LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

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

Copyright statement

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

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

Liability statement

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

Access Agreement

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

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

THE GENETICS OF AMYOTROPHIC LATERAL SCLEROSIS

Kevin Kenna

A thesis submitted to the University of Dublin
for the degree of Doctor of Philosophy.

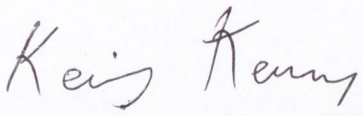
Smurfit Institute of Genetics
University of Dublin, Trinity College.
2013

TRINITY COLLEGE
5 JAN 2015
LIBRARY DUBLIN

DECLARATION

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

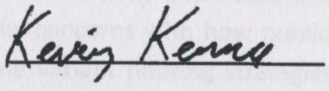
I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed: 

DECLARATION

I hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other University and that, unless otherwise stated, the work described is entirely my own. I agree that Trinity College Library may lend or copy this thesis upon request.

Signed,



Kevin Kenna
30/09/2013

SUMMARY

Amyotrophic lateral sclerosis, commonly referred to as motor neuron disease or simply ALS, is a neurological disorder which involves the degeneration of upper and lower motor neurons. The disease is associated with a progressive loss in voluntary muscle function and usually proves fatal within 3 years of symptomatic onset. There is no cure and the most effective therapy to date is believed to extend survival by only a few months. Multiple genetic causes have been identified but these account for only a minority of cases and in most instances disease onset cannot be explained. Furthermore, the true pathogenicity of many reported disease mutations has not been validated.

The work described in this thesis is centred around 5 primary analyses, each of which was intended to improve understanding of the genetic factors that do and do not play a role in the aetiology of ALS. Three of these analyses take advantage of recent developments in DNA sequencing technologies and a significant proportion of the text is therefore dedicated to the implementation of these technologies and quality control of the data they can be used to generate. There is also some discussion about additional works currently in progress and possible directions for future research.

The first analysis discussed in this work represents a statistical evaluation of the pathogenicity of previously reported disease variants using data mined from public archives of genetic variation and large population resequencing projects. The results cast serious doubt over the pathogenicity of at least 50 reported ALS mutations, in addition to mutations implicated in other disorders. Perhaps more importantly, they also highlight serious concerns with how public genetic data is commonly used in disease gene mapping and outline variant filtering strategies which achieve controllable type I error rates. The second analysis, concerns the *in silico* functional profiling of 9 known ALS genes. This involves extensive mining of biological databases and the implementation of various network and similarity analyses for the purposes of identifying potentially disease related gene features. The results revealed numerous properties of the 9 training genes which could be used to achieve significant enrichments of the genome for other ALS related genes. These features were then used to guide the selection of promising candidate genes for genetic investigation. The third analysis involved the screening of a population based cohort of ALS patients and matched controls for variation across 33 of the most well established disease loci. This represented the most comprehensive survey of relative ALS gene variant frequencies to date and the first reported application of multiplexed targeted high-throughput sequencing in ALS research. The results revealed that only 10% of Irish ALS cases could be explained by known disease mutations but that at least a further 7% carried possible disease variants. There was also evidence to indicate that in certain cases disease susceptibility represents an oligogenic process with certain ALS mutations potentially only causing disease in specific genetic backgrounds. Finally, comparison of the results with those reported from related analyses of other population based patient cohorts revealed significant population-based heterogeneity in ALS aetiology, a phenomenon with potentially important consequences for international collaborations, replication studies and the generalization of research findings across ethnic backgrounds. The

fourth analysis represented a genetic follow up on the candidate genes selected through *in silico* functional profiling. The investigation involved a very limited number of samples meaning that there was not sufficient power to definitively establish the effects of even known ALS variants. However, an extensive battery of single variant and aggregate analyses revealed numerous highly promising findings, and further analysis of certain candidate genes may yet provide important new insights into ALS pathobiology. The final analysis involved screening of the mitochondrial genome for ALS related variation using off-target reads generated during previous targeted resequencing experiments. Mutations of mitochondrial DNA have previously been reported to cause rare motor neuron disorders, however our investigation identified only one of these mutations in a cohort of over 400 ALS patients and it was observed to occur with equal frequency among cases and controls. Nonetheless, more comprehensive analyses of case-control mutation rates revealed highly suggestive evidence for an association between disease risk and rare nonsynonymous variants of the gene *ND4*. Replication studies will be required to determine the authenticity of this signal but if validated the finding could represent an important new insight into the nature of ALS pathogenesis.

In summary the work described in this thesis has called into question the pathogenicity of many mutations previously associated with ALS or other disorders, highlighted potential consequences of the inappropriate use of reference genetic data in disease gene mapping, demonstrated the potential utility of *in silico* functional profiling in the prioritization of undiscovered disease genes, characterized the relative and cumulative contributions of previously reported ALS genes to disease burden in Ireland, revealed evidence to support the relevance of oligogenic susceptibility in ALS aetiology, validated the existence of significant population based heterogeneity in ALS aetiology and revealed several promising candidate disease loci for follow up in future work.

ACKNOWLEDGEMENTS

The work described in this thesis primarily concerns the genetics of the neurodegenerative disease Amyotrophic Lateral Sclerosis and involved the use of clinical data and DNA donated by healthy volunteers and persons afflicted by this devastating disorder. Without their kind and invaluable participation most of the described work would not have been possible. It would also not have been possible without the contributions and guidance of the project supervisors Prof. Orla Hardiman and Prof Daniel Bradley and funding provided by the Health Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 259867, the Irish Health Research Board , the charities Trinity Foundation and Research Motor Neurone and both the individuals who donate to these charities and fundraise on their behalf. All data relating to family histories for patients from the Republic of Ireland was kindly provided by Dr. Susan Byrne while all data relating to patient cognitive status was kindly provided by Dr Marwa Elamin. All other clinical information for patients from the Republic Ireland was obtained from the Irish ALS register while DNA for patients and controls from the Republic of Ireland was obtained from the associated DNA bank. All clinical data and DNA samples relating to individuals from Northern Ireland were obtained from the ALS register of Northern Ireland, courtesy of Dr Collette Donaghy. Preparation of libraries for targeted resequencing using "capture 2" was performed by Dr Russell McLaughlin.

My utmost thanks go to Prof Hardiman and to Prof Bradley for all their time, effort and advice and for allowing me the opportunity to get involved with ALS research in the first place. I would also like to thank those that I have worked with over the past few years, particularly Susan Byrne, Marwa Elamin, Peter Bede, Russell Mc Laughlin, Catherine Lynch, Mark Heverin and Dominique Plante from Prof Hardiman's team and Cristina Gamba, Emma Finlay, Eppie Ruth Jones, Frauke Stocke, Ian Richardson, Matthew Teasdale, Rui Martiniano, Sarah Connell, Valeria Mattiangeli, Victoria Mullin and Yonas Hirutu from Prof Bradley's lab. Thanks first of all for just being nice people and secondly for having always been willing to provide help if needed. I would like to extend a special thanks to Dr. Matthew Teasdale, for helping me get started (and continue) up a rather steep learning curve in computer programming and to Dr. Russell Mc Laughlin, for always being willing to hear out and scrutinize any idea, no matter how undercooked.

Finally, I could not go without also acknowledging those who I've relied upon most, throughout or prior to this work, namely extended family and friends, my grandfather Kevin, my brothers Stephen and Brendan, my father Pat and mother Cathy and in particular the extremely patient woman with whom I've been lucky enough to share the past 9 years (and without whom I would be completely lost), Aoife.

ACRONYMS

AD	Alzheimer's disease
AFR	African
ALS	Amyotrophic lateral sclerosis
ALSg	ALS related gene
ALSoD	Amyotrophic lateral sclerosis online database
ALSp	Protein encoded by an ALS related gene
AMR	South/ Central American
ASN	Asian
BQ	Base quality
BQSR	Base quality score recalibration
CI	Confidence interval
CMT	Charcot-Marie-Tooth disease
CNV	Copy number variant
ESP	National Heart Lung and Blood institute Exome Sequencing Project
EUR	European
fALS	Familial ALS
FTD	Frontotemporal dementia
GATK	Genome Analysis Toolkit
GO	Gene Ontology
GQ	Genotype quality
HGVS	Human genome variation society
HWE	Hardy Weinberg Equilibrium
IBMPFD	Inclusion body myopathy with Paget's disease and frontotemporal dementia
LD	Linkage disequilibrium
LSDB	Locus specific database
MAF	Minor allele frequency
MQ	Mapping quality
MT	Mitochondrial
OMIM	Online Mendelian Inheritance in Man
OR	Odds ratio
p	Probability
PCR	Polymerase chain reaction
PD	Parkinson's disease
QC	Quality Control
QQplot	Quantile-Quantile plot
rpPCR	Repeat primed PCR
rvALS	Reduced validation ALS gene set
sALS	Sporadic ALS
SNP	Single nucleotide polymorphism
SV	Structural variation
tALS	Training ALS gene set

UBI Ubiquitinated inclusion
vALS Validation ALS gene set

CONTENTS

DECLARATION i
SUMMARY ii
ACKNOWLEDGEMENTS iv
ACRONYMS v
CONTENTS vii
1. INTRODUCTION 1
1.1 ALS: Clinical features and aetiology 2
1.1.1 Clinical features 2
1.1.2 Genes versus Environment 2
1.1.3 Familial ALS 3
1.1.4 Sporadic ALS 3
1.1.5 Uncertainties in the clinical relevance of identified ALS gene variants 8
1.1.6 Undiscovered genetic risk factors 8
1.1.7 Other unanswered questions in ALS aetiology 11
1.2 DNA sequencing 15
1.2.1 Assaying genetic variation 15
1.2.2 DNA sequencing 16
1.2.3 Target enrichment 17
1.3 Research objectives and thesis structure 19
2. USING REFERENCE DATABASES OF GENETIC VARIATION TO EVALUATE VARIANT PATHOGENICITY 20
2.1 Introduction 21
2.2 Materials and Methods 22
2.2.1 Mapping of PhenCode mutations 22
2.2.2 Mapping of ALS mutations 22
2.2.3 Mapping to reference variation databases 22
2.2.4 Case-control analyses 22
2.2.5 Data analysis 23
2.3 Results 23
2.3.1 Estimating the proportion of reported disease mutations captured by dbSNP, the 1000 genomes project and the ESP 23

2.3.2 Estimating the proportion of ALS-associated variants captured by dbSNP, the 1000 genomes project and the ESP 23
2.3.3 Analysis of ALS associated variants identified in the 1000 Genomes and ESP 24
2.3.4 Analysis of ALS-associated variants identified in dbSNP 28
2.4 Discussion 28
2.5 Conclusion 30
3. *In silico* functional profiling of ALS related genes 32
3.1 Introduction 33
3.2 Materials and methods 34
3.1 Gene sets 34
3.2. Retrieval of molecular interaction data 35
3.2.3 Processing of interaction data 35
3.2.4 Protein sequence similarity 36
3.2.5 Functional enrichment analyses 36
3.2.5 Disease annotation 36
3.2.6 Quantification of relative network importance 36
3.2.7 Quantification of gene annotation similarity 37
3.2.8 Data analysis 37
3.3.2 Sequence similarity 41
3.3.3 Candidate gene prioritization 42
3.3.4 Evaluation of network based prioritization 44
3.3.5 Evaluation of gene annotation based prioritization 45
3.4 Discussion 46
3.5 Conclusion 47
4. TARGETED RESEQUENCING 48
4.1 Introduction 49
4.2 Materials and Methods 49
4.2.1 Target Enrichment Kit Design 49
4.2.2 Study participants 49
4.2.3 Targeted resequencing 50
4.2.4 Sequence alignment 50
4.2.5 Additional data processing and analysis 52
4.3 Results 53
4.3.1 Target enrichment kit design 53
4.3.2 Pilot study 53

4.3.3 Extended study 56
4.3.4 SNP array genotype concordance 57
4.4 Discussion 59
4.5 Conclusion 60
5. ANALYSIS OF KNOWN ALS GENES 61
5.1 Introduction 62
5.2 Material and methods 62
5.2.1 Study participants and targeted resequencing 62
5.2.2 Estimation of variant detection power 62
5.2.3 Variant calling, annotation and quality control 62
5.2.4 Single variant analyses 63
5.2.5 Gene burden analyses 63
5.2.6 Analysis of variant co-occurrence 64
5.2.7 Statistical analyses 64
5.3 Results 64
5.3.1 Targeted resequencing 64
5.3.2 ALS related variants 67
5.3.3 Patient Phenotypes 71
5.3.4 Co-occurrences of ALS gene variants 71
5.3.5 Variation in the frequency of ALS variants across populations 72
5.4 Discussion 74
5.5 Conclusion 76
6. CANDIDATE GENE ANALYSIS 77
6.1 Introduction 78
6.2 Materials and Methods 78
6.2.1 Variant calling and annotation 78
6.2.2 Variant quality control 78
6.2.3 Genotype QC 79
6.2.4 Power calculations 80
6.2.5 Single variant association testing 80
6.2.6 Aggregate association analyses 80
6.2.7 Data analyses 81
6.3 Results 83
6.3.1 Variant quality control 83
6.3.2 Patient variants 85

6.3.3 Analysis of independent variant effects 89
6.3.4 Analysis of aggregated variant effects 96
6.4 Discussion 98
6.5 Conclusion 99
7. WHOLE MITOCHONDRIAL GENOME ANALYSIS 100
7.1 Introduction 101
7.2 Materials and Methods 101
7.2.1 Study participants 101
7.2.2 Data analysis 101
7.3 Results 102
7.3.1 Sequence coverage 102
7.3.2 Variant callset 102
7.3.3 Power analyses 102
7.3.4 Association testing 105
7.4 Discussion 113
7.5 Conclusion 114
8. DISCUSSION AND FUTURE WORK 115
8.1 Extending statistical analyses of known and candidate ALS genes 116
8.2 Biochemical characterization of candidate disease variants 116
8.3 Candidate gene studies – Pros and cons 117
8.4 Screening for ALS related structural variation 119
8.5 Design of future gene mapping efforts 120
BIBLIOGRAPHY 122

1. INTRODUCTION

1.1 ALS: Clinical features and aetiology

1.1.1 Clinical features

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease, and the most common of what are more generally referred to as motor neuron disorders. This disease is primarily of late onset and is diagnosed on the basis of evidence for an ongoing loss of upper and lower motor neurons. Afflicted individuals exhibit a progressive paralysis and muscle wasting that can be associated with the loss of virtually all voluntary muscle function. This can represent not only a loss of mobility but also losses of the capacity to ingest food and to perform basic communication skills like speech and writing. The disease can also involve significant extra motor pathology, manifesting as cognitive and behavioural decline and in many cases individuals initially diagnosed with ALS subsequently develop frontotemporal dementia (FTD) and vice versa. Death usually occurs within 3 years of symptomatic onset and is most commonly the result of respiratory failure. The condition is associated with extensive clinical heterogeneity and age of onset, symptomatic progression and disease severity can vary dramatically from patient to patient. No cure has been discovered but a single disease modifying therapy (Riluzole) is available and is believed to extend patient survival by a median of 2-3 months[1].

