared with 4	0.002	2.268	(1.357 - 3.790)
Entry at stage 3 compared with 4	< 0.001	3.023	(2.055 - 4.447)

Table 5. Effect of variables on time spent in Stage 4, by Cox regression

Variables were included step-wise in the model and removed if there was no significant improvement in the model fit. Variables tested were treatment arm, stage at trial entry, interaction between treatment arm and stage at trial entry, age, and sex. Only treatment arm and stage at trial entry were retained in the model.

Sensitivity testing

Combining treatment groups made no difference to the result. Analysing all treatment groups as a whole against placebo showed a significant prolongation of Stage 4 in the treatment groups (P = 0.01), as did limiting the analysis to the two higher doses against placebo (P = 0.006, Figure 14).

Altering the vital capacity threshold defining Stage 4 to ≤80% did not change the findings (P = 0.014), although reducing it to ≤70% meant that only 39 fulfilled respiratory criteria for Stage 4 and the effect was no longer seen (P = 0.18, Supplementary Figure 14). The findings were unchanged when treatment groups were combined, regardless of the definition of Stage 4: vital capacity ≤70%, two higher doses vs placebo P = 0.037, all doses vs placebo P = 0.067, and ≤80%, two higher doses vs placebo P = 0.002, all doses vs placebo P = 0.002.

Three people are recorded as not taking trial medication, one randomized to 50mg Riluzole treatment, and two randomized to 100mg. It is therefore unlikely that the findings are confounded by differences in ability to take medication. All 217 participants in the LiCALS trial entered at Stage 1. Treatment with lithium did not prolong the duration of any stage (Stage 1, P = 0.98; Stage 2, P = 0.73, Stage 3, P = 0.082 and Stage 4, P = 0.25).

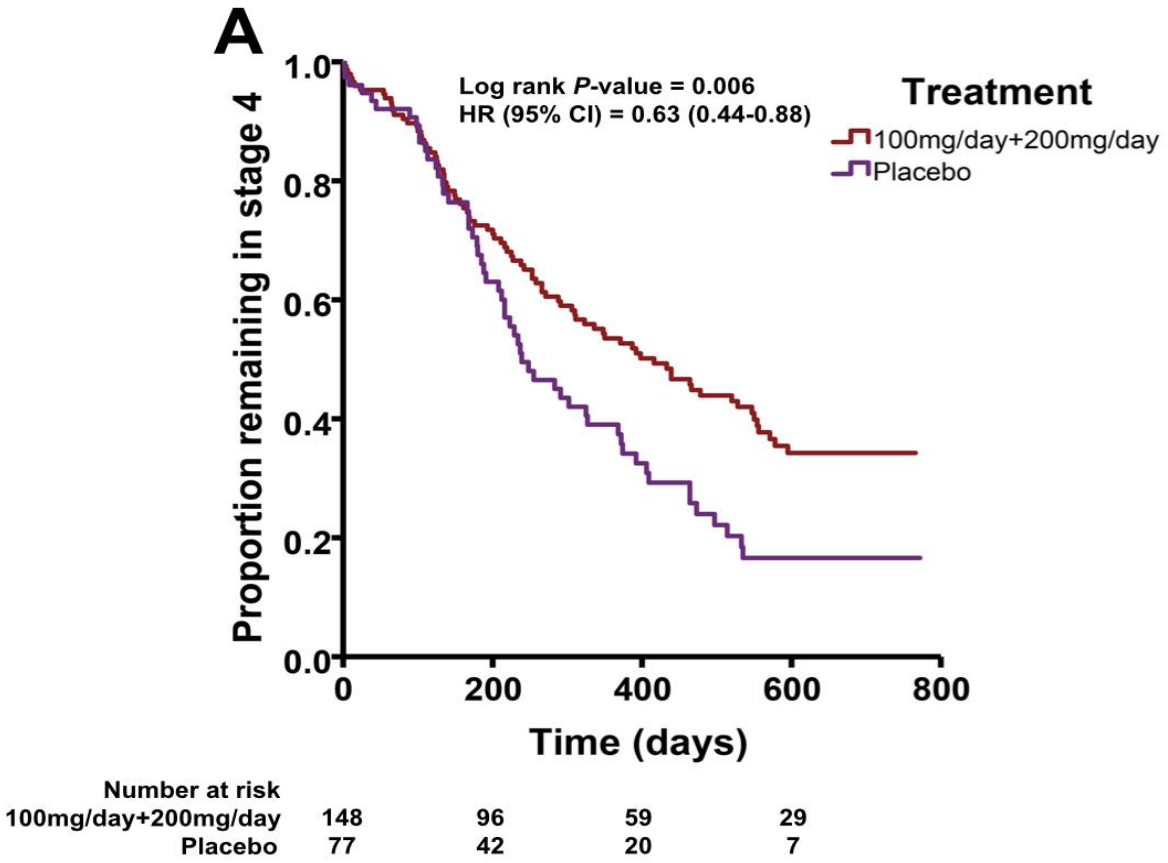


Figure 14. Patients progressing from each stage of amyotrophic lateral sclerosis with 100 mg plus 200 mg Riluzole or placebo.

Figure 14-A. Treatment with higher doses significantly prolonged time in stage 4 compared with placebo (p=0·006).

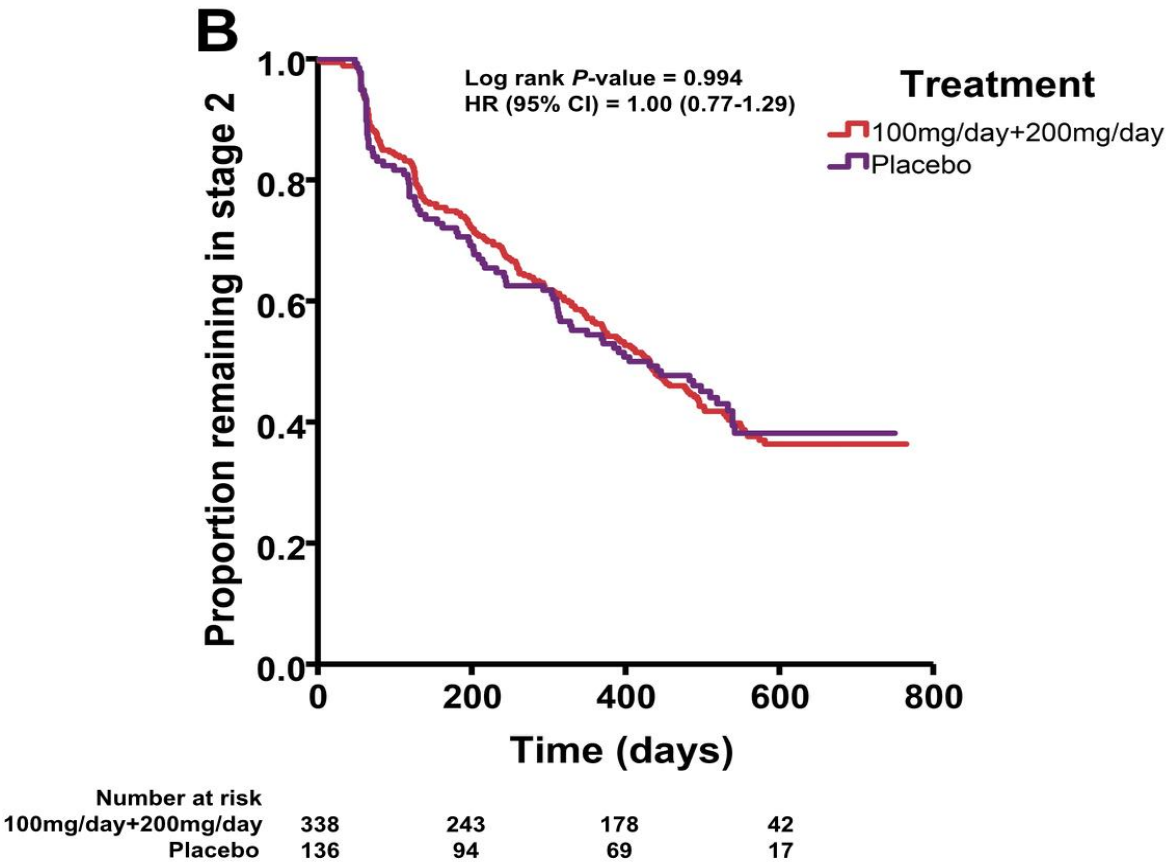


Figure 14-B. Patients progressing from each stage of amyotrophic lateral sclerosis with 100 mg plus 200 mg Riluzole or placebo. Treatment with higher doses did not prolong time in stage 2 compared with placebo (p=0·994).

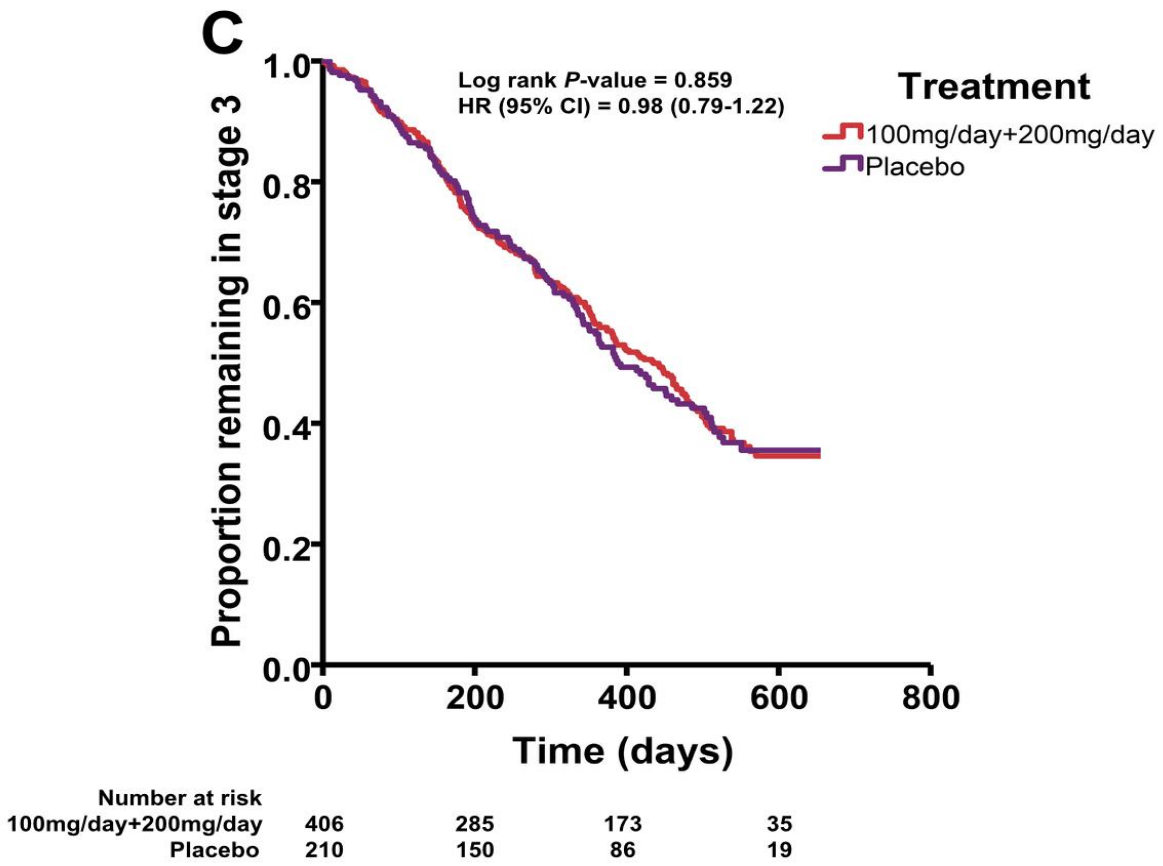


Figure 14-C. Patients progressing from each stage of amyotrophic lateral sclerosis with 100 mg plus 200 mg Riluzole or placebo. Treatment with higher doses did not prolong time in stage 3 compared with placebo (p=0·859). HR=hazard ratio.

Multi-state Outcome Analysis of Treatment

MOAT analysis confirmed the findings of the Kaplan-Meier approach, showing that those treated with Riluzole had a significantly longer Stage 4 than those on placebo, but the same duration of other stages (Table 6), although there was a suggestion that Stage 3 was shortened in those receiving Riluzole.

To exclude an artifactual explanation for findings, we did comparison tests using data from the LiCALS trial in which all 217 participants entered at stage 1, of which 214 were randomly assigned, 107 to treatment and 107 to placebo. Treatment with lithium did not prolong the duration of any stage (for stage 1, HR 1·00, 95% CI 0·84–1·19, p=0·98; stage 2, 1·04, 0·83–1·30, p=0·73, stage 3, 1·40, 0·96–2·04, p=0·082, and stage 4, 1·51, 0·74–3·05, p=0·25). MOAT analysis confirmed the findings of the Kaplan-Meier approach, showing that those treated with Riluzole had a longer stage 4 than those on placebo (table 3).

Stage change	50mg/day		100mg/day		200mg/day		Placebo	
	Mean number of days	95% CI	Mean number of days	95% CI	Mean number of days	95% CI	Mean number of days	95% CI
2-3	109	99 - 118]	70	60 - 81]	100	89 - 110]	82	72 - 91]
3-4	38	29 - 48]	52	43 - 61]	30	23 - 37]	69	61 - 78]
4-5	207	195 219]	234	222 246]	226	215 237]	198	186 209]

Table 6. Multi State Outcome Analysis of Treatment (MOAT).

Analysis of the mean number of days to transition from one stage to the next by treatment group. Time in Stage 4 is prolonged on Riluzole at higher doses compared with placebo.

3.4. Discussion

We have shown that treatment with Riluzole prolongs time in Stage 4 in ALS, and that this result is robust to the method of analysis, and independent of the stage at which treatment is started. This finding implies that the survival benefit of Riluzole is achieved by extending Stage 4, not by prolonging Stages 2 or 3, or generally slowing disease.

Until now, patients counselled about Riluzole have been told that it extends life, but not at which point, since this was not clear from the original study. The original analysis of the dose-ranging Riluzole trial showed no overall effect on function[150], which led to the conundrum of how to explain an improvement in survival without a concomitant effect on function. Our finding that the extension of life is due to an extension of Stage 4 helps to resolve this, since function at this stage is limited, and a flattening of the slope of functional decline would be very hard to detect. This difficulty is compounded by the observation that the ALSFRS-R slope change with time is curvilinear and therefore flatter at the beginning and end[53]. Although the benefit may not seem to be desirable on one level, since the prolongation of life is at a late disease stage when disability is high, rather than an early stage, when the person with ALS is functionally well, all our other current treatments that extend life act at the last stage of disease. Non-invasive ventilation for example, has been shown to prolong life and improve quality[160], and is used at Stage 4. The take-up of non-invasive ventilation (NIV) is high among patients[155], suggesting that prolongation of life at later stages of disease is not undesirable in itself, especially as Riluzole seems to be well tolerated in advanced ALS[161]. Similarly, gastrostomy is used to support those with nutritional failure due to dysphagia, which improves quality of life[162] and is also applied at Stage 4. A direct clinical implication of our findings is that patients can be told that Riluzole extends the later stages of ALS, but it is important to note that it may also extend Stage 1, since we have no information on this stage from the trial data.

When Riluzole was first identified as a beneficial treatment for ALS, its use in various health systems was controversial, because the survival benefit was seen as small, while the drug cost was seen as high. A combination of health economics analysis and pressure from patient groups led to its widespread adoption[54], although in some countries approval was delayed[163–165]. In the UK, the National Institute for Health and Clinical Excellence (NICE) approved Riluzole for the treatment of ALS following a detailed cost-benefit analysis that included the concept of Quality Adjusted Life Years[166,167]. Although our finding of a prolongation of Stage 4 might affect such analyses, Riluzole is no longer on patent, and its cost is now small compared with other treatments.

Despite increasing scientific effort to understand the mechanism of therapeutic benefit of Riluzole, the mechanism of action remains unclear. Several hypotheses have been suggested such as central anti-glutaminergic modulation of excitotoxic effect, mitochondrial dysfunction, and peripheral axonal effects on persistent sodium channel function with potentiation of calcium-dependent potassium currents. We have shown that 100 mg/day of Riluzole was associated with longer survival in the last clinical stage of ALS before death (stage 4) compared with placebo. These findings suggest that the disease-modulatory effects and survival benefits of Riluzole occur in an advanced stage of disease. Furthermore, it has been suggested that Riluzole treatment is most effective in individuals with ALS with impaired respiratory function. Benefit was observed in individuals with reduced vital capacity, or in patients on non-invasive positive pressure ventilation with reduced maximal inspiratory pressure, consistent with our observation that Riluzole benefits people with stage 4 ALS. Previous work from our lab has shown that feeding failure, defined as swallowing difficulty, resulting in 5 to 10% premorbid weight loss is often coincident with clinically significant respiratory failure,4, 5 so it would therefore be important to establish whether the beneficial metabolic effect targets

weight loss, respiratory function, or both, given the coincidental timing of these. Future work should include studying the characteristics of King's stage 4, which is the defined by the need for specific clinical intervention. The findings from the MIROCALS trial on the effect of Riluzole on respiratory function, weight, blood biomarkers, and cerebrospinal fluid biomarkers will be important in this regard.

A strength of this study is the use of clinical staging to analyse clinical trial data in a neurodegenerative disease, allowing an examination of when any benefit occurs in a way that is easily understood by clinicians and patients. Thus, as an outcome measure, clinical staging has an important role to play in future trial design in ALS, and other neurodegenerative diseases. In cancer, another group of diseases which if untreated lead to progressive disability and death, trials routinely use staging to decide on the appropriate treatment and to assess outcome[168]. A further benefit is that successful treatments can be shown to reverse the progression through clinical stages.

There are several important weaknesses of this study. It is a post-hoc analysis, and as such cannot be regarded as providing the same level of evidence as a pre-specified analysis, since the study design did not take into account staging in the calculation of statistical power, or in the assessment criteria. Furthermore, the criteria for Stage 4 mean that 153 (16%) of the 959 patients were in Stage 4 at enrolment. This would not usually be the case in a modern trial in ALS. Clinical stage was estimated from trial data. We have previous experience in this process, and have successfully applied an algorithm to the ALSFRS-R to derive clinical stage[169]. In this study, we could not use the ALSFRS-R because such a scale did not exist when the trial data were collected. As a result, we had to generate a new algorithm to estimate clinical stage. There is no way to validate this new algorithm, since one of the scales it uses, the Norris scale, is no

longer in use. To overcome this, we have applied the same logical process to the data that was used to generate the ALSFRS-R algorithm for staging. Furthermore, adjusting the criteria defining the clinical stages does not change the findings of the study, and using two entirely different analytical approaches reached the same conclusions.

A further limitation of this study arises from the strict inclusion criteria of the original trial, which was restricted to people with El Escorial probable or definite ALS[150]. This prevents our study from analysing the treatment effects of Riluzole in Stage 1 of disease. However, some studies have suggested that the effects of Riluzole may be transient[170,171] and support treatment in early stages of ALS[171]. To determine whether Riluzole extends Stage 1 will require a specific trial to address this issue. Although it is ethically difficult to perform new studies exclusively on Riluzole, studies of Riluzole embedded within other clinical trials have been completed, for example in the study of Dexpramipexole in ALS (EMPOWER, NCT01281189)[172], or are underway, for example in a study of low-dose interleukin 2 in ALS (MIROCALS, NCT03039673)[173]. Such an approach could potentially address our findings within a prospective study design, or retrospectively confirm these findings using similar techniques to ours within existing data.

Riluzole is currently the only treatment shown to prolong life in ALS, or indeed in any neurodegenerative disease. We have shown that it acts by prolonging Stage 4 ALS rather than by slowing the entire disease course or prolonging early stages. Similar methods should be employed in future clinical trials where survival is an endpoint, to show where benefit is accrued, and to allow a full discussion of effects when counselling patients about treatment.

3.5 Conclusions

In conclusion, we showed that Riluzole prolongs survival in the last clinical stage of ALS; this finding needs to be confirmed in a prospective study, and treatment effects at stage 1 still need to be analysed. The ALS stage at which benefit occurs is important for counselling of patients before starting treatment. Staging should be used in future ALS clinical trials to assess the stage at which survival benefit occurs, and a similar approach could be used for other neurodegenerative diseases.

Chapter 4. Developing bioinformatics pipelines for the analysis of next-generation-sequencing data

4.1 DNAscan: a fast, computationally and memory efficient bioinformatics pipeline for the analysis of DNA next-generation-sequencing data

4.1.1 Introduction

To understand the role of gene variation in disease progression, methods are needed to assay that variation. This chapter describes an analysis method designed as a pipeline from raw data to report, applied to researchers and then to patients and clinicians. Next generation sequencing technologies (NGS) play a key role in human genetic research. The effort needed to sequence a whole genome has reduced from about 15 years of work at a cost of $3 billion in 2003[174] to hours for ~$1000 in 2011

Producing sequencing data, whether it is whole genome sequencing (WGS), whole exome sequencing (WES) or targeted gene panels, is common practice for the study of genetic bases of biological processes. In biomedical research, NGS data are widely used to investigate the genetic causes of disease, allowing for the study of genomic variants including single nucleotide variations, small insertions or deletions of a few bases, as well as structural variants.

On a large scale, international collaborations are forming sequencing consortia to study the genetic landscape of thousands of individuals. Examples of such consortia are Project MinE [74] and The Cancer Genome Atlas[175]. Project MinE is an international consortium seeking to obtain sequencing data from 15,000 Amyotrophic Lateral Sclerosis cases and 7,500 matched controls. TCGA is a rich dataset of sequencing data of over 11,000 individuals affected by 33 different tumour types. On an individual scale, NGS data are also being investigated for their use in diagnostic medicine and so called Precision Medicine[176,177], whose aim is to tailor medical treatments to patient genetics.

There are several practical challenges when processing NGS data. For example, 40x WGS data for one sample produced on the Illumina Hiseq 2000, one of the most popular sequencers, is about 400 gigabytes in its raw format (fastq format)[91]. This size can be reduced to approximately one fourth when the data is compressed down to about 100 gigabytes, using lossless formats such as fastq.gz (gzip-compressed version of fastq) and bam[92]. Such big files are not easy to handle for the average non-specialised scientist or lab, since they require sophisticated tools, bioinformatics skills and high-performance computing for their analysis. Indeed, as an example, consider mapping one of these files, typically about 1 billion 150-base-pair long reads, to the human genome, a key process in the analysis of WGS data. Assuming that a standard midrange desktop computer with 4 cpus and 16 gigabytes of RAM is used with The Burrow Wheeler Aligner (BWA)[93], probably the most widely used mapper, aligning this data to the human reference genome would take about 1 day, and this would only be the first step of an NGS data analysis pipeline. Faster mappers exist. For example SNAP[94] would only take about 4 hours to complete the same job, using the same number of cpus, but it requires about 65Gb RAM, making it an unsuitable choice if large memory High-Performance-Computing (HPC) facilities are not available.

For big collaborations and projects, collecting NGS data from thousands of individuals, powerful and expensive HPC facilities, as well as highly specialised staff are needed. To handle such data, the Project MinE consortium makes use of SURFsara, the Dutch national HPC facility, and the TCGA invested millions of dollars in HPC infrastructures and e-infrastructures (https://cancergenome.nih.gov) making use of, among others, Amazon Web Services (AWS) and Seven Bridges (https://www.sevenbridges.com).

Further challenges are represented by the large number of bioinformatics tools available for NGS analysis. Omictools[178], a web database where most available tools are listed and reviewed, lists over 7000 bioinformatics NGS tools, and given the great interest in this field, new tools

are frequently released. Therefore, designing a bioinformatics pipeline for the analysis of NGS data, taking into account both the available computing facilities and the study aim, is not trivial, and requires specialised expertise.

Here we describe DNAscan, an extremely fast, accurate and computationally light bioinformatics pipeline for the analysis, annotation and visualisation of DNA NGS data. DNAscan is designed to provide a powerful and easy-to-use tool for applications in biomedical research and diagnostic medicine, at minimal computational cost. DNAscan can analyse 40x WGS data in 8 hours using 8 threads and 16 Gb of RAM and WES data in 1 hour using 4 threads and 10 Gb of RAM, enabling the processing of NGS data to be carried out on most midrange computers and the minimisation of computational costs. The pipeline can detect SNVs, small indels, SVs, repeat expansions and viral genetic material (or any other organism). Its results are annotated using a variety of databases including ClinVar[179], EXAC[180] , dbSNP[181] and dbNSFP[182], made available for a local deployment of the gene iobio platform for an on-the-fly visualisation and user-friendly quality control (QC) reports are generated. DNAscan also allows the user to restrict the analysis to any subregion of the human genome, including the whole exome or a set of genes or gene panel, speeding up the processing time and generating region specific reports. DNAscan is available on GitHub[183] and Docker and Singularity[184] images are also available for fast and reliable deployment. The extent to which this is my work is explicitly stated in the Appendix. I am the second author on two published manuscripts describing the work in this chapter.

4.1.2 Motivation

The current genomics initiatives include a variety of projects ranging from large scale international sequencing projects such as ProjectMinE and ADNI[185], collecting thousands of whole-genome-sequencing samples, to medical studies designing gene panels for diagnostic

purposes. In this context, research and medical workers, with different scientific backgrounds and levels of bioinformatics skills, deal with NGS data constantly. The pipeline requirements were the following:

1) Speed: Some studies for which the pipeline is being used, require the analysis to be concluded within working hours.

2) Usable on personal laptops and desktops: Even if most academic institutions have HPC facilities available for their staff, this does not necessarily mean everything can be processed on them. In many circumstances, many factors, e.g. the informatics skills of the research workers, privacy and ownership policies, technical obstacles, etc., might make using local machines necessary. This is common in a medical research environment.

3) Annotation and visualisation: For non-specialised users, e.g. physicians, wet lab biologists, etc., an automatic annotation of the results and user-friendly interface for their visualisation can be fundamental.

4) Screen for microbial presence: NGS data are widely used to investigate the role of microbes, such as viruses and bacteria, in many diseases. Viral and bacterial metagenomics studies are also common.

5) Specific known repeats: Many repeat expansions have a crucial role in the development of several diseases. E.g. C9orf72 repeat[186] in Amyotrophic Lateral Sclerosis (ALS).

6) Region restricted analysis: Use of shared datasets for heterogeneous research purposes often focuses on subregions of the genome. E.g. screening NGS samples for particular variants or over a disease specific panel of genes.

7) Reproducibility and easy and fast deployment: The pipeline must be easily deployable and its results and analyses reproducible on any machine. This is to favour both reproducibility of our research and collaborations through the analysis pipeline sharing.

8) Easy to use: It must be suitable for a wide range of users with different level of informatics expertise.

4.1.3 Methods

4.1.3.1 Hardware

The tests using four threads were performed on a single machine with 16GB RAM and an Intel i7-670 processor. The tests using more than four threads were performed on a Dell PowerEdge R630 server.

4.1.3.2 Performance profiling

Memory usage was recorded using the following command lines:

$ nohup bash -c 'while true; do (echo "%CPU %MEM ARGS $(date)" && ps -e -o pcpu,pmem,args -- sort=pcpu | cut -d" " -f1-5 | tail) >> ps.log; sleep 5; done' &

Peak memory usage was parsed by:

$ cat ps.log | grep -e $proc_name -e ARGS | awk 'BEGIN{max=0}{ if (SUM>max) max=SUM; if ($3=="ARGS") print SUM, SUM=0; else SUM+=$2}END{print max}'

4.1.3.3 Variant calling assessment

To assess the performance of DANscan in calling SNVs and indels, we used the Illumina Genome Analyzer II whole exome sequencing of NA12878 (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/ technical/working/20101201_cg_NA12878/NA12878.ga2.exome.maq.raw.bam). Illumina platinum calls (ftp://platgene_ro@ussd-ftp.Illumina.com/) were used as true positives.

GATK BPW calls were generated using default parameters and following the indications on https:// software.broadinstitute.org/gatk/ (https://software.broadinstitute.org/gatk/best-practices/ bp_3step.php?case=GermShortWGS) for germline snvs and indels calling. These include the Pre- processing (https://software.broadinstitute.org/gatk/best-practices/bp_3step.php? case=GermShortWGS&p=1) and variant discovery steps for single sample, i.e. skipping the Merge and Join Genotype steps (https://software.broadinstitute.org/gatk/best-practices/bp_3step.php? case=GermShortWGS&p=2).