A population based register of Irish ALS patients has been in place since 1995[2]. This represents the longest running register of its kind and has facilitated detailed characterization of the Irish ALS population[2-7]. Features of the disease in Ireland appear similar to what has been reported across other countries. Lifetime risk is approximately 1/233 for men and 1/387 for women[6], the median age of onset is ~66 yrs and median survival from first symptom onset is ~16.4 months[7]. In 58% of cases disease is said to be of "spinal" onset, meaning that first symptoms relate to the arms or legs[7] and in a further 35% disease is said to be of "bulbar" onset, meaning that first symptoms relate to difficulties with speech or swallowing. In the final 7% of cases, disease is said to either be of "respiratory" onset, in which case patients present with difficulty breathing or else is referred to as "generalized". Approximately 13.8% of Irish cases fulfil criteria for a diagnosis of FTD while a further 34.1% fulfil criteria for some other level of cognitive impairment[5].

1.1.2 Genes versus Environment

As with cancer, diabetes, schizophrenia, Alzheimer's disease and the vast majority of common or relatively common disease phenotypes, no single common causative factor determines ALS susceptibility. Certain cases appear to stem from highly penetrant single gene defects while others appear to result from exposure to environmental toxins. However, in most instances even individual ALS cases are believed to be multifactorial in origin, involving multiple genetic and/or environmental components. Studies of disease concordance rates in affected monozygotic and dizygotic twinships have indicated that 53-84% of the variance in ALS occurrence (referred to as phenotypic variance), is accounted for by genetic susceptibility while the remaining 16-47% is accounted for by environment and gene-environment interactions[8]. Considerable progress has been made in elucidating the genetic component of disease risk but the majority remains unexplained while next to nothing is known about the environmental factors that modulate disease expression.

1.1.3 Familial ALS

In a very small proportion of cases, ALS is passed down generation to generation within families exhibiting Mendelian inheritance patterns. In the majority of such instances, transmission appears autosomal dominant, though X-linked and autosomal recessive inheritance have also been observed. There is also anecdotal evidence to support the potential relevance of mitochondrial inheritance, namely an elevated tendency for patient family histories to lie on the maternal as opposed to paternal lineage[9]. As Mendelian inheritance does not appear to be a feature in most instances, it has become convention to divide ALS into familial (fALS) and sporadic (sALS) forms. The 2 subtypes are clinically indistinguishable with the division describing no more than whether a patient presents with a readily identifiable family history of the disease. The distinction has nonetheless proved important as the majority of well established genetic risk loci have been established through analyses of fALS pedigrees.

In all, 5-14% of ALS cases are classified as fALS[10]. Seldom is the associated history sufficient to establish clear inheritance patterns, but there is nonetheless a tendency to assume onset is generally attributable to single gene defects. Variable combinations of gene-mapping techniques including linkage, homozygosity, association, candidate gene and whole exome analyses have lead to the reporting of 12 fALS or "Mendelian ALS" genes (Table 1.1, Box1.1). These genes include *SOD1, ALS2, DCTN1, SETX, ANG, VAPB, TARDBP, FUS, OPTN, DAO, VCP, SIGMAR1, C9orf72, UBQLN2* and *PFN1* and with the notable exception of *C9orf72,* it is generally believed that multiple variants of a given gene are capable of triggering disease onset. As ALS is associated with both mutations of different genes and different mutations within a given gene, the condition is said to exhibit both locus and allelic heterogeneity (collectively referred to as genetic heterogeneity).

A notable feature of many fALS associated genes is that they have also been implicated in the pathogenesis of other disease phenotypes. In certain situations it has been that different mutations of the same gene associate with different phenotypes, a phenomenon referred to as pleiotropy. In others, it has been that a given mutation appears to trigger one of multiple potential phenotypes, a phenomenon referred to as variable expressivity. Examples of pleiotropic ALS genes include *DCTN1*, *SETX* and *OPTN*, which in addition to ALS have been implicated in Perry syndrome[11], ataxia oculomotor apraxia[12] and glaucoma[13] respectively. Notable examples of ALS associated mutations that exhibit variable expressivity include a large intronic hexanucleotide repeat expansion within the *C9orf72* gene that associates with ALS, FTD and potentially other neurodegenerative disorders[14], and certain mutations of *VCP* that appear to cause ALS and either inclusion body myopathy with Paget's disease of bone (IBMPFD)[15] or facioscapulohumeral muscular dystrophy (FSHD)[16]. It has also been observed that the family histories of many ALS patients appear to exhibit an enrichment for morbidities such as psychosis[10], suggesting that the degree of variable expressivity among ALS mutations may be even greater than is currently appreciated.

1.1.4 Sporadic ALS

Although not associated with clear inheritance patterns, genetics is also believed to play a highly important role in sALS. The heritability of sALS, meaning the proportion of sALS phenotypic variance attributable to genetic factors, is believed to lie between 38 and 78%[8]. Unlike fALS, the relative importance of individual sALS risk factors varies considerably and the aetiology of sALS is

anticipated to exhibit even greater heterogeneity than that of fALS. In certain cases, it appears that sALS can in fact also be attributable to single gene defects. One example of this concerns a P525L mutation within the fALS gene *FUS*[17]. The mutation has been documented to occur *de novo* as opposed to being inherited and causes a juvenile onset and highly aggressive disease. The severity of the associated phenotypes means that probability of it being transmitted from one generation to the next is extremely small and that most associated cases should hence be expected to occur sporadically. However, it also appears that a significant proportion of sALS cases may result from germline inheritance of what are believed to represent at least reasonably penetrant ALS variants. In these cases, the absence of an observable family history appears likely to simply reflect ascertainment bias. For example, fALS associated *SOD1* mutations are often believed to act with penetrances >=70%. At 70% penetrance and assuming a typical family size of 2 parents and 2 children, it would be anticipated to only observe disease recurrence within only ~50% of nuclear families, meaning the remaining 50% may appear sporadic[18]. This is in-keeping with what has been reported in the literature and highlights that though the division of ALS into fALS and sALS has proven useful in certain research settings, it is to a large extent artificial.

Most sALS associated mutations are not anticipated to act with high penetrance. Perhaps the most well characterized of this class of variants represents an exonic CAG repeat expansion within the gene *ATXN2*. Intermediate size expansions (27-33 repeats) have shown replicable associations with ALS but are observed within 2.4% of the general population[19]. Based on their frequency among cases (~5.5%), the expansions appear to increase the odds of disease ~2.3 fold. Different length expansions at the exact same locus have been implicated in spinocerebellar ataxia and schizophrenia, demonstrating that as in fALS, the genetic architecture of sALS is believed to have important overlaps with other disorders. Other robust and replicable associations with sALS risk have been reported for variants within *UNC13A* and *DPP6* and in addition to these a very large number of candidate disease variants (>500) have been defined on the basis of tentative associations with either disease risk or other clinical parameters like patient survival (Table 1.2). Given the lack of clear inheritance patterns, the mapping of sALS associated variants has generally been achieved using different methodologies to those which have been employed in fALS. These involve population as opposed to family based analyses, with the most commonly employed including candidate gene analyses and genome-wide association studies (GWAS) (Table 1.2, Box 1.1).

Table1.1: Genes associated with familial ALS

Gene	Year	Mapping	Reference	Mutations	Model	Phenotype
ALS2	2001	Linkage mapping and candidate gene resequencing	20,21	23	AR	jALS, jPLS, iHSP
ANG	2004	Candidate gene analysis	22	29	AD	ALS, PD
C9orf72	2011	Linkage mapping and targeted resequencing	23,24	1	AD	ALS, ALS-FTD, PD
DAO	2010	Linkage mapping and candidate gene resequencing	25	2	AD	ALS, SCA2
DCTN1	2003	Linkage mapping and candidate gene resequencing	26	6	AD	DHMN-7B, PS, ALS, ALS-FTD
FUS	2009	Linkage/ homozygosity mapping and candidate gene resequencing	27,28	66	AD, AR	ALS, ALS-FTD, jALS, ETM4
OPTN	2010	Homozygosity mapping and candidate gene resequencing	29	31	AD, AR	ALS, GLC1E
PFN1	2012	Family based exome-resequencing	30	4	AD	ALS
SETX	2004	Linkage mapping and candidate gene resequencing	31	7	AD	jALS, AOA2
SIGMAR1	2011	Homozygosity mapping and candidate gene resequencing	32	1	AR	jALS
SOD1	1993	Linkage mapping and candidate gene resequencing	33	168	AD, AR	ALS
TARDBP	2008	Linkage mapping and candidate gene resequencing	34	50	AD	ALS, ALS-FTD
UBQLN2	2011	Linkage mapping and candidate gene resequencing	35	5	XD	ALS, ALS-FTD
VAPB	2004	Linkage mapping and candidate gene resequencing	36	2	AD	ALS, SMA
VCP	2010	Family based exome-resequencing	37	4	AD	ALS, IBMPFD

AD Autosomal dominant; AOA2 Ataxia oculomotor apraxia type 2; AR Autosomal recessive; DHMN-7B Distal hereditary motor neuropathy type 7b; ETM4 Essential tremor type 4; GLC1E Glaucoma type 1e; IBMPFD Inclusion body myopathy with Paget's disease and frontotemporal dementia; iHSP infantile hereditary spastic paraplegia; jALS juvenile ALS; jPLS juvenile primary lateral sclerosis; PD Parkinson's disease; PS Perry syndrome; SCA Spinocerebellar ataxia type 2; SMA Spinal

Muscular Atrophy; XD X-linked dominant. A more extensive account of ALS related mutations and associated clinical feature can be found on the Amyotrophic lateral sclerosis online database (ALSoD; http://alsod.iop.kcl.ac.uk/)[38] and Online Mendelian Inheritance in Man (OMIM; http://www.omim.org).Mutation counts were based on mutations reported within ALSoD.

Table 1.2: Genes associated with sporadic ALS

Gene	Year	Mapping	Reference	Details	Replicated
ATXN2	2010	Candidate gene association testing	19	Intermediate length polyglutamine expansions associated with increased risk of ALS. Longer expansions associated with SCA	Y
CHMP2B	2006	Candidate gene resequencing	39,40	Rare heterozygous mutations identified among ALS patients	-
DPP6	2008	SNP array based genome wide association study	41-43	Association with disease susceptibility in single marker and CNV analyses	Y
ELP3	2009	Microsatellite based genome wide association study	44	Association with disease susceptibility in single marker analyses. Supportive evidence from a Drosophila mutagenesis screen	N
FIG4	2009	Candidate gene resequencing	45	Rare heterozygous mutations identified among ALS patients	-
FGGY	2007	SNP array based genome wide association study	46	Association with disease susceptibility in single marker analyses	N
ITPR2	2007	SNP array based genome wide association study	47	Association with disease susceptibility in single marker analyses	N
KIFAP3	2009	SNP array based genome wide association study	48	Association with patient survival in single marker analyses	N
MAPT	2001	Candidate gene association analysis	49	Association with susceptibility to the Guam ALS-parkinsonism-	N

				dementia complex	
NIPA1	2010	SNP array based genome wide association study and candidate gene resequencing	41,50	Association between disease susceptibility and deletions/ polyalanine repeat expansions	N
NEFH	1999	Candidate gene resequencing	51	Association between disease risk and tail domain indels. Supported by additional reports of mutation carriers among ALS cases	N
PARK7	2005	Candidate gene resequencing	52	Homozygous mutation carriers identified among patients with a Parkinsonism-Dementia-ALS complex	N
PON1 *PON2* *PON3*	2006	Candidate gene association analyses	53-57	Associations with disease risk but original studies did not account for multiple testing	N
SPG4	2005	Candidate gene resequencing	58	Heterozygous mutation identified among an individual with atypical ALS	N
SPG11	2010	Candidate gene resequencing	59	Mutations observed in homozygous configuration among patients with AR jALS	N
SQSTM1	2011	Candidate gene resequencing	60,61	Excess in the frequency of rare variants among cases. Supported by additional reports of mutation carriers among ALS cases	N
TAF15	2011	Candidate gene resequencing	62	Mutations observed among fALS cases	N
UNC13A	2009	SNP array based genome wide association	63	Associations with disease risk and patient survival.	Y

For brevity, only a selected subset of genes potentially classifiable as ALS associated are described. A more extensive account of ALS associated genes and related mutations can be found on the ALSoD and OMIM websites(http://alsod.iop.kcl.ac.uk/, http://www.omim.org). Replicated, denotes whether or not a statistically significant association signal has been replicated in an independent sample cohort. Genes reported in this table have to date only been associated with non-Mendelian ALS, however this does not infer that certain genes may also play a role in Mendelian disease. ALS AR jALS Autosomal recessive, juvenile onset ALS, CMT4j Charcot-Marie-Tooth disease type 4j; CNV Copy number variation, HSP Hereditary spastic paraplegia; PD Parkinson's disease, SCA Spinocerebellar ataxia;

1.1.5 Uncertainties in the clinical relevance of identified ALS gene variants

Although a large number of genes (>100) and an even larger number of mutations (>500) have been associated with ALS, the true pathogenic relevance of most remains uncertain. This uncertainty is primarily a consequence of ALS being a disease of relatively low frequency but extensive aetiological heterogeneity. This means that observations of ALS due to any one particular cause are generally quite rare, making the demonstration of robust and replicable disease associations extremely difficult. Another difficulty, particularly relevant in the case of genes mapped on the basis of association studies, has been that of distinguishing correlation from causation. For example, the rs12608932 single nucleotide variant (SNV) in *UNC13A,* have been demonstrated to exhibit a strong and replicable association with ALS risk but it is not yet known whether the variant itself exerts a toxic effect or whether it is merely in linkage disequilibrium (LD, Box 1.1) with a causal variant. Findings reported in certain studies have also been dependent on the use of what may represent inaccurate assumptions about disease architecture or in certain cases may have be confounded by failure to adequately account for relevant biases. Examples of this, could include the use of assumptions regarding the selective pressures against ALS mutations in exome resequencing analyses[30,37], the use of differing genotyping methodologies for cases and controls in association mapping studies or failures to account for the potential existence of population structure (differing ethnic composition among case-control cohorts).