SpeedSeq calls were generated running the "align" and "var" commands as described on Github (https://github.com/hall-lab/speedseq)

RTG Tools ("vcfeval" command) was used to evaluate the calls (https://github.com/ RealTimeGenomics/rtg-tools).

F-measure, Precision and Sensitivity were defined as in Equation 1.

Equation 1: $Precision = \frac{T_p}{T_p + F_p}$; $Sensitivity = \frac{T_p}{T_p + F_n}$; $Accuracy = \frac{T_p + T_n}{T_p + T_n + F_n + F_p}$

Equation 1 Precision and Sensitivity

C9orf72 repeat primed PCR

C9orf72 gene Hexanucleotide Expansions were determined using a Repeat Primed PCR (RP-PCR), previously published by DeJesus-Hernandez et al. [58].

Simulated reads

150,000,000 Illumina 100-bases-long paired-end human reads were generated using pIRS [59] with default parameters and hg19.

Whole-Genome-Sequencing data of ALS patients

Venous blood was drawn from patients and controls from which genomic DNA was isolated using standard methods. DNA concentrations set at 100ng/ul as measured by a fluorometer with the PicoGreen® dsDNA quantitation assay. DNA integrity was assessed using gel electrophoresis. All samples were sequenced using Illumina's FastTrack services (San Diego, CA, USA) on the Illumina Hiseq 2000 platform. Sequencing was 150bp paired-end performed using PCR-free library preparation and yielded ~40x coverage across each sample.

Miseq ALS gene panel

The ALS gene panel was designed using Illumina TruSeq Custom Amplicon and implemented on an Illumina Miseq platform. This utilises polymerase chain reaction amplicon-based target enrichment and screens for variants in 10 ALS disease genes: *BSCL2*, *CEP112*, *FUS*, *MATR3*, *OPTN*, *SOD1*, *SPG11*, *TARDBP*, *UBQLN2*, and *VCP*. For these genes full exon sequencing was examined. These genes were selected based on their association with ALS, so that all the chief causal genes were included as well as several risk factors and a selection of variants with an uncertain relationship to ALS, either through a lack of evidence or through the gene more commonly causing a related disease. This panel of genes were sequenced in a study was initiated before the discovery of the ALS genes *KIF5A*, *MOBP*, *C21orf2*, *CHCHD10*, *MATR3*, *NEK1*, *TBK1* and *TUB4A*.

Whole-Genome-Sequencing of HIV infected human cells

Genomic libraries were prepared using the TruSeq® DNA Sample Prep kit V2 (Illumina) following the manufacturer's instructions. Briefly, 1 µg of genomic DNA was sheared with the Covaris 2 system (Covaris). The DNA fragments were then end-repaired, extended with an 'A' base on the 3′ end, ligated with indexed paired-end adaptors and PCR amplified. PCR amplification was carried out as follows: initial denaturation at 98°C for 30 sec, followed by 8 cycles consisting of 98°C for 10 sec, 60°C for 30 sec and 72°C for 30 sec, then a final elongation at 72°C for 5 min. Four different genomic libraries were pooled and sequenced in one lane of an Illumina Hiseq 2000 sequencer using a 2 x 95bp paired end indexing protocol. Demultiplexed fastq files were obtained for each sample using the Illumina CASAVAv1.8.1 software. The complete high throughput sequencing dataset was downloaded from the Sequence Read Archive (SRA) [60] under accession number SRA056122.

Viral database

In this paper DNAscan makes use of the whole non-redundant NCBI database of complete viral genomes (9334 genomes). These can be downloaded both as a multi sequence fast file, together with its fai index and HISAT2 index from our GitHub repository (link) and directly from the NCBI database ftp server (ftp.ncbi.nlm.nih.gov/refseq/release/viral)

DNAscan can also be used to screen for the DNA of other organisms including bacteria or fungi in which case the user can download the preferred database from the NCBI ftp server or from our GitHub where the corresponding index files can also be found (e.g. ftp.ncbi.nlm.nih.gov/refseq/ release/bacteria or ftp.ncbi.nlm.nih.gov/refseq/release/)

Software availability

DNAscan is available on GitHub (https://github.com/KHP-Informatics/DNAscan). Docker and Singularity images are also available for fast deployment and reproducibility (see instructions on Github).

4.1.3.4 List of tools

- Genome Analysis Toolkit 3.8 - BWA 0.7.15
- Picard 2.2.1
- Samtools 1.5
- HISAT2 2.1.0
- Bcftools 1.5
- RTG Tools 3.6.2
- Python 3.5
- MultiQC 1.2
- FastQC v0.11.7
- Docker 1.7.1
- Docker-compose 1.4.2
- Freebayes 1.0.2
- ExpansionHunter 2.0.9
- Manta 1.0.3
- Annovar "Version: $Date: 2016-02-01 00:11:18 -0800 (Mon, 1 Feb 2016)”
- Bedtools2 2.25

4.1.4 Results

4.1.4.1 Pipeline description

Dr Iacoangeli and I conceived and planned the study. Code writing was done by Dr Iacoangeli. All the data in this chapter was generated by me, and I contributed to the analysis, and to the writing of the paper. I am second author on a paper in The BMC Bioinformatics journal describing the work in this chapter.

DNAscan pipeline consists of four stages (Figure 15), Alignment, Analysis, Annotation and Report generation, and can be run in three modes, Fast, Normal and Intensive, according to user requirements. Fast mode minimises the RAM and time required while Normal and Intensive modes improve the variant calling performance by performing an alignment and small indel calling refinement stage respectively.

The user can tailor the DNAscan pipeline to their needs by performing any subset of the available analyses or restricting them to a subregion of the human genome. e.g. if variants have already been called (this is commonly the case when the sequencing is provided by companies such as Illumina, Novagen, etc) and we are only interested in a particular set of genes, the snv/indel calling steps would be skipped and DNAscan can be used to call SVs only and results visualisation on that set of genes. DNAscan accepts genome regions both as a bed file or a list of gene names. Optionally, the analysis can be restricted to the whole exome.

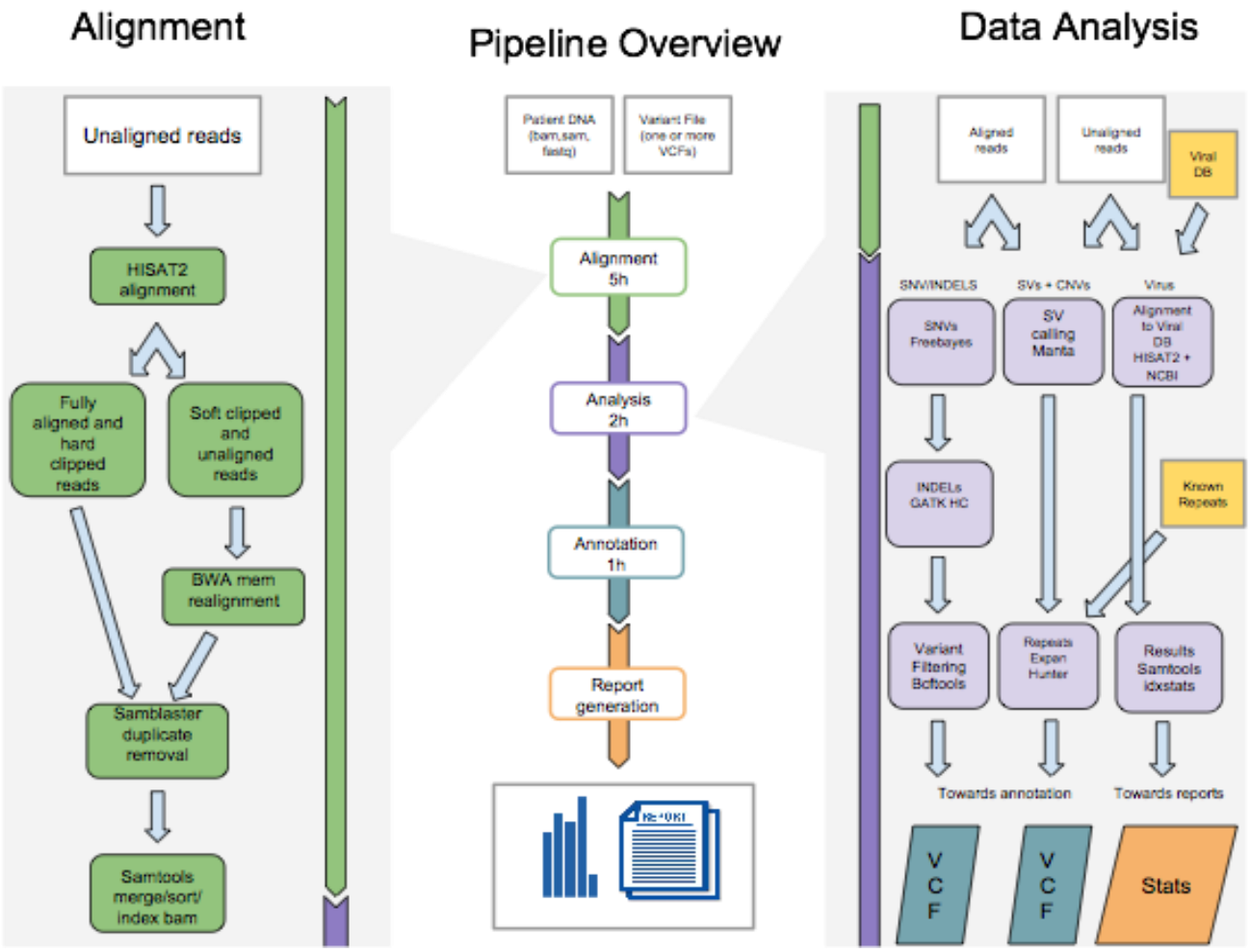


Figure 15. Central panel: Pipeline overview.

DNAscan accepts sequencing data, and optionally variant files. The pipeline firstly performs an alignment step (details in the left panel), followed by a customisable data analysis protocol (details in the right panel). Finally, results are annotated and a user-friendly QC report is generated. Right panel: detailed description of the post alignment analysis pipeline (intensive mode). Aligned reads are used by the variant calling pipeline (Freebayes + GATK HC + Bcftools), both aligned and unaligned reads are used by Manta and ExpensionHunter (for which no repeat description files have to be provided) to look for structural variants. The unaligned reads are mapped to a database of known viral genomes (NCBI database) to screen for their DNA in the input sequencing data. Left panel: Alignment stage description. Raw reads

are aligned with HISAT2. Resulting soft-clipped reads and unaligned reads are realigned with BWA mem and then merged with the others using Samtools.

4.1.4.2 Alignment

DNAscan accepts sequencing data both in raw fastq (and its .gz compressed version) and as a Sequence Alignment Map (SAM) file (and its compressed version BAM). HISAT2[187] and BWA are used to map the reads to the reference genome (Figure 15, left panel). This step is skipped if the user provides ready-aligned data in SAM or BAM formats. HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads to a reference genome. Based on an extension of BWT[188] for graphs[189], HISAT2 implements a large set of small graph FM indexes (GFM)[190] that collectively cover the whole genome. These "local" indexes, combined with several alignment strategies, enable rapid and accurate alignment of sequencing reads. This new indexing scheme is called a Hierarchical Graph FM index (HGFM). Thanks to this novel approach HISAT2 can guarantee a high performance, comparable to state-of-the-art tools, in approximately one quarter of the time of BWA and Bowtie2[191].

Variant calling pipelines based on HISAT2 generally perform poorly on indels[192]. To address this issue, DNAscan uses BWA to realign soft-clipped and unaligned reads. This alignment refinement step is skipped if DNAscan is run in fast mode.

Samblaster[193] is used to mark duplicates during the alignment step and Sambamba[194] to sort the aligned reads. Both the variant callers, Freebayes[195] and GATK Haplotype Caller (HC)[196] used in the following step, are duplicate-aware, meaning that they automatically ignore reads marked as duplicate. The user can optionally exclude it from the workflow. This might be

necessary in some studies, e.g. when an intensive PCR amplification of small regions is required.

4.1.4.3 Pipeline Analysis

Various analyses are performed on the mapped sequencing data (Figure 15, right panel): SNV and small indel calling is performed using Freebayes, whose reliability is well reported[197,198] . However, taking advantage of the documented better performance of GATK HC in small indel calling we decided to add a customised indel calling step to DNAscan, called Intensive mode. This step firstly extracts the genome positions for which an insertion or a deletion is present on at least one read, and secondly calls indels using GATK HC on these selected positions. The reduced number of positions where this occurs allows for a targeted
use of GATK HC, limiting the required computational effort and time. Resulting SNV and small indel calls are finally hard filtered with Bcftools[199].

Two Illumina developed tools, Manta[88] and Expansion Hunter[200] are used for detecting medium and large structural variants (> 50bp) including insertions, deletions, translocations, duplications and known repeat expansions. These tools are optimised for high speed and can analyse a 40x WGS sample in about 30 minutes using 8 threads, maintaining a very high performance.

DNAscan also has options to scan the sequencing data for microbial genetic material. It performs a computational subtraction of human host sequences to identify sequences of infectious agents including viruses, bacteria or fungi, by aligning the non-human or unaligned reads to the whole NCBI database[201,202] of known viral, bacterial or any custom set of microbial genomes and reporting the number of reads aligned to each non-human genome, its length and the number of bases covered by at least one read.

4.1.4.4 Annotation

Variant calls are then annotated using Annovar[203]. The annotation includes the use of databases such as ClinVar[179], Exac[180], dbSNP[181] and dbNSFP[182] (more information about how to customise the annotation, e.g. by selecting alternative databases and/or focusing on specific genome regions, are available on GitHub).

4.1.4.5 Report generation

Finally, a user friendly, readable and customisable report is generated. Bcftools, Samtools and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) are used to generate statistics on the variants called by the pipeline, the quality of the sequencing data and its alignment. MultiQC[204] is used to wrap up these stats and make them available as html report. A local deployment of a set of iobio services including the gene.iobio platform (gene.iobio.io) are available for an on-the-fly visualisation of the result variant files (gene.iobio and vcf.iobio) and sequencing data (bam.iobio[205]).

4.1.4.6 Pipeline (DNAscan) benchmarking

Benchmarking every DNAscan component is beyond the aim of my project since a range of literature is available[88,197,200,206]. However, to our knowledge, none exists assessing HISAT2 either for small DNA read mapping or as part of DNA variant calling pipelines. In the following section, we both assess the performance of HISAT2 with BWA and Bowtie2 mapping 1.25 billion WGS reads sequenced with the Illumina Hiseq 2000 and 150 million simulated reads (see Methods) and compare our SNV/indel calling pipeline in Fast, Normal and Intensive modes with the Genome Analysis Toolkit best practices workflow and SpeedSeq[207] over the whole exome sequencing of NA12878. Illumina platinum calls are used as true positives.

We show how DNAscan represents a powerful tool for medical and scientific use by analysing real DNA sequence data from two ALS patients and of HIV infected human cells. For the ALS patients we use both a gene panel of 10 ALS-related genes, whose feasibility for diagnostic medicine has been previously investigated[177], and their equivalent WGS data from the Project MinE sequencing dataset. DNAscan is used to look for SNVs, small indels, SVs, and known repeat expansions. The WGS of an HIV infected human cell sample[208] is used to test DNAscan for virus detection.

4.1.4.7 The HISAT2 aligner assessment

To assess the performance of the HISAT2 aligner we used two datasets: 1.25 billion WGS reads of a human WGS DNA sample sequenced with an Illumina Hiseq 2000 (see Methods), and 150 million simulated human reads (see Methods). In this assessment we took into consideration the memory footprint (RAM), the time needed to complete the alignment, percentage of reads mapped to the reference genome, and the percentage of uniquely mapped reads and properly paired reads. The performance of HISAT2 was compared with the BWA aligner (mem algorithm[209]) and Bowtie2, which are two of the most widely used small reads aligners. Table 7 shows the results from this test.

On this real dataset, HISAT2 uses 4.2 gigabytes of RAM, slightly larger than Bowtie2 (3.8Gb RAM), while BWA has the biggest memory usage (9.1Gb RAM). In terms of speed, HISAT2 completes the mapping in 4 hours while both the other aligners take about 4-5 times longer (19 hours 27 minutes for BWA and 17 hours 50 minutes for Bowtie2). In terms of percentage of uniquely mapped reads, HISAT2 closely compares with BWA (86.17% and 86.80% respectively) outperforming Bowtie2 (75.14%).

On the simulated dataset, all aligners perform well, however HISAT2 is over 4 times faster than the others, although uniquely aligning slightly fewer reads (97.75% versus BWA-MEM 100%). These results highlight how HISAT2 performs comparably to BWA and Bowtie2 while keeping a low memory footprint (4.2Gb RAM) and the highest speed (over 4 times faster than the other aligners on the real dataset). All tests were run using 4 threads on a machine with 16GB RAM and two Intel Xeon E5-2670 processors.

BOWTIE2	Simulated reads			Real reads		
	HISAT2	BWA-MEM	BOWTIE2	HISAT2	BWA-MEM	BOWTIE2
Number of reads (Millions)	150	150	150	1250	1250	1250
Time (minutes)	32	130	115	245	1167	1070
Memory fingerprint (Gigabytes)	4.2	6.6	3.8	4.2	9.1	3.8
Aligned reads (%)	99.82	100	99.98	90.14	96.22	93.89
Uniquely aligned reads (%)	97.7	100	99.98	86.17	86.80	75.14
Properly paired reads (%)	99.46	100	52.48	85.61	95.60	62.55

Table 7. DNAscan Alignment assessment results.

HISAT2, BWA and Bowtie2 were tested on 150 million simulated Illumina paired end human reads and 1.250 billion real Illumina paired end human reads. For the three aligners on the two dataset the table shows the time taken, their memory fingerprint and the percentage of aligned-one-or-more-times reads, aligned-only-once reads and properly pared. All tests were run using 4 threads.

4.1.4.8 Variant calling assessment

To assess the performance of the DNAscan variant calling pipeline with the GATK best practice workflow (GATK BPW) and SpeedSeq, we used the exome of the well-studied NA12878 sample and the Illumina platinum calls[210] as a gold standard (our set of true calls). GATK BPW consists of using the BWA aligner, the removal of duplicates with Picard, a base recalibration step and variant calling with GATK-HC. The SpeedSeq pipeline uses BWA for alignment, Sambamba and Samblaster to sort reads and to remove duplicates, and Freebayes for variant calling. Considering the overlap in the software used by DNAseq and SpeedSeq, assessing their performance is therefore of interest. Figure 16A shows the results from this test. DNAscan in Fast mode performs comparably with both the GATK BPW and the SpeedSeq on SNVs. Their F-measure (F_m), a harmonic mean of precision and sensitivity defined in equation 1 (see Methods), is 0.923, 0.911 and 0.928 respectively.

For indels, DNAscan in Fast mode performs poorly (F_m= 0.570). In Normal mode DNAscan shows improvements, reaching an indel calling precision and sensitivity comparable to SpeedSeq (F_m equal to 0.610 and 0.620 respectively). The better performance of the Normal mode is driven by a major increase in sensitivity, which reaches 0.734 from 0.596. However, GATK BPW outperforms SpeedSeq on indels (GATK BPW F_m = 0.815). DNAscan, in Intensive mode, can perform comparably to GATK BPW also on indels with an F_m of 0.820.

Figure 16B shows a comparison of the time needed by the tested pipelines and their memory usage. DNAscan in Fast mode completes the analysis in just 63 minutes while SpeedSeq takes over twice the time (132 minutes) and GATK BPW 5 times longer (310 minutes). DNAscan in both Normal and Intensive mode completes the analysis in a reasonable time (77 and 98 minutes respectively). In terms of memory, DNAscan uses as little as 10GB of RAM in Fast mode, 12GB in Normal and Intensive mode, while GATK BPW 15GB and SpeedSeq over 25GB.

4.1.4.9 ALS Miseq and Whole-Genome-Seq test cases

Using DNAscan in Fast mode, we analysed real DNA sequence data from two ALS patients (case A and case B). Case A carries a non-synonymous mutation in the *FUS* gene[66] variant (NC_00001610:g.[31191418C>G\])[66] known to be a cause of ALS (ClinVar id RCV000017609.27,RCV000017611.25). Case B is a confirmed *C9orf72* expansion carrier (see Methods), this repeat expansion is known to be strongly associated with ALS. A panel of 10 ALS related genes was sequenced with the Illumina Miseq platform for both cases, while 40x WGS data was generated with the Illumina Hiseq 2000 for case B only. The Miseq gene panel was designed and tested for diagnostic purposes[66] (see Methods). For these 10 genes (*BSCL2*, *CEP112*, *FUS*, *MATR3*, *OPTN*, *SOD1*, *SPG11*, *TARDBP*, *UBQLN2*, and *VCP*), the full exon set was sequenced, generating over 825,000 222-base-long paired reads. The WGS sample (paired reads, read length = 150, average coverage depth = 40) was sequenced as part of Project MinE sequencing dataset. For both the samples, SNVs, indels, and SVs were called. For the WGS sample, DNAscan also looked for the *C9orf72* repeat.

For the WGS sample we ran DNAscan on the whole genome. However, both for practical reasons and to simulate a specific medical diagnostic interest, we focus our analysis report on the 126 ALS related genes reported on the ALSoD webserver[211] (http://alsod.iop.kcl.ac.uk/misc/dataDownload.aspx#C5).

Frontotemporal Dementia (FTD) is a neurodegenerative disease which causes neuronal loss predominantly involving the frontal or temporal lobes. Considering its genetic and electrophysiological overlap with ALS[212,213] we report variants linked to FTD as well as to ALS in the following results

Table 8 shows the results from this analysis. For the Case A Miseq DNA gene panel, DNAscan detected 13 SNVs reported to be related to amyotrophic lateral sclerosis and 4 to FTD on ClinVar, 6 non-synonymous variants and 6 variants with a deleteriousness CADD phred score[214] equal to or higher than 13, meaning that they are predicted to be in the top 5% most deleterious substitutions. Finally, the known pathogenic *FUS* SNV rs121909670 was detected. No SVs were found. The whole analysis was performed in ~30 minutes using 4 threads. Since no ALS related repeat expansions are known in these genes, DNAscan was not used to look for any repeat expansions in this analysis.

On the WGS data of Case B, for the selected 126 genes, DNAscan identified 33 SNVs reported to be related to amyotrophic lateral sclerosis and 3 to frontotemporal dementia on ClinVar, 64 non- synonymous variants, 748 variants with a deleteriousness CADD phred score equal to or higher than 13, one 60-base-pair long insertion, 3 over 100,000-base-pair long deletions and 1 tandem duplication. DNAscan was also able to detect the known *C9orf72* expansion.

Table 8 also shows the WGS findings restricted to the same regions sequenced with Miseq (table 8 - WGS panel genes column). Considering these results together with the intersection between the WGS and Miseq results of the same ALS patient, thus case B (Miseq ∩ WGS), we can see how all the variants reported to be linked to ALS/FTD on ClinVar were also detected in the WGS data and no novel variants, among the classes considered in table 8, were spotted. The whole analysis was performed in 8 hours using 8 threads.

4.1.4.10 Virus scanning

We used DNAscan to detect the presence of viral genetic material in a whole genome sequencing sample of HIV infected human cells (see Methods). The DNA sequencing data was produced using the Illumina Hiseq 2000 sequencer generating about 350 million 95-base-long paired reads (see Methods). Following the well-established approach of computational

subtraction of human host sequences to identify sequences of infectious agents like viruses, the human reads (91%, Figure 17-a) were subtracted by mapping the sequencing data to the reference human genome using HISAT2. To screen our sequencing sample for the presence of known viral DNA, HISAT2 was then used to map the unmapped reads from the initial mapping phase of the pipeline (9%, Figure 17-a) to all the viral genomes available on the NCBI virus database.

Figure 18a shows in descending order, the 20 viral genomes to which the highest number of reads were mapped. They show both the presence of HIV DNA and bacterial DNA in our sample. Indeed, 4,412,255 reads mapped to the human immunodeficiency virus (NCBI id NC_001802.1) and only the Escherichia virus phiX174 (NCBI id NC_001422.1), a bacterial virus, presented a comparable number of reads (4,834,017 reads). This phage is commonly found in Illumina sequencing protocols[214].

Figure 18b shows a logarithmic representation of the number of reads aligned to the viral genomes, highlighting a smaller (3-4 orders of magnitude) number of reads belonging to other viruses. The disproportion between the presence of the first two hits (phiX174 and HIV) and the rest of the viruses is also shown in Figure 18b. Discussing this heterogeneous viral genetic material in this sample is beyond the aim of this article. The complete results with the list of the whole set of viruses (120 viruses) for which at least one read was aligned can be found on Github (https://alfredokcl.github.io/sample_report/Virus_test.pdf). The whole screening was performed by DNAscan using 4 threads in ~2 hours

4.1.4.11 Reports and visualisation utilities

DNAscan produces a wide set of QC and result reports and provides utilities for visualisation and interpretation of the results. MultiQC is used to wrap up and visualise QC results of the

sequencing data (FastQC), its alignment (Samtools) and variant calls (Bcftools). An example is available at the following link:
https://alfredokcl.github.io/sample_report/multiqc_report.html

A tab delimited file including all variants found within a selected region is also generated. This report would include all annotations performed by Annovar in a format that is easy to handle with any Excel-like software by users of all levels of expertise. An example is available at the following link: https://alfredokcl.github.io/sample_report/sample_variant_report.txt
Three iobio services are locally provided with the pipeline allowing for the visualisation of the alignment file (bam.iobio[205]), the called variants (vcf.iobio) and for a gene based visualisation and interpretation of the results (gene.iobio)
These utilities are also available for evaluation at the following links: http://bam.iobio.io
http://vcf.iobio.io
http://gene.iobio.io

Various analyses are performed on the mapped sequencing data (Figure 15, right panel): snv and small indel calling is performed using Freebayes, whose reliability is well reported[88,200] However, taking advantage of the documented better performance of GATK HC in small indel calling we decided to add a customised indel calling step to DNAscan, called intensive mode. This step firstly extracts the genome positions for which an insertion or a deletion is present on at least one read, and secondly calls indels using GATK HC on these selected positions. The reduced number of positions where this occurs allows for a targeted use of GATK HC, limiting the required computational effort and time. Resulting SNV and small indel calls are finally hard filtered with Bcftools (see methods).

Two Illumina developed tools, Manta and Expansion Hunter are used for detecting medium and large structural variants (> 50bp) including insertions, deletions, translocations, duplications and known repeat expansions. These tools are optimised for high speed and can analyse a 40x WGS sample in about 30 minutes using 8 threads, maintaining a very high performance.

DNAscan also has options to scan the sequencing data for microbial genetic material. It performs a computational subtraction of human host sequences to identify sequences of infectious agents including viruses, bacteria or fungi, by aligning the non-human or unaligned reads to the whole NCBI database[201] of known viral, bacterial or any custom set of microbial genomes and reporting the number of reads aligned to each non-human genome, its length and the number of bases covered by at least one read.

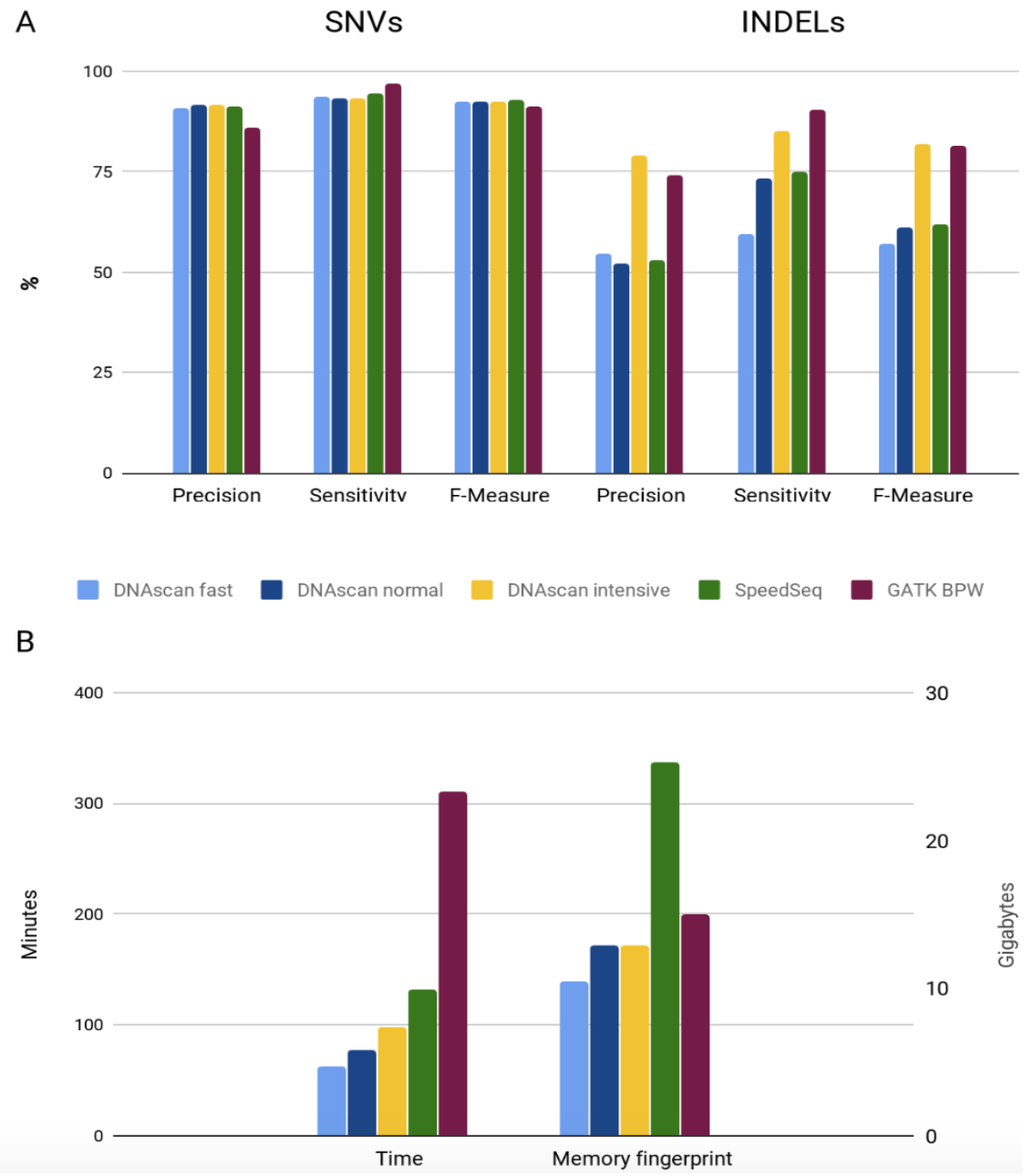


Figure 16.

DNAscan Variant calling assessment.

Graph A shows the precision, sensitivity and F-measure of DNAscan in fast, normal and intensive mode, SpeedSeq and GATK best practice workflow (BPW) in calling SNVs and small indels over the whole exome sequencing of NA12878. Illumina platinum calls were used as true positives. The first three columns show the results for SNVs and the last three columns for indels. Graph B shows the time needed and the memory fingerprint for the same pipelines.

	MISEQ		Miseq ∩ WGS Case B	WGS	
	Case A	**Case B**		**gene panel Case B**	**ALS genes Case B**
Analysis time (minutes)	30	30	---	---	460
Data size (MBs)	40	40	---	---	70000
N. of ALS-related variants	13	11	11	11	33
N. of FD-related variants	4	1	1	1	3
N. of non-synonymous variants	6	7	3	3	64
N. of variants with CADD>13	6	9	4	4	748
N. long insertions	0	0	0	0	1
N. long deletions	0	0	0	0	3
N. Duplications	0	0	0	0	1
N. Inversions	0	0	0	0	0
C9orf72 expansion	---	---	---	---	Yes
rs121909670	Yes	---	---	---	---

Table 8. WGS panel genes column.

Table 8 shows the WGS findings restricted to the same regions sequenced with Miseq (table 8 - WGS panel genes column). Considering these results together with the intersection between the WGS and Miseq results of the same ALS patient, thus case B (Miseq ∩ WGS), we can see how all the variants reported to be linked to ALS/FD on ClinVar were also detected in the WGS data and no novel variants, among the classes considered in table 8, were spotted. The whole analysis from fastq to annotated vcf file and human readable summary results was performed in 8 hours using 8 threads.

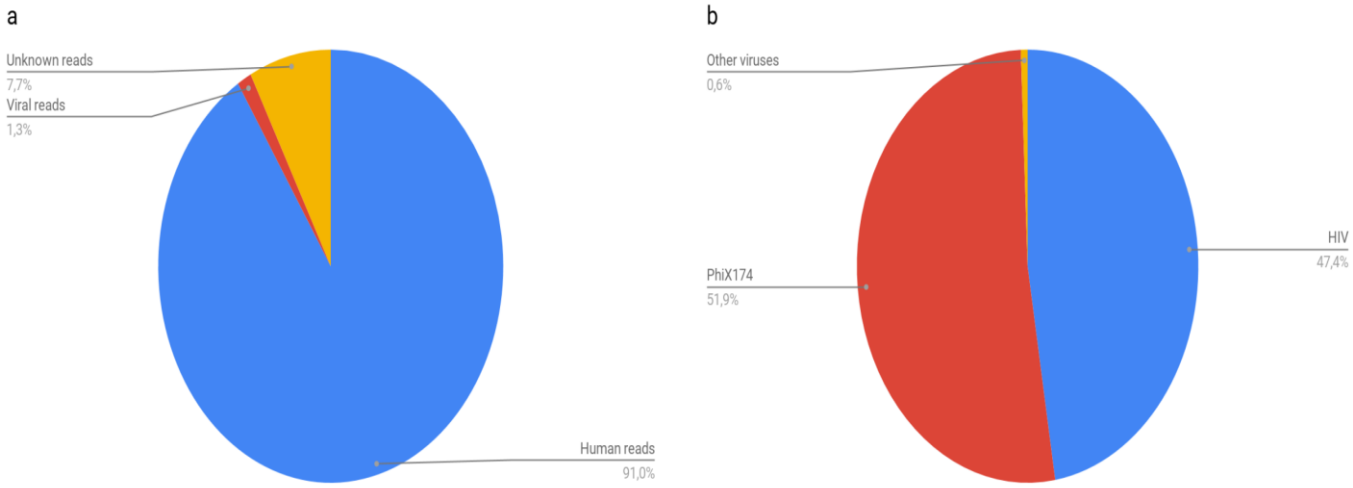


Figure 17. Comparison of Human and Virus reads captured by DNAscan.

Pie chart shows the proportion of human reads (blue), viral reads (red) and unknown reads (yellow). Pie chart b shows the proportion for viral reads belonging to HIV (blue), PhiX174 (red) and to other viruses (yellow). Human reads are defined as reads which aligned to the human reference genome, viral reads as the reads which did not align to the human reference genome but aligned to at least one of the NCBI viral genomes and unknown reads as the reads which did not align neither to the human nor to any viral reference genomes.

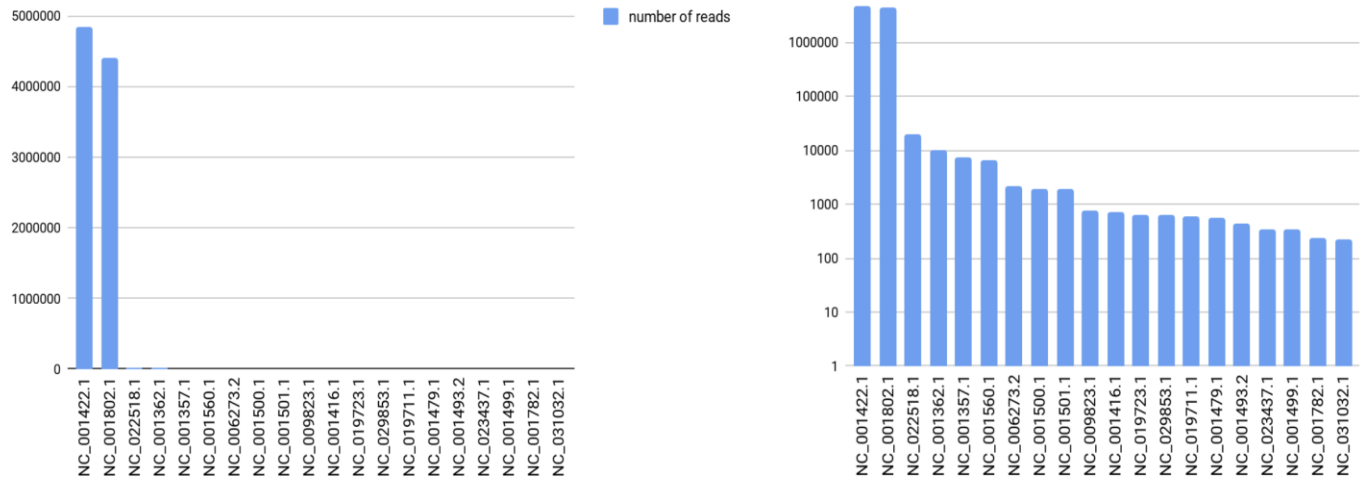


Figure 18. numbers of aligned reads aligned using NCBI database of viral genomes.

The reads which HISAT2 failed to align to the human reference genome were aligned to whole NCBI database of viral genomes. In the graphs we plotted numbers of aligned reads in linear (a) and logarithmic (b) scale, for the 20 viral genomes to which the highest number of reads was aligned.

4.1.5 Discussion

DNAscan is an extremely fast and computationally efficient pipeline for the analysis, annotation and visualisation of NGS DNA sequencing data. DNAscan can analyse 40x WGS data in 8 hours and WES data in one hour on a mid-range commercial computer. DNAscan also provides utilities to favour a user friendly visualisation and interpretation of its outcomes.

After the assessment showing a positive outcome of HISAT2 versus BWA and Bowtie2, we showed the performance of DNAscan is comparable to the widely used GATK BPW. Its three running modes (fast, normal and intensive) allow the user to tailor the pipeline to their needs, while reducing the time and RAM needed compared to the current GATK BPW and SpeedSeq.

I also reported a few use cases such as the analysis of Miseq and WGS data of an ALS case for diagnostic purposes and the virus screening of HIV infected human cells. In the ALS test I showed how with both technologies DNAscan detected a range of reported ALS related variants in a reasonable time (half an hour for the Miseq panel and 8 hours for the WGS) reporting the presence of both the C9orf72 expansion and rs121909670. In the HIV test, DNAscan detected the expected viral presence by finding both the HIV virus and a phage such as PhiX174 commonly present in Illumina NGS DNA data. It is important to highlight that contamination and unknown sequence reads can come from multiple sources including but not limited to sequencing controls such as PhiX, cloning vectors, adapters, PCR primers, nucleic acid impurities present in reagents required for sample isolation and preparation, sequencing artifacts and human error[215].

Other DNA NGS data analysis pipelines exist. Omictools currently lists 101 such tools. Most of which do not cover the whole data analysis, annotation and visualisation process and are computationally more intensive. Among these SpeedSeq and GATK BPW are two of the most

popular ones. SpeedSeq is a framework for fast genome analysis and interpretation. Analysing a 40x WGS sample with SpeedSeq would take 10 hours on a machine with 32 CPU and 126G of RAM. GATK BPW is a pipeline designed and developed by the Broad institute (https://www.broadinstitute.org) which aims to provide the community with best practice standard pipelines and software for the analysis of NGS data. At present it can be used for SNV and small indel calling and analysing a WGS sample with the GATK BPW would take about 24 hours on a machine with 32 CPU and 126G of RAM. Compared to these pipelines, DNAscan requires a substantially lower computational effort, provides user friendly utilities for the visualisation and interpretation of results and allows the user to screen the NGS data for microbial genetic material and detect known repeat expansions. Taking into consideration the well reported involvement of microbes and the role of repeat expansions in several diseases of genetic interest[215,216], we believe both these analyses to be valuable research tools.

Cloud computing and storage services offer practically unlimited computational power and storage. However, this has a cost and its optimisation, in particular for large scale sequencing projects, is not of secondary importance. Amazon Web Services (AWS) is one of the most popular cloud computing services. Performing the alignment, variant calling and annotation using DNAscan fast mode on an EC2 instance (https://aws.amazon.com/ec2/pricing/on-demand/) would cost about $2.96 (8 hours of usage of a t2.2xlarge machine with 8 cpus). The same analysis using SpeedSeq would cost about $22.28 (10 hours of usage of a c3.8xlarge machine with 32 cpus). These prices do not take into account the storage, were updated on 28th of January 2017 and take into consideration the cheapest machines available in the US East (Ohio) region matching the pipeline computational requirements (8 cpus and 16Gb of RAM for DNAscan and 32 cpus and 128Gb of RAM for SpeedSeq[207].

DNAscan is also available as a docker and a singularity image. These allow the user to deploy quickly and reliably the pipeline on any machine where either of these softwares is available. Singularity also allows for the deployment of the pipeline on environments for which the user does not have root permission. This could be particularly useful for users working on shared HPC facilities.

4.1.6 Conclusions

DNAscan is a novel and promising pipeline which can be used for genetic research as well as medical research. It is suitable for both small and large-scale analysis. Covering the whole end-to-end analysis process, from sequencing data in fastq format to the results visualisation, generating user friendly reports and providing result navigation utilities. DNAscan aims to be suitable for a wide audience of users, ranging from research and medical workers with basic command line usage knowledge to expert bioinformaticians.

4.2 ALSgeneScanner: a pipeline for the analysis and interpretation of DNA NGS data of ALS patients

4.2.1 Introduction

Many gene variants have been identified that drive the degeneration of motor neurons in ALS, increase susceptibility to the disease or influence the rate of progression[95]. The ALSoD webserver[65] lists more than 120 genes and loci which have been associated with ALS, although only a subset of these have been convincingly shown to be ALS-associated[3], demonstrating one of the challenges of dealing with genetic data–interpretation of findings. Next-generation sequencing provides the ability to sequence extended genomic regions or a whole-genome relatively cheaply and rapidly, making it a powerful technique to uncover the genetic architecture of ALS[66]. However, there remain significant challenges, including interpreting and prioritizing the found variants and setting up the appropriate analysis pipeline to cover the necessary spectrum of genetic factors, which includes expansions, repeats, insertions/deletions (indels), structural variants and point mutations. For those outside the immediate field of ALS genetics, a group that includes researchers, hospital staff, general practitioners, and increasingly, patients who have paid to have their genome sequenced privately, the interpretation of findings is particularly challenging.

The problem is exemplified by records of *SOD1* gene variants in ALS. More than 180 ALS-associated variants have been reported in *SOD1*[65]. In most cases, the basis of stating the variant is related to ALS is simply that it is rare and found in *SOD1*. Neither of these conditions is sufficient for such a statement to be made. The p.D91A variant, for example, reaches polymorphic frequency in parts of Scandinavia, and yet has been convincingly shown to be causative of ALS. A few variants have been modelled in transgenic mice, shown to segregate with disease or have other strong evidence to support their involvement[59,217] but most do not

have such support. Rare variation can be expected to occur by chance, and its existence in a gene is not evidence of relationship to a disease, making interpretation of sequencing findings difficult. Although various tools are available to predict the pathogenicity of a protein-changing variant, they do not always agree, further compounding the problem.

We therefore developed ALSgeneScanner, an ALS-specific framework for the automated analysis and interpretation of DNA sequencing data. The tool is specifically targeted for use by people with knowledge outside genetics.

4.2.2 Methods

4.2.2.1 Gene and loci prioritization

ALSgeneScanner groups genes and loci associated with ALS into three classes: i) genes for which a manual scientific literature review identified a strong and significant association with the disease or influence on the phenotype in ALS (see Table 1), ii) genes in which variants of clinical significance have been reported on ClinVar[179] and for which no contradictory interpretation is present, and iii) genes for which any association evidence has been submitted to ALSoD[218]. The union of these three sets (available on Github) of genes is used to restrict the genome analysis. However, ALSgeneScanner allows the user to use a custom list of genes.

4.2.2.2 Variant prioritization

The pathogenicity prediction programmes, SIFT, Polyphen2 HDIV, Polyphen2 HVAR, LRT, MutationTaster, MutationAssessor, Fathmm, PROVEAN, Fathmm-MKL coding, MetaSVM and CADD are used to prioritize variants. A variant is scored X where X is equal to the number of tools which predict it to be pathogenic. A higher priority is given to variants which are reported to be “likely pathogenic” or “pathogenic” on ClinVar. For each tool we used the

authors' recommendations for the categorical interpretation of the variants. For each variant, the score ranges between 0 and 11 according to the number of computational tools (11 in total) that predict it to be pathogenic.

4.2.2.3 Whole-genome sequencing

The whole-genome sequencing (WGS) sample used to assess the computational performance of ALSgeneScanner was sequenced as part of Project MinE. Venous blood was drawn from patients and controls and genomic DNA was isolated using standard methods. DNA integrity was assessed using gel electrophoresis. All samples were sequenced using Illumina's FastTrack services (San Diego, CA, USA) on the Illumina Hiseq X platform. Sequencing was 150bp paired-end performed using PCR-free library preparation and yielded ~40x coverage across each sample.

4.2.2.4 Whole-exome sequencing

To assess the computational performance of ALSgeneScanner we also used the Illumina Genome Analyzer II whole exome sequencing of NA12878 (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20101201_cg_NA12878/NA12878.ga2.exome.maq.raw.bam)

4.2.2.5 VariBench and ClinVar datasets

To assess our variant prioritization approach, we used a set of non-synonymous variants from the VariBench dataset for which the effect is known, and all ALS-associated non-synonymous variants stored in ClinVar (71 benign and 121 pathogenic). The VariBench variants are not ALS genes specifically, but because they are all annotated depending on whether or not they are deleterious, the general principles of the method could be tested. The dataset includes VariBench protein tolerance dataset 1

(http://structure.bmc.lu.se/VariBench/tolerance_dataset1.php) comprising 23,683 human non synonymous coding neutral SNPs and 19,335 pathogenic missense mutations [18].

4.2.2.6 Evaluation of performance

Receiver Operating Characteristic (ROC) curves and their corresponding area under the curve statistic (AUC) were calculated using *easyROC*[219]. Accuracy, Precision and Sensitivity are defined as in Equation 1 where T_p is true positives, F_p false positives, F_n false negatives and T_n true negatives.

Equation 1: $$Precision = \frac{T_p}{T_p + F_p} \; ; \; Sensitivity = \frac{T_p}{T_p + F_n} \; ; \; Accuracy = \frac{T_p + T_n}{T_p + T_n + F_n + F_p}$$

4.2.2.7 Hardware

All tests were performed on a single, mid-range, commercial computer with 16GB RAM and an Intel i7-670 processor.

4.2.2.8 Output

Resulting variants are reported in a tab delimited format to favour practical use of worksheet software such as iWork Number, Microsoft Excel or Google Spreadsheets.

4.2.3.9 Software

ALSgeneScanner is available on GitHub (https://github.com/KHP-Informatics/DNAscan). The repository provides detailed instructions for tool usage and installation. A bash script for an automated installation of the required dependencies is also provided as well as Docker[220] and Singularity[184] images for a fast and reliable deployment. A Google spreadsheet with the

complete list of genes and loci used by ALSgeneScanner is publicly available to visualize and comment (see Github repository).

4.2.3 Results

Dr Iacoangeli and I conceived planned the study. Code writing was done by Dr Iacoangeli. All the data in this chapter was generated by me, and I contributed to the analysis, and to the writing of the paper. I am second author on a paper in The Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration journal describing the work in this chapter.

Manual literature review identified 36 genes and 2 loci (Table 9) with strong and reproducible supporting evidence of association with ALS or influence on phenotype. ClinVar reported SNVs and small indels in 44 genes and 7 structural variants ranging in size from 3 to 50 million base pairs. ALSoD reported variants in 126 genes and loci. The union of these sets of genes contained 150 genes and loci. The Venn diagram in Figure 20 shows the overlap between the three sets.

Using a midrange commercial computer (4 CPUs and 16 gigabytes of RAM), (Figure 21) ALSgeneScanner could analyse 40x WGS data of one individual in about 7 hours using 12.8 Gb of RAM, and whole-exome sequencing data in 1 hour and 20 minutes using 8.5 Gb of RAM.

We tested the computational score that the tool used to rank variants on two datasets. The VariBench dataset which includes ~40,000 variants of known effect (pathogenic and benign), and on the ALS associated ClinVar entries. Figure 22 shows the results on the two datasets and Table 2 precision, sensitivity and accuracy of the method in function of the chosen threshold.

The ROC curve for the VariBench dataset (Figure 22, AUC = 0.90) suggests a cut-off equal to 9 which maximises the accuracy (0.827) however, a lower or higher cut-off can be chosen to reach a better precision or sensitivity according to the user's needs. For example, for diagnostics a higher sensitivity is generally required and a cut-off equal to 5 would increase the sensitivity to 0.90 (Table 10). The ROC curve for the ClinVar variants suggests a cut-off equal to 7. Comparing to the VariBench variants, ClinVar ALS variants are more difficult to assess. The AUC for such variants is 0.82 (Figure 22) and the accuracy for the ideal cut-off is 0.75 (Table 10). This performance drop can be explained by the following factors: first the uncertainty in the ClinVar entries. ClinVar provides the community with an infrastructure to allow researchers to store their clinical observations, but the quality checks are very limited and the only filter we have adopted in this study to select the variants was the absence of contradictory entries. The second is the difficulty that available computational tools have in assessing the effect of ALS related variants[3,108], in part because the mechanism of ALS is unknown, and in part because at least some of the variants result in a toxic gain of function that is difficult to understand or model.

Gene	Associated ND		Key Reference
ANG	ALS/PD		10.1038/srep41996.
ANXA11	ALS		10.1126/scitranslmed.aad9157
APOE		Longer survival	10.1212/WNL.58.7.1112
ATXN2	ALS		10.1038/nature09320
CAMTA1		Shorter survival	10.1001/jamaneurol.2016.1114
C21orf2	ALS		10.1038/ng.3622
C9orf72	FTD/ALS	Primarily bulbar onset	10.1016/j.neuron.2011.09.011.
CCNF	FTD/ALS		10.1038/ncomms11253
CHCHD10	FTD/ALS		10.1093/brain/awu138
DAO	ALS		10.1073/pnas.0914128107
DCTN1	ALS		10.1212/01.WNL.0000134608.83927.B
EPHA4		Longer survival	10.1038/nm.2901
FIG4	ALS		10.1038/ejhg.2016.186
FUS	FTD/ALS	Early age of onset and shorter survival	10.1126/science.1165942
HNRNPA1	ALS		10.1038/nature11922
IDE		Shorter survival	10.1001/jamaneurol.2016.1114
MATR3	ALS		10.1038/nn.3688
MOBP	ALS		10.1038/ng.3622
NEK1	ALS		10.1038/ng.3626
OPTN	ALS		10.1126/science.aaa3650
PFN1	ALS	Limb onset	10.1038/nature11280
PGRN	FTD/ALS		10.1126/science.1077209
SARM1	ALS		10.1038/ng.3622
SCFD1	ALS		10.1038/ng.3622
SOD1	ALS	Limb onset, early age of onset and shorter survival	10.1038/362059a0
SPG11	ALS		10.1126/science.aaa3650
SQSTM1	FTD/ALS		10.1001/archneurol.2011.250
SETX	ALS		10.1086/421054
TAF15	ALS		10.1073/pnas.1109434108
TARDBP	FTD/ALS		10.1126/science.1154584
TBK1	ALS		10.1038/ng.3622
TUBA4A	FTD/ALS		10.1016/j.neuron.2014.09.027
UBQLN2	FTD/ALS		10.1038/nature10353
UNC13A	ALS	Shorter survival	10.1038/ng.3622
VAPB	ALS		10.1086/425287
VCP	FTD/ALS		10.1016/j.neuron.2010.11.036
8p23.2	ALS		10.1038/ng.3622
1p34- rs3011225		Late age of onset	10.1016/j.neurobiolaging.2012.07.017

Table 9. List of ALS genes identified by literature review

	VariBench variants			*ClinVar ALS variants*		
Score	Precision	Sensitivity	Accuracy	Precision	Sensitivity	Accuracy
0	0.430	1	0.430	0.612	1	0.613
1	0.507	0.990	0.581	0.659	0.957	0.670
2	0.549	0.978	0.644	0.707	0.949	0.728
3	0.580	0.950	0.682	0.745	0.949	0.770
4	0.618	0.928	0.721	0.754	0.889	0.754
5	0.653	0.900	0.751	0.798	0.812	0.759
6	0.692	0.875	0.779	0.832	0.761	0.759
7	0.736	0.841	0.801	0.863	0.701	0.749
8	0.783	0.796	0.817	0.911	0.615	0.728
9	0.845	0.731	0.827	0.926	0.538	0.691
10	0.919	0.635	0.819	0.931	0.462	0.649
11	0.954	0.436	0.748	0.925	0.316	0.565

Table 10. ALSgeneScanner variant prioritization performance

4.2.4 Discussion

We have developed ALSgeneScanner, a fast, efficient and complete pipeline for the analysis and interpretation of DNA sequencing data in ALS, targeted for use by non-geneticists. The method is able to distinguish pathogenic from non-pathogenic variants with 83% accuracy and reports findings in a simple format, able to be exported for further analysis. With the decreasing costs and increasing availability of next generation sequencing, health care professionals and motivated patients are progressively more likely to have whole genome sequence data available, without the tools to interpret findings. An automated system to provide a meaningful report therefore has a potentially important part to play in giving patients ownership of their data and arming them with the knowledge to understand it.

Omictools[178], a web database where available bioinformatics tools are listed and reviewed, lists over 7000 such tools for next generation sequencing, including more than 100 pipelines; given the great interest in this field, new tools are frequently released. As a result, designing a bioinformatics pipeline for the analysis of next generation sequencing data, keeping the system simple to use on a standard computer and translating the output into a format that is easily understood, is not trivial, and requires specialised expertise. The computational effort and the informatics skills required to use typical pipelines can dramatically limit the use of next generation sequencing data. Adequate high-performance computing facilities and staff specialised in informatics are not always present in medical and research centres. Furthermore, the use of cloud computing facilities, which could theoretically provide unlimited resources, is not always possible due to privacy and ownership issues, cost and the expertise required for their use. To this end, ALSgeneScanner is computationally light as it can run on a midrange commercial computer, is easy to use since it performs sophisticated analyses using few command lines and is comprehensive, including the necessary analyses to identify all known

ALS associated genetic factors. Finally, a tab-delimited output, in which the analysis results are enriched with information from several widely used databases such as ClinVar and ALSoD, as well as information from our manual literature review, and the graphical visualization utilities integrated in the pipeline as part of DNAscan, favour an easily accessible interpretation of the results.

Our table of sensitivity, specificity and accuracy (Table 10) means that the appropriate cut-off can be used to interrogate data, depending on whether the aim is the exclusion of potentially harmful variants, or the detection of definitely harmful variants.

4.2.5 Conclusion

ALSgeneScanner puts a powerful bioinformatics tool, able to exploit the potentialities of next generation sequencing data in the hands of patients, ALS researchers and clinicians.