Another important issue which is of relevance to even the most well established of identified ALS mutations is a lack of understanding regarding their relative clinical importance within the individual, the local population and the global population. This issue can be subdivided into 2 components, the first of which concerns limited understandings of variant effect sizes and the second of which concerns limited understandings of variant frequencies. Despite that there is often a total lack of supporting evidence, many fALS and even sALS associated variants are often reported and interpreted as though they account for the entirety of an individual's disease risk. This assumption can have important consequences within several clinical and research settings such as genetic counselling, estimating explained and unexplained trait heritability, design of disease model systems, allocation of research/ clinical resources and patient stratification in clinical trials or epidemiological analyses. Gene variant frequencies have of course been reported in a variety of studies but often these have involved analyses of unrepresentative patient cohorts and seldom have they allowed for comparison of variant frequencies across different genes. In addition to understanding disease structure in local populations, an understanding of how ALS variant frequencies vary across populations can serve to understood how well research findings are likely to translate across ethnicities and guide the conduct of collaborative and replication orientated studies.

1.1.6 Undiscovered genetic risk factors

As already outlined ALS is known to exhibit considerable heritability (53-84% of phenotypic variance). Due to the reasons discussed in section 1.1.5, the exact proportion of this which is accounted for by identified risk factors is uncertain. However, it is nonetheless anticipated that the majority, remains unexplained. To date well established ALS genetic risk factors are primarily believed to represent rare moderate-high penetrance disease variants. As ALS is seldom associated with Mendelian inheritance, high penetrance disease variants are not expected to account for a large fraction of disease incidence. A possible exception to this, may be within

juvenile onset ALS, wherein such mutations could occur *de novo* and family histories are not observed because the severity of the phenotype precludes of the passing of causal variants to subsequent generations. Notably, this has proved relevant in other conditions that would be anticipated to associate with comparable levels of negative selective, including rare disorders like Duchenne muscular dystrophy (http://omim.org/entry/310200) but also more common phenotypes like autism and schizophrenia[64].

Even at moderate or reasonably high levels of penetrance, the transmission of genetic mutations can be obscured considerably, particularly with a late onset disease or in the event of variable expressivity. Based on this it would seem that moderately penetrant disease variants could theoretically account for a sizeable proportion of ALS risk. If this were the case, the absence of corresponding association signals in recent GWAS[42,46,47,63,65] would suggest that either such variants are individually quite rare or else that they have recurred multiple times on different haplotypic backgrounds. However, disease concordance rates among monozygotic twins suggest that on average, ALS variants exhibit a maximal penetrance of only 4-20%[8]. Therefore, while moderately penetrant disease variants may account for a proportion of undiscovered risk, they should not be expected to represent the dominant category.

If high and moderately penetrant disease variants do not dominate genetic risk, then it stands to reason that low or perhaps even very low penetrance variants do. Such variants could be quite common yet have gone undetected by current GWAS due to deficits in statistical power. Results from one of the most recent GWAS suggests that this in fact likely to be the case, and that common variation accounts for ~12% of ALS phenotypic variance[66]. However, this may still leave an important component of ALS heritability to be accounted for by rare-low penetrance variant that individually make only miniscule contributions to disease burden. Another possibility is that a significant component of ALS risk could be accounted for by genetic interactions. This refers to situations where the combined effects of 2 or more variants exceeds the sum of the independent effects. Under this setting it could be that certain ALS variants are toxic only in certain genetic backgrounds, and that independent effects are non-existent. There are no documented examples of this as yet but very few studies have explored the potential relevance of anything other than single variant effects. Perhaps the most relevant finding reported to date concerns a reported excess in the frequency with which fALS patients harbour variants of multiple ALS related genes[67]. The study was not designed to assess the relative importance of independent versus interaction effects but does at least demonstrate the need for exploration of polygenic disease models.

Box 1.1: Concepts in disease gene mapping

Genetic markers: Known and assayable genetic variants that can be used to distinguish the state of chromosomal segments across individuals.

Genetic linkage: A phenomenon describing to the tendency of alleles occurring on the same ancestral chromosome to be inherited together. Generally, the shorter the physical distance between alleles, the lower the probability of separation during recombination and the stronger the linkage.

Linkage disequilibrium (LD): A phenomenon referring to the non-random association of different alleles. Most commonly a result of linkage but correlations can be negative as well as positive and alleles need not necessarily occur on the same chromosome.

Linkage mapping: An approach wherein the location of causative mutations is inferred by testing genetic markers for linkage with a trait of interest.

Association mapping: An approach wherein the location of causative mutations is determined by testing genetic markers for association with a trait of interest.

Candidate gene analyses: Screening of hypothesized or "candidate" disease genes for disease associated mutations.

Genome wide association study: Antype of study wherein markers are distributed across the entire genome are assayed for association with a trait of interest. Most commonly conducted using SNP arrays but also performed using whole genome sequencing and other methodologies.

Homozygosity mapping: An approach wherein the locations of mutations responsible for recessively inherited traits are determined by identifying segments of the genome which occur in a homozygous state across a series of afflicted individuals.

Exome resequencing: An approach wherein the entire-protein coding sequence of the genome (the exome) is resequenced. Putative disease variants are often subsequently identified by applying a variable assortment of hypothesis based variant filtration steps. These steps can include the removal of variants which violate expected segregation/inheritance patterns, which represent known genetic polymorphisms or which are not bioinformatically predicted to have severe impacts on protein function.

1.1.7 Other unanswered questions in ALS aetiology

Many fundamental questions concerning ALS aetiology remain unanswered. Undoubtedly one of the most important issues represents the nature of environmental risk factors, namely how many of them are there? how diverse are they? what are they? and how extensively do they influence disease risk individually and collectively? So far, the most well established environmental association with ALS concerns cyanotoxin exposure. This is believed to account for a spike in disease incidence on the pacific island of Guam[68]. The evidence for this effect includes a positive correlation between recent declines in the magnitude of this spike and reduced consumption of cycad seeds containing the toxin beta-N-methylamino-L-alanine. Moreover, feeding of this toxin to Macaques was found to induce a phenotype reminiscent of human ALS[68]. Other reported associations have concerned level of physical activity and chemical weapon exposure, with most widely cited pieces of supporting evidence concerning potential increases in disease frequency among Italian soccer players[69] and gulf war veterans[70]. In a more recent study, an association between ALS risk and heightened leisure time physical activity was in fact confirmed, though no dose-response type relationships or correlations between disease risk and occupational activity were observed. The authors of this study therefore speculated that the observed effect may concern a feature of at risk individuals as opposed to an explanation for why certain individuals are at risk.

Arguably of considerable more importance than the identification of causal risk factors is an understanding of the mechanisms by which contribute to disease expression. An understanding of these mechanisms could prove invaluable to the development of therapeutic strategies and clinical management plans but also potentially guide the identification of undiscovered risk factors. One obvious possibility with several fALS associated mutations is that they may simply disrupt the normal functioning of the genes in which they occur. Quite often, Mendelian disorders associated with loss of gene function effects exhibit recessive inheritance patterns. This is because expression of the phenotype requires a near complete ablation of normal gene function through disruption of both alleles. Examples of this include muscular dystrophies (http://omim.org/entry/310200), phenylketonuria (http://omim.org/entry/261600) and perixisome biogenesis disorders (http://omim.org/entry/614870) among many others. There are of course several instances where the presence of one normal allele is not sufficient to allow adequate gene functioning. This is referred to as haploinsufficiency and examples include Ehlers-Danlos (http://omim.org/entry/130050) and Marfan (http://omim.org/entry/154700) syndrome, which despite being attributed to loss of gene function effects are associated with autosomal dominant inheritance. Identified ALS genes share varying assortments of notable biological properties but as a whole, even the most well established genes do not exhibit any clearly defining features. By in large they tend to be ubiquitously expressed and participate in a diverse array of biological processes (Fig 1.1, Table 1.3). This could infer that as with the distribution of genetic risk, there is considerable heterogeneity in the underlying mechanisms of ALS pathogenesis. However, it could also be that the toxicity of ALS mutations is more a function of the mutations themselves as opposed to the genes in which they occur or simply that the crucial and unifying feature of ALS related genes remains to be discovered. The fact that Mendelian ALS generally exhibits autosomal dominant inheritance suggests that unless haploinsufficiency is a feature of most fALS related genes, the disease is not simply a matter of loss in normal gene functioning. *In vitro* and *in vivo* models of various human ALS mutations also suggest that toxicity is usually more likely to reflect gain of function than loss of function effects. For example, many

ALS mutations of *SOD1* do not appear to influence the normal functioning of the encoded protein at all and knocking out of the gene in mice does not recapitulate the ALS like phenotypes observed with transgenic over-expression of certain human mutations[71]. As indicated, it may even be the case that with certain mutations the nature of the "host" gene is of little or no consequence. For example the *C9orf72* repeat expansion represents the most common ALS variant identified to date but despite extensive screening it is the only variant of the gene shown to associate with ALS. This suggests that most mutations disrupting *C9orf72* function do not contribute to ALS aetiology. There is also considerable evidence to suggest that toxicity of the expansion simply results from the generation of generic repetitive mRNA sequences during transcription[72]. Also in-keeping with this idea is the striking fact that the proteins encoded by many ALS associated genes, including *TARDBP, SOD1, FUS, OPTN* and, *UBQLN2*, often localize to ubiquitinated neuronal inclusions within afflicted variant carriers[73]. This could be an indication that toxicity is simply the result of generic gains in protein aggregation propensity. In this case, despite being irrelevant, some level of correlation with normal gene biology might still be expected, as certain classes of protein are inherently more aggregation prone than others. Furthermore, it could be that only aggregates of genes with specific molecular interaction or gene expression profiles can trigger disease, or that actually disease can involve both loss and gain of function effects. Likewise it should also be noted that despite certain tantalizing lines of evidence, protein aggregation may not represent a primary disease process in ALS pathobiology. Moreover, ALS2, which is associated with mutations of the *ALS2* gene and is recessively inherited, is believed to involve loss of function effects. Taken together it should therefore perhaps be anticipated that ALS pathogenesis will involve a diverse array of mechanisms, that loss and gain of function effects will be relevant but that gain of function effects may prove the most important. As to the existence of final common disease pathways, it currently remains too early to do more than speculate but the tight genotype-phenotype correlations observed with certain gene mutations (Table 1.1) argue some heterogeneity should be expected.

Table 1.3: Major ALS gene functions

Gene	Molecular Function	Biological Function
ALS2	Guanine nucleotide exchange factor	Modulator of endosomal dyanmics
ANG	Ribonuclease	Implicated in angiogenesis and regulation of gene expression
ATXN2	?	?
C9orf72	?	?
DAO	Oxidation of D-amino acids	May act as a detoxifying agent to removes D-amino acids that accumulate during aging
DCTN1	Largest subunit of the Dynactin complex	Mediates binding of cell organelles to dynein for microtuble transport
FUS	RNA binding protein	RNA processing, regulation of gene expression
OPTN	Regulator of TNFα signalling pathways	May function in apoptosis, inflammation, vasoconstriction, cellular morphogenesis and vesicular trafficking
PFN1	Actin binding protein	Regulates actin polymerization
SETX	DNA/RNA helicase	DNA/ RNA processing
SIGMAR1	Transmembrane receptor	Modulation of intracellular Ca2+ release
SOD1	Super oxide dismutase	Detoxification of superoxide radicals
TARDBP	DNA/ RNA binding protein	RNA processing, regulation of gene expression
UBQLN2	Links ubiquitin ligases and proteasomes	Protein homeostasis
VAPB	Membrane protein	Vesicular trafficking
VCP	ATP binding protein	Vesicular transport

Functional descriptions retrieved from summaries provided by entrez gene (http://www.ncbi.nlm.nih.gov/gene).

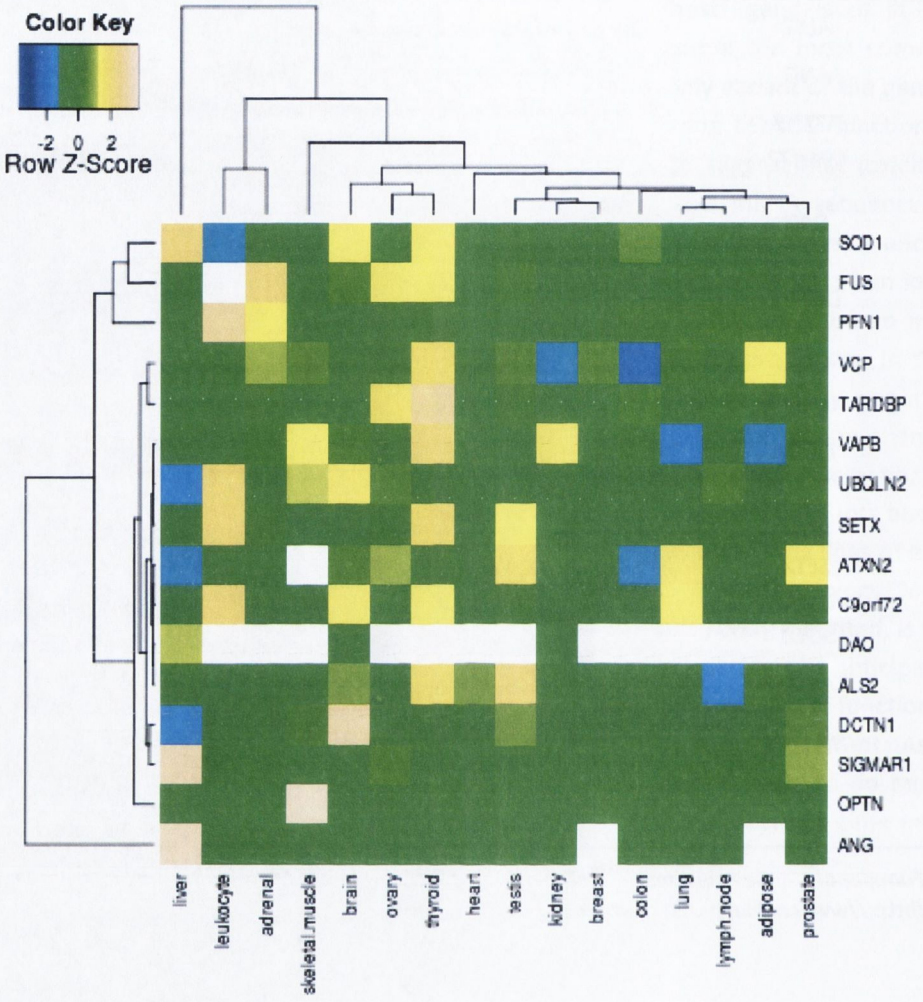


Figure 1.1: Major ALS gene tissue expression profiles. Generated using data retrieved from the ensembl expression atlas (http://www.ebi.ac.uk/gxa/experiments/E-MTAB-513?ref=aebrowse). Expression values are normalized on a per gene basis with white patches denoting an absence of detectable expression.

1.2 DNA sequencing

1.2.1 Assaying genetic variation

Genetic variation is conventionally assayed through either DNA sequencing or custom genotyping assays (Figure 1.1). Custom assays, come in many forms and usually have the advantages of being relatively simple and inexpensive. They can be used for the investigation of single variants or incorporated into highly parallelized platforms like SNP arrays, which currently allow for the genotyping of over 1 million sites in a single experiment. The major limitation with custom assays is that they can only be used to genotype known genetic variation. DNA sequencing can be used to both genotype known genetic variants and discover novel variation but often represents a more costly option than an approach based on custom assays. However, as discussed in section 1.4.2, DNA sequencing technologies are going through a period of rapid development.

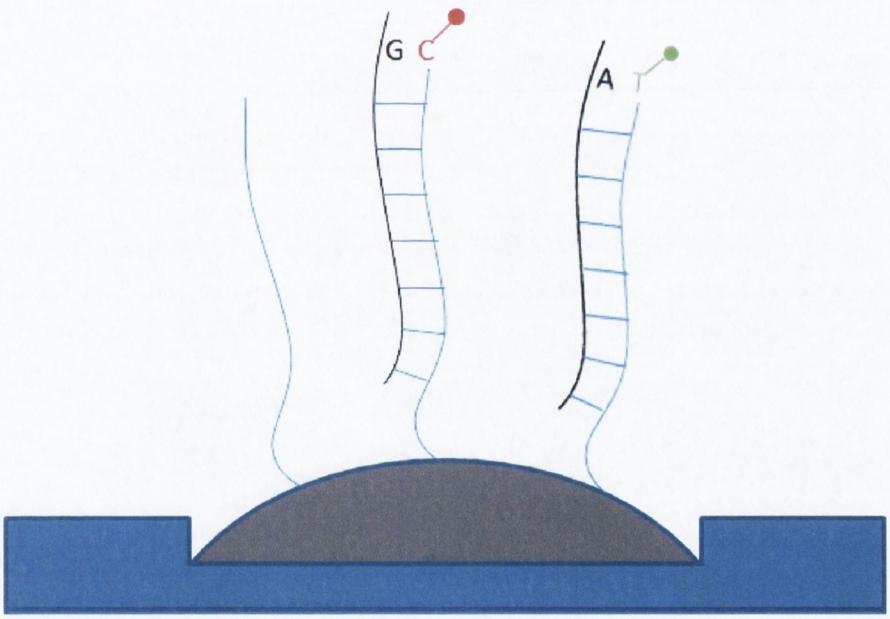


Figure 1.2: The Illumina Infinium Assay. Oligonucleotide probes (light blue) complementary to the genomic sequence preceding a SNP of interest are immobilized onto silica beads(grey) at specific coordinates on the surface of a DNA microarray. The probes are let hybridize with amplified sample DNA (black) and fluorescently labelled nucleotides. complementary to the sample allele at the SNP position, are incorporated in a single base extension reaction. Sample genotypes are then inferred based on the relative fluorescent signals generated at each microarray position. In this example, the relative strength of red and green fluorescence will result in the sample being designated as heterozygous for an A>G SNP.

1.2.2 DNA sequencing

Current DNA sequencing technologies can be divided into 1st, 2nd and 3rd generation systems. The mainstay of 1st generation systems has been the chain terminator method developed by Sanger *et al.* in the mid 1970's [74,75] and automated by Applied Biosystems in the mid 1980's. The method, commonly referred to as Sanger sequencing, represented the primary workhorse of the Human Genome project[76] and is still in wide-spread use today. In the context of disease genetics, it is commonly used for resequencing individual or limited numbers of known/candidate disease loci and for validating results from high-throughput sequencing experiments. The technology has not scaled exceptionally well however and is not commonly used for interrogating large amounts of genomic sequence. In 2005 the first 2nd generation system, was launched by 454. This was subsequently followed by Illumina's Genome Analyzer, HiSeq and MiSeq systems and Applied Biosystem's SOLiD system. The inner working of these sequencers differ, but all essentially involve anchoring many short DNA molecules to fixed locations for massively parallel resequencing through sequential "wash and scan" cycles (Figure 1.2)[77]. This massive parallelization allowed for immense gains in sequencing capacity. For example a comparison study published in 2012 reported that while the AB 3730xl Sanger sequencing system could generate 1.9~84Kb per run, the 454 GS FLX could generate 0.7Gb per run, the HiSeq2000 could generate 600Gb per run and the SOLiDv4 could generate 120Gb per run[78]. Moreover, the cost of sequencing 1 million bases was estimated as $2400 with the AB 3730xl, $10 with the 454 GS FLX, $0.07 with the HiSeq2000 and $0.13 with the SOLiDv4[78]. Offering both the highest throughput and lowest per base cost the Illumina systems have unsurprisingly emerged as the most popular. However, each system retains its own advantages and can represent both the most efficient and inexpensive strategy for specific applications. Third generation sequencers have not yet come into widespread use but are expected to offer further reductions in sequencing costs in addition to numerous other technical advantages such as increased read lengths, elimination of the need for sample amplification and real time data processing.

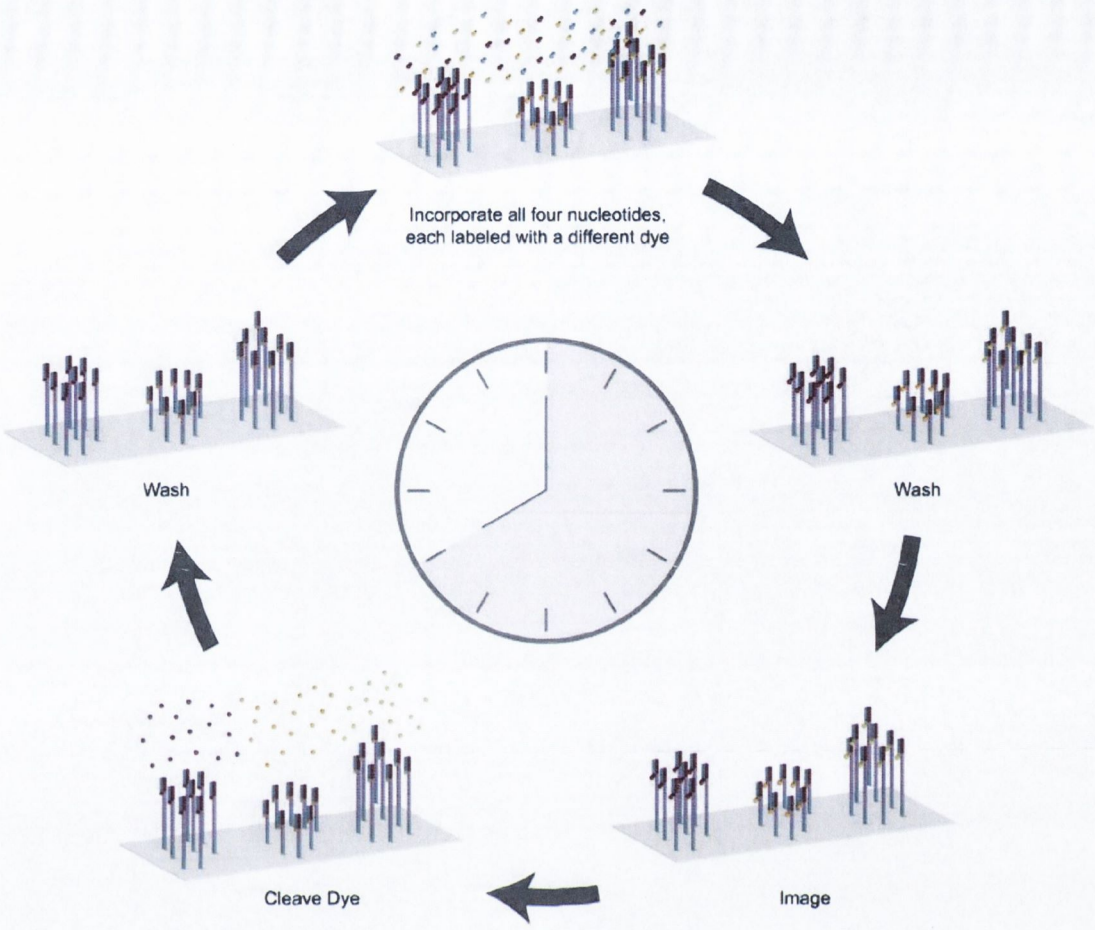


Figure 1.3: Illumina High-Throughput sequencing. DNA molecules are anchored to the surface of a sequencer flow cell through hybridisation of ligated adapter sequences to immobilized oligonucleotide probes. Molecules are then locally amplified to form "clusters". Labelled nucleotides are then incorporated in a single base extension reaction, excess reagents are washed away and the image is scanned to determine which bases have been incorporated at each cluster. The labels are then cleaved from the newly incorporated nucleotides and washed away to facilitate the next round of extension. As the process progresses, asycnhronicity develops in the extension of molecules within clusters. This results in a progressive reduction in base call quality and limits read length. Image taken from [77].

1.2.3 Target enrichment

Second generation sequencers have provided substantial reductions in the cost of sequencing, however whole-genome sequencing still remains quite expensive. Furthermore, it is surplus to requirement in situations where only specific target intervals are of interest. With first generation systems, targeted resequencing was generally conducted by independently amplifying each interval by polymerase chain reaction (PCR) and subsequently sequencing each pool of PCR amplicons. Conventional PCR methodologies do not scale very well, so alternative strategies were needed for more extensive applications like exome-resequencing. This motivated the development of sequence enrichment systems like Agilent's SureSelect, Nimbelgen's SeqEZ Library and Illumina's TruSeq. These work by using various types of labelled oligonucleotide probes to enrich processed DNA libraries for fragments originating from genomic intervals of

interest (Figure 1.3). The most popular application of high-throughput target enrichment systems has probably been exome-resequencing, which usually costs upwards of $2000 per sample. However, they can also be customized for more focused applications such as candidate gene resequencing, resulting in significantly lower costs per sample.