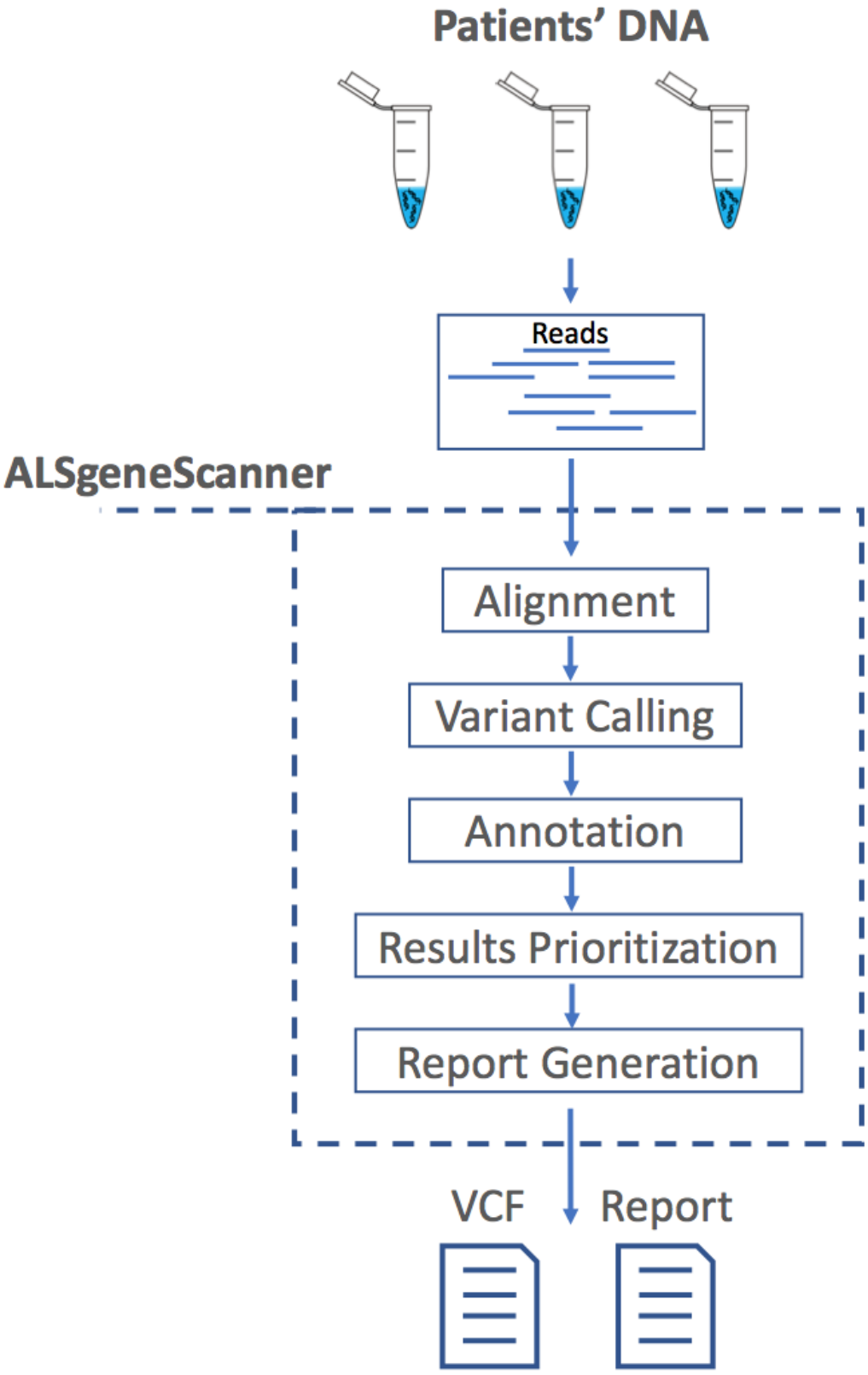


Figure 19. ALSgeneScanner pipeline main steps. From sequencing data in fastq format to the report generation of the results.

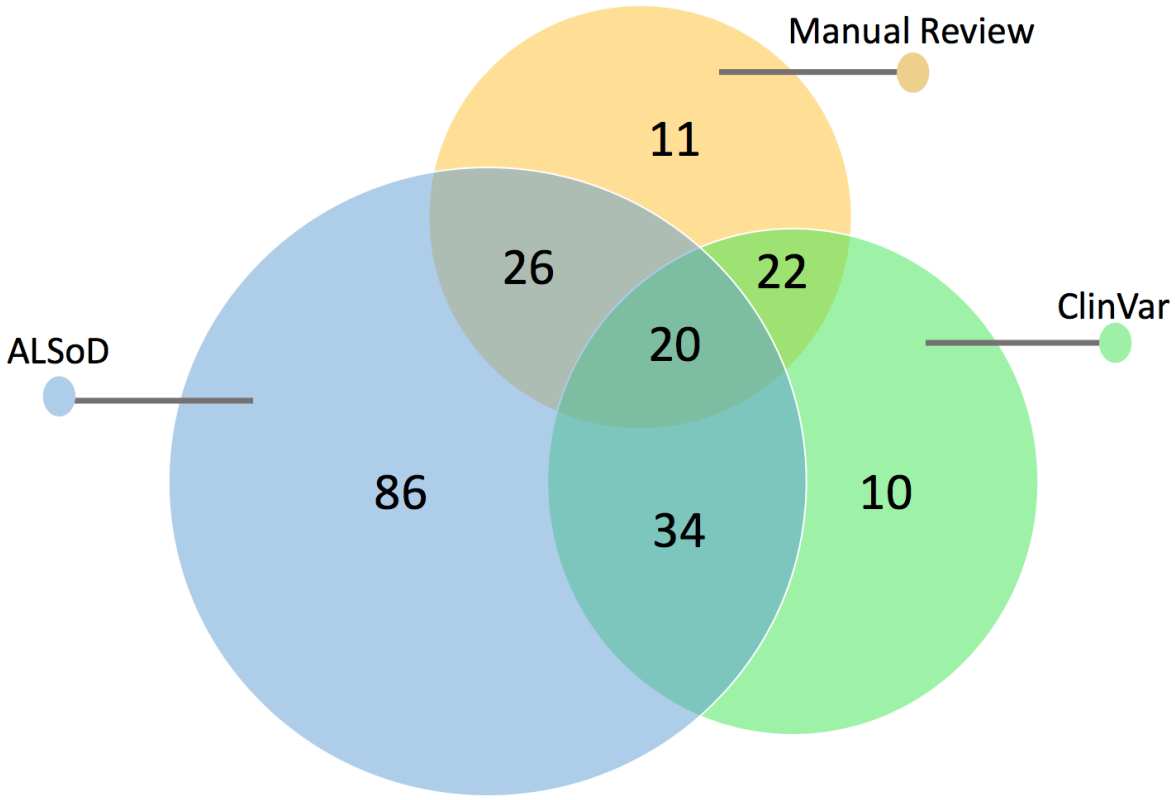


Figure 20. Number of ALS selected genes used in ALSgeneScanner. Venn diagram of the ALS related genes that we selected in our literature review, found in the ALSoD webserver and in the ClinVar database.

Figure 21. Computational performance of the ALSgeneScanner pipeline.

Computational performance of the ALSgeneScanner pipeline to process whole genome sequence and whole exome sequence data from fastq file to the generation of the final result report.

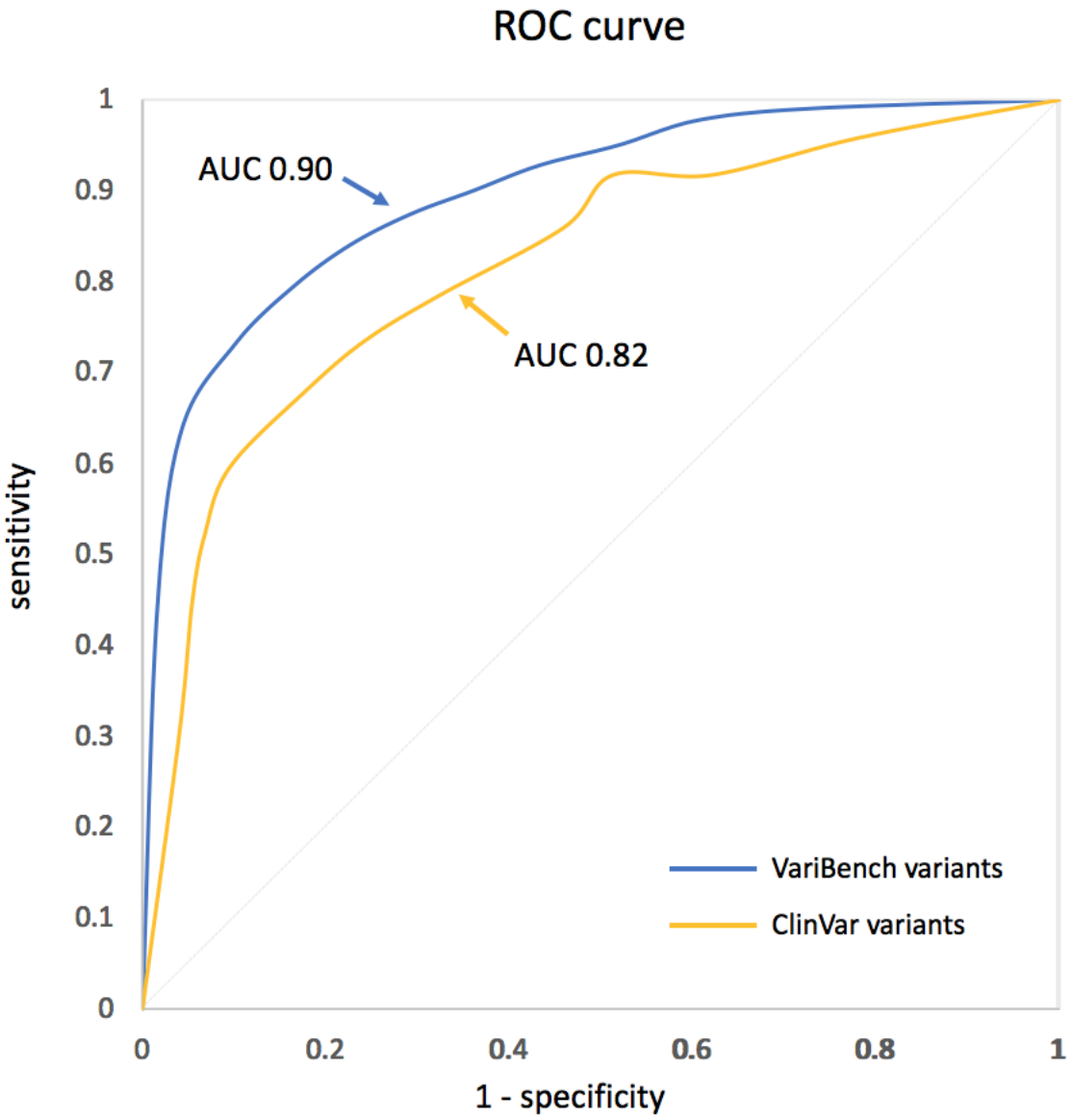


Figure 22. Comparison of performance of ALSgeneScanner in VariBench and ClinVar datasets. ROC curve of the performance of ALSgeneScanner in VariBench and ClinVar datasets.

Chapter 5. Telomere length analysis in ALS

5.1 Next Generation Sequence Analysis of Telomere length in ALS

5.1.1 Introduction

Although the heritability of ALS is about 60%[5], the heritability explained by common gene variants is only about 11% [120] suggesting that other forms of genetic variation play an important role. Such variants may also play a part in controlling disease progression, but are not well studied. One such sequence variation is the telomere.

Telomeres are repeated DNA sequences located at the ends of chromosomes and exist to maintain DNA integrity during cellular replication; chromosome ends tend to shorten with replication, and the repeat region protects against the loss of important gene sequences because loss of repeats can be tolerated[221]. As such, telomeres shorten naturally with age as repeats are lost during replication cycles[222]. Natural variation in telomere length exists in the population, with women on average having longer telomeres than men[223]; shorter telomeres are associated with an increased risk of cancer[224].

A major risk factor for ALS is age[4], and ALS is also more common in men than women[104]. Furthermore, there are some similarities between ALS and cancer, such as evidence for a multistep process in pathogenesis[102,140]. All these factors are related to telomere length. We therefore investigated telomere length in ALS. The extent of my contribution to the work in this chapter is explicitly described in the Appendix. I am first author of a paper describing this work.

5.1.2 Methods

5.1.2.1 Whole-genome sequencing

Samples were from multiple centres across the UK contributing to the international Project MinE whole genome sequencing initiative[74].

DNA was isolated from venous blood using standard methods. The DNA concentrations were set at 100ng/ul as measured by a fluorometer with the PicoGreen® dsDNA quantitation assay. DNA integrity was assessed using gel electrophoresis. All samples were sequenced using Illumina's FastTrack services (San Diego, CA, USA) on the Illumina Hiseq 2000 platform[225]. Sequencing was 100bp paired-end performed using PCR-free library preparations and yielded ~40x coverage across each sample. Binary sequence alignment/map formats (BAM) were generated for each individual.

5.1.2.2 Determination of Telomere Length

TelSeq[226] was used to quantify telomere length using data from whole genome sequences. Telomere lengths were estimated from reads, defined as repeats of more than seven TTAGGG motifs.

5.1.2.3 Assessment of nine loci affecting mean telomere length and their association with ALS

We selected nine SNPs, reported in multiple genome-wide association studies (GWAS) as associated with mean telomere length in European-derived populations[227–230]. The selected SNPs were rs6772228-*PXK*[227], rs9419958-*OBFC1*[229], rs9420907-*OBFC1*[228], rs4387287-*OBFC1*[230], rs3027234-*CTC1*[229], rs8105767-*ZNF208*[228], rs412658-*ZNF676*[229], rs6028466-*DHX35*[229], and rs755017-*ZBTB46*[229].

5.1.2.4 Statistical Analysis

The effects of telomere length on ALS were tested using a generalised linear regression model, which included total telomere length, age and sex to predict disease affected status. To assess the model, Pearson's chi-squared test was used.

Because telomere length correlates with age, we performed an additional test to examine the possibility that survival bias could affect the results. To do this, we also performed the analysis restricted to the subgroup of people with ALS onset below the median cohort age (62 years). Although such an analysis would halve our sample and therefore greatly reduce statistical power, the direction of effect should be observable.

To evaluate SNP effects on telomere length we calculated Nagelkerke's R^2 from the results of a generalised linear model using the value of telomere length, age, gender, and nine SNPs selected for having been previously shown to associate with telomere length.

To assess the effect of covariates on telomere length affecting survival, we used Cox regression, controlling for age, gender and site of disease onset (bulbar or spinal).

To assess the association of genes with ALS we used the SNP-set Sequence Kernel Association Test (SKAT)[231], which is a test for association between a set of rare and common variants and continuous/dichotomous phenotypes using kernel machine methods.

Statistical tests were performed using IBM SPSS Statistics 24.0 (SPSS Inc., Illinois)[156], RStudio 1.0.143 (RStudio Inc., Boston), R Foundation for Statistical Computing 3.4.1 (R Core Team, V)[157]

5.1.2.5 Ethical approval

Informed consent was obtained from all volunteers included in this project. Generation of whole genome sequences was approved by the Trent Research Ethics Committee 08/H0405/60

5.1.3 Results

I conceived the study, planned the work, collected the data, performed the analyses, and wrote the paper. I am a first author on a paper in The Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration journal describing the work in this chapter

There were 1241 people with apparently sporadic ALS and 335 controls. The median age for people with ALS was 62.5 years and for controls, 60.1 years, with a male-female ratio of 62:38 Table 11.

	ALS	Controls
Total n	1241	335
Male:Female ratio	766:475 (62% male)	124:211 (37% male)
Mean age	62.9 (SD 11.08)	60.1 (SD 11.47)

Table 11. Demographics of the UK sample.

The mean telomere length in people with ALS was 3.95kb, and in controls, 3.80kb, not taking into account gender or age (Figure 23). Generalized linear regression accounting for these covariates showed a mean 9% (95% CI 3%, 15%) increase of telomere length in people with ALS compared to age and gender-matched controls ($p = 0.008$). Covariate analysis showed that females ($p = 0.03$) and younger people ($p = 2 \times 10^{-16}$) had on average longer telomeres (Table 12), confirming the results of earlier studies that telomere length reduces with age and females have on average longer telomeres.

	Estimate	SE of estimate	p-value
Age (per year)	-1%	0.1%	2×10^{-16}
Gender (Male vs Female)	-5%	2%	0.03
Case-Control status (Controls vs Cases)	-9%	3%	0.008

Table 12. Telomere length comparison between people with ALS and healthy controls using a generalized linear model.

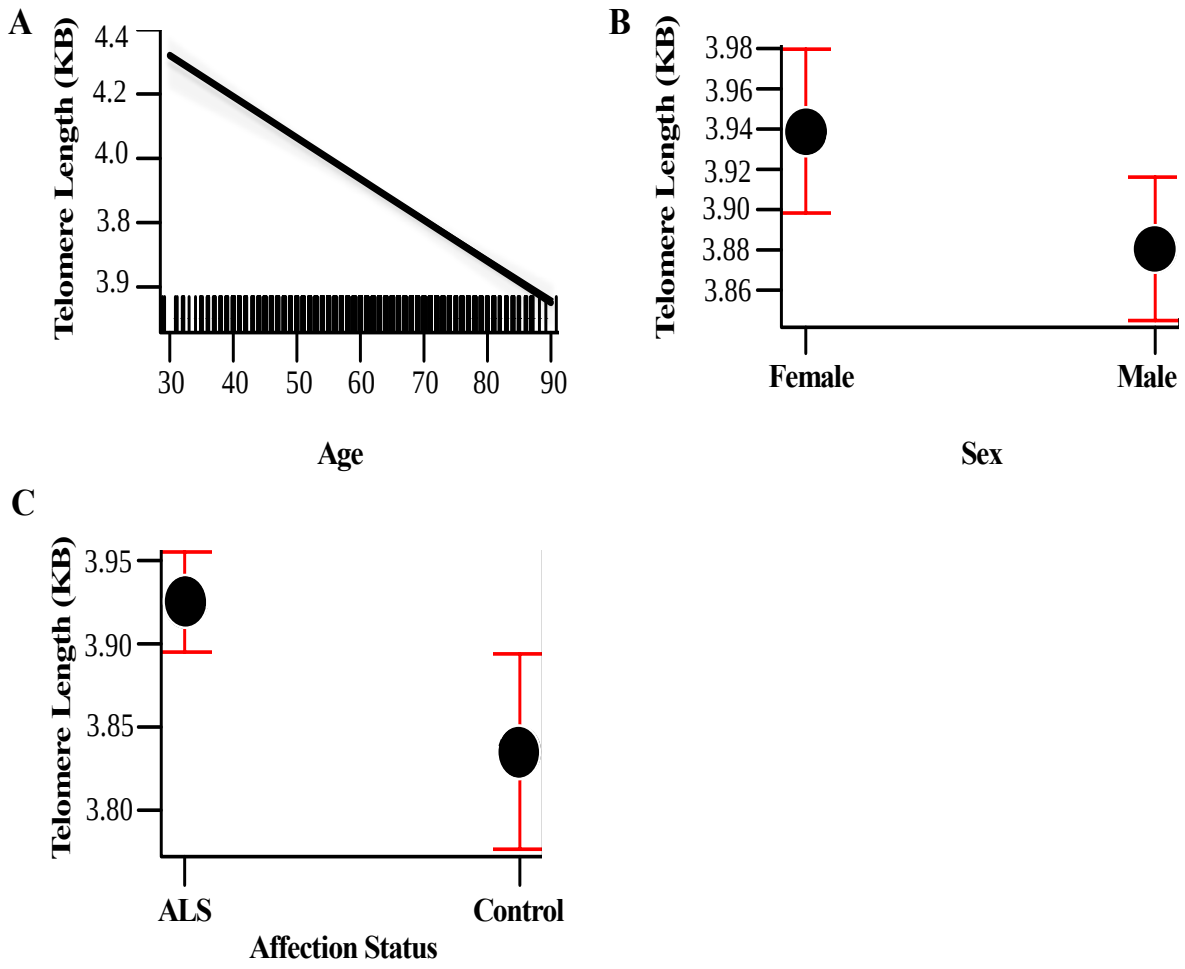


Figure 23.Plotting general linear model covariates in telomere length analysis.

A- Estimate of average telomere length plotted against age in years. B- Telomere length plotted using gender. C- Comparison between telomere length between ALS and controls. All average length estimates in kilobases.

There was no association between telomere length and site of disease onset (p = 0.7), or with *C9orf72* expansion status (p = 0.24). In the analysis exploring survival bias as an explanation

for our results, in which we restricted testing to those younger than the median age, the same direction of effect was observed, although as expected, because of the greatly reduced sample size, this did not reach statistical significance (p = 0.08).

Cox regression analysis showed that in the ALS group, those with longer telomeres had a 16% increase in median survival (hazard ratio 0.81 (95% CI 0.72–0.91), p = 0.001).

The generalised linear regression model showed that of the nine SNPs associated with telomere length, two were also associated with ALS: rs8105767 near the *ZNF208* gene ($p = 1.29 \times 10^{-4}$, MAF=0.03) and rs6772228, which is in an intron for the *PXK* gene (p = 0.001, MAF=0.03; Table 13), but the SKAT test did not show an association of overall variant burden in these genes with ALS after correction for multiple testing (*ZNF208*, p = 0.81 and *PXK*, p = 0.03). Nagelkerke's R^2 test showed that the nine selected SNPs contributed 3% to the variance in total telomere length.

SNP	Risk genotype	Beta	Std. Error	p-value
rs6772228	**A/T**	**0.83**	**0.37**	**0.001**
rs6772228	**A/A**	**0.33**	**0.37**	**0.001**
rs9419958	T/T	0.61	0.56	0.036
rs9419958	C/T	0.59	0.54	0.031
rs9420907	C/C	0.59	0.53	0.021
rs9420907	A/A	0.6	0.53	0.024
rs4387287	A/A	0.71	0.16	0.014
rs4387287	A/C	0.63	0.07	0.067
rs3027234	G/G	0.62	0.07	0.009
rs3027234	A/G	0.63	0.07	0.011
rs8105767	**G/A**	**0.66**	**0.08**	**1.29×10^{-4}**
rs8105767	**G/G**	**0.91**	**0.07**	**4.89×10^{-4}**
rs412658	C/C	0.73	0.07	0.008
rs412658	C/T	0.69	0.06	0.058
rs6028466	G/G	0.36	0.22	0.087
rs6028466	A/G	0.37	0.22	0.088
rs755017	A/A	0.45	0.63	0.054
rs755017	A/G	0.45	0.63	0.052

Table 13. Assessment of telomere associated SNPs. Investigation of causal effect of telomere length on ALS by using 9 lead single-nucleotide polymorphisms (SNPs) identified by telomere length GWAS. SNP rs8105767 (PXK) and rs6772228 (ZNF208) are associated with ALS in a UK cohort.

5.1.4 Discussion

We have shown that longer telomeres are associated with ALS and with longer survival in ALS. In keeping with previous studies, we found that mean telomere length was longer in females and shortened with increasing age. Of a panel of nine SNPs known to be associated with telomere length, two showed association with ALS, one in *ZNF208*, and the other in *PXK*.

Although both these SNPs, rs6772228 and rs8105767, are known to be associated with telomere length, no association with ALS was seen in a previous large genome-wide association study[7], suggesting that either there is a population-specific effect, or that the telomere length itself is driving the association, and other factors that influence it have a larger effect than these SNPs. Another possibility is that the difference in results is because the analysis performed was different, as we have tested genotypic association, whereas the genome-wide association study used linear mixed modelling of alleles.

Telomeres have largely been investigated for their roles in cancer and ageing, shorter telomeres being associated with disease pathology and death. Surprisingly, telomere elongation is also seen in about 15% of cancers, such as adenocarcinoma of the lung and pancreas[232], and in general, cancers with long telomeres are resistant to therapy and carry a poor prognosis[233]. Telomere elongation phenomena are well documented but far less well understood than telomere shortening phenomena[233–237].

A study of telomere length in ALS brains found a trend to longer telomeres in glial cells[238] consistent with our results, but is in contrast to an earlier small study of 50 people with ALS and 50 controls, finding that shorter telomeres are associated with ALS[239]. Microglial cells

have been implicated in ALS disease progression[240], although their exact role in ALS is still not known. Previously, it has been shown that microglia utilise telomerase to regulate telomere length in *vivo*, this might indicate a higher proliferative activity of glial cell in ALS[241].

Our study has some strengths and weaknesses. Although ALS is a disease of the central nervous system, our telomere data are derived from leukocyte DNA, since our DNA source was whole blood. The relationship between leukocyte telomeres, which can be expected to shorten with age as leukocytes undergo mitosis, and telomeres in neurons, which are post-mitotic, is not clear[242], but glial and other cells that do undergo mitosis are probably involved in ALS pathogenesis, and provide a possible mechanism. Furthermore, we did not directly measure telomere length using Southern blotting, but estimated it using whole genome sequence data. However, our findings have the advantage of a large sample size of more than 1200 cases, compared with previous reports of 50 or fewer. Furthermore, our examined cohort is more homogeneous in genetic background, and the sequencing technology used was the same across the entire cohort. However, one limitation of our method is that we cannot draw firm conclusions about the exact length of a telomere. The method we have used, Telseq, correlates with results from Southern blotting[226], and Q-PCR[243] and is in widespread use[242,244]. Nevertheless, different sequencing technologies will generate different telomere length estimates because of differences in library preparation and platform [245,246]. To overcome this potential weakness, we have used the same industry-leading sequencing platform for all samples, as well as designing the study to minimize batch effects by having cases and controls sharing the same sequencing plate.

We found that longer telomeres were associated both with ALS and with increased survival in ALS. It is possible that telomere length does not associate with ALS risk but only with survival, and that our cohort was biased in such a way that those with longer survival were more likely to be genotyped. In that case, we would also observe an apparent association with risk, but the driver would be the actual association with increased survival. While this possibility cannot be completely excluded, the cohort tested was an incident cohort, collected from a population rather than a specialist clinic, reducing the likelihood of this explanation. Furthermore, we assessed survival bias by testing the relationship between telomere length and ALS in the younger half of the sample. We found the direction of association of longer telomeres with ALS was still present, although as expected, the statistical power was reduced due to the smaller number of young controls (<175). Replicating these findings in a bigger cohort such as the entire Project MinE sample, is an important future step.

Telomere elongation is maintained by two pathways: telomerase activation or a telomerase-independent mechanism termed alternative lengthening of telomeres (ALT). Using Project MinE data browser[247] and the latest GWAS summary statistics[75] the telomerase gene shows no association with ALS, which suggests involvement of the ALT telomere elongation mechanism. Indeed, our results are consistent with the involvement of ALT for the following reasons: in ALT telomere elongation, telomere length usually results in a heterogeneous length, while in telomerase-based elongation a consistent length is usually observed throughout the study cohorts[232,248]. Furthermore, a common observation in the ALT alerted pathway is extreme elongation in net telomere length (>5kb) in a subset of samples, which is also a consistent with our results[234]. Thus, further investigation of involvement of ALT activation and involvement of telomerase pathway inhibition might be an effective strategy in understanding the observed longer telomere length phenomena in ALS.

The mechanisms of ALT pathway activation are not clear. It has been suggested that nuclear receptors are involved in these recombination processes[232,249,250]. Furthermore, it has been suggested that the exposed ssDNA 3′ telomere overhangs can invade into the sister chromatid or extra-chromosomal telomeric DNA repeats that can be used as templates and can then elongate their ends using the DNA replication machinery resulting in net elongation in telomere length[232,251]

There are multiple methods available for telomere length analysis, including terminal restriction fragmentation, quantitative fluorescence in situ Hybridization (Q-FISH) [252], polymerase chain reaction-based techniques (PCR) and southern blotting. These techniques have the disadvantage of lengthy protocols and limitations, such as the requirement that DNA is extracted from fresh blood, or that chromosomes are individually stained, which is a time-consuming process[245,246,252,253]. Differences in applying these techniques between laboratories can create measurement differences[254]. Thus, for large scale analyses, whole genome sequence data that can be processed using a standard bioinformatics pipeline can standardize measurements and overcome many of these issues[255]. We have shown that measuring telomere length in a UK cohort is feasible using a bioinformatics tool such as TelSeq, and that this is fast and cost-effective. Estimating the telomere length with TelSeq on a single 40x whole genome sequence takes about 90 minutes using 4 threads on a midrange computer, which would translate to about 100 days for our entire dataset. Since high performance computing access is now straightforward, and multiple computers are able to run the same analysis in parallel, the analysis time can easily be shortened significantly.

5.1.5 Conclusion

In this large study of telomere length and ALS, we have shown that longer telomeres in leukocytes are associated with ALS, and with increased survival in those with ALS.

5.2 Gene-based association analysis of rare variants with telomere length in ALS.

5.2.1 Introduction

With growing evidence of the role of genetics in the pathology of complex disease, low frequency variants (minor allele frequency 0.5-5%) and rare variants have become the next target for ALS researchers [66,67]. Rare and low frequency variants usually have large effect sizes which can help find causal genes but generally not tagged by GWAS microarrays (microarray technology profiles genes according to gene expression/functional activity)[66,68,256]. As a result, the application of next generation sequencing, which includes whole genome sequencing and whole exome sequencing, have become powerful tools to understand the genetic aetiology of complex traits[257].