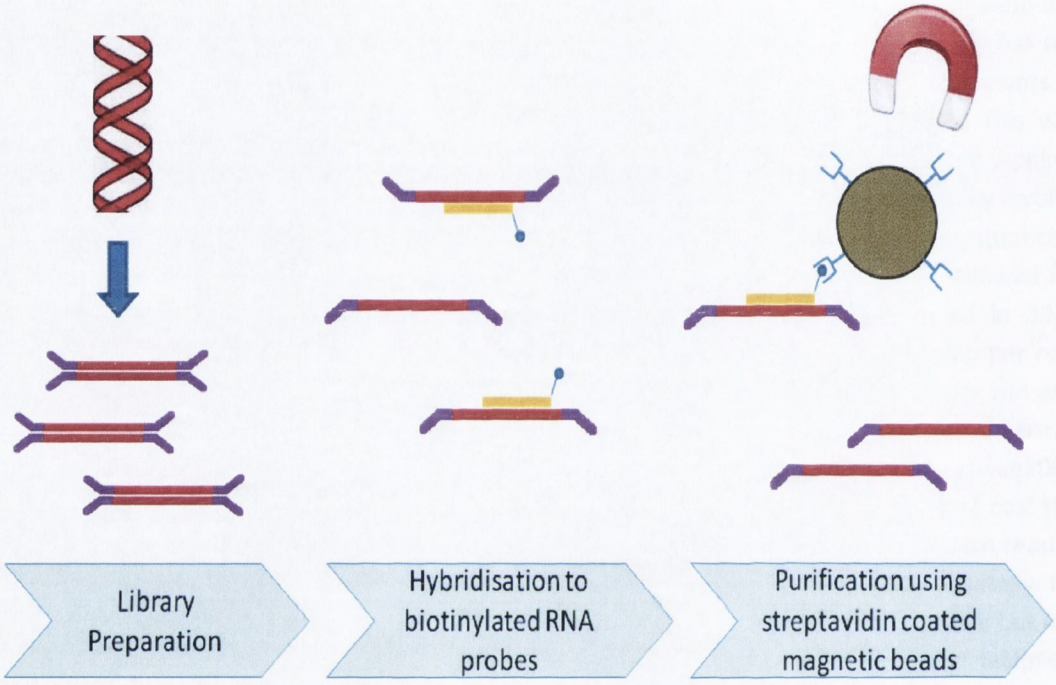


Figure 1.4: SureSelect Target Enrichment. Genomic DNA is processed to sequencer ready libraries (discussed in section 4.2). Libraries are incubated with biotinylated RNA probes that hybridise with fragments containing sequences from target intervals. Target sequences are retained in solution through binding of RNA probes to streptavidin coated magnetic beads as non-target sequences are washed away.

1.3 Research objectives and thesis structure

The primary research objectives of the works described in this thesis were to scrutinize the pathogenicity of previously reported disease variants, functionally profile identified ALS genes for the purposes of identifying disease related features, establish the relative and cumulative importance of variants across previously reported disease loci and to screen various candidate genomic intervals for disease related variation. Other secondary objectives were also pursued and are discussed throughout the work as relevant.

In this section, I have provided an introduction to the clinical and epidemiological features of ALS, what is known and unknown regarding the underlying aetiology of ALS, and DNA sequencing technologies. The next section will represent an exploration of the use of public databases of genetic variation in evaluating the pathogenicity of putative disease variants. This is primarily centred around an evaluation of previously reported ALS mutations but also involves more generalized analyses of reported disease variants and an assessment of the risks associated with variant filtering strategies commonly employed in candidate gene and exome-resequencing based studies. In section 3, I discuss the *in silico* functional profiling of 9 previously reported ALS genes and the subsequent use of seemingly disease related features in candidate gene selection. In section 4, I discuss the design and implementation of a multiplexed targeted resequencing strategy for the analysis of 33 known and 718 candidate ALS genes. I also discuss the analyses and quality control involved in preparing the resulting data for use in variant calling. In section 5, I discuss the analysis of the targeted resequencing data relating to the 33 known ALS genes. This involves characterization of the relative and cumulative frequencies of disease variants and a comparison of the genetic architecture of ALS in Ireland with that reported across other European populations. I also discuss results from investigations into the potential importance of oligogenic susceptibility and the notion that certain variants only cause disease on specific genetic backgrounds. In section 6, I discuss the analysis of the targeted resequencing data relating to the 718 candidate genes. Specifically, I describe the issues involved in controlling for the confounding influences of sequencing and analysis artefacts as well as the results from a variety of statistical analyses designed to test the independent and aggregate effects of observed variation on disease risk and patient phenotype. In section 7, I discuss the use of "off-target" data, generated during the previous targeted resequencing experiments, in characterizing the role of mitochondrial variation in ALS susceptibility and clinical course. Once again this involves the implementation of various single variant and aggregate analyses as well as screens for variants previously implicated in motor neuron disorders. Finally, section 8, includes a brief discussion of certain limitations of the described work, potential options for following up on candidate disease loci the nature of some works in progress and suggestions for future study design.

2. USING REFERENCE DATABASES OF GENETIC VARIATION TO EVALUATE VARIANT PATHOGENICITY

Kenna *et al*, Human Mutation, 2013

2.1 Introduction

ALS exhibits considerable locus and allelic heterogeneity. To date, over 100 genes and 500 mutations have been associated with the disorder. Several of these mutations were identified through comprehensive linkage analyses and have been observed to segregate with disease in families exhibiting Mendelian inheritance. Most were identified in isolated cases during screening of known or candidate ALS genes in moderately large case-control or case-only cohorts. For the latter, and hence the vast majority of reported ALS mutations, very little if anything is known about variant effect sizes. With limited sample sizes and usually only 1 or 2 observations per cohort, it has not been possible for reporting studies to conduct formal association analyses. Instead such mutations have either been reported as possible or probable ALS mutations, or else simply been assumed to represent ALS mutations. There has also been a strong tendency to assume that such variants cause disease in a Mendelian fashion. A consequence of this is that individuals carrying these variants are often excluded from any further aetiological investigation. Additionally, such variants have also been used to generate disease models and taken into extensive clinical and biochemical despite a lack of any formal evidence to support pathogenic relevance. Determining the true pathological relevance of these mutations is therefore of critical importance to a variety of clinical and research efforts.

Over the years, an immense amount of data concerning human genetic variation has been generated. A notable fraction of this has been produced by recent large population resequencing projects like the 1000 genomes project[79] and National Heart Lung and Blood Institute Exome Sequencing Project (ESP), and much of whatever data has been made publically available has been actively curated within dedicated archives like dbSNP[80]. These resources are often used to evaluate the pathogenicity of variants identified in disease orientated resequencing projects, with the most commonly employed strategy being to simply disregard the potential relevance of any previously reported variant. This approach has proved particularly popular among studies applying exome resequencing to what are assumed to represent Mendelian disorders, as it is simple and drastically reduces the candidate variant search space[30,37,81-83]. A potentially major limitation of this strategy is that it fails to account for the possibility that relevant disease variants may actually be represented in these resources. There is therefore an unquantified risk of falsely rejecting variant pathogenicity, potentially leading to highly erroneous conclusions.

In this chapter, I firstly explore the frequency with which disease mutations reported across a variety of phenotype orientated locus specific databases (LSDBs) are captured by dbSNP, the 1000 genomes project and the ESP. To explore whether and to what extent ALS variants captured by these resources might influence disease risk, I then apply an assortment of case-control analyses to explore the applicability of various disease models.

2.2 Materials and Methods

2.2.1 Mapping of PhenCode mutations

The details of mutations reported across the PhenCode network of LSDBs were downloaded from the PhenCode website ('http://phencode.bx.psu.edu/dist/phencode/database/gvPosHg19.txt','http://phencode.bx.psu.edu/dist/phencode/database/gvAttr.txt')[84]. Genomic positions, reference DNA sequences and alternate DNA sequences were established for point mutations annotated with a disease attribute value other than 'not phenotype-associated'. Where codon positions, reference amino acids and alternate amino acids were provided in place of DNA descriptions, and where only a single point mutation within the reference codon would result in the reported amino acid substitution, genomic positions, reference alleles and alternate alleles were inferred.

2.2.2 Mapping of ALS mutations

The details of reported ALS mutations and the associated mutation carriers were downloaded from the ALSoD website ('http://alsod.iop.kcl.ac.uk/misc/dataDownload.aspx#C1')[38,85]. Reported gene names, amino acid positions, reference DNA sequences, alternate DNA sequences, reference amino acids and alternate amino acids were used to establish HGVS DNA descriptions[86] for all nonsynonymous point mutations relative to the relevant transcripts from Ensembl 69[87]. Amino acid positions for mutations within *SOD1 (*MIM# 147450) and *ANG* (MIM# 105850) were adjusted to correspond to the unprocessed peptide sequence. Genomic positions relative to GRCh37 were established from HGVS DNA descriptions using variant effect predictor ('http://www.ensembl.org/Homo_sapiens/UserData/UploadVariations')[88].

2.2.3 Mapping to reference variation databases

Reference and alternate alleles established for PhenCode and ALS mutations were standardized to the "+" strand. Tabix[89] was used to retrieve records with positional matches in the ALS or PhenCode datasets from the human component of dbSNP build137 ('ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/00-All.vcf.gz'), the March 2012 variant release from the 1000 genomes project ('ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/') and the ESP6500 variant release from the NHLBI exome sequencing project (Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [Accessed 2012 Jul 18]). Records exhibiting positional but not allelic matches to ALS or PhenCode mutations were discarded from further analysis. Submitter handles for dbSNP records were retrieved using the NCBI's "Batch Entrez" service ('http://www.ncbi.nlm.nih.gov/sites/batchentrez').

2.2.4 Case-control analyses

Allele count based Fisher exact tests were used to test reported ALS variants for disease association. Reported case allele frequencies were retrieved from the biomedical literature using links provided by ALSoD. Reported control allele frequencies were retrieved from the 1000 genomes and ESP6500 data releases described above. One-tailed binomial tests were then used

to explicitly test reported ALS variants for an absence of association with disease status under predefined assumptions of variant penetrance and population risk. Specifically, it was investigated whether the counts of mutation carriers among controls significantly exceeded expectation under models where all observed carriers develop ALS (full penetrance). The expected carrier frequencies for these models were taken as the product of disease risk (1/426 as per Armon *et al*[90]) and the reported patient carrier frequencies. For binomial analyses, carrier counts referred to the summed count of heterozygous and homozygous carriers where dominant models were tested and to the count of homozygous carriers where recessive models were tested. Alternatively alternate allele counts were used in place of carrier counts, in which case the expected carrier frequencies represented the MAF derived from the reported patient carrier frequencies and an assumption of Hardy-Weinberg equilibrium (HWE).

2.2.5 Data analysis

Unless otherwise indicated, all analyses were performed using R version 2.14.1 'http://www.r-project.org/' and python version 2.6.5 'http://www.python.org/'.

2.3 Results

2.3.1 Estimating the proportion of reported disease mutations captured by dbSNP, the 1000 genomes project and the ESP

Details of reported disease mutations are available across a wide range of phenotype orientated LSDBs and a useful integration of many of these is provided by the PhenCode project [84]. To obtain a general estimate of the frequency with which previously reported disease mutations were represented in dbSNP, the 1000 genomes project and the ESP, the details of 93,092 mutations were downloaded from the PhenCode website. Genomic coordinates were established for 14,258 distinct point mutations annotated as pathogenic in the context of 2,027 disease phenotypes. 7191 (50.4 %) of these were identified within dbSNP, 686 (4.8 %) were identified within the 1000 genomes project and 1459 (10.2 %) were identified within the ESP. Collectively the 3 databases contained 7391 of the 14,258 mapped mutations (51.8 %), representing a combined total of 1682 clinical phenotypes.

2.3.2 Estimating the proportion of ALS-associated variants captured by dbSNP, the 1000 genomes project and the ESP

The details of 436 ALS mutations and the 692 associated mutation carriers were downloaded from ALSoD [38,85]. Genomic coordinates were established for 342 nonsynonymous point mutations and each mutation was assigned to one or more meta-population groupings based on the reported ethnicity of carriers (Supp: Table S1). To be compatible with the groupings used by the 1000 genomes project and ESP, these meta-populations included "African" (AFR), "south/ central American" (AMR), "east Asian" (ASN), "European" (EUR) and "United States" (US). 146 mutations were matched with records in either dbSNP (142, 41.5%), the 1000 Genomes project (28, 8.2%) or the ESP (56, 16%) (Supp. Table S1).

2.3.3 Analysis of ALS associated variants identified in the 1000 Genomes and ESP

The data releases from the 1000 genomes project and ESP were generated by high throughput resequencing of individuals from fixed and well defined sample cohorts. In the case of the 1000 genomes project, this included 1092 individuals sampled from 14 geographically defined populations while in the case of the ESP it included 4300 European Americans and 2203 African Americans sampled from phenotypically defined populations in the United States. Assuming that the cohorts analyzed by the 2 projects can be taken as representative, the published lifetime risk of ALS would suggest that 12-26 (95% CI) of the 7595 individuals assayed will develop ALS. Furthermore, if it is assumed that a given ALS mutation accounts for <10% of cases (justification provided in section 5), no more than 4 carriers of any variant would be expected assuming a full penetrance autosomal dominant disease model. However, for the 57 ALS associated variants observed in either the 1000 genomes or ESP, the collective carrier frequencies ranged from 1-409 individuals (Fig 2.1). Simple case control comparisons were performed to test for associations with disease status. As genotype frequencies can vary dramatically across populations[79,91], these analyses were firstly performed where only individuals originating from the same meta-population as the relevant ALS patients were considered. However, as consideration of larger sample numbers may boost power for variants that are not population specific, the analyses were also performed using the full 1000 genomes/ ESP cohorts. The results revealed strong association signals for only 2 of variants; *SOD1*:p.Asp91Ala ($p=7.0x10^{-76}$), *SPG11*:p.Val2053Met ($p=1.4x10^{-04}$)(Table 2.1). These variants had both been implicated in recessively inherited ALS and were both estimated to exhibit high effect sizes in the comparison (*SOD1*:p.Asp91Ala OR=75 – Inf; *SPG11*:p.Val2053Met OR=22 - Inf). Other variants including *SETX*:p.Cys1554Gly ($p=2.6x10^{-02}$), *SETX*:p.Ile2547Thr ($p=4.5x10^{-02}$) and *DAO*:p.Arg38His ($p=5.6x10^{-02}$) exhibited tentative associations. However these were borderline and it is also important to note that under this analysis, certain variants may exhibit inflated associations due to ascertainment bias or population stratification.

For the remainder of variants no notable associations with disease status were observed. However, the case cohorts within which these were identified tended to be relatively small and the upper bounds of the effect size 95% confidence intervals invariable stretched to infinity. The tests were therefore underpowered and the absence of detectable associations could simply reflect limited sample sizes as opposed to the absence of true effects. It was therefore decided to extend the analysis to include an investigation of whether or not these variants are likely to cause ALS in a Mendelian fashion (full penetrance, see section 2.2.4 for details). Under this model the background population carrier frequency should equal the product of disease frequency and the frequency of carriers among afflicted individuals. To gain power, reported patient carrier frequencies were taken to be both representative and accurate. This limits the accuracy of the results but due to ascertainment bias, reported carrier frequencies should generally be over as opposed to underestimated and hence the tests should generally be biased towards failing to reject tested disease models. At a significance threshold of $p<=0.05$, a null hypothesis of full penetrance could be rejected for 43 mutations under the population matched based analyses and 49 under the full cohort based analyses (Table 2.1). Consideration of the results from both sets of analyses introduces a multiple testing element. Therefore while 51 mutations could be rejected with $p<=0.05$ by at least one of the analyses, only 47 could be rejected upon consideration of both with Bonferroni adjustment of the significance requirements. However for the purposes of this

study, tested disease models were rejected for any variant exhibiting an association at p <=0.05 in either analysis. Of the 6 variants found to represent credible full penetrance disease mutations, 4 only just failed to meet the rejection criteria. The remaining 2 were once again *SOD1*:p.Asp91Ala and *SPG11*:p.Val2053Met however these variants were tested under recessive disease models but neither had actually been observed by the 1000 genomes project/ ESP in the homozygous state. An allele count based test which assumed of Hardy Weinberg equilibrium within the reference populations was therefore also performed (section 2.2.1). This test achieves a gain in power by using information provided by heterozygous carriers but once again the proposed disease models was not rejected for either variant (matched cohort: p=1, p=1; full cohort: p=1, p=1). The results suggested that moderate penetrance were entirely plausible for several other variants including *DAO*:p.Arg38His (95% CI=$1.7X10^{-02}$ - 0.74) and *OPTN*:p.Ala481Val ($1.1X10^{-02}$ - 0.49) but that for most, if effects on disease risk exist then penetrance is most likely <10%.

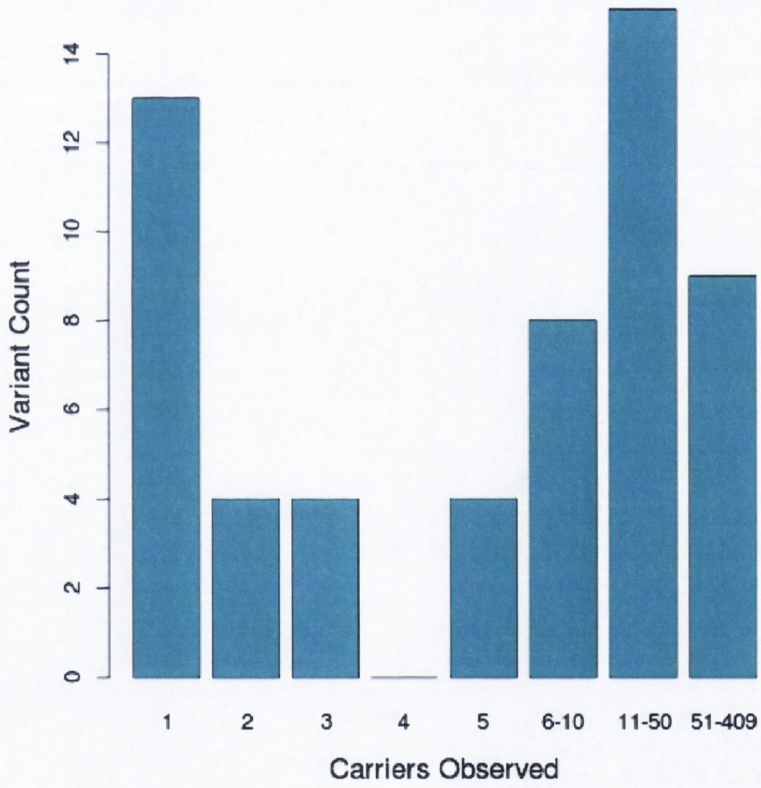


Fig 2.1: Frequencies of ALS associated variants within the combined 1000 genomes and ESP datasets. For the purposes of presentation the 285 variants with carrier counts of 0 have been omitted.

Table 2.1: Analysis of ALS related mutations.

Mutation	Ref	ALS	Matched	Global	Binomial		Fisher	
		(Carrier Freq)	(AF)	(AF)	P	Penetrance	OR	p
ANG:p.Pro21Ser	[92]	1/737 (EUR)	2/8600	2/13006	9.3x10-05	2.1X10-03 - 3.3X10-02	0.1 - Inf	0.38
ANG:p.Lys41Ile	[22]	2/1629 (EUR)	23/9356	27/15188	1.3x10-73	4.0X10-04 - 9.0X10-04	4.2X10-02 - Inf	0.99
ANG:p.Ile70Val	[22]	3/1629 (EUR)	5/8600	7/13006	2.7x10-15	1.7X10-03 - 9.8X10-03	0.33 - Inf	0.38
ANG:p.Lys78Glu	[93]	1/581 (EUR)	1/8600	1/13006	1.7x10-02	3.7X10-03 - 0.16	0.19 - Inf	0.22
ANG:p.Pro136Leu	[94]	1/298 (US)	1/13006	1/13006	5.0x10-02	1.1X10-02 - 0.48	0.57 - Inf	8.6x10-02
CDH13:p.Arg65Cys	[95]	1/190 (EUR)	1/8366	1/12418	5.0x10-02	1.1X10-02 - 0.48	0.57 - Inf	8.5x10-02
CDH13:p.Gly113Arg	[95]	1/190 (EUR)	17/9014	19/14308	7.2x10-38	2.1X10-03 - 5.4X10-03	6.8X10-02 - Inf	0.52
CDH13:p.Arg246Trp	[95]	1/190 (EUR)	1/8258	1/12098	5.0x10-02	1.1X10-02 - 0.47	0.56 - Inf	8.6x10-02
CDH13:p.Glu367Gln	[95]	1/190 (EUR)	13/8336	13/12250	2.7x10-27	2.4X10-03 - 7.1X10-03	8.0X10-02 - Inf	0.46
CDH22:p.Glu92Lys	[95]	1/190 (EUR)	1/9356	38/15188	5.9x10-82	1.8X10-03 - 3.4X10-03	5.2X10-02 - Inf	0.62
CDH22:p.Arg533His	[95]	3/190 (EUR)	31/9356	36/15188	2.9x10-62	4.0X10-03 - 8.1X10-03	0.62 - Inf	0.15
CRIM1:p.Asn406Ser	[95]	1/190 (EUR)	4/9356	5/15188	5.6x10-08	6.1X10-03 - 4.3X10-02	0.25 - Inf	0.18
CRYM:p.Arg169Cys	[95]	1/190 (EUR)	1/8600	1/12998	5.2x10-02	1.1X10-02 - 0.49	0.59 - Inf	8.3x10-02
DAO:p.Arg38His	[96]	1/126 (EUR)	1/8600	1/13006	7.7x10-02	1.7X10-02 - 0.74	0.89 - Inf	5.6x10-02
DAO:p.Arg199Trp	[25]	1/222 (EUR)	5/8600	6/13006	1.4x10-10	4.2X10-03 - 2.4X10-02	0.17 - Inf	0.26
DCTN1:p.Arg785Trp	[97]	2/250 (EUR)	7/8600	9/13006	1.5x10-14	5.9X10-03 - 2.6X10-02	0.74 - Inf	8.4x10-02
DCTN1:p.Thr1249Ile	[97]	1/250 (EUR)	39/9356	44/15188	1.1x10-105	8.4X10-04 - 1.6X10-03	2.4X10-02 - Inf	0.88
DIAPH3:p.Ile234Thr	[95]	1/190 (EUR)	5/8128	5/11738	2.6x10-09	4.6X10-03 - 2.6X10-02	0.18 - Inf	0.24
DIAPH3:p.Pro578Leu	[95]	1/190 (EUR)	7/9016	27/14304	2.9x10-57	2.3X10-03 - 4.9X10-03	6.9X10-02 - Inf	0.52
DIAPH3:p.Pro588Leu	[95]	9/190 (EUR)	376/9028	417/14334	0	1.2X10-03 - 1.5X10-03	0.29 - Inf	0.98
DIAPH3:p.Pro596Leu	[95]	1/190 (EUR)	49/9062	53/14366	2.0x10-123	8.8X10-04 - 1.5X10-03	2.4X10-02 - Inf	0.87
DIAPH3:p.Arg1042His	[95]	1/190 (EUR)	5/8874	5/13898	3.9x10-09	5.0X10-03 - 2.9X10-02	0.2 - Inf	0.22
FEZF2:p.Gly188Asp	[95]	1/190 (EUR)	5/8586	6/12986	3.5x10-10	4.8X10-03 - 2.8X10-02	0.19 - Inf	0.23
FIG4:p.Tyr647Cys	[45]	1/473 (EUR)	2/8598	2/13004	2.2x10-04	3.3X10-03 - 5.1X10-02	0.16 - Inf	0.27
FUS:p.Gly191Ser	[98]	1/1009 (EUR)	1/8600	1/12994	1.0x10-02	2.1X10-03 - 9.3X10-02	0.11 - Inf	0.34
GRB14:p.His507Tyr	[95]	1/190 (EUR)	17/8600	17/13006	5.6x10-37	2.0X10-03 - 5.2X10-03	6.4X10-02 - Inf	0.54

LUM:p.Leu199Pro	[95]	4/190 (EUR)	42/9356	47/15188	2.4x10-80	4.1X10-03 - 7.5X10-03	0.77 - Inf	0.1
NEFH:p.Lys867Asn	[95]	1/190 (EUR)	0/8600	28/13006	6.4x10-61	2.0X10-03 - 4.2X10-03	6.0X10-02 - Inf	0.57
NEFH:p.Glu918Gly	[95]	1/190 (EUR)	0/756	3/2182	4.0x10-07	1.7X10-03 - 1.6X10-02	7.4X10-02 - Inf	0.47
NETO1:p.Gly401Arg	[95]	1/190 (EUR)	26/9356	31/15188	1.5x10-66	1.5X10-03 - 3.3X10-03	4.7X10-02 - Inf	0.66
OMA1:p.His69Tyr	[95]	2/190 (EUR)	20/9356	124/15188	7.7x10-293	1.3X10-03 - 1.8X10-03	0.11 - Inf	0.82
OMA1:p.Glu272Gly	[95]	1/190 (EUR)	15/8600	20/13006	4.7x10-41	2.2X10-03 - 6.1X10-03	7.3X10-02 - Inf	0.5
OMA1:p.Asp365Tyr	[95]	1/190 (EUR)	44/9356	51/15188	2.0x10-119	9.9X10-04 - 1.8X10-03	2.8X10-02 - Inf	0.83
OPTN:p.Gln314Leu	[99]	1/274 (EUR)	3/8600	3/13006	8.1x10-06	4.6X10-03 - 4.4X10-02	0.2 - Inf	0.22
OPTN:p.Ala481Val	[100]	1/190 (EUR)	1/8600	1/13006	5.2x10-02	1.1X10-02 - 0.49	0.59 - Inf	8.3x10-02
SETX:p.Cys1554Gly	[101]	1/3 (EUR)	40/9356	47/15188	1.1x10-27	6.8X10-02 - 0.13	2 - Inf	2.6x10-02
SETX:p.Ile2547Thr	[101]	1/3 (EUR)	71/9356	76/15188	4.2x10-63	4.1X10-02 - 6.6X10-02	1 - Inf	4.5x10-02
*SOD1:p.Asp91Ala	[102]	34/451 (EUR)	9/9356	9/15188	1	1 - 1	75 - Inf	7.0x10-76
SOX5:p.Gln362Pro	[95]	1/190 (EUR)	2/8600	2/13006	1.4x10-03	8.3X10-03 - 0.13	0.39 - Inf	0.12
*SPG11:p.Val2053Met	[59]	1/25 (AMR)	0/0	3/12992	1	1 - 1	22 - Inf	1.4x10-04
SQSTM1:p.Ala33Val	[60]	3/520 (AMR,EUR)	6/7388	6/10962	2.1x10-11	4.1X10-03 - 2.0X10-02	0.77 - Inf	8.9x10-02
SQSTM1:p.Val153Ile	[60]	2/513 (EUR)	9/8600	9/13006	6.0x10-19	2.4X10-03 - 8.8X10-03	0.29 - Inf	0.33
SQSTM1:p.Pro228Leu	[60]	1/513 (EUR)	4/8600	5/13006	1.9x10-10	2.1X10-03 - 1.5X10-02	8.6X10-02 - Inf	0.43
SQSTM1:p.Arg321Cys	[60]	1/513 (EUR)	5/9336	159/15144	0.0e	1.9X10-04 - 2.6X10-04	4.7X10-03 - Inf	1
SQSTM1:p.Ser370Pro	[60]	1/10 (AFR)	29/4406	29/13006	1.6x10-38	1.3X10-02 - 2.6X10-02	0.37 - Inf	0.13
SQSTM1:p.Pro392Leu	[60]	3/513 (EUR)	22/9356	29/15188	3.3x10-60	2.0X10-03 - 4.5X10-03	0.32 - Inf	0.46
SQSTM1:p.Gly411Ser	[60]	1/513 (EUR)	0/8600	1/13006	2.9x10-02	6.4X10-03 - 0.28	0.33 - Inf	0.14
SYT9:p.Val154Met	[95]	12/190 (EUR)	295/9348	388/15176	0.0e	2.1X10-03 - 2.7X10-03	0.57 - Inf	0.54
SYT9:p.Leu353Val	[95]	3/190 (EUR)	100/9348	110/15176	5.9x10-240	1.4X10-03 - 2.1X10-03	0.2 - Inf	0.77
TAF15:p.Asp386Asn	[62]	1/1447 (GLO)	1/13006	1/13006	1.1x10-02	2.3X10-03 - 9.8X10-02	0.12 - Inf	0.33
TAF15:p.Arg388His	[62]	2/1447 (GLO)	7/13006	7/13006	3.6x10-16	1.5X10-03 - 6.7X10-03	0.19 - Inf	0.51
TAF15:p.Arg395Gln	[62]	2/1447 (US)	2/13002	2/15184	3.0x10-04	3.3X10-03 - 5.1X10-02	0.49 - Inf	0.15
TARDBP:p.Ala90Val	[34]	1/148 (GLO)	3/13006	3/13006	1.7x10-04	1.3X10-02 - 0.12	0.57 - Inf	8.6x10-02
TARDBP:p.Asn267Ser	[103]	1/666 (EUR)	1/756	1/2182	1.3x10-03	2.9X10-04 - 1.2X10-02	1.5X10-02 - Inf	0.87
TARDBP:p.Asn390Ser	[104]	1/200 (EUR)	1/8600	1/13006	4.9x10-02	1.1X10-02 - 0.47	0.56 - Inf	8.7x10-02