Current thinking is that common diseases are the consequence of the additive effects of small increases in risk from multiple common variations (polygenic), and rare diseases are the consequence of single gene variants that are themselves rare but have a large effect on the probability of disease (monogenic). ALS lifetime prevalence that is far greater than is typical for a monogenic disease but far less than a common disease[258], and it is perhaps, therefore, to be expected that its genetic architecture also seems to sit somewhere between polygenic effects and monogenic high-penetrance disease.[4,140,259]

We have previously shown that longer telomeres are associated with ALS and with longer survival in ALS. In keeping with previous studies, we found that mean telomere length was longer in females and shortened with increasing age. Of a panel of nine SNPs known to be associated with telomere length, two showed association with ALS, one in *ZNF208*, and the

other in *PXK*. The effect of telomere length in health and disease is well documented, but genetic factors that influence telomere length are not well understood. Therefore, we tested the relationship between the length of telomere and rare variations in ALS using a large sample set from the UK.

5.2.2 Methods

5.2.2.1 Whole-genome sequencing

Samples were from multiple centres across the UK contributing to the international Project MinE whole genome sequencing initiative[74].

DNA was isolated from venous blood using standard methods. The DNA concentrations were set at 100ng/ul as measured by a fluorometer with the PicoGreen® dsDNA quantitation assay. DNA integrity was assessed using gel electrophoresis. All samples were sequenced using Illumina's FastTrack services (San Diego, CA, USA) on the Illumina HiSeq 2000 platform[225]. Sequencing was 100bp paired-end performed using PCR-free library preparations and yielded ~40x coverage across each sample. Binary sequence alignment/map formats (BAM) were generated for each individual.

5.2.2.2 Determination of Telomere Length

TelSeq[226] was used to quantify telomere length using data from whole genome sequences. Telomere lengths were estimated from reads, defined as repeats of more than seven TTAGGG motifs.

5.2.2.3 Quality Control and rare variants extraction

VCF Annotation was performed using ANNOVAR Nov2014[203] and Variant Effect Predictor (VEP v84) tools[260] and compared against the ExAC database of genetic variants[261] (http://exac.broadinstitute.org/), to remove common variants, we filtered variants with minor allele frequency more than 0.01. Any locus with significant missingness between cases and controls was excluded using PLINK v1.09[262]. An in-house coverage analysis determined that 92% of the desired regions were covered by at least 5 reads. Variants called with 10–20 reads were flagged to be visually inspected to remove false positives. Furthermore, we excluded variants with a poor genotyping quality, defined as variants with sequencing quality score less than 20 (out of 100) as well as variants with minimal read depth less than 5 base pairs.

Patients were later classified into three groups: group 1, ALS; group 2, ALS with bulbar site of onset ALS; group 3, ALS with spinal site of onset ALS.

5.2.2.4 Statistical Analysis

To compare the rare-variant burden in cases and controls -we used the Variable Threshold (VT) test. This test is based on the assumption that the minor allele frequencies of causal rare variants is different from those of non-functional rare variants[263]. To assess the model, Pearson's chi-squared test was used.

The idea behind this approach is that there exists some (unknown) threshold for which variants with a minor allele frequency are more likely to be functional than are variants with a minor allele frequency above the certain threshold. This is because selection pressure prevents a rare deleterious variant from becoming frequent while a non-functional variant could drift to high frequencies. Variable Threshold works by finding the maximum z-score across all possible values for the threshold.

Because rare variation is too infrequent to test statistically in a sample of this size, we used a region-based test comparing the rare-variant burden in people with ALS with that in controls. We included intronic, exonic, synonymous, coding, and known causal variants only if they were novel or had a minor allele frequency of at most 0.01 and were of high-quality with no bias in the proportion of missing data between cases. To calculate the Variable Threshold, we used R package (AssotesteR) using the following function:

VT(y, X, maf, perm)

y is a numeric vector with phenotype status: 0 = controls, 1 = cases and no missing data is allowed. X is a numeric matrix or data frame with genotype data coded as 0, 1, 2. Missing data is allowed. maf is a numeric value indicating the minor allele frequency threshold for rare variants which must be a positive number between 0 and 1, and is 0.01 by default. Perm is a positive integer indicating the number of permutations (100 by default). Using gene window size, we assessed Variable MAF Threshold with different strategies for weighting variants. To restrict the analysis to rare variants, MAF threshold was set not to exceed the 0.01 frequency.

5.2.3 Results

I conceived and designed the study, planned the work, collected the data, created the pipeline for analysis described, performed the analyses, and wrote the paper. I am a first author on a paper to be submitted describing the work in this chapter.

There were 1128 people with ALS, 330 with bulbar onset (154 males, 176 females) and 798 with spinal onset (534 males, 264 females). The variable threshold analysis showed an associated between telomere length and rare variation on chromosome 5 in ALS (rs1051688193_T/C; $p = 8.6 \times 10^{-9}$; Figure 24), with the T variant associated with longer telomeres. No other variant passed the multiple testing correction threshold $p = 5.0 \times 10^{-8}$ (Table 14) [256,264].

Figure 24 Summary of Telomere Rare Variations identified in ALS individuals

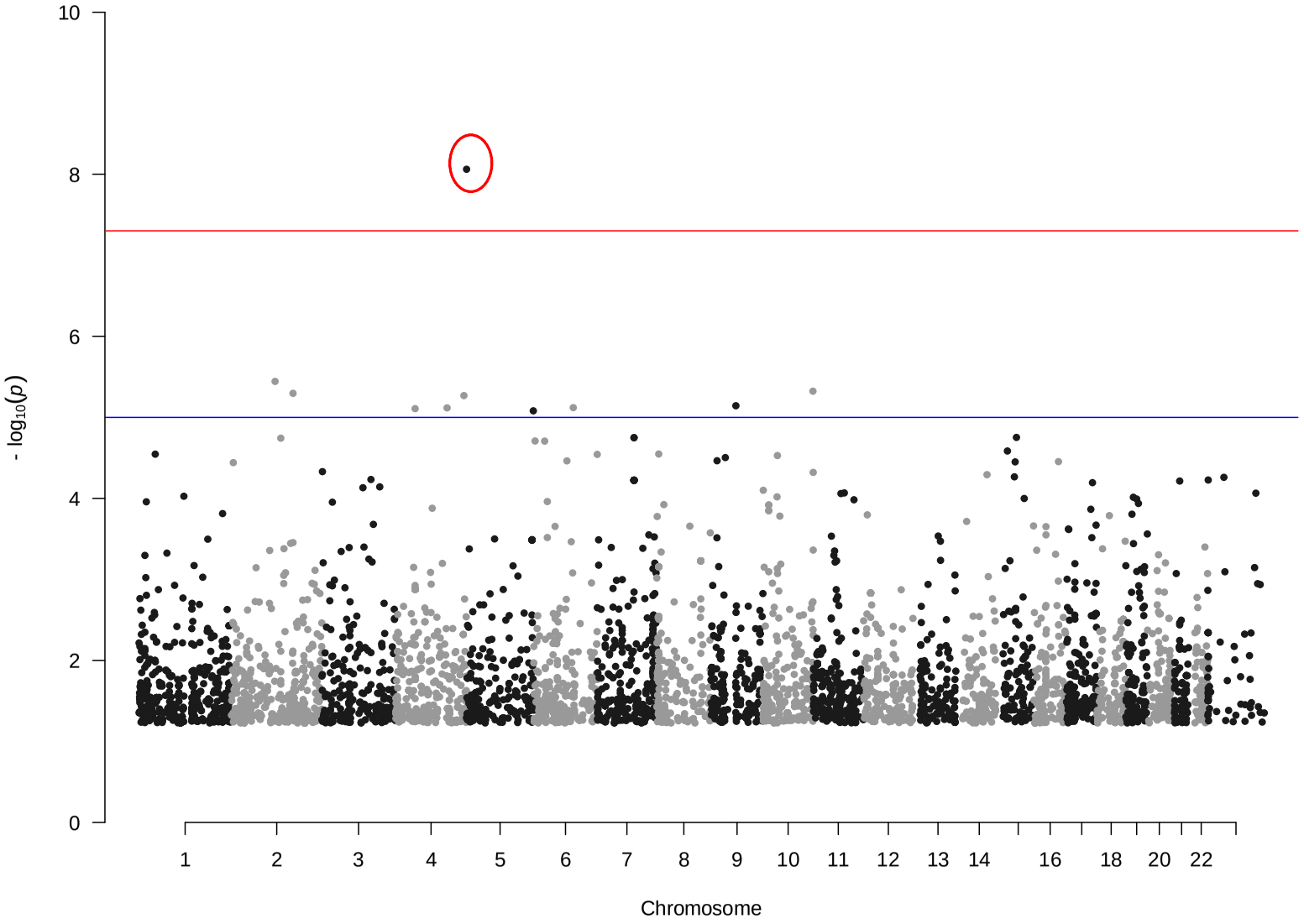


Figure24; Manhattan plot showing the summary of telomere associated rare variations identified in ALS individuals. Variants (5:352349_T/C) is

Summary of Variable Threshold analysis; Top 20 variants

CHROM	BEG	END	MARKER_ID	STAT	PVALUE	BETA	SEBETA	R2
5	352349	352349	5:352349_T/C	5.8015	8.67E-09	1.187	0.2046	0.03078
2	117587211	1.18E+08	2:117587211_G/A	4.6581	3.60E-06	1.463	0.3141	0.02006
10	134192379	1.34E+08	10:134192379_A/T	4.5991	4.75E-06	1.246	0.271	0.01956
2	165876796	1.66E+08	2:165876796_C/T_rs28581396	-4.5858	5.06E-06	-0.8797	0.1918	0.01945
4	184207363	1.84E+08	4:184207363_T/C_rs528195621	4.5726	5.39E-06	1.246	0.2726	0.01934
9	66833568	66833568	9:66833568_G/A_rs199728833	4.5107	7.18E-06	1.099	0.2437	0.01883
6	106301893	1.06E+08	6:106301893_C/T	4.4988	7.59E-06	1.415	0.3146	0.01874
4	138878036	1.39E+08	4:138878036_A/G_rs559567647	4.4974	7.64E-06	1.005	0.2234	0.01872
4	53105994	53105994	4:53105994_T/C_rs11133432	-4.4922	7.82E-06	-0.9212	0.2051	0.01868
5	179580231	1.8E+08	5:179580231_G/A_rs11249676	-4.4787	8.32E-06	-1.4	0.3126	0.01857
15	55927260	55927260	15:55927260_T/C_rs578045011	4.3122	1.77E-05	1.172	0.2717	0.01724
7	98688693	98688693	7:98688693_T/C_rs2176021	-4.31	1.78E-05	-1.045	0.2424	0.01722
7	98706012	98706012	7:98706012_T/C_rs13237374	-4.31	1.78E-05	-1.045	0.2424	0.01722
2	133027652	1.33E+08	2:133027652_G/T_rs796842260	4.3069	1.81E-05	1.356	0.3148	0.0172
6	3687277	3687277	6:3687277_G/A_rs181437615	4.2887	1.96E-05	0.9551	0.2227	0.01706
6	29520988	29520988	6:29520988_T/C_rs754441399	4.2884	1.96E-05	1.348	0.3144	0.01705
15	31508791	31508791	15:31508791_T/C_rs8035866	-4.2241	2.61E-05	-1.033	0.2445	0.01655
8	5906552	5906552	8:5906552_A/G_rs6559085	-4.2047	2.84E-05	-1.321	0.3142	0.01641
1	44463198	44463198	1:44463198_A/G_rs201807166	4.204	2.84E-05	1.148	0.2731	0.0164
6	170739317	1.71E+08	6:170739317_C/T_rs4710825	4.2022	2.87E-05	1.019	0.2426	0.01639

Table 14 Table: Summary of Variable Threshold analysis; Top 20 variants

5.2.4 Discussion

We have examined the relationship between telomere length and rare genetic variation in the whole genome in 1128 individuals with ALS. SNP rs1051688193 is in an intron in the Human Aryl Hydrocarbon Receptor Repressor (*AHRR*) gene. No previous association has been reported between this rare variant and ALS or telomere length and telomere common variants, athough multiple studies have shown an association of *AHRR* gene variants with environmental pollutants such as dioxin (2,3,7,8, tetrachlorodibenzo-p-dioxin). Dioxins are toxic environmental pollutants known to induce an acute inflammatory response, immune suppression and carcinogenesis[265]. Dioxins have been suggested as a possible risk factor for ALS and were found to induce up to a 3-fold increase in TDP-43 protein in human neuronal cell lines (BE-M17 cells), motor neuron differentiated iPSCs, and in murine brain[266].

There are over 30 genes associated with telomere length identified by multi-population genome wide association studies, perhaps targeting rare variants in these genes is an effective strategy to understand how telomere length is maintained[233,267,268].

Our study has several weaknesses. Although this study was done in over 1000 people, it is still underpowered for small effects or rare variants. Furthermore, in rare variant analysis, small sampling differences can result in a very small p-value. Therefore, unless the sample size is very large, the result should be always interpreted with caution. For example, to detect a rare variant with frequency of 0.1% with odds ratio 0.2 and 80% power would need a sample size of 60,000 cases and an equal number of controls[269]. For this reason, this result will need to be replicated in a larger cohort.

5.2.5 Conclusion

In conclusion, we have investigated the association between telomere length and rare genetic variants in ALS. We have identified one rare variant that has an effect on telomere length. Further studies using a larger sample is required to confirm the results.

Chapter 6. Structural variation analysis in ALS

6.1 Introduction

Genome wide association studies have shown considerable success in identifying ALS risk genes through SNP associations. However, these variants tend to have very small effect sizes and can explain only a small portion of heritability. Therefore, in order to understand the disease risk conveyed by genetic variation, it is crucial to consider other genomic variants such as genomic structural variations.

Structural variants (including copy number variations) comprise different forms of genomic imbalance including insertions, deletions, inversions, duplications and inter-chromosomal translocations[80]. Structural variants have been associated with multiple psychiatric and neurological diseases such as Charcot-Marie-Tooth neuropathy[84], schizophrenia[85] and autism[270,271]. Multiple attempts to understand the relationship of structural variations including copy number variation to ALS have been limited by sequencing technology, computational burden, and the small number of samples[272,273]. Measuring the intensity of signals derived from a genotyping array is the most used method in detecting structural variants[274,275], but advances in next generation sequencing technologies and increased computing power have made studying structural variation feasible[276]. As a consequence, as part of the Project MinE initiative, a genomic structural variation working group was created to study the implications of structural variations in ALS risk and phenotype. In this chapter I report the analysis of structural variations in known ALS genes using 6580 whole genome sequences and test genotype-phenotype correlation using the rich Project MinE deep phenotype dataset.

6.2 Methods

6.2.1 Data sources

Samples were from multiple centres across the world contributing to the international Project MinE whole genome sequencing initiative. A total of 7 sources spanning the United States and Europe, including Ireland, Belgium, the Netherlands, Spain, Turkey, and the United Kingdom were included in this study. All cases met the revised El Escorial criteria.

6.2.2 Whole-genome sequencing

DNA was isolated from venous blood using standard methods. The DNA concentrations were set at 100ng/ul as measured by a fluorometer with the PicoGreen® dsDNA quantitation assay. DNA integrity was assessed using gel electrophoresis. All samples were sequenced using Illumina's FastTrack services (San Diego, CA, USA) on the Illumina HiSeq 2000 and HiSeqX platforms[2225]. Sequencing was 100bp and150bp paired-end performed according to the sequencing platform, using PCR-free library preparations and yielded ~40x coverage across each sample. Binary sequence alignment/map formats (BAM) were generated for each individual.

6.2.3 Quality Control

There were 6321 samples (4374 cases and 1947 controls) which passed quality control from a total of 6580 whole genome sequences. Quality control was preformed separately on genotyped data of each population according to the Project MinE methods published previously[7].

Sample mismatch was tested using sex checks, where genetic sex was compared to reported gender.

Quality control was preformed separately on genotyped data of each population according to the Project MinE methods published previously[7]. After quality control, the full set of genomic

Variant Call Formats (gVCFs) were merged together by first converting the gVCFs to Plink format and then merging all files together. This generated a single dataset containing all variant sites across all individuals. Non-autosomal chromosome and multi-allelic variants were excluded from pilot analyses. Sample and SNP quality control were performed using Plink[1,2] and VCFtools[3]. To begin sample quality control, missingness by sample was calculated on a per-chromosome basis; all samples had missingness < 10% across all 22 chromosomes and no samples were removed at this step.

All other sample quality control steps were performed on a set of high-quality biallelic SNPs that had minor allele frequency minor allele frequency 10%, missingness < 0.1%, were linkage disequilibrium pruned at an r^2 threshold of 0.2, were not A/T or C/G SNPs, did not lie in the major histocompatibility complex or lactase gene locus, and did not occur in the inversions on chromosome 8 or chromosome 17. The ~30,000 SNPs overlapping this set of SNPs and HapMap 3 were used to calculate principal components projecting the ALS cases and controls onto the HapMap 3 samples. Samples of non-European ancestry, defined as further than 10 standard deviations from the European-ancestry population principal components in HapMap 3 (CEU, people of Northern and Western European ancestry living in Utah; TSI, Tuscans in Italy), were excluded from analysis to ensure an ancestrally homogeneous group of samples for association testing. Samples with an inbreeding coefficient > 3 standard deviations from the mean of the distribution were excluded, as were unexpectedly related samples. Genotypes available from genotyping on the Illumina Omni 2.5M array were compared to sequencing genotypes, and samples with < 95% concordance were dropped from the analysis. Lastly, samples with discordant sex information (comparing chromosome X genotypes and phenotype information) were excluded.

For variant quality control, variants with missingness > 5% were removed, as were variants out of Hardy-Weinberg equilibrium in controls ($p < 1 \times 10^{-6}$) and monomorphic variants (induced by sample exclusions). Differential missingness between cases and controls was checked and variants with $p < 1 \times 10^{-6}$ were removed. Variants with extreme low or extreme high depth of coverage (> 6 standard deviations from the mean of the total depth distribution) were also excluded. Finally, the mitochondrial, X and Y chromosomes were excluded from analysis (but will be included in later analyses as sample sizes in Project MinE continue to grow). Approximately 10 million sites were lost during variant quality control.

Principal components analysis

To calculate principal components, a pruned set of high-quality SNPs with a genotyping rate 0.98 were included and SNPs within the major histocompatibility complex or lactase gene were excluded.

Identity-by-descent analysis

All non-singleton variants were phased using SHAPEIT2[5]. Subsequently BEAGLE4[23] was used to detect likely runs of identity by descent between individuals. The hg19 recombination map obtained from the 1000 Genomes Project was used to transform genetic positions from basepairs to centimorgans (cM). Presumed identity by descent segments shorter than one cM were excluded and regions with excessive identity by descent were excluded after visual inspection.

6.2.4 Determination of pathogenic ALS gene variants

A panel of 25 ALS genes was tested (*ALS2, ANG, ATXN2, C9orf72, CHCHD10, DAO, ERBB4, FUS, HNRNPA1, MOBP, NEK1, OPTN, PFN1, SCFD1, SETX, SOD1, SPG11, SQSTM1,*

TARDBP, TBK1, TUBA4A, UBQLN2, UNC13A, VAPB and *VCP*)[3,58]. These 25 genes were selected based on their association with ALS, so that all the chief causal genes were included as well as several risk factors, and genes harbouring large-effect, rare, Mendelian ALS gene variants.

Manta[88] by Illumina was used for extracting the structural variants, variant assembly and scoring the whole genome sequence. A VCF file then was generated for each participant.
To calculate the number of structural variations type in each gene, an in-house pipeline was used to filter the variants according to quality score, size and type of structural variants. Insertions with size less than 200bp were excluded as recommended by the Manta protocol. Variants that passed quality control were recorded.

Repeat primed PCR and Expansion Hunter[200] were used to assay the hexanucleotide repeat expansion in the *C9orf72* gene.

To determine the effect of structural variation, age of onset and age of death was compared in people with sporadic ALS carrying the structural gene variants and those not.
To determine the effect of family history, individuals with ALS were classified into three groups: group 1, familial ALS; group 2, apparently sporadic ALS with structural variation; group 3, apparently sporadic ALS without an identified structural variation.

6.2.5 Statistical Analysis

The effects of structural variants on ALS were tested using a generalised linear regression model, which included the total number of structural variations to predict disease affection status. To account for different sequencing platforms and population stratification, principal

components and technology platform were added as covariates to the generalised linear model. To assess the model, Pearson's chi-squared test was used.

As the age of onset and the age of death are not normally distributed, the mean age of onset and age of death between the groups was compared with the non-parametric Mann-Whitney U test. Uncorrected p values are reported. To estimate the size of any ascertainment bias observed, the median time between symptom onset and diagnosis was compared between those with familial ALS and those with apparently sporadic ALS in a Mann-Whitney U test. Analyses were performed in R.

6.2.6 Ethical Approval

All participants gave written informed consent.

6.3 Results

This work is almost exclusively mine, including designing the study and creating the pipeline for the analysis. I conceived and designed the study, planned the work, collected the data, created the pipeline for analysis described, performed the analyses, and wrote the paper. It forms part of large collaborative project where the phenotypic data and the principal components data were provided by the Project MinE phenotyping working group.

After quality control there were 4315 people with ALS and 1880 healthy controls included in the study, 4236 with apparently sporadic ALS and 79 with familial ALS. The male-female ratio was 2:1. Additionally, there were 31 people with ALS with cognitive impairment and 20

individuals with ALS-FTD. In this study 4,287 individuals were sequenced using the HiSeqX Illumina platform, and 1,908 sequenced using the HiSeq2000 platform (Table 15)

Cohort	Sample	Case	Control	Female	Male
Belgium	548	368	180	209	339
Ireland	403	267	136	161	242
Netherlands	2894	1859	1035	1182	1712
Spain	338	233	105	145	193
Turkey	223	148	75	87	136
United Kingdom	1402	1124	278	603	799
United States	387	316	71	153	234
Total	6195	4315	1880	2540	3655

Table 15.Project MinE study demography

The generalised linear regression model showed that of the 25 genes with known association with ALS, only three genes were affected with structural variants events: *C9orf72* gene expansion ($p = 2 \times 10^{-16}$), inversion in the *VCP* gene ($p = 2 \times 10^{-4}$) and *ERBB4* gene insertion ($p = 6 \times 10^{-4}$) (Table 16). All passed the multiple testing correction threshold (p=0.002) (Table 16).

Gene	p-value	SV-type	Frequency in cases	Frequency in control
C9orf72	2 x 10^{-16}	Expansion	244	4
VCP	2 x 10^{-4}	Inversion	2430	669
ERBB4	6 x 10^{-4}	Insertion	2001	476

Table 16. Structural variation results in sporadic ALS. A general linear model was used which included the total number of structural variations to predict disease affected status.

To assess structural variation burden on age of onset, we assessed those with structural variants in these genes, and those without. The mean age of onset in people with *C9orf72* gene expansion was 2.7 years younger than those with no *C9orf72* gene expansion ($p=8.8\times10^{-8}$, 95% CI for the difference 1.2 to 4.2 years). The mean age of onset in people with *VCP* gene inversion was 3 years younger than for people with no *VCP* gene inversion ($p=4.2\times10^{-13}$, 95% CI for the difference 2.2 to 3.7 years). Additionally, the mean age of onset in those with *ERBB4* gene insertion was one year younger than for those with no *ERBB4* insertion (p=0.003, 95% CI for the difference 0.25 to 1.72 years) (Table 17).

Gene	Age of onset (No SV)	Age of onset (With SV)	p-value	Difference in years
C9orf72	62.0	58.8	8.8×10^{-8}	3.2 (4.31-1.96)
VCP	62.6	59.7	4.2×10^{-13}	2.97(2.22 - 3.72)
ERBB4	61.2	60.2	0.003	1.00(0.25-1.72)

Table 17 Structural variation burden for age of onset.

The assessment of structural variation burden on age at death shows that people with *C9orf72* gene expansion died on average 3.8 years younger than people with no *C9orf72* gene expansion ($p=2.3\times10^{-9}$, 95% CI for the difference 2.6 to 5.1 years). People with *VCP* gene inversion died on average 1.8 years younger than those with no *VCP* gene inversion ($p=1.4\times10^{-5}$, 95% CI for the difference 1.0 to 2.5 years). No difference in age at death was observed between people with *ERBB4* gene insertion and those with no *ERBB4* gene insertion (p=0.1) (Table 18).

Gene	Age of death (No SV)	Age of death (With SV)	p-value	Difference in years
C9orf72	66.0	62.2	2.3×10^{-9}	3.8(2.64-5.10)
VCP	66.7	64.8	1.4×10^{-5}	1.8(1.04- 2.58)
ERBB4	65.9	65.0	0.1	NA

Table 18.Structural variation burden for age of death.

The assessment of the effect of having an ALS family history shows that individuals with familial form of ALS (group 1) had on average 4 years younger age of onset than those with sporadic ALS with no structural variants (group 3) (p=0.02, 95% CI for the difference 3.75 to -0.18 years). Meanwhile age of death was 4.5 years younger in individuals with an ALS family history (group 1) when compared to sporadic ALS group with no structural variants in *C90rf72, VCP* and *ERBB4* genes (group 3) (p=0.01, 95% CI for the difference 1.1 -7.8 years) (Table 19). No difference in age of onset or age of death was observed when we compared the ALS group with family history (group 1) against those with sporadic ALS with structural variants in *C90rf72, VCP* and *ERBB4* genes (group2) (Table 19).

	Familial	**Sporadic with no SV**	**p-value**	**Difference in years**
Age of onset (mean)	57.2	61.3	0.02	4.1(3.75 -0.18)
Age of death (mean)	61.7	66.2	0.01	4.5(-7.83- 1.16)
	Familial	**Sporadic with SV**	**p-value**	**Difference in years**
Age of onset (mean)	57.2	59.1	0.13	NA
Age of death (mean)	61.7	64.3	0.09	NA

Table 19. Comparison between age of onset and age at death between familial ALS and sporadic ALS with and without structural variation.

We extended the analysis to assess the association of gene structural variants with site of onset. The generalised linear regression model showed an association between individuals with *C9orf72* repeats expansion with bulbar site of onset (p=0.01). I Inversion in the *VCP* gene

associated with bulbar onset ($p= 3.5 \times 10^{-12}$) and FTD ($p=1.1 \times 10^{-4}$). Further analysis showed that insertion in the *ERBB4* gene was associated with respiratory onset in 13 of the 56 people carrying this variant (p=0.005) and with cognitive changes in 13 of the 30 people have *ERBB4* gene insertion (p=0.001).

6.4 Discussion

We have shown that genomic structural variants in the *C9orf72*, *VCP*, and *ERBB4* genes are associated with ALS risk, highlighting the importance of structural variation events in ALS. These types of variants are difficult to study without advances in sequencing technologies and international collaborative projects such as Project MinE.

We have shown that repeat expansions, insertions and inversions are the most common structural variation types associated with ALS.

Previous studies of *C9orf72* repeat expansion and onset age have led to conflicting results.[9,277–279] Also, the correlation between repeat size and diagnosis is poorly understood in apparently sporadic ALS, as most of the studies have been done on familial ALS.[280–282]. Our results provide a confirmation that *C9orf72* carriers have a younger age of onset and age at death than non-expanded carriers in large dataset from multiple populations.

We have shown that inversion in the *VCP gene* and repeat expansion in the *C9orf72* gene are markers for younger age of onset and death in apparently sporadic ALS. Comprehensive information on genetic susceptibility could contribute importantly to ALS risk stratification. Previously genetic testing was offered for patients with known familial history. Although family history has long been identified as a risk factor for ALS, genetic screening for younger patients might help in understanding disease progression and survival during clinical trials.

Additionally, carriers of *VCP* and *ERBB4* gene structural variants are expected to have a higher age of death than *C9orf72*. Thus, treatment response in clinical trials should account for these genomic variants, as a biased analysis could otherwise occur either due to lower or higher age at death. Moreover, *VCP* and *ERBB4* genes can be used as an early marker for cognitive impairment and ALS-FTD. Approximately 35% of patients with ALS experience cognitive or behavioural impairment, with an additional 15% having frontotemporal dementia. The association of *VCP* and *ERBB4* genes with FTD and cognitive impairment is reported[283–285].

Studies have estimated that around 12% of the human genome is susceptible to inversion[80 286]. Our study shows that inversions in the *VCP* gene associated with bulbar onset and FTD. Moreover, insertion in the *ERBB4* gene was associated with thoracic site of onset and cognitive changes confirming multiple studies linking *VCP* and *ERBB4* genes with FTD and cognitive/behaviour changes[283,287–290].

Furthermore, we have shown there were associations in genes previously identified from family-based studies (*C9orf72, ERBB4* and *VCP*) supporting the notion that familial and sporadic ALS are not mutually exclusive categories but rather a spectrum[4,102,140,291].