For both binomial and fisher analyses, reported p-values and effect sizes describe the result for the most significant effect observed across either the matched or global analysis.
**Denotes reported relevance of recessive disease models. EUR European; US United States; AMR South/ Central American*

2.3.4 Analysis of ALS-associated variants identified in dbSNP

dbSNP records are based on externally generated genetic data which has been submitted to the project for curation. Records therefore do not relate to a single well defined sample cohort, and often manual review and extensive follow up is required before. genotype frequencies can be established for clearly defined reference populations. However, in many cases minor allele frequency data from the 1000 genomes project is available and database content such as the associated submitter handles can be used to determine whether an observed variant is likely to be of interest. Of the 142 ALS mutations identified in dbSNP, 55 had also been reported by the 1000 genomes project or the ESP. The remaining 87 had primarily been imported from the OMIM online catalogue of human genes and genetic disorders ('http://omim.org/') or NCBI's "GeneReviews" collection of peer reviewed disease descriptions (http://www.ncbi.nlm.nih.gov/sites/GeneTests/review) (Supp Table S2). Review of the OMIM and GeneReviews documentation associated with ALS mutations, revealed that most had simply been identified within individuals afflicted with ALS. However, some were reported in the context of additional disease phenotypes. For example c.1175C>T(p.P392L) in *SQSTM1* (MIM# 601530) was described only in the context of Paget's disease (MIM# 602080, MIM# 601530.0001) while c.547C>T(p.Arg183*) in *FIG4* (MIM# 609390), c.464G>A(p.Arg155His) in *VCP* (MIM# 601023) and c.572G>A(p.Arg191Gln) in *VCP* were recorded both as causes of ALS and either Charcot-Marie Tooth disease (MIM# 611228, MIM# 609390.0003), inclusion body myopathy with early onset Paget's disease and frontotemporal dementia (IBMPFD; MIM# 167320, MIM# 601023.0001), and a facioscapulohumeral muscular dystrophy like phenotype (MIM# 158900, MIM# 601023.0006). These mutations may therefore not represent highly penetrant ALS mutations in the strictest sense but may still constitute more broad categories of high penetrance disease mutations.

2.4 Discussion

The genetic variation reported within dbSNP, the 1000 genomes project and the ESP ultimately reflects the cumulative analysis of a very large series of individuals. By chance alone, these resources would therefore be expected to include mutations contributing to a wide array of disease phenotypes. Furthermore, unlike the 1000 genomes project or ESP, a given dbSNP record may reflect the observation of a variant within highly selected sample populations, including disease based cohorts. Causative disease variants may therefore be overlooked by studies attempting to establish the basis of a genetic disorder if any variant reported by these resources is dismissed without due consideration. This in turn can lead to misinterpretation of the likely disease relevance of other variants. For example using exome sequencing, Johnson *et al.* identified the c.572G>A(p.Arg191Gln) mutation in *VCP* as the underlying cause of disease within a family presenting with dominantly inherited ALS [37]. The variant was selected from a list of 4 possible disease mutations, with the potential relevance of all other exonic variants having been excluded through segregation analyses, *in silico* predictions of pathogenicity and occurrence within dbSNP and the 1000 genomes. The *VCP*:c.572G>A(p.Arg191Gln) variant has subsequently

been imported into dbSNP from OMIM, where it is indicated to have independently been reported in the context of additional disease phenotypes (MIM# 601023.0006) [15,16]. A re-analysis of the ALS kindred investigated by Johnson *et al.* with a more recent build of dbSNP could therefore result in the selection of one of the other 3 potential disease variants identified as the most credible basis for disease.

In this study we observed that 52% of mutations reported as pathogenic across the PhenCode network of LSDBs can be observed within one of the 3 major reference databases of human genetic variation. These mutations had been implicated in a highly diverse array of 1682 clinical phenotypes which included Alzheimer's disease [MIM# 104300], breast cancer [MIM# 114480], deafness [MIM# 612644], familial hypertrophic cardiomyopathy [MIM# 192600], phenylketonuria [MIM# 261600] and retinitis pigmentosa [MIM# 180100] among many others. The results from our analysis of ALS mutations could be taken to suggest that certain of these mutations may well represent high penetrance disease variants but also that many are likely to represent either non-pathogenic or else low penetrance variants. The vast majority of ALS variants identified in dbSNP but not the 1000 genomes project or ESP, had simply been imported from the OMIM database or NCBI's GeneReviews. Given the disease orientated nature of these resources it would be anticipated that variants imported from these sources would generally have been encountered in the context of disease phenotypes. Review of the corresponding OMIM/ GeneReviews records revealed that most of the imported ALS mutations had simply been identified among individuals afflicted with ALS but that some had been reported in the context of additional disease phenotypes. This latter observation is particularly noteworthy as it highlights the potential importance of allowing for variable expressivity when evaluating the credibility of high penetrance disease models. Analysis of reported genotype data for the 57 ALS associated mutations identified in the 1000 genomes project/ ESP revealed generally revealed an absence of notable evidence to support disease associations. Conversely the evidence arguing against involvement in at least Mendelian disease tended to be considerably stronger. Exceptions to this included the *SOD1*:p.Asp91Ala and *SPG11*:p.Val2053Met mutations which were found to have MAF entirely compatible with fully penetrant recessive disease expression. This is notable as not only is it a demonstration that variants responsible for rare Mendelian disorders can be encountered by the 1000 genomes project/ ESP, but it highlights that where recessively inherited conditions are concerned such mutations can be identified without necessarily having to assay individuals expected to develop the disease.

Thirty two of the mutations rejected in binomial analyses as full penetrant ALS variants occurred within genes that would generally be regarded as candidate ALS genes (*CDH13* [MIM# 601364], *CDH22* [MIM# 609920], *CRIM1* [MIM# 606189], *DIAPH3* [MIM# 614567], *FEZF2* [MIM# 607414], *GRB14* [MIM# 601524], *LUM* [MIM# 600616], *NETO1* [MIM# 607973], *OMA1, SOX5* [MIM# 604975], *SQSTM1* [MIM# 601530], *SYT9* [MIM# 613528]). The remaining 25 occurred within genes that would be regarded as established or at least probable ALS genes (*ANG*, *DAO* [MIM# 124050], *DCTN1* [MIM# 601143], *FIG4* [MIM# 609390], *FUS* [MIM# 137070], *NEFH* [MIM# 162230], *OPTN* [MIM# 602432], *SETX* [MIM# 608465], *SOD1, TARDBP* [MIM# 605078], *TAF15* [MIM# 601574]). The published evidence supporting the pathogenicity of these mutations varied extensively. Some such as c.1169A>G(p.Asn390Ser) in *TARDBP* or c.571G>A(p.Gly191Ser) in *FUS* had simply been encountered during screening of patient cohorts [98,104]. Others had been identified through patient screening but subsequently been shown to have potentially important functional effects, such as

c.407C>T(p.Pro136Leu) in *ANG* which was shown to reduce ribonucleolytic activity and abolish angiogenic activity[94]. Perhaps the most notable of the variants was the c.595C>T(p.Arg199Trp) in *DAO*. This mutation had been identified through linkage analysis in a 3 generational ALS kindred and it was found that expression of the mutated protein in neuronal cells decreased cell viability while increasing the presence of UBI [25]. However, our analysis not only failed to detect notable evidence for disease association but also indicated that if the variant does influence ALS risk, it acts with a penetrance in the order of 0.42-2.4%. If on the otherhand, c.595C>T(p.Arg199Trp) does not represent an ALS mutation, the causative variant within the original kindred remains to be ascertained and the pathogenic relevance of subsequently reported *DAO* mutations becomes more questionable. However it should be noted that the binomial modelling of variant penetrance performed here has certain limitations. Firstly, the 1000 genomes and ESP cohorts were treated as an unbiased sample of the general population. However, age, gender or potentially other biases are likely to have been relevant in sample ascertainment, limiting the accuracy of the lifetime risk estimate employed. Additionally the approach did not cater for the inaccuracies in reported patient carrier frequencies. In general such inaccuracies would be anticipated to bias towards a failure to reject a tested disease model, as due to ascertainment bias, carrier frequencies may represent overestimates more often than underestimates but the reverse is quite possible. A certain degree of caution in interpreting those results based on very few observations (<=2 occurrences in the control cohort) should therefore be employed. Secondly it should also be re-iterated that the analyses consider only full penetrance disease models and that penetrance refers only to the proportion of carriers that develop ALS. As discussed in section 1.1.6, such variants are anticipated to account for only a small proportion of genetic risk and the potential importance of those variants attributed smaller effects should therefore not be overlooked. Likewise, the fact that a variant does not cause ALS with high penetrance does not infer that it does not trigger some aspect of a disease spectrum that includes ALS with high penetrance.

2.5 Conclusion

Our results highlight the lack of formal evidence to support the pathogenicity of a substantial proportion of reported ALS mutations. Moreover they also reveal that a considerable proportion of these mutations are likely to only act with low penetrance (<10%), if they do indeed influence disease risk at all. Based on our analysis of the PhenCode network of LSDBs, these findings may prove generalizable across a wider array of disease phenotypes. Our results also demonstrate the potential consequences of utilizing simplistic variant filtering strategies that do not consider the content of database records before using them to exclude variant pathogenicity in gene mapping. In the case of the 1000 genomes project and the ESP, records can easily be used to establish genotype frequencies within well defined reference populations. These can then be incorporated into simple binomial tests for the purposes of evaluating the credibility of disease causing effects as shown here. Moreover the whole process can be incorporated into a simple and automatable analysis pipeline. In the case of dbSNP, tailored sub-releases and minor allele frequency data imported from projects such as the 1000 genomes can be used to flag variants which represent

known polymorphisms but in many cases optimal variant classification may require a certain degree of manual review.

3. *In silico* functional profiling of ALS related genes

3.1 Introduction

The clinical manifestations of disease mutations are ultimately mediated through effects on the functioning of normal biological pathways. As might be expected, mutations which influence related biological processes can cause related disease phenotypes. For example, the pathogenesis of Alzheimer's disease (AD) is at least partially driven by the accumulation of pathogenic isoforms of the amyloid beta peptide (Aβ) and familial forms of AD are caused by mutations that influence Aβ synthesis (Figure 3.1). The molecular mechanisms driving ALS pathogenesis have not yet been determined but studies of known ALS genes have suggested the potential relevance of oxidative stress and dysfunctions in protein homeostasis, vesicular transport, RNA metabolism and mitochondrial homeostasis [105].

The existence of clear functional relationships between certain disease genes highlights the potential for identifying candidate disease loci through functional profiling and this basic idea has already contributed to the discovery of at least 2 ALS genes[22,28]. Numerous *in silico* prediction tools have been developed to conduct such profiling, including POCUS[106], PROSPECTR[107], Endeavour[108], GeneWanderer[109], SUSPECTS[110] ToppGene[111] and ToppNet[112] among others (http://homes.esat.kuleuven.be/~bioiuser/gpp/). The inner workings of these tools vary but all utilize one or more categories of biological data to rank a set of test genes by their relevance to either a set of provided key words or a training set of known disease genes.

In this section I describe, the *in silico* functional profiling of 9 known ALS genes (*ALS2, ANG, DCTN1, FUS, OPTN, SETX, SOD1, TARDBP, VAPB*) and the subsequent use of potentially disease related features in candidate gene prioritization. I also discuss results from a series of validation analyses performed using a set of 93 additional disease genes.