Limitation

In this study we used two tools for structural variation calling. Perhaps using a panel of structural variants callers will provide further insight to the study. The strength of using Expansion Hunter is that it was developed primarily and tested to measure gene repeat expansions in ALS[200]. Expansion Hunter is a result of collaboration between Illumina and the Project MinE consortium. Furthermore, in this research we report the structural variation in known ALS genes, extending the analysis to the entire genome will give a comprehensive view of the implications of this type of genomic variation in ALS. Moreover, ALS is a disease of

the central nervous system, but our WGS data are derived from leukocyte DNA, since our DNA source was whole blood. Thus, replication of this study using brain tissue is of interest and perhaps could give insight into the relationship between leukocyte DNA and post-mitotic neurons. Another weakness of this study is that we used WGS technology. Using targeted sequencing technology such as PCR, PacBio or Nanopore to study the structural variation in these known ALS genes might provide superior sequencing data. However, our findings have the advantage of a large sample size of more than 4300 cases. A further limitation of this study is that we cannot exclude the possibility that these gene carrier groups might harbour other important undiscovered ALS genes and the age effect that we observe is not specific. However, all the included cases have a structural variant in the reported gene supporting our results.

6.5 Conclusion

In conclusion, we have shown that structural variations are important genomic events and can be used as genomic markers to understand ALS disease trajectory. Further study is needed to understand the implications of structural variants on ALS risk and survival.

Chapter 7. Final Discussion

7.1 Discussion

ALS is a terminal illness. Often after diagnosis, a person will ask "how long do I have?", and "what will happen?". Understanding disease trajectories is a key to answering both questions. Thinking in terms of these trajectories provides a predictive map of disease course from diagnosis to death[292,293].

Trajectories are predictive, though not determinative, meaning that these predictive factors are multifactorial in nature and dependent on each other. For example, it is known that patients who are involved in clinical trials tend to receive a better quality of health care, therefore it is expected that this will contribute to clinical trial outcome. Although the effect is very small, it is predictable[294].

Understanding trajectories of illness will help in preparation and advanced planning of clinical care. In doing so, clinicians can plan and deliver appropriate care according to specific patients' needs. For example, in ALS, bulbar site of onset is a good indicator of short survival therefore, one to one nursing is most likely needed within 12 months (or less) from diagnosis and for a short period of time. Meanwhile, disease progression is known to be slower in ALS with a spinal site of onset, and one-to-one nursing would be expected to take place at a later stage than for bulbar onset, and for a longer period. Thus, understanding disease trajectory is crucial for health care professionals, clinicians, patients and their carers to gain a better understanding of disease trajectories with direct implication for service planning and health economics. We therefore aimed to understand factors that affect illness trajectories in ALS and draw out key clinical implications.

Disease staging allows a simple description of the extent of physical or functional involvement in an affected person and guides management. Clinical stage in amyotrophic lateral sclerosis can be assigned using King's staging with a simple protocol based on the number of CNS regions involved and the presence of significant nutritional or respiratory failure. There are many clinical staging systems in ALS, with two currently in widespread use. We decided to investigate and compare both these ALS staging systems, King's clinical staging and Milano-Torino (MiToS) functional staging, using data from the LiCALS phase 3 clinical trial (EudraCT 2008-006891-31). The disease stage was derived retrospectively for each system from the ALS Functional Rating Scale-Revised sub-scores using standard methods. The two staging methods were then compared for timing of stages using box plots, correspondence using chi-square tests, agreement using a linearly weighted kappa coefficient and concordance using Spearman's rank correlation. We found that for both systems, progressively higher stages occurred at progressively later proportions of the disease course, but the distribution differed between the two methods. King's stage 3 corresponded to MiToS stage 1 most frequently, with earlier King's stages 1 and 2 largely corresponding to MiToS stage 0 or 1. The Spearman correlation was 0.54. There was fair agreement between the two systems with a kappa coefficient of 0.21. We concluded that the distribution of timings shows that the two systems are complementary, with King's staging showing greatest resolution in early to mid-disease corresponding to clinical or disease burden, and MiToS staging having higher resolution for late disease, corresponding to functional involvement. We therefore propose using both staging systems when describing ALS.

In conclusion, MiToS and King's ALS staging systems are complementary, with the King's clinical staging system having a higher resolution in early-mid diseases stages, and the MiToS system having a higher resolution in late disease stages. In the King's staging system there is more homogeneity between patients in the same stage, and a greater discrimination between patients in different disease stages.

In order for the King's staging system to be applied by different health care professionals and varying levels of experience working in ALS, as for example may be the case in a multicentre clinical trial, we designed a Standard Operating Procedure (SOP) for the use of the King's system. We then investigated whether it could be used by a variety of health care professionals.

We wrote case vignettes representative of ALS patients at different disease stages. During two workshops, we taught health care professionals how to use the SOP, then asked them to stage the vignettes using the SOP. We measured the extent to which SOP staging corresponded with correct clinical stage.

We found that the reliability of staging using the SOP was excellent, with a Spearman's Rank coefficient of 0.95 ($p<0.001$) and was high for different groups of health care professionals, and for those with different levels of experience in ALS. The limits of agreement between SOP staging and actual clinical stage lie within a single stage, confirming that there is a clinically acceptable level of agreement between staging using the SOP and actual King's clinical stage. There were also no systematic biases of the SOP over the range of stages, and therefore no over-staging or under-staging

We have demonstrated that the staging Standard Operating Procedure provides a reliable method of calculating clinical stages in ALS patients and can be used by a range of health care professionals.

The next step was to validate the King's ALS staging system. It is important that the assigned clinical stage matches expectations, and generally corresponds with how a health care professional would intuitively stage the patient. We therefore investigated the relationship between King's clinical ALS stage and ALS stage as intuitively assigned by health care professionals. We wrote 17 case vignettes describing people with ALS at different disease stages from very early limited disease involvement through to severe, multi-domain disease. During two workshops, we asked health care professionals to intuitively stage the vignettes and compared the answers with the actual King's clinical ALS stage. There was a good correlation between King's clinical ALS stage and intuitively assigned stage, with a Spearman's Rank correlation coefficient of 0.64 ($p<0.001$). There was no difference in the intuitive stages assigned by practitioners of different types or at different levels of experience.

Across a spectrum of ALS scenarios, King's clinical ALS stage corresponded to intuitive ALS stage as assigned by a range of health care professionals

In the next step, we used the Riluzole trial data to assess the use of King's stage as an endpoint for clinical trials analysis, and to determine if additional trial information could be obtained by a different analysis, such as the timing of any benefit of Riluzole.

Riluzole is the only drug to prolong survival for amyotrophic lateral sclerosis (ALS) and, at a dose of 100 mg, is associated with a 35% reduction in mortality in a clinical trial. A key question is whether the survival benefit occurs at an early stage of disease, late stage, or is

spread throughout the course of the disease. To address this question, we used the King's clinical staging system to do a retrospective analysis of data from the original dose-ranging clinical trial of Riluzole. It is important to know whether the survival benefit of Riluzole in patients with amyotrophic lateral sclerosis occurs early, late, or throughout the course of the disease to enable proper counselling of patients. We searched PubMed for reports published at any date up to July 31, 2017, using the terms “Riluzole”; “amyotrophic lateral sclerosis”, “motor neuron disease”, “motor neurone disease”, “ALS”, or “MND”; and “stage” or “staging”. We included randomised, placebo-controlled trials in patients with ALS that involved Riluzole alone and studies of clinical staging in ALS. We identified two trials, both of which were randomised, placebo-controlled studies and one of which included sufficient data for retrospective staging. We approached the commercial owners of the Riluzole clinical trial data for full academic access to the trial database.

In the original dose-ranging trial, patients were enrolled between December 1992, and November 1993, in Belgium, France, Germany, Spain, Canada, the USA, and the UK if they had probable or definite ALS as defined by the El Escorial criteria. The censor date for the Riluzole survival data was set as the original study end date of Dec 31, 1994. For this analysis, King's clinical ALS stage was estimated from the electronic case record data of the modified Norris scale, UK Medical Research Council score for muscle strength, El Escorial category, vital capacity, and gastrostomy insertion data. The lowest allocated stage was 2 because the original trial only included patients with probable or definite ALS. We used a $\chi 2$ test to assess the independence of stage at trial enrolment and treatment group, Kaplan-Meier product limit distribution to test the transition from each stage to subsequent stages, and Cox regression to confirm an effect of treatment group on time in stage, controlling for covariates. We did sensitivity analyses by combining treatment groups, using alternative strategies to stage,

stratifying by stage at trial enrolment, and using multistate outcome analysis of treatments (MOAT).

We analysed the case records of all 959 participants from the original dose-ranging trial, 237 assigned to 50 mg/day Riluzole, 236 to 100 mg/day, 244 to 200 mg/day, and 242 to daily placebo. Clinical stage at enrolment did not significantly differ between treatment groups (p=0·22). Time in stage 4 was longer for patients receiving 100 mg/day Riluzole than for those receiving placebo (hazard ratio [HR] 0·55, 95% CI 0·36–0·83; log-rank p=0·037). Combining treatment groups and stratifying by stage at enrolment showed a similar result (HR 0·638, 95% CI 0·464–0·878; p=0·006), as did analysis with MOAT where the mean number of days spent in stage 4 was numerically higher for patients given Riluzole at higher doses compared with patients receiving placebo. Time from stages 2 or 3 to subsequent stages or death did not differ between Riluzole treatment groups and placebo (p=0·83 for stage 2 and 0·88 for stage 3).

We showed that Riluzole prolongs survival in the last clinical stage of ALS; this finding needs to be confirmed in a prospective study, and treatment effects at stage 1 still need to be analysed. The ALS stage at which benefit occurs is important for counselling of patients before starting treatment. Staging should be used in future ALS clinical trials to assess the stage at which survival benefit occurs, and a similar approach could be used for other neurodegenerative diseases.

By use of the King's clinical ALS staging system, we showed in a retrospective analysis that Riluzole prolonged stage 4 ALS in a dose-dependent manner, with no apparent prolongation of stages 2 or 3. We were unable to determine if there was an effect on stage 1. The timing of any benefit from Riluzole affects the information that needs to be given to patients, because

they are likely to interpret the benefit of prolongation of a later stage of disease as different from the benefit of prolonging an early stage, or prolongation of the disease course in general. The timing of benefit also has implications for health economics because the later stages of ALS are associated with higher costs than earlier stages, and therefore prolonging stage 4 is more costly than prolonging stages 1 or 2. Further studies are needed to determine if there is a survival benefit of Riluzole in stage 1 ALS.

We have shown that clinical staging is a powerful tool to understand disease trajectory and staging analyses should be used in future clinical trials of treatments in patients with ALS and other neurodegenerative diseases.

We have shown that clinical staging is a powerful tool to understand disease trajectory. These staging systems correlate with a decline in functional measures, health utility and quality of life scores, and an increase in socioeconomic costs, comprising costs of healthcare and loss of productivity[25,146,295,296]. ALS stages have validity as meaningful outcome measures in ALS clinical trials, as they help to account for the inherent heterogeneity in patient populations, and can facilitate the development of drugs which differing efficacies throughout disease progression[28,29]. King's staging system is now used in multiple clinical trials and progression to a higher ALS disease stage is now being utilised as a primary outcome measure in MIROCALS a Phase 2 randomised controlled trial and EDARAVONE clinical trial. I have shown that it can be used to determine the stage at which Riluzole prolongs survival in ALS in a retrospective study. Staging can be used to select patients for clinical studies, for example King's staging system was used to investigate neuroimaging biomarkers in early disease

stages[298]. Furthermore, King's staging has utility in mapping to neuroimaging and biochemical biomarkers in ALS patients, correlating with reduction in white matter integrity in ALS patients with repeat expansions in the *C9orf72* gene on neuroimaging[2997], and with higher levels of CSF neurofilament light chain[300,301]. Staging correlates with other important disease parameters, including sustained and forced vital capacity[296] and energy expenditure[302]. Mitochondrial dysfunction is involved in the pathogenesis of ALS[303], and mitochondrial activity detected in patient peripheral blood mononuclear cells decreases with increasing disease stage[304], indicating that staging may correlate with underlying aetiopathogenic mechanistic markers.

ALS is on a clinical, genetic and pathological spectrum with frontotemporal dementia (FTD). Up to 15% of people with ALS have a diagnosis of FTD, and cognitive impairment occurs in about 50% of people with ALS[305,306]. Measures of behaviour and cognition are reduced in later ALS disease stages[145,307]. Moreover, specific staging systems to measure the extent of cognitive involvement in ALS have been developed[37,308], and have been used recently in parallel with the King's ALS clinical staging system[32].

Clinical methods of understanding disease progression such as staging, can be made more powerful if we also understand biological factors that influence it. We therefore used next generation sequencing data coupled with survival information to achieve this. To process the next generation sequencing data there were significant challenges. A huge number of bioinformatics tools exist; it is therefore difficult to design an analysis pipeline; next generation sequencing analysis is computationally intensive, requiring expensive infrastructure which can be problematic given that many medical and research centers do not have adequate high-

performance computing facilities and the use of cloud computing facilities is not always possible due to privacy and ownership issues. We have therefore developed a fast and efficient bioinformatics pipeline that allows for the analysis of DNA sequencing data, while requiring little computational effort and memory usage. We achieved this by exploiting state-of-the-art bioinformatics tools. DNAscan can analyse raw, 40x whole genome NGS data in 8 hours, using as little as 8 threads and 16 Gigabytes of RAM, while guaranteeing a high performance. DNAscan can look for SNVs, small indels, structural variations, repeat expansions and viral (or any other organism's) genetic material. Its results are annotated using a customisable variety of databases including ClinVar, EXAC and dbSNP, and a local deployment of the gene.iobio platform is available for an on-the-fly result visualisation.

DNAscan is a novel and promising pipeline which can be used for genetic research as well as medical research. It is suitable for both small and large-scale analysis. Covering the whole end-to-end analysis process, from sequencing data in fastq format to the results visualisation, generating user friendly reports and providing result navigation utilities.

With the increasing availability of next generation sequencing data in ALS, non-specialists, including health care professionals and patients, are obtaining their genomic information without a corresponding ability to analyse and interpret it. Furthermore, the relevance of novel or existing variants in ALS genes is not always apparent. Therefore, we developed ALSgeneScanner, a tool that is easy to install and use, able to provide an automatic, detailed,

ALSgeneScanner puts a powerful bioinformatics tool, able to exploit the potentialities of next generation sequencing data in the hands of patients, ALS researchers and clinicians.

annotated report, on a list of ALS genes from whole genome sequence data in a few hours and whole exome sequence data in about one hour on a readily available mid-range computer. This will be of value to non-specialists and aid in the interpretation of the relevance of novel and existing variants identified in DNA sequencing data.

Although the heritability of ALS is about 60%, only about 15% is explained by common gene variants, suggesting that other forms of genetic variation are important. Telomeres maintain DNA integrity during cellular replication and shorten naturally with age. Gender and age are risk factors for ALS and also associated with telomere length. We therefore investigated telomere length in ALS.

We estimated telomere length by applying a bioinformatics analysis to whole genome sequence data of leukocyte-derived DNA from people with ALS and age and gender-matched matched controls in a UK population. We tested the association of telomere length with ALS and ALS survival.

There were 1241 people with ALS and 335 controls. The median age for ALS was 62.5 years and for controls, 60.1 years, with a male-female ratio of 62:38. Accounting for age and sex, there was a 9% increase of telomere length in ALS compared to matched controls ($p = 0.008$). Those with longer telomeres had a 16% increase in median survival ($p = 0.001$). Of nine SNPs associated with telomere length, two were also associated with ALS: rs8105767 near the *ZNF208* gene ($p = 1.29 \times 10^{-4}$) and rs6772228, which is in an intron for the *PXK* gene ($p = 0.001$).

We have shown that longer telomeres in leukocyte-derived DNA are associated with ALS, and with increased survival in those with ALS.

Subsequently, we decided to test the relationship between the length of telomere and rare variations in ALS using large scale cohort from the UK. Using the Variable Threshold analysis, we shown an association between telomere length and rare variation in chromosome 5 (5:352349_T/C) in ALS individuals ($p=8.6\times10^{-9}$). Although this association is significant, validation in a bigger cohort is required. This variant could potentially be used as a single test for telomere length in ALS.

We have shown that telomere length is associated with rare variation in ALS

The genetic aetiology of ALS is multifactorial, and a significant proportion of genetic variance remains unexplained. In this project, we studied the implication of structural variations in the largest single disease whole genome sequencing dataset in the world, and found that genomic structural variations are implicated in ALS. The generalised linear regression model showed that of the 25 genes harbouring single nucleotide variants with known association with ALS risk, only three harboured structural variants, and these were associated with ALS risk as well: *C9orf72* gene expansion ($p = 2 \times 10^{-16}$), inversion in the *VCP* gene ($p = 2 \times 10^{-4}$) and *ERBB4* gene insertion ($p = 6 \times 10^{-4}$). All passed the multiple testing correction threshold (p=0.002). Additionally, we have shown that *VCP* and *ERBB4* genes are associated with cognitive impairment and ALS-FTD. Furthermore, in this study, we found that the mean age of onset in familial ALS was 4.5 years younger than for apparently sporadic ALS (p=0.01, 95% CI for the difference 1.1 to 7.8 years), confirming our previous study that familial ALS has a younger age of onset[291].

ALS-causing GGGGCC hexanucleotide repeat expansion of *C9orf72* is known to account for the largest burden of both sporadic (~7%) and familial (~40%) ALS. In this study I have

demonstrated that next generation sequencing technology such as WGS provides a powerful tool to investigate the frequency of important risk factors such as *C9orf72* in a large cohort such as Project MinE. Furthermore, I have shown that WGS gives researchers the opportunity to investigate other genomic variants such as genomic structural variation. The results suggest that these events are more common than previously thought in *C9orf72, VCP* and *ERBB4* which supports the hypothesis of oligogenic inheritance in ALS and the notion that these genes should be regarded as risk genes rather than Mendelian genes.

We have shown that structural variations are important genomic events and can be used as genomic markers to understand ALS disease trajectory. Further study is needed to understand the implications of structural variations on ALS risk and survival.

In conclusion, in this thesis I showed that clinical staging systems and factors influencing survival in ALS including structural genomic variation, telomere length and rare genomic variants can help in mapping disease progression.

7.2 Research Obstacles and future direction

There are several limitations to this work. We used clinical trial data to compare King's and MiToS ALS staging systems and to understand the relationship between the survival benefit of Riluzole and disease stage, rather than using clinic or population data. However, this may be advantageous, as results are more likely to be relevant to other clinical trials. We have previously shown that clinical trial data show a shift towards a greater proportion of disease course passed for a given stage[28]. This occurrence is likely a result of left censoring due to the population being selected for trial participation and sourced from a biased clinic population.

In the telomere length analysis, one limitation of our method is that we cannot draw firm conclusions about the exact length of a telomere, as different sequencing technologies will generate different telomere length estimates because of differences in library preparation and platform [245,246]. To overcome this potential weakness, we have used the same industry-leading sequencing platform for all samples, as well as designing the study to minimize batch effects by having cases and controls sharing the same sequencing plate.

In the structural variations study, we used two tools for structural variation calling, perhaps using a panel of SV callers will provide further insight to the study. The strength of using Expansion Hunter is that it was developed primarily and tested to measure gene repeat expansions in ALS[200]. Expansion Hunter is a result of collaboration between Illumina and the Project MinE consortium.

The biggest obstacle was the limited number of structural variations studied in ALS. This was because of the computational burden, as this type of analysis requires access to cluster computing and fluency in programming languages such as Python, Java and Awk.

Bioinformatics tools such as DNAscan and ALSgeneScanner can to some extent provide an alternative solution to the computational burden for ALS researchers. Moreover, whole genome sequencing is a cutting edge and evolving area where the best methods for analysis, both in data generation and statistical analysis, are currently unknown and the exact power of such studies cannot easily be estimated because no formal method currently exists.

Future Direction

In the last decade, our understanding of the genetic basis of ALS has significantly improved, yet we do not know if the results are applicable to all world populations. In order to create new clinical tools or applicable treatments for all ALS, much more focus is needed to engage underrepresented minority groups in genomics studies such as Project MinE. Understanding ALS genetic architecture in non-European populations would help demonstrate the validity of future clinical tests or treatments across populations. There is a growing demand in the genomic research community to diversify scientific research, but it is most likely that this sampling bias will continue. Thus, reviewers and funding organisations should demand a new strategy in study design and funding priority to create a balanced ethnic representation.

Other gene hunting strategies are needed. The balance between innate and adaptive immunity is important to maintain a healthy cycle of protection and repair. Our knowledge of the interactions of immune and nervous systems has changed significantly during the last decade. Evidence of neuroinflammation and neuroimmune activation have been shown to play a role in the aetiology of a variety of neurological disorders. The implication of immune reactions in ALS are well documented in the literature, but our understanding of the genetic influences and variations in the ALS neuro-immune axis is still at an early stage compared to other neurological, autoimmune, complex diseases, such as multiple sclerosis or systemic lupus

erythematosus. This is not a new topic, as evident by the associations identified after the discovery of the *SOD1* gene. When modelled in mice, they develop an ALS-like disease but there is a striking role of the nervous system immunological cells, astroglia and microglia[309,310]. Understanding the mechanisms of the underlying dynamic genetic variations is becoming crucial for understanding the pathophysiology of disease and developing personalized treatment. With only one publication under the name 'Immunogenetics and Amyotrophic Lateral Sclerosis', immuno-genetics might be a good strategy for gene hunting[311].

Genetics can help to stratify ALS groups, but studying the interaction of a clinical trial with genetics might give us in-depth insight into the robustness of proposed pathological networks in ALS. Circulating immune complexes and IgG are important to measure the immune response. As an example, IgG are a biomarker, as studies have shown that IgG remains high throughout the ALS course[312]. Subsequently, higher circulating immune complexes indicate an early inflammatory marker in the disease course, and therefore studying these observations might guide us in locating ALS genes[312]. There are multiple studies in immunogenetics where an increase has been documented in the expression of over 25 immune related genes found during the early stages of ALS. Interestingly, 15 other implicated immunogenes are found to decrease their expression, indicating a complex immune reaction[313].

In a Twin ALS epigenetics study, methylated fragments in immunity-related genes were documented, such as *EGFR*, *TGFb1*, and *TNFRSF11A* and a high concentrations of IL-6, TNF-a, and IL-1 in the same twin samples[314]. The major histocompatibility complex (MHC) is an extremely gene-dense region with long-range LD and hundreds of immunologically active genes[315]. Analysing the mRNA expression of a panel of 14 genes that encode MHCI receptors led to a report that the killer cell immunoglobulin-like receptor, three IgG domains and long cytoplasmic tail 2 (KIR3DL2), which encodes an inhibitory MHCI receptor, was uniquely expressed in all human ALS astrocytes[316]. Studying the genetic basis of the immune reaction

might explain the direct effect of epigenetics on the phenotype, severity, and progression of the ALS course, similarly to other autoimmune complex diseases. Fine mapping and the usage of the genetic Immunochip has narrowed the association signals to differing extents across the various autoimmune complex diseases[315]. New genetics editing tools such as CRISPR can then be tested when there is a gene of interest. Currently, the field of ALS has several studies using a panel of biomarkers in assessment of neuroinflammation and also examining Treg cell immuno-modulation therapy.'

References

1. Hardiman, O., van den Berg, L. H. & Kiernan, M. C. Clinical diagnosis and management of amyotrophic lateral sclerosis. *Nat. Rev. Neurol.* **7,** 639–649 (2011).
2. van Es, M. A. *et al.* Amyotrophic lateral sclerosis. *Lancet* (2017). doi:10.1016/S0140-6736(17)31287-4
3. Brown, R. H. & Al-Chalabi, A. Amyotrophic Lateral Sclerosis. *N. Engl. J. Med.* **377,** 162–172 (2017).
4. Al-Chalabi, A. & Hardiman, O. The epidemiology of ALS: a conspiracy of genes, environment and time. *Nat. Rev. Neurol.* **9,** 617–28 (2013).
5. Al-Chalabi, A. *et al.* An estimate of amyotrophic lateral sclerosis heritability using twin data. *J. Neurol. Neurosurg. Psychiatry* **81,** 1324–1326 (2010).
6. Chiò, A. *et al.* UNC13A influences survival in Italian amyotrophic lateral sclerosis patients: A population-based study. *Neurobiol. Aging* **34,** (2013).
7. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48,** 1043–1048 (2016).
8. Diekstra, F. P. *et al.* C9orf72 and UNC13A are shared risk loci for amyotrophic lateral sclerosis and frontotemporal dementia: A genome-wide meta-analysis. *Ann. Neurol.* **76,** 120–133 (2014).
9. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72,** 257–268 (2011).
10. Fogh, I. *et al.* Association of a Locus in the CAMTA1 Gene With Survival in Patients With Sporadic Amyotrophic Lateral Sclerosis. *JAMA Neurol.* **73,** 812–820 (2016).
11. Al-Chalabi, A. *et al.* Amyotrophic lateral sclerosis: moving towards a new classification system. *Lancet Neurol.* **15,** 1182–1194 (2016).
12. Wolf, J. *et al.* Variability and prognostic relevance of different phenotypes in amyotrophic lateral sclerosis - Data from a population-based registry. *J. Neurol. Sci.* **345,** 164–167 (2014).
13. Wijesekera, L. *et al.* Natural history and clinical features of the flail arm and flail leg ALS variants.
14. Ravits, J. M. & La Spada, A. R. Als motor phenotype heterogeneity, focality, and spread: Deconstructing motor neuron degenerationsymbol. *Neurology* (2009). doi:10.1212/WNL.0b013e3181b6bbbd
15. Ekhtiari Bidhendi, E. *et al.* Two superoxide dismutase prion strains transmit amyotrophic lateral sclerosis-like disease. *J. Clin. Invest.* (2016). doi:10.1172/JCI84360
16. Clarke, G., Lumsden, C. J. & McInnes, R. R. Inherited neurodegenerative diseases: the one-hit model of neurodegeneration. *Hum. Mol. Genet.* (2001). doi:10.1093/hmg/10.20.2269

17. Nandedkar, S. D., Barkhaus, P. E. & StÅlberg, E. V. Motor unit number index (MUNIX): principle, method, and findings in healthy subjects and in patients with motor neuron disease. *Muscle Nerve* **42,** 798–807 (2010).
18. Geevasinga, N. *et al.* Diagnostic criteria in amyotrophic lateral sclerosis: A multicenter prospective study. *Neurology* **87,** 684–90 (2016).
19. Brooks, B. R. *et al.* El Escorial World Federation of Neurology criteria for the diagnosis of amyotrophic lateral sclerosis. *J. Neurol. Sci.* **124,** 96–107 (1994).
20. de Carvalho, M. *et al.* Electrodiagnostic criteria for diagnosis of ALS. *Clinical Neurophysiology* **119,** 497–503 (2008).
21. Miller, R. G., Munsat, T. L., Swash, M. & Brooks, B. R. Consensus guidelines for the design and implementation of clinical trials in ALS. *J Neurol Sci* **169,** 2–12 (1999).
22. Schrooten, M., Smetcoren, C., Robberecht, W. & Van Damme, P. Benefit of the Awaji diagnostic algorithm for amyotrophic lateral sclerosis: A prospective study. *Ann. Neurol.* **70,** 79–83 (2011).
23. Cedarbaum, J. M. *et al.* The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *J. Neurol. Sci.* **169,** 13–21 (1999).
24. Roche, J. C. *et al.* A proposed staging system for amyotrophic lateral sclerosis. *Brain* **135,** 847–852 (2012).
25. Chiò, A., Hammond, E. R., Mora, G., Bonito, V. & Filippini, G. Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatry* 38–44 (2013). doi:10.1136/jnnp-2013-306589
26. Fang, T. *et al.* Comparison of the King's and MiToS staging systems for ALS. *Amyotroph. Lateral Scler. Front. Degener.* (2016).
27. Balendra, R. *et al.* Estimating clinical stage of amyotrophic lateral sclerosis from the ALS Functional Rating Scale. *Amyotroph. Lateral Scler. Frontotemporal Degener.* **15,** 279–84 (2014).
28. Hardiman *et al.* Use of clinical staging in amyotrophic lateral sclerosis for phase 3 clinical trials. *J. Neurol. Neurosurg. Psychiatry* jnnp-2013-306865 (2014). doi:10.1136/jnnp-2013-306865
29. Ferraro, D. *et al.* Amyotrophic lateral sclerosis: a comparison of two staging systems in a population-based study. *Eur J Neurol* (2016). doi:10.1111/ene.13053
30. T., F., F., J. & A., A.-C. Nonmotor Symptoms in Amyotrophic Lateral Sclerosis: A Systematic Review. *International Review of Neurobiology* (2017). doi:10.1016/bs.irn.2017.04.009
31. Montuschi, A. *et al.* Cognitive correlates in amyotrophic lateral sclerosis: A population-based study in Italy. *J. Neurol. Neurosurg. Psychiatry* **86,** 168–173 (2015).
32. Crockford, C. *et al.* ALS-specific cognitive and behavior changes

associated with advancing disease stage in ALS. *Neurology* 10.1212/WNL.0000000000006317 (2018). doi:10.1212/WNL.0000000000006317

33. Rippon, G. A. *et al.* An observational study of cognitive impairment in amyotrophic lateral sclerosis. *Arch. Neurol.* **63,** 345–52 (2006).
34. Swinnen, B. & Robberecht, W. The phenotypic variability of amyotrophic lateral sclerosis. *Nat. Publ. Gr.* **10,** 661–70 (2014).
35. Lomen-Hoerth, C., Anderson, T. & Miller, B. The overlap of amyotrophic lateral sclerosis and frontotemporal dementia. *Neurology* **59,** 1077–1079 (2002).
36. Strong, M. J. The syndromes of frontotemporal dysfunction in amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler.* **9,** 323–338 (2008).
37. Abrahams, S. *et al.* Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration Screening for cognition and behaviour changes in ALS. *Amyotroph. Lateral Scler. Front. Degener.* **15,** 1–2 (2014).
38. Crockford, C. *et al.* ALS-specific cognitive and behavior changes associated with advancing disease stage in ALS. *Neurology* 10.1212/WNL.0000000000006317 (2018). doi:10.1212/WNL.0000000000006317
39. Niven, E. *et al.* Validation of the Edinburgh Cognitive and Behavioural Amyotrophic Lateral Sclerosis Screen (ECAS): A cognitive tool for motor disorders. *Amyotroph. Lateral Scler. Frontotemporal Degener.* **16,** 172–9 (2015).
40. Tazen, S. *et al.* Amyotrophic lateral sclerosis and spinocerebellar ataxia type 2 in a family with full CAG repeat expansions of ATXN2. *JAMA Neurol.* **70,** 1302–4 (2013).
41. Etemadifar, M., Abtahi, S.-H., Akbari, M. & Maghzi, A.-H. Multiple sclerosis and amyotrophic lateral sclerosis: is there a link? *Mult. Scler.* **18,** 902–4 (2012).
42. Howland, R. H. Schizophrenia and amyotrophic lateral sclerosis. *Compr. Psychiatry* **31,** 327–36 (1990).
43. Cooper-Knock, J., Shaw, P. J. & Kirby, J. The widening spectrum of C9ORF72-related disease; Genotype/phenotype correlations and potential modifiers of clinical phenotype. *Acta Neuropathologica* **127,** 333–345 (2014).
44. Neuenschwander, A. G., Thai, K. K., Figueroa, K. P. & Pulst, S. M. Amyotrophic lateral sclerosis risk for spinocerebellar ataxia type 2 ATXN2 CAG repeat alleles: a meta-analysis. *JAMA Neurol.* **71,** 1529–34 (2014).
45. Carreiro, A. V. *et al.* Prognostic models based on patient snapshots and time windows: Predicting disease progression to assisted ventilation in Amyotrophic Lateral Sclerosis. *J. Biomed. Inform.* (2015).

doi:10.1016/j.jbi.2015.09.021
46. Polkey, M. I. *et al.* Respiratory Muscle Strength as a Predictive Biomarker for Survival in Amyotrophic Lateral Sclerosis. *Am. J. Respir. Crit. Care Med.* **195,** 86–95 (2017).
47. Chiò, A. *et al.* Amyotrophic Lateral Sclerosis Outcome Measures and the Role of Albumin and Creatinine: A Population-Based Study. *JAMA Neurol.* **71,** 1–9 (2014).
48. Calvo, A. *et al.* Factors predicting survival in ALS: a multicenter Italian study. *J. Neurol.* (2017). doi:10.1007/s00415-016-8313-y
49. Elamin, M. *et al.* Executive dysfunction is a negative prognostic indicator in patients with ALS without dementia. *Neurology* **76,** 1263–1269 (2011).
50. Marin, B. *et al.* Population-Based Evidence that Survival in Amyotrophic Lateral Sclerosis is Related to Weight Loss at Diagnosis. *Neurodegener. Dis.* **16,** 225–34 (2016).
51. Del Aguila, M. A., Longstreth, W. T., McGuire, V., Koepsell, T. D. & Van Belle, G. Prognosis in amyotrophic lateral sclerosis: A population-based study. *Neurology* (2003). doi:10.1212/01.WNL.0000049472.47709.3B
52. Wolf, J. *et al.* Factors predicting survival in ALS patients - Data from a population-based registry in Rhineland-palatinate, Germany. *Neuroepidemiology* (2015). doi:10.1159/000381625
53. Czaplinski, A., Yen, A. A. & Appel, S. H. Forced vital capacity (FVC) as an indicator of survival and disease progression in an ALS clinic population. *J. Neurol. Neurosurg. Psychiatry* (2006). doi:10.1136/jnnp.2005.072660
54. Knibb, J. A. *et al.* A clinical tool for predicting survival in ALS. *J. Neurol. Neurosurg. Psychiatry* 1–7 (2016). doi:10.1136/jnnp-2015-312908
55. Shoesmith, C. L., Findlater, K., Rowe, A. & Strong, M. J. Prognosis of amyotrophic lateral sclerosis with respiratory onset. *J. Neurol. Neurosurg. Psychiatry* (2007). doi:10.1136/jnnp.2006.103564
56. Czaplinski, A., Yen, A. A. & Appel, S. H. Amyotrophic lateral sclerosis: Early predictors of prolonged survival. *J. Neurol.* **253,** 1428–1436 (2006).
57. Diekstra, F. P. *et al.* UNC13A is a modifier of survival in amyotrophic lateral sclerosis. *Neurobiol. Aging* **33,** (2012).
58. Gaastra, B. *et al.* Rare genetic variation in UNC13A may modify survival in amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Front. Degener.* (2016). doi:10.1080/21678421.2016.1213852
59. Andersen, P. M. & Al-Chalabi, A. Clinical genetics of amyotrophic lateral sclerosis: what do we really know? *Nat Rev Neurol* **7,** 603–615 (2011).
60. Andersen, P. M. *et al.* Amyotrophic lateral sclerosis associated with homozygosity for an Asp90Ala mutation in CuZn-superoxide dismutase. *Nat. Genet.* **10,** 61–6 (1995).
61. Cudkowicz, M. E. *et al.* Epidemiology of mutations in superoxide

dismutase in amyotrophic lateral sclerosis. *Ann. Neurol.* (1997). doi:10.1002/ana.410410212
62. Westeneng, H.-J. *et al.* Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. *Lancet Neurol.* (2018). doi:10.1016/S1474-4422(18)30089-9
63. Sato, Y. *et al.* Prediction of prognosis of ALS: Importance of active denervation findings of the cervical-upper limb area and trunk area. *Intractable Rare Dis. Res.* (2015). doi:10.5582/irdr.2015.01043
64. Hardiman, O. *et al.* Amyotrophic lateral sclerosis. *Nat. Rev. Dis. Prim.* **3,** 17071 (2017).
65. Web Administrator: Olubunmi Abel | PI: Prof Ammar Al-Chalabi. ALSoD: Amyotrophic Lateral Sclerosis Online Genetics Database. Available at: http://alsod.iop.kcl.ac.uk/home.aspx.
66. Morgan, S. *et al.* A comprehensive analysis of rare genetic variation in amyotrophic lateral sclerosis in the UK. *Brain* (2017). doi:10.1093/brain/awx082
67. Panoutsopoulou, K., Tachmazidou, I. & Zeggini, E. In search of low frequency and rare variants affecting complex traits. *Hum. Mol. Genet.* 1–18 (2013). doi:doi: 10.1093/hmg/ddt376
68. Kosmicki, J. A., Churchhouse, C. L., Rivas, M. A. & Neale, B. M. Discovery of rare variants for complex phenotypes. *Human Genetics* **135,** 625–634 (2016).
69. Lacey, S., Chung, J. Y. & Lin, H. A comparison of whole genome sequencing with exome sequencing for family-based association studies. *BMC Proc.* **8,** S38 (2014).
70. Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A. & Jabado, N. What can exome sequencing do for you? *J. Med. Genet.* **48,** 580–589 (2011).
71. Spielmann, M. & Mundlos, S. Structural variations, the regulatory landscape of the genome and their alteration in human disease. *BioEssays* **35,** 533–543 (2013).
72. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Brief. Funct. Genomics* **13,** 371–7 (2014).
73. Barnett, I. J., Lee, S. & Lin, X. Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies. *Genet. Epidemiol.* **37,** 142–151 (2013).
74. Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *Eur. J. Hum. Genet.* 1 (2018). doi:10.1038/s41431-018-0177-4
75. Nicolas, A. *et al.* Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* (2018). doi:10.1016/j.neuron.2018.02.027
76. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-

Linked FTD and ALS. *Neuron* **72,** 245–256 (2011).
77. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475,** 163–165 (2011).
78. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–54 (2006).
79. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315,** 848–53 (2007).
80. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics* (2016). doi:10.1038/nrg.2015.25
81. Roses, A. D. *et al.* Structural variants can be more informative for disease diagnostics, prognostics and translation than current SNP mapping and exon sequencing. *Expert Opin. Drug Metab. Toxicol.* (2016). doi:10.1517/17425255.2016.1133586
82. McCarroll, S. A. & Altshuler, D. M. Copy-number variation and association studies of human disease. *Nat. Genet.* (2007). doi:10.1038/ng2080
83. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews Genetics* (2013). doi:10.1038/nrg3373
84. Lupski, J. R. *et al.* Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy. *N. Engl. J. Med.* (2010). doi:10.1056/NEJMoa0908094
85. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* (2017). doi:10.1038/ng.3725
86. Xie, C. & Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10,** 80 (2009).
87. Chen, X. *et al.* CONSERTING: integrating copy-number analysis with structural-variation detection. *Nat. Methods* **12,** 527–30 (2015).
88. Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btv710
89. Samad, A., Huff, E. F., Cai, W. & Schwartz, D. C. Optical mapping: a novel, single-molecule approach to genomic analysis. *Genome Res.* **5,** 1–4 (1995).
90. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17,** 333–351 (2016).
91. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38,** 1767–1771

(2010).
92. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
93. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* (2010). doi:10.1093/bioinformatics/btp698
94. Zaharia, M. *et al.* Faster and More Accurate Sequence Alignment with SNAP. (2011).
95. Al-Chalabi, A., Van Den Berg, L. H. & Veldink, J. Gene discovery in amyotrophic lateral sclerosis: Implications for clinical management. *Nature Reviews Neurology* (2017). doi:10.1038/nrneurol.2016.182
96. Chiò, A. *et al.* Global epidemiology of amyotrophic lateral sclerosis: A systematic review of the published literature. *Neuroepidemiology* **41,** 118–130 (2013).
97. Alonso, A. & Hernán, M. A. Temporal trends in the incidence of multiple sclerosis: A systematic review. *Neurology* (2008). doi:10.1212/01.wnl.0000316802.35974.34
98. Mitchell, J. D. *et al.* Timelines in the diagnostic evaluation of people with suspected amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND)--a 20-year review: can we do better? *Amyotroph. lateral Scler.* **11,** 537–41 (2010).
99. Lee, J. R. J., Annegers, J. F. & Appel, S. H. Prognosis of amyotrophic lateral sclerosis and the effect of referral selection. *J. Neurol. Sci.* **132,** 207–215 (1995).
100. Sorenson, E. J., Mandrekar, J., Crum, B. & Stevens, J. C. Effect of referral bias on assessing survival in ALS. *Neurology* **68,** 600–602 (2007).
101. Logroscino, G. *et al.* Descriptive epidemiology of amyotrophic lateral sclerosis: new evidence and unsolved issues. *J. Neurol. Neurosurg. Psychiatry* **79,** 6–11 (2008).
102. Al-Chalabi, A. *et al.* Analysis of amyotrophic lateral sclerosis as a multistep process: A population-based modelling study. *Lancet Neurol.* **13,** 1108–1113 (2014).
103. Mehta, P. *et al.* Prevalence of Amyotrophic Lateral Sclerosis — United States, 2015. *MMWR. Morb. Mortal. Wkly. Rep.* **67,** 1285–1289 (2018).
104. McCombe, P. A. & Henderson, R. D. Effects of gender in amyotrophic lateral sclerosis. *Gend. Med.* (2010). doi:10.1016/j.genm.2010.11.010
105. Al-Chalabi, A. & Lewis, C. M. Modelling the effects of penetrance and family size on rates of sporadic and familial disease. *Hum. Hered.* **71,** 281–288 (2011).
106. Majounie, E. *et al.* Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: A cross-sectional study. *Lancet Neurol.* **11,** 323–330 (2012).

107. Chiò, A. *et al.* Clinical characteristics of patients with familial amyotrophic lateral sclerosis carrying the pathogenic GGGGCC hexanucleotide repeat expansion of C9ORF72. *Brain* **135,** 784–793 (2012).
108. Kenna, K. P. *et al.* NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.* (2016). doi:10.1038/ng.3626
109. Williams, K. L. *et al.* Novel TBK1 truncating mutation in a familial amyotrophic lateral sclerosis patient of Chinese origin. *Neurobiol. Aging* **36,** 3334.e1-3334.e5 (2015).
110. Rosen, D. R. *et al.* Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362,** 59–62 (1993).
111. Kaur, S. J., McKeown, S. R. & Rashid, S. Mutant SOD1 mediated pathogenesis of Amyotrophic Lateral Sclerosis. *Gene* (2016). doi:10.1016/j.gene.2015.11.049
112. Bruijn, L. I. & Cleveland, D. W. Mechanisms of selective motor neuron death in ALS: insights from transgenic mouse models of motor neuron disease. *Neuropathol. Appl. Neurobiol.* **22,** 373–387 (1996).
113. Lagier-Tourenne, C. *et al.* Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat. Neurosci.* **15,** 1488–1497 (2012).
114. Van Deerlin, V. M. *et al.* TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: a genetic and histopathological analysis. *Lancet Neurol.* **7,** 409–416 (2008).
115. Sreedharan, J. *et al.* TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science* (2008). doi:10.1126/science.1154584
116. Morita, M. *et al.* A locus on chromosome 9p confers susceptibility to ALS and frontotemporal dementia. *Neurology* (2006). doi:10.1212/01.wnl.0000200048.53766.b4
117. Laaksovirta, H. *et al.* Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: A genome-wide association study. *Lancet Neurol.* (2010). doi:10.1016/S1474-4422(10)70184-8
118. Shatunov, A. *et al.* Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: A genome-wide association study. *Lancet Neurol.* (2010). doi:10.1016/S1474-4422(10)70197-6
119. van Es, M. A. *et al.* Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat. Genet.* (2009). doi:10.1038/ng.442
120. McLaughlin, L. R., Vajda, A. & Hardiman, O. Heritability of amyotrophic lateral sclerosis insights from disparate numbers. *JAMA Neurology* (2015). doi:10.1001/jamaneurol.2014.4049
121. Alonso, A., Logroscino, G. & Hernán, M. A. Smoking and the risk of

amyotrophic lateral sclerosis: a systematic review and meta-analysis. *J. Neurol. Neurosurg. Psychiatry* **81,** 1249–52 (2010).
122. Lacorte, E. *et al.* Physical activity, and physical activity related to sports, leisure and occupational activity as risk factors for ALS: A systematic review. *Neuroscience and Biobehavioral Reviews* **66,** 61–79 (2016).
123. Scarmeas, N., Shih, T., Stern, Y., Ottman, R. & Rowland, L. P. Premorbid weight, body mass, and varsity athletics in ALS. *Neurology* **59,** 773–775 (2002).
124. Beard, J. D. *et al.* Military service, deployments, and exposures in relation to amyotrophic lateral sclerosis etiology. *Environ. Int.* **91,** 104–115 (2016).
125. Abhinav, K., Al-Chalabi, A., Hortobagyi, T. & Leigh, P. N. Electrical injury and amyotrophic lateral sclerosis: a systematic review of the literature. *J. Neurol. Neurosurg. Psychiatry* (2006). doi:10.1136/jnnp.2006.104414
126. Fischer, H. *et al.* Occupational exposure to electric shocks and magnetic fields and amyotrophic lateral sclerosis in Sweden. *Epidemiology* (2015). doi:10.1097/EDE.0000000000000365
127. Bozzoni, V. *et al.* Amyotrophic lateral sclerosis and environmental factors. *Functional Neurology* **31,** 7–19 (2016).
128. Delzor, A. *et al.* Searching for a link between the L-BMAA neurotoxin and amyotrophic lateral sclerosis: a study protocol of the French BMAALS programme. *BMJ Open* **4,** e005528 (2014).
129. Rooney, J. *et al.* No association between soil constituents and amyotrophic lateral sclerosis relative risk in Ireland. *Environ. Res.* **147,** 102–107 (2016).
130. Malek, A. M. *et al.* Exposure to hazardous air pollutants and the risk of amyotrophic lateral sclerosis. *Environ. Pollut.* **197,** 181–186 (2015).
131. Sutedja, N. a *et al.* Exposure to chemicals and metals and risk of amyotrophic lateral sclerosis: a systematic review. *Amyotroph. Lateral Scler.* (2009). doi:10.3109/17482960802455416
132. Wang, M.-D., Little, J., Gomes, J., Cashman, N. R. & Krewski, D. Identification of risk factors associated with onset and progression of amyotrophic lateral sclerosis using systematic review and meta-analysis. *Neurotoxicology* (2016). doi:10.1016/j.neuro.2016.06.015
133. Evans, M. C., Couch, Y., Sibson, N. & Turner, M. R. Inflammation and neurovascular changes in amyotrophic lateral sclerosis. *Mol. Cell. Neurosci.* **53,** 34–41 (2013).
134. Zhao, W., Beers, D. R. & Appel, S. H. Immune-mediated mechanisms in the pathoprogression of amyotrophic lateral sclerosis. *Journal of Neuroimmune Pharmacology* **8,** 888–899 (2013).
135. Henkel, J. S. *et al.* Regulatory T-lymphocytes mediate amyotrophic lateral sclerosis progression and survival. *EMBO Mol. Med.* **5,** 64–79 (2013).

136. Mantovani, S. *et al.* Immune system alterations in sporadic amyotrophic lateral sclerosis patients suggest an ongoing neuroinflammatory process. *J. Neuroimmunol.* **210,** 73–79 (2009).
137. Rentzos, M. *et al.* Alterations of T cell subsets in ALS: A systemic immune activation? *Acta Neurol. Scand.* **125,** 260–264 (2012).
138. McCormick, A. L., Brown, R. H., Cudkowicz, M. E., Al-Chalabi, A. & Garson, J. A. Quantification of reverse transcriptase in ALS and elimination of a novel retroviral candidate. *Neurology* **70,** 278–283 (2008).
139. Li, W. *et al.* Human endogenous retrovirus-K contributes to motor neuron disease. *Sci. Transl. Med.* **7,** 307ra153 (2015).
140. Chiò, A. *et al.* The multistep hypothesis of ALS revisited: The role of genetic mutations. *Neurology* 10.1212/WNL.0000000000005996 (2018). doi:10.1212/WNL.0000000000005996
141. Cedarbaum, J. M. *et al.* The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *J. Neurol. Sci.* (1999). doi:10.1016/S0022-510X(99)00210-5
142. Pupillo, E., Messina, P., Logroscino, G. & Beghi, E. Long-term survival in amyotrophic lateral sclerosis: A population-based study. *Ann. Neurol.* **75,** 287–297 (2014).
143. Westeneng, H. J., Al-Chalabi, A., Hardiman, O., Debray, T. P. & van den Berg, L. H. The life expectancy of Stephen Hawking, according to the ENCALS model. *The Lancet Neurology* (2018). doi:10.1016/S1474-4422(18)30241-2
144. Fang, T. *et al.* Stage at which riluzole treatment prolongs survival in patients with amyotrophic lateral sclerosis: A retrospective analysis of data from a dose-ranging study. *The Lancet Neurology* (2018). doi:10.1016/S1474-4422(18)30054-1
145. Trojsi, F. *et al.* Neuropsychological assessment in different King's clinical stages of amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Front. Degener.* **17,** 228–235 (2016).
146. Jones, A. R. *et al.* Health utility decreases with increasing clinical stage in amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Front. Degener.* (2014). doi:10.3109/21678421.2013.872149
147. UKMND-LiCALS Study Group *et al.* Lithium in patients with amyotrophic lateral sclerosis (LiCALS): a phase 3 multicentre, randomised, double-blind, placebo-controlled trial. *Lancet. Neurol.* **12,** 339–45 (2013).
148. Edge, S. B. & Compton, C. C. The american joint committee on cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. *Annals of Surgical Oncology* (2010). doi:10.1245/s10434-010-0985-4
149. Cedarbaum, J. M. *et al.* The ALSFRS-R: a revised ALS functional rating

scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). *J. Neurol. Sci.* **169,** 13–21 (1999).
150. Balendra, R. *et al.* Use of clinical staging in amyotrophic lateral sclerosis for phase 3 clinical trials. *J. Neurol. Neurosurg. Psychiatry* (2015). doi:10.1136/jnnp-2013-306865
151. Lacomblez, L., Bensimon, G., Leigh, P. N., Guillet, P. & Meininger, V. Dose-ranging study of riluzole in amyotrophic lateral sclerosis. Amyotrophic Lateral Sclerosis/Riluzole Study Group II. *Lancet* (1996). doi:10.1016/S0140-6736(96)91680-3
152. Balendra, R. *et al.* Use of clinical staging in amyotrophic lateral sclerosis for phase 3 clinical trials. *J. Neurol. Neurosurg. Psychiatry* **86,** 45–9 (2015).
153. Berry, J. D. *et al.* The Combined Assessment of Function and Survival (CAFS): A new endpoint for ALS clinical trials. *Amyotroph. Lateral Scler. Front. Degener.* **14,** 162–168 (2013).
154. (NICE), I. for H. and C. E. Motor neurone disease: assessment and management | recommendations | Guidance and guidelines | NICE. *Inst. Heal. Care Excell.* (2016).
155. Bowden, C. L., Mintz, J. & Tohen, M. Multi-state outcome analysis of treatments (MOAT): application of a new approach to evaluate outcomes in longitudinal studies of bipolar disorder. *Mol. Psychiatry* **21,** 237–242 (2016).
156. IBM Corp. IBM SPSS Statistics for Macintosh, Version 24.0. *2016* (2016).
157. RStudio Team. RStudio: Integrated Development for R. . *RStudio, Inc* (2015). doi:10.1007/978-81-322-2340-5
158. R Development Core Team, R. *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing* **1,** (2011).
159. SAS Institute Inc. *SAS Institute Inc. JMP Pro 12, SAS Institute Inc. Cary, NC* (2015).
160. NICE. Motor Neurone Disease: Assessment and Management (NG42). *Natl. Inst. Heal. Care Excell. Guidel.* 1–48 (2016).
161. Riviere, M., Meininger, V., Zeisser, P. & Munsat, T. An Analysis of Extended Survival in Patients With Amyotrophic Lateral Sclerosis Treated With Riluzole. *Arch. Neurol.* **55,** 526 (1998).
162. Gordon, P. H. *et al.* Progression in ALS is not linear but is curvilinear. *J. Neurol.* (2010). doi:10.1007/s00415-010-5609-1
163. Martin, N. H. *et al.* Psychological as well as illness factors influence acceptance of non-invasive ventilation (NIV) and gastrostomy in amyotrophic lateral sclerosis (ALS): A prospective population study. *Amyotroph. Lateral Scler. Front. Degener.* (2014). doi:10.3109/21678421.2014.886700
164. Johnson, J. *et al.* Eating-derived pleasure in amyotrophic lateral sclerosis

as a predictor of non-oral feeding. *Amyotroph. Lateral Scler.* (2012). doi:10.3109/17482968.2012.704925
165. Bensimon, G. *et al.* A study of riluzole in the treatment of advanced stage or elderly patients with amyotrophic lateral sclerosis. *J. Neurol.* (2002). doi:10.1007/s004150200071
166. Seibold, H., Zeileis, A. & Hothorn, T. Model-Based Recursive Partitioning for Subgroup Analyses. *Int. J. Biostat.* (2016). doi:10.1515/ijb-2015-0032
167. Wokke, J. Riluzole. *Lancet* **348,** 795–799 (1996).
168. Guidance on the use of Riluzole (Rilutek) for the treatment of Motor Neurone Disease | Guidance and guidelines | NICE.
169. Bensimon, G., Lacomblez, L. & Meininger, V. A controlled trial of riluzole in amyotrophic lateral sclerosis. ALS/Riluzole Study Group. *N. Engl. J. Med.* **330,** 585–91 (1994).
170. Stewart, A. *et al.* The clinical effectiveness and cost-effectiveness of riluzole for motor neurone disease: A rapid and systematic review. *Health Technology Assessment* (2001). doi:10.3310/hta5020
171. *AMYOTROPHIC LATERAL SCLEROSIS (ALS)-OPPORTUNITY ANALYSIS AND FORECASTS TO 2018*.
172. Eggermont, A. M. M. *et al.* Prolonged Survival in Stage III Melanoma with Ipilimumab Adjuvant Therapy. *N. Engl. J. Med.* (2016). doi:10.1056/NEJMoa1611299
173. Traynor, B. J., Alexander, M., Corr, B., Frost, E. & Hardiman, O. An outcome study of riluzole in amyotrophic lateral sclerosis--a population-based study in Ireland, 1996-2000. *J. Neurol.* (2003). doi:10.1007/s00415-003-1026-z
174. Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: Lessons from large-scale biology. *Science* (2003). doi:10.1126/science.1084564
175. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature Genetics* (2013). doi:10.1038/ng.2764
176. Dong, L. *et al.* Clinical Next Generation Sequencing for Precision Medicine in Cancer. *Curr. Genomics* **16,** 253–63 (2015).
177. Morgan, S. *et al.* Investigation of next-generation sequencing technologies as a diagnostic tool for amyotrophic lateral sclerosis. *Neurobiol. Aging* (2015). doi:10.1016/j.neurobiolaging.2014.12.017
178. Henry, V. J., Bandrowski, A. E., Pepin, A. S., Gonzalez, B. J. & Desfeux, A. OMICtools: an informative directory for multi-omic data analysis. *Database (Oxford).* (2014). doi:10.1093/database/bau069
179. Landrum, M. J. *et al.* ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkv1222
180. Karczewski, K. J. *et al.* The ExAC browser: Displaying reference data

information from over 60 000 exomes. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw971
181. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* (2001). doi:10.1093/nar/29.1.308
182. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* (2011). doi:10.1002/humu.21517
183. Dabbish, L., Stuart, C., Tsay, J. & Herbsleb, J. Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository. in *Proceedings of the 2012 ACM Conference on Computer Supported Cooperative Work* (2012). doi:10.1145/2145204.2145396
184. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS One* (2017). doi:10.1371/journal.pone.0177459
185. Frisoni, G. B. *et al.* The pilot European Alzheimer's Disease Neuroimaging Initiative of the European Alzheimer's Disease Consortium. *Alzheimer's Dement.* (2008). doi:10.1016/j.jalz.2008.04.009
186. Van Blitterswijk, M., Dejesus-Hernandez, M. & Rademakers, R. How do C9ORF72 repeat expansions cause amyotrophic lateral sclerosis and frontotemporal dementia: Can we learn from other noncoding repeat expansion disorders? *Current Opinion in Neurology* (2012). doi:10.1097/WCO.0b013e32835a3efb
187. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* (2015). doi:10.1038/nmeth.3317
188. Lai, Y. A block-sorting lossless data compression algorithm. in *IMID 2009* (2009). doi:10.1.1.37.6774
189. Sirén, J., Välimäki, N. & Mäkinen, V. Indexing graphs for path queries with applications in genome research. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* (2014). doi:10.1109/TCBB.2013.2297101
190. Simpson, J. T. & Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* (2010). doi:10.1093/bioinformatics/btq217
191. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* (2012). doi:10.1038/nmeth.1923
192. Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P. & Kocher, J. P. A. Indel detection from RNA-seq data: Tool evaluation and strategies for accurate detection of actionable mutations. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbw069
193. Faust, G. G. & Hall, I. M. SAMBLASTER: Fast duplicate marking and structural variant read extraction. in *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu314
194. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P.

Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv098
195. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3432
196. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* (2010). doi:10.1101/gr.107524.110
197. Sandmann, S. *et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci. Rep.* (2017). doi:10.1038/srep43169
198. Smith, H. E. & Yun, S. Evaluating alignment and variant-calling software for mutation identification in C. elegans by whole-genome sequencing. *PLoS One* (2017). doi:10.1371/journal.pone.0174446
199. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr330
200. Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* (2017). doi:10.1101/gr.225672.117
201. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral Genomes resource. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gku1207
202. Agarwala, R. *et al.* Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw1071
203. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).
204. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw354
205. Miller, C. A., Qiao, Y., Disera, T., D'Astous, B. & Marth, G. T. Bam.iobio: A web-based, real-time, sequence alignment file inspector. *Nature Methods* (2014). doi:10.1038/nmeth.3174
206. Baruzzo, G. *et al.* Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods* (2016). doi:10.1038/nmeth.4106
207. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* (2015). doi:10.1038/nmeth.3505
208. De Iaco, A. *et al.* TNPO3 protects HIV-1 replication from CPSF6-mediated capsid stabilization in the host cell cytoplasm. *Retrovirology* (2013). doi:10.1186/1742-4690-10-20
209. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
210. Eberle, M. A. *et al.* A reference data set of 5.4 million phased human

variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* (2017). doi:10.1101/gr.210500.116

211. Abel, O., Powell, J. F., Andersen, P. M. & Al-Chalabi, A. ALSoD: A user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. *Hum. Mutat.* (2012). doi:10.1002/humu.22157
212. Rippon, G. A. *et al.* An observational study of cognitive impairment in amyotrophic lateral sclerosis. *Arch. Neurol.* (2006). doi:10.1001/archneur.63.3.345
213. Synofzik, M., Otto, M., Ludolph, A. & Weishaupt, J. H. Genetische Architektur der amyotrophen Lateralsklerose und frontotemporalen Demenz. *Nervenarzt* **88,** 728–735 (2017).
214. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* (2014). doi:10.1038/ng.2892
215. Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C. & Pati, A. Large-scale contamination of microbial isolate genomes by illumina Phix control. *Stand. Genomic Sci.* (2015). doi:10.1186/1944-3277-10-18
216. Daly, G. M. *et al.* Host subtraction, filtering and assembly validations for novel viral discovery using next generation sequencing data. *PLoS One* (2015). doi:10.1371/journal.pone.0129059
217. Liu, J. *et al.* Toxicity of familial ALS-linked SOD1 mutants from selective recruitment to spinal mitochondria. *Neuron* (2004). doi:10.1016/j.neuron.2004.06.016
218. Wroe, R., Wai-Ling Butler, A., Andersen, P. M., Powell, J. F. & Al-Chalabi, A. ALSOD: the Amyotrophic Lateral Sclerosis Online Database. *Amyotroph. Lateral Scler.* **9,** 249–50 (2008).
219. Goksuluk, D., Korkmaz, S., Zararsiz, G. & Karaagaoglu, A. E. easyROC: An Interactive Web-tool for ROC Curve Analysis Using R Language Environment. *R J.* (2016).
220. Dirk Merkel. Docker Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.* (2014). doi:10.1097/01.NND.0000320699.47006.a3
221. Muzumdar, R. & Atzmon, G. Telomere Length and Aging. in *Reviews on Selected Topics of Telomere Biology* 3–30 (2012). doi:10.5772/2329
222. Kong, C. M., Lee, X. W. & Wang, X. Telomere shortening in human diseases. *FEBS Journal* **280,** 3180–3193 (2013).
223. Gardner, M. *et al.* Gender and telomere length: Systematic review and meta-analysis. *Exp. Gerontol.* (2014). doi:10.1016/j.exger.2013.12.004
224. Xu, L., Li, S. & Stohr, B. A. The Role of Telomere Biology in Cancer. *Annu. Rev. Pathol. Mech. Dis.* **8,** 49–78 (2013).
225. Illumina. HiSeq™ 2000 Sequencing System. *Specif. Sheet Illumina® Seq.* (2010).

226. Ding, Z., Mangino, M., Aviv, A., Spector, T. & Durbin, R. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res.* **42,** (2014).
227. Pooley, K. A. *et al.* A genome-wide association scan (GWAS) for mean telomere length within the COGS project: Identified loci show little association with hormone-related cancer risk. *Hum. Mol. Genet.* **22,** 5056–5064 (2013).
228. Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nat. Genet.* **45,** 422–427 (2013).
229. Mangino, M. *et al.* Genome-wide meta-analysis points to CTC1 and ZNf676 as genes regulating telomere homeostasis in humans. *Hum. Mol. Genet.* (2012). doi:10.1093/hmg/dds382
230. Levy, D. *et al.* Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. *Proc. Natl. Acad. Sci.* **107,** 9293–9298 (2010).
231. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **92,** 841–853 (2013).
232. Min, J., Wright, W. E. & Shay, J. W. Alternative lengthening of telomeres can be maintained by preferential elongation of lagging strands. *Nucleic Acids Res.* **45,** 2615–2628 (2017).
233. Haycock, P. C. *et al.* Association between telomere length and risk of cancer and non-neoplastic diseases a mendelian randomization study. *JAMA Oncol.* **3,** 636–651 (2017).
234. Bryan, T. M., Englezou, A., Gupta, J., Bacchetti, S. & Reddel, R. R. Telomere elongation in immortal human cells without detectable telomerase activity. *EMBO J* **14,** 4240–4248 (1995).
235. Arora, R. & Azzalin, C. M. Telomere elongation chooses TERRA ALTernatives. *RNA Biol.* **12,** 938–941 (2015).
236. Blackburn, E. H., Greider, C. W. & Szostak, J. W. Telomeres and telomerase: The path from maize, Tetrahymena and yeast to human cancer and aging. *Nature Medicine* (2006). doi:10.1038/nm1006-1133
237. Cesare, A. J. & Reddel, R. R. Alternative lengthening of telomeres: Models, mechanisms and implications. *Nature Reviews Genetics* (2010). doi:10.1038/nrg2763
238. Linkus, B. *et al.* Telomere shortening leads to earlier age of onset in ALS mice. *Aging (Albany. NY).* **8,** 382–393 (2016).
239. De Felice, B. *et al.* Telomerase expression in amyotrophic lateral sclerosis (ALS) patients. *J. Hum. Genet.* **59,** 555–561 (2014).
240. Mesci, P. *et al.* System xC- is a mediator of microglial function and its deletion slows symptoms in amyotrophic lateral sclerosis mice. *Brain* (2015). doi:10.1093/brain/awu312
241. Flanary, B. E. & Streit, W. J. Effects of axotomy on telomere length,

telomerase activity, and protein in activated microglia. *J. Neurosci. Res.* (2005). doi:10.1002/jnr.20636
242. Barthel, F. P. *et al.* Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* (2017). doi:10.1038/ng.3781
243. Cook, D. E. *et al.* The genetic basis of natural variation in Caenorhabditis elegans telomere length. *Genetics* (2016). doi:10.1534/genetics.116.191148
244. Cai, N. *et al.* Molecular signatures of major depression. *Curr. Biol.* (2015). doi:10.1016/j.cub.2015.03.008
245. Aviv, A. *et al.* Impartial comparative analysis of measurement of leukocyte telomere length/DNA content by Southern blots and qPCR. *Nucleic Acids Res.* **39,** e134–e134 (2011).
246. Ma, T. S. Applications and limitations of polymerase chain reaction amplification. *Chest* **108,** 1393–1404 (1995).
247. Consortium, P. M. *et al.* The Project MinE databrowser: bringing large-scale whole-genome sequencing in ALS to researchers and the public. *bioRxiv* 377911 (2018). doi:10.1101/377911
248. Cesare, A. J. & Griffith, J. D. Telomeric DNA in ALT Cells Is Characterized by Free Telomeric Circles and Heterogeneous t-Loops. *Mol. Cell. Biol.* (2004). doi:10.1128/mcb.24.22.9948-9957.2004
249. Conomos, D. *et al.* Variant repeats are interspersed throughout the telomeres and recruit nuclear receptors in ALT cells. *J. Cell Biol.* (2012). doi:10.1083/jcb.201207189
250. Marzec, P. *et al.* Nuclear-Receptor-Mediated Telomere Insertion Leads to Genome Instability in ALT Cancers. *Cell* (2015). doi:10.1016/j.cell.2015.01.044
251. Episkopou, H. *et al.* Alternative Lengthening of Telomeres is characterized by reduced compaction of telomeric chromatin. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gku114
252. Baerlocher, G. M., Mak, J., Tien, T. & Lansdorp, P. M. Telomere length measurement by fluorescence in situ hybridization and flow cytometry: Tips and pitfalls. *Cytometry* **47,** 89–99 (2002).
253. Eastmond, D. A., Schuler, M. & Rupa, D. S. Advantages and limitations of using fluorescence in situ hybridization for the detection of aneuploidy in interphase human cells. *Mutat. Res. Lett.* **348,** 153–162 (1995).
254. Aviv, A., Valdes, A. M. & Spector, T. D. Human telomere biology: pitfalls of moving from the laboratory to epidemiology. *Int. J. Epidemiol.* **35,** 1424–1429 (2006).
255. Barrett, J. H., Iles, M. M., Dunning, A. M. & Pooley, K. A. Telomere length and common disease: study design and analytical challenges. *Hum. Genet.* (2015). doi:10.1007/s00439-015-1563-4
256. Moutsianas, L. *et al.* The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About

Complex Disease. *PLoS Genet.* (2015). doi:10.1371/journal.pgen.1005165
257. Meynert, A. M., Ansari, M., FitzPatrick, D. R. & Taylor, M. S. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* **15,** 247 (2014).
258. Johnston, C. A. *et al.* Amyotrophic lateral sclerosis in an urban setting. *J. Neurol.* **253,** 1642–1643 (2006).
259. Martin, S., Al Khleifat, A. & Al-Chalabi, A. What causes amyotrophic lateral sclerosis? *F1000Research* **6,** 371 (2017).
260. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* (2016). doi:10.1186/s13059-016-0974-4
261. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* (2016). doi:10.1038/nature19057
262. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
263. Price, A. L. *et al.* Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am. J. Hum. Genet.* (2010). doi:10.1016/j.ajhg.2010.04.005
264. Greenwood, C. M. T., Xu, C. & Ciampi, A. Significance Thresholds for Rare Variant Signals. in *Assessing Rare Variation in Complex Traits* 169–183 (Springer New York, 2015). doi:10.1007/978-1-4939-2824-8_12
265. Watanabe, T. *et al.* Human arylhydrocarbon receptor repressor (AHRR) gene: genomic structure and analysis of polymorphism in endometriosis. *J. Hum. Genet.* (2001). doi:10.1007/s100380170070
266. Ash, P. E. A. *et al.* Dioxins and related environmental contaminants increase TDP-43 levels. *Mol. Neurodegener.* (2017). doi:10.1186/s13024-017-0177-9
267. Pooley, K. A. *et al.* A genome-wide association scan (GWAS) for mean telomere length within the COGS project: Identified loci show little association with hormone-related cancer risk. *Hum. Mol. Genet.* (2013). doi:10.1093/hmg/ddt355
268. Levy, D. *et al.* Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. *Proc. Natl. Acad. Sci.* (2010). doi:10.1073/pnas.0911494107
269. Kiryluk, K. Challenges in Rare Variant Association Studies for Complex Kidney Traits: CFHR5 and IgA Nephropathy. *J. Am. Soc. Nephrol.* **27,** 2547–51 (2016).
270. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *J. Hum. Genet.* (2008). doi:10.1016/j.ajhg.2007.12.009.
271. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* (2010). doi:10.1038/nature09146
272. Blauw, H. M. *et al.* A large genome scan for rare CNVs in amyotrophic

lateral sclerosis. *Hum. Mol. Genet.* (2010). doi:10.1093/hmg/ddq323
273. Wain, L. V. *et al.* The role of copy number variation in susceptibility to amyotrophic lateral sclerosis: Genome-wide association study and comparison with published loci. *PLoS One* (2009). doi:10.1371/journal.pone.0008175
274. Gambin, T. *et al.* Identification of novel candidate disease genes from de novo exonic copy number variants. *Genome Med.* (2017). doi:10.1186/s13073-017-0472-7
275. Dhami, P. *et al.* Exon Array CGH: Detection of Copy-Number Changes at the Resolution of Individual Exons in the Human Genome. *Am. J. Hum. Genet.* (2005). doi:10.1086/429588
276. Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res.* **22,** 1525–1532 (2012).
277. Hsiung, G. Y. R. *et al.* Clinical and pathological features of familial frontotemporal dementia caused by C9ORF72 mutation on chromosome 9p. *Brain* (2012). doi:10.1093/brain/awr354
278. Gijselinck, I. *et al.* A C9orf72 promoter repeat expansion in a Flanders-Belgian cohort with disorders of the frontotemporal lobar degeneration-amyotrophic lateral sclerosis spectrum: A gene identification study. *Lancet Neurol.* (2012). doi:10.1016/S1474-4422(11)70261-7
279. Van Mossevelde, S. *et al.* Clinical Evidence of Disease Anticipation in Families Segregating a *C9orf72* Repeat Expansion. *JAMA Neurol.* **74,** 445 (2017).
280. Van Mossevelde, S., van der Zee, J., Cruts, M. & Van Broeckhoven, C. Relationship between C9orf72 repeat size and clinical phenotype. *Current Opinion in Genetics and Development* (2017). doi:10.1016/j.gde.2017.02.008
281. Gijselinck, I. *et al.* The C9orf72 repeat size correlates with onset age of disease, DNA methylation and transcriptional downregulation of the promoter. *Mol. Psychiatry* (2016). doi:10.1038/mp.2015.159
282. van der Zee, J. *et al.* A Pan-European Study of the C9orf72 Repeat Associated with FTLD: Geographic Prevalence, Genomic Instability, and Intermediate Repeats. *Hum. Mutat.* (2013). doi:10.1002/humu.22244
283. Dols-Icardo, O. *et al.* Analysis of known amyotrophic lateral sclerosis and frontotemporal dementia genes reveals a substantial genetic burden in patients manifesting both diseases not carrying the C9orf72 expansion mutation. *J. Neurol. Neurosurg. Psychiatry* (2018). doi:10.1136/jnnp-2017-316820
284. Neumann, M. *et al.* TDP-43 in the ubiquitin pathology of frontotemporal dementia with VCP gene mutations. *J. Neuropathol. Exp. Neurol.* (2007). doi:10.1097/nen.0b013e31803020b9
285. Rosso, S. M. Familial frontotemporal dementia with ubiquitin-positive inclusions is linked to chromosome 17q21-22. *Brain* (2001).

doi:10.1093/brain/124.10.1948
286. Dittwald, P. *et al.* Inverted Low-Copy Repeats and Genome Instability-A Genome-Wide Analysis. *Hum. Mutat.* (2013). doi:10.1002/humu.22217
287. Kimonis, V. E. *et al.* Clinical studies in familial VCP myopathy associated with paget disease of bone and frontotemporal dementia. *Am. J. Med. Genet. Part A* (2008). doi:10.1002/ajmg.a.31862
288. Watts, G. D. J. *et al.* Inclusion body myopathy associated with Paget disease of bone and frontotemporal dementia is caused by mutant valosin-containing protein. *Nat. Genet.* (2004). doi:10.1038/ng1332
289. Ju, J. S. & Weihl, C. C. Inclusion body myopathy, Paget's disease of the bone and fronto-temporal dementia: A disorder of autophagy. *Hum. Mol. Genet.* (2010). doi:10.1093/hmg/ddq157
290. Weihl, C. C., Pestronk, A. & Kimonis, V. E. Valosin-containing protein disease: Inclusion body myopathy with Paget's disease of the bone and fronto-temporal dementia. *Neuromuscular Disorders* (2009). doi:10.1016/j.nmd.2009.01.009
291. Mehta, P. R. *et al.* Younger age of onset in familial amyotrophic lateral sclerosis is a result of pathogenic gene variants, rather than ascertainment bias. *J. Neurol. Neurosurg. Psychiatry* jnnp-2018-319089 (2018). doi:10.1136/jnnp-2018-319089
292. Murray, S. A. Illness trajectories and palliative care. *BMJ* (2005). doi:10.1136/bmj.330.7498.1007
293. Ballentine, J. M. *The Five Trajectories Supporting Patients During Serious Illness The Five Trajectories: Supporting Patients During Serious Illness*. (2018).
294. Nijjar, S. *et al.* Participation in clinical trials improves outcomes in women's health: a systematic review and meta-analysis. *BJOG An Int. J. Obstet. Gynaecol.* **124,** 863–871 (2017).
295. Oh, J. *et al.* Socioeconomic costs of amyotrophic lateral sclerosis according to staging system. *Amyotroph. Lateral Scler. Front. Degener.* (2015). doi:10.3109/21678421.2014.999791
296. Pinto, S. & de Carvalho, M. Comparison of slow and forced vital capacities on ability to predict survival in ALS. *Amyotroph. Lateral Scler. Front. Degener.* (2017). doi:10.1080/21678421.2017.1354995
297. Tramacere, I. *et al.* The MITOS system predicts long-term survival in amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatry* (2015). doi:10.1136/jnnp-2014-310176
298. Trojsi, F. *et al.* High angular resolution diffusion imaging abnormalities in the early stages of amyotrophic lateral sclerosis. *J. Neurol. Sci.* (2017). doi:10.1016/j.jns.2017.07.039
299. Floeter, M. K., Danielian, L. E., Braun, L. E. & Wu, T. Longitudinal diffusion imaging across the C9orf72 clinical spectrum. *J. Neurol. Neurosurg. Psychiatry* (2018). doi:10.1136/jnnp-2017-316799

300. Gaiani, A. *et al.* Diagnostic and prognostic biomarkers in amyotrophic lateral sclerosis: Neurofilament light chain levels in definite subtypes of disease. *JAMA Neurol.* (2017). doi:10.1001/jamaneurol.2016.5398
301. Puentes, F. *et al.* Immune reactivity to neurofilament proteins in the clinical staging of amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatry* (2014). doi:10.1136/jnnp-2013-305494
302. Lee, J., Baek, H., Kim, S. H. & Park, Y. Association between estimated total daily energy expenditure and stage of amyotrophic lateral sclerosis. *Nutrition* (2017). doi:10.1016/j.nut.2016.06.007
303. Ehinger, J. K., Morota, S., Hansson, M. J., Paul, G. & Elmér, E. Mitochondrial dysfunction in blood cells from amyotrophic lateral sclerosis patients. *J. Neurol.* **262,** 1493–1503 (2015).
304. Faes, L. & Callewaert, G. Mitochondrial dysfunction in familial amyotrophic lateral sclerosis. *Journal of Bioenergetics and Biomembranes* (2011). doi:10.1007/s10863-011-9393-0
305. Ringholz, G. M. *et al.* Prevalence and patterns of cognitive impairment in sporadic ALS. *Neurology* (2005). doi:10.1212/01.wnl.0000172911.39167.b6
306. Ling, S. C., Polymenidou, M. & Cleveland, D. W. Converging mechanisms in als and FTD: Disrupted RNA and protein homeostasis. *Neuron* (2013). doi:10.1016/j.neuron.2013.07.033
307. Burke, T. *et al.* A Cross-sectional population-based investigation into behavioral change in amyotrophic lateral sclerosis: subphenotypes, staging, cognitive predictors, and survival. *Ann. Clin. Transl. Neurol.* (2017). doi:10.1002/acn3.407
308. Lulé, D. *et al.* Cognitive phenotypes of sequential staging in amyotrophic lateral sclerosis. *Cortex* **101,** 163–171 (2018).
309. Pompl, P. N. *et al.* A therapeutic role for cyclooxygenase-2 inhibitors in a transgenic mouse model of amyotrophic lateral sclerosis. *FASEB J.* **17,** 725–7 (2003).
310. Kettenmann. Neuroglia. *Oxford, UK. Oxford Univ. Press.* **Third Edit,**
311. Antel JP, Richman DP, A. B. Immunogenetics and amyotrophic lateral sclerosis. *UCLA Forum Med Sci* **(19):151-7,** (1976).
312. Bataveljic, D., Milosevic, M., Radenovic, L. & Andjus, P. Novel molecular biomarkers at the blood-brain barrier in ALS. *BioMed Research International* **2014,** (2014).
313. Butovsky, O. *et al.* Modulating inflammatory monocytes with a unique microRNA gene signature ameliorates murine ALS. *J. Clin. Invest.* **122,** 3063–3087 (2012).
314. Lam, L. *et al.* Epigenetic changes in T-cell and monocyte signatures and production of neurotoxic cytokines in ALS patients. *Fed. Am. Soc. Exp. Biol.* 1–13 (2016). doi:10.1096/fj.201600259RR
315. Parkes, M., Cortes, A., van Heel, D. A. & Brown, M. A. Genetic insights

into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.* **14,** 661–73 (2013).
316. Song, S. *et al.* Major histocompatibility complex class {I} molecules protect motor neurons from astrocyte-induced toxicity in amyotrophic lateral sclerosis. *Nat. Med.* **22,** 397–403 (2016).

Appendix

Supplementary materials

Clinical vignettes

Case 1: A 59 year old man describes a worsening history of difficulty with keys, turning door handles, and manipulating buttons over the last year. His arm has become noticeably thinner in that time, and occasionally the muscles twitch. He cannot straighten the fingers of his right hand easily. He notices his walking has also slowed down, so that whereas he could previously walk a mile in about 15 minutes, it now takes about 30. There is no problem with speaking or swallowing, and there are no respiratory symptoms. Clinical examination confirms distal weakness of the right upper limb and proximal wasting and weakness of the lower limbs, with brisk reflexes in some muscles.

Case 2: A 65 year old man developed problems with walking two years ago. He initially noticed his right foot was dragging, and occasionally he would trip over. Eighteen months ago, he started finding it increasingly difficult to lift heavy objects and to dress himself. Over the past 6 months he has found his speech has become slurred, although he has not had difficulty with chewing or swallowing food. He was previously able to walk for a mile without becoming short of breath, but now after 10 minutes he feels breathless. He also becomes short of breath when lying flat at night and he wakes up with a headache. His clinical examination reveals weakness of both lower limbs, primarily on the right side, and some proximal weakness in his upper limbs, with pathologically brisk reflexes. His peripheral capillary oxygen saturation on room air is 93% and an arterial blood gas test reveal that he is hypercarbic with pCO_2 of 6.3 kPa.

Case 3: A 60 year old woman reports a two and a half year history of difficultly walking, and she finds that recently this has become worse. Her main problem is that her left leg feels weak, and she finds it drags while she is walking. She also notices twitching in her leg muscles. She now is unable to walk for longer than 10 minutes. A few weeks ago, she started having weakness in her right hand, finding that she cannot easily use her right thumb when picking up small objects. She has not experienced any difficulty swallowing or with speech, or with her breathing. Clinical examination reveals she has distal wasting and weakness of both lower limbs and her right hand and a pathologically brisk jaw jerk reflex.

Case 4: A 71 year old man reports a seven month history of pain and weakness in his right leg making walking difficult. He has not noticed any weakness in his arms, or any problems swallowing or with his speech. His breathing has also been normal. Clinical examination confirms he has proximal right lower limb weakness. Several fasciculations are noted in his right upper limb and he has pathologically brisk reflexes in all limbs with bilateral finger jerks and Hoffmann's jerks elicited.

Case 5: A 63 year old man has a fourteen month history of slurred speech, finding this has become worse over the past four months. His speech is now difficult to understand, and his voice has become softer, so he has to repeat himself frequently. He notices he is having excessive saliva, and finds it difficult to swallow this, so he frequently drools. Over this same period of time he finds himself choking whilst eating solid food, so he now can only manage a food of a very soft consistency or liquid food, but occasionally chokes on this also. His clinicians recommend that he should now consider having a gastrostomy inserted, but he decides against this intervention.

Case 6: A 75 year old man describes a six month history of difficulty using his right arm. He finds he cannot carry heavy objects and now finds it problematic combing his hair and brushing his teeth with his right arm. He sometimes finds it difficult to cut up food and notices handwriting is slower and clumsier. He has no difficulty with walking or with his speech, swallowing or breathing. Clinical examination reveals proximal wasting and weakness of his right upper limb with pathologically brisk upper limb reflexes.

Case 7: A 41 year old man noticed cramps in his left arm nine months ago with weakness, and now his right arm and hand are also weak. He can no longer work as a builder as he cannot lift his building materials and tools. He sometimes also finds it difficult to tie his shoelaces and fasten zips and buttons. He is still able to walk to the local train station one mile away, and does not feel that his legs are weak, that his walking has slowed down or that he has difficulty climbing stairs. On clinical examination his strength is reduced in the upper limbs and normal in the lower limbs with pathologically brisk knee and ankle reflexes and crossed adductor reflexes. Plantars are downgoing.

Case 8: A 66 year old woman has a four year history of worsening weakness of her left leg, leading to difficulty walking. She now uses a walking stick around the house, and before her symptoms started, she could walk a mile, but can now no longer do this. Over the past two years she has also noticed difficulty lifting her grandson and finds it challenging to cut up food and dress herself. Over the past few months she occasionally gets breathless at night and on exertion. She does not sleep well, finding she is sleepy during the day. Her peripheral capillary oxygen saturation on room air is 95% and her other respiratory tests are within normal limits.

Case 9: A 77 year old man has had a two year history of progressive difficulty with walking. He used to be able to walk a mile in 20 minutes, but now cannot walk that far without his legs feeling stiff. He has not had any weakness in his arms, and has not had difficulty swallowing food, with his speech or breathing. His clinical examination confirms proximal weakness, increased tone and pathologically brisk reflexes in his lower limbs but is otherwise normal. He had a gastrostomy tube inserted four months ago due to oropharyngeal cancer, which is being treated with surgery, chemotherapy and radiotherapy.

Case 10: An 84 year old woman has progressive dysarthria and excessive salivation for seven months. On examination she has a wasted tongue, a pathologically brisk jaw jerk and bilateral wasting of the small muscles of her hands, including the first dorsal interossei. The rest of the examination is normal.

Case 11: A 72 year old woman has a fourteen month history of dragging her right foot while walking. Her left leg also became affected three months ago and she has no other weakness. She is found by her doctor to have pathologically brisk patellar reflexes and brisk biceps and supinator reflexes in her right arm. All other examination findings are normal.

Case 12: A 55 year old woman describes a ten month history of weakness and stiffness in both hands. A few months later she starts to limp with her right foot slapping on the ground. She has no reported problems with her speech or swallowing, and no respiratory symptoms. On examination she has a wasted, fasciculating tongue and distal wasting and weakness in her limbs. During the clinic visit her voice becomes quieter and she slurs her sentences. Respiratory testing shows the vital capacity is 75% of predicted.

Case 13: A 46 year old man has a nine month history of progressive weakness of the legs. On examination he has increased tone in all four limbs with wasting and fasciculation in his lower limbs and crossed adductor reflexes.

Case 14: A 57 year old man noticed muscle wasting in his right hand two years ago. Last year his right foot started to flap when he was walking. Over the past two months his speech has become slurred after talking for a long time. Examination reveals a wasted and fasciculating tongue, weakness of the right hand and a right foot drop.

Case 15: A 70 year old man has a year long history of difficulty swallowing, choking and excessive salivation. He becomes very fatigued when eating and suffers from frequent chest infections despite a modified diet. His weight has dropped from 65 Kg to 55 Kg.

Case 16: A 54 year old woman reports weakness in her left leg slowly getting worse over the last three years. Over the past year her arms have become weaker. For the past few months she has been feeling short of breath and sleeping poorly. Her forced vital capacity (FVC) is less than 50% of the predicted value.

Case 17: A 70 year old woman reports a four year history of difficulty walking requiring walking aids with no other symptoms. Examination reveals increased tone in the all the limbs, more so on the right side. She has a pyramidal pattern of weakness in all limbs, and is unable to lift her arms. She has pathologically brisk reflexes in her limbs. There are no bulbar signs, no weight loss, and respiratory testing is normal.