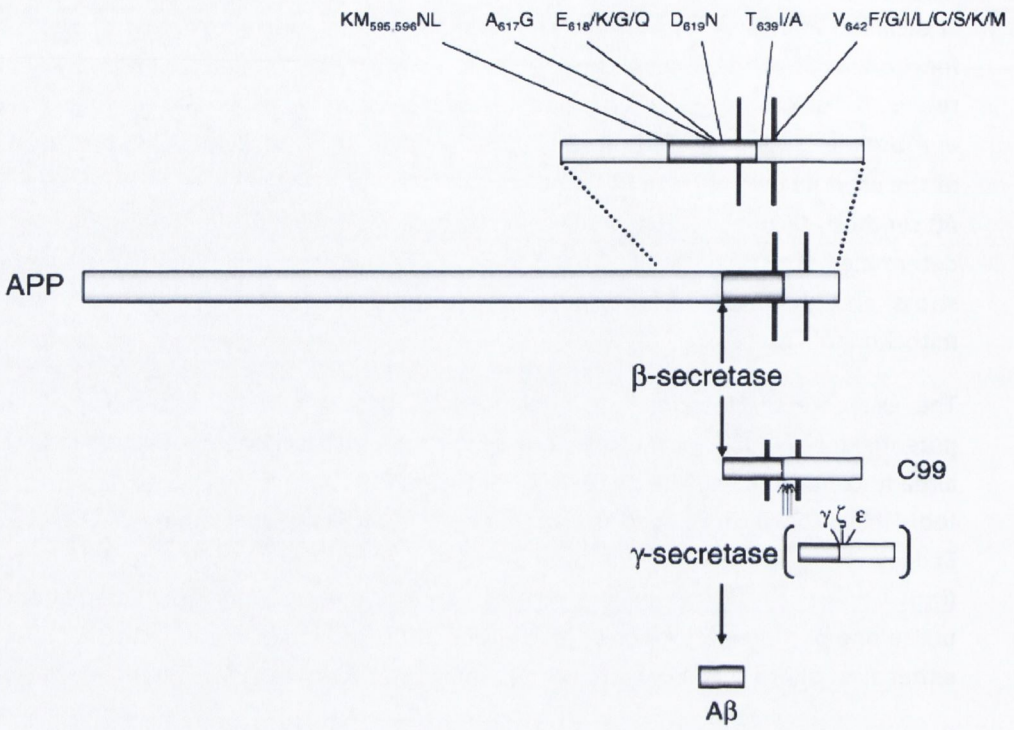


Figure 3.1 Aβ synthesis. Aβ is a normal product of the amyloid precursor protein (APP) processing pathway. It is generated by proteolytic cleavage of APP by β and γ–secretase. Mutations of the *APP* gene or the genes that encode the active site components of γ-secretase (*PSEN1* and *PSEN2*) increase the production of pathological Aβ species and cause early onset familial AD. Image taken from [113].

3.2 Materials and methods

3.1 Gene sets

Four gene sets are considered in this chapter. The first is referred to as the "training set" (tALS) and represents the panel of 9 known ALS genes (*ALS2, ANG, DCTN1, FUS, OPTN, SETX, SOD1, TARDBP, VAPB*) around which all functional profiling is centred. The second is referred to as the "validation set" (vALS). This represents the genes used to evaluate the performance of ALS gene enrichment methods and includes 93 known or potential ALS genes not represented within the training set (*AGT, ALAD, APEX1, APOE, AR, ATXN2, B4GALT6, BCL11B, BCL6, C1orf27, C9orf72, CCS, CDH13, CDH22, CHMP2B, CNTF, CNTN4, CNTN6, CRIM1, CRYM, CSNK1G3, CST3, CYP2D6, DAO, DIAPH3, DISC1, DOC2B, DPP6, DYNC1H1, EFEMP1, ELP3, EWSR1, FEZF2, FGGY, FIG4, GARS, GRB14, HEXA, HFE, IFNK, ITPR2, KDR, KIFAP3, LIF, LIPC, LOX, LUM, MAOB, MAPT, MOBKL2B, MTND2, NAIP, NEFH, NETO1, NT5C1A, OGG1, OMA1, PCP4, PGRN, PON1, PON2, PON3, PRPH, PSEN1, PVR, RAMP3, RBMS1, RNF19A, SCN7A, SELL, SEMA6A, SIGMAR1, SLC1A2, SLC39A11, SMN1, SMN2, SNCG, SOD2, SOX5, SPAST, SPG11, SPG7, SUSD1, SYT9, TAF15, UBQLN2, UNC13A, VCP, VDR, VEGFA, VPS54, ZFP64, ZNF746*). The third gene set is referred to as the "reduced validation" set

(rvALS). This represented the set used to evaluate the potential influence of discovery biases and included genes not know to represent ALS genes at the time of the original analyses (*C9orf72, VCP, ATXN2, C9orf72, TAF15, UBQLN2, VCP*). The final set is referred to as the "test set". This includes the 5980 genes potential candidate genes selected on the basis of protein sequence and/or interaction data as described in section 3.3.2.

3.2. Retrieval of molecular interaction data

Experimentally demonstrated molecular interactions were retrieved from the Biomolecular Interaction Network Database (BIND)[114], the Biological General Repository for Interaction Datasets (BioGRID)[115] and the Human Protein Reference Database (HPRD)[116] via entrez gene[117] and from IntAct[118], the Molecular INTeraction database (MINT)[119], DIP[120], Reactome[121] and the Pathway Interaction Database (PID)[122] directly. PubMed searches were used to retrieve publications which included the names of known ALS genes within the title or abstract. These publications were then screened for reports of protein-protein interactions.

3.2.3 Processing of interaction data

The retrieved interaction data was integrated into a single database of binary interactions. All interactors were mapped to entrez gene IDs and interactions where a participant could not be mapped to an entrez gene ID were discarded. Interactions were also removed where one or both participants did not represent proteins, where the only reported evidence for the interaction was an observation of "colocalization" or computational prediction, where neither participant corresponded to a human protein sequence or where 1 participant corresponded to a human sequence but the other exhibited less than 90% sequence identity to a human ortholog. Complexes reported in XML formats were converted to binary interactions by "spoke expansion" unless the reported experimental evidence justified use of "matrix expansion" (Figure 3.2)

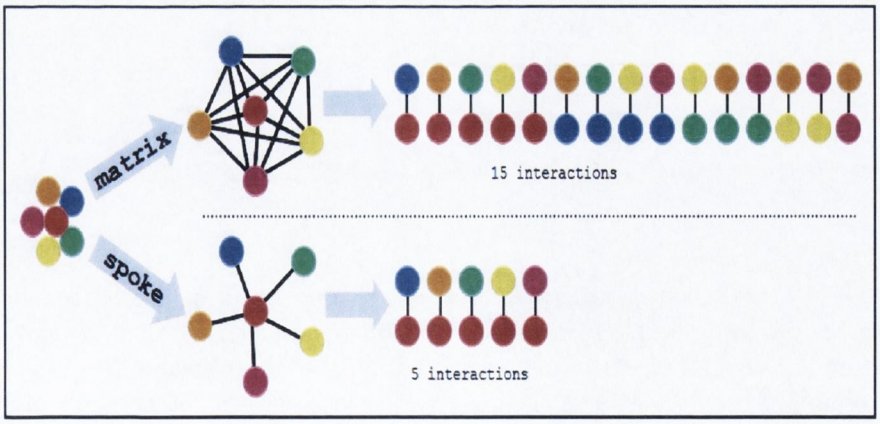


Figure 3.2: Resolving complexes. Under matrix expansion, binary interactions are inferred between all components of a reported complex. Under spoke expansion, binary interactions are inferred only between a "prey" protein and each of the associated "bait" proteins. Image taken from "http://www.ebi.ac.uk/intact/".

3.2.4 Protein sequence similarity

Analyses of protein sequence similarity were conducted using the Basic Local Alignment Search Tool (BLAST)[123]. The "blastp" algorithm was used for all analyses and only sequences from the NCBI Reference Sequence Database (RefSeq)[124] were considered.

3.2.5 Functional enrichment analyses

Functional enrichment analyses were performed using ToppFun (http://toppgene.cchmc.org/)[125]. Enabled categories included GO: Molecular Function, GO: Biological Process, GO: Cellular Component, Human Phenotype, Mouse Phenotype, Domain, Pathway, Pubmed, Transcription factor binding site, Gene Family, Co-expression, Computational, MicroRNA and Drug. Details on the data-sources mined by ToppFun can be found on the project website. P-values for reported enrichments were estimated from hypergeometric distributions. Multiple testing correction was performed using the Bonferroni method.

3.2.5 Disease annotation

Disease annotations were either retrieved directly from the literature or from the morbid map download provided by OMIM (http://www.omim.org/downloads).

3.2.6 Quantification of relative network importance

Relative network importance was quantified using a custom implementation of the PageRank-priors algorithm[126] (Script S1). This represents a generalization of the Google PageRank algorithm developed by Larry Page and Sergey Brin. It computes, the probability of traversing each node in a network during a random walk that originates at one of a set of seed nodes and restarts stochastically with back probability β (Equation 3.1). Based on initial testing and published recommendations[126], the β parameter was set to 0.3 for all analyses. For all analyses the prior bias assigned to test nodes was 0 while the prior bias assigned to seed nodes (training genes) was the inverse of the count of seed nodes.

$$\pi(v)^{(i+1)} = \beta p_v + (1-\beta) \sum_{u=1}^{d_{in}(v)} p(v|u)\pi^{(i)}(u)$$

Equation 3.1

p_v denotes the prior bias assigned to a node v. This represents the probability of a random walk starting at that node. β denotes the probability that on a given step a random walk jumps back to a seed node before continuing. Where β=0, the algorithm calculates the global importance of each node in the network as opposed to the importance of each node relative the set of seed nodes. This special case represents the original PageRank algorithm. Where β = 1, a random walk never leaves the set of seed nodes and the importance of all other nodes in the network remains at 0. The β parameter therefore facilitates fine tuning of the weighting given to seed node proximity. $p(v|u)$ denotes the probability of arriving at node v when taking a random step from node u. If node u is not a direct neighbour of node v, this value will be 0. Otherwise, this value will be given by the inverse of the number of neighbours of node u. $\pi(v)^{(i+1)}$ and $\pi^{(i)}(u)$ represent the importance of nodes v and u on the i^{th}+1 and i^{th} iteration of the algorithm. For each iteration, the

importance of a node is therefore determined by the relative importance of its neighbours and the number of nodes those neighbours interact with.

3.2.7 Quantification of gene annotation similarity

The ToppGene webtool (http://toppgene.cchmc.org/)[125] was used to rank test genes by their similarity to the set of training genes. As with functional enrichment analyses performed using the related ToppFun webtool (section 3.2.5), enabled features included GO: Molecular Function, GO: Biological Process, GO: Cellular Component, Human Phenotype, Mouse Phenotype, Domain, Pathway, Pubmed, Transcription factor binding site, Gene Family, Co-expression, Computational, MicroRNA and Drug. For categorical annotations test genes were scored based on their similarity to terms over-represented among the set of training genes, with over-representation being assessed by modelling under a hypergeometric distribution. Similarity was assessed using Sugeno fuzzy measures[127]. This allowed the informativeness of a term to be taken into account during scoring, meaning that lower weightings could be assigned to more general terms while higher weightings could be assigned to more specific terms. The methodology also allowed for the comparison of non-identical terms in hierarchical ontologies such as the gene ontology (GO), based on distance to a common parental term (Figure 3.3). For numerical annotations, including gene expression profiles, test genes were scored based on the similarity of various expression values to the averages of those across the set of training genes. In this case, similarity was measured using Pearson's correlation coefficient. Similarity scores were calculated for each category independently. The significance of each score was then assessed through scoring a set of randomly sampled genes and significance values were combined to an overall score using Fisher's inverse chi-square method (Equation 3.2)

$$\chi^2 = -2\sum_{i=1}^{k} log_e(pi)$$

Equation 3.2

pi denotes the i^{th} of k p-values. The distribution of the resulting χ^2 statistic can be approximated by a chi-square distribution with 2k degrees of freedom.

3.2.8 Data analysis

The probabilities of apparent enrichments for vALS proteins within the local interaction neighbourhoods of tALS proteins and among the sequence similarity hits of tALS proteins were estimated using one-tailed hypergeometric tests. These were performed using the following R command;

```
p=dhyper(q,m,n,k)+phyper(q,m,n,k,lower.tail=F)
```

For network enrichment analyses; q denoted the number of vALS proteins occurring within a given network distance of tALS proteins, m denoted the number of vALS genes represented in the network, n denoted the number of non-vALS genes represented in the network and k denoted the total number of genes occurring within a given network distance of tALS proteins. For sequence similarity hit enrichment analyses; q denoted the number of vALS genes returned as protein sequence similarity hits for tALS genes, m denoted the total number of vALS genes, n denoted the

total number of genes annotated as protein coding by entrez gene (26805) and k denoted the total number of sequence similarity hits returned for tALS genes. The probability of the observed overlap between potential candidate genes identified through molecular interaction and sequence similarity analyses was also calculated using a one-tailed hypergeometric test. To reduce the potential influence of study bias the considered gene space was limited to genes represented within the interaction data. All comparisons of vALS and non-vALS gene rankings were conducted using one-tailed Wilcoxon ranksum tests. When assessing the significance of vALS gene rankings through permutation, p-values were computed for each vALS gene independently and then combined using Fisher's inverse chi-squared method (equation 3.2).

Network diagrams were generated using cytoscape version 3.0.2 (http://www.cytoscape.org/). Unless otherwise indicated all data processing and analysis was conducted using custom scripts written for python 2.6.5 (http://www.python.org/), R version 2.14.1 (http://www.r-project.org/) or bash version 4.2.24 (http://www.gnu.org/software/bash/).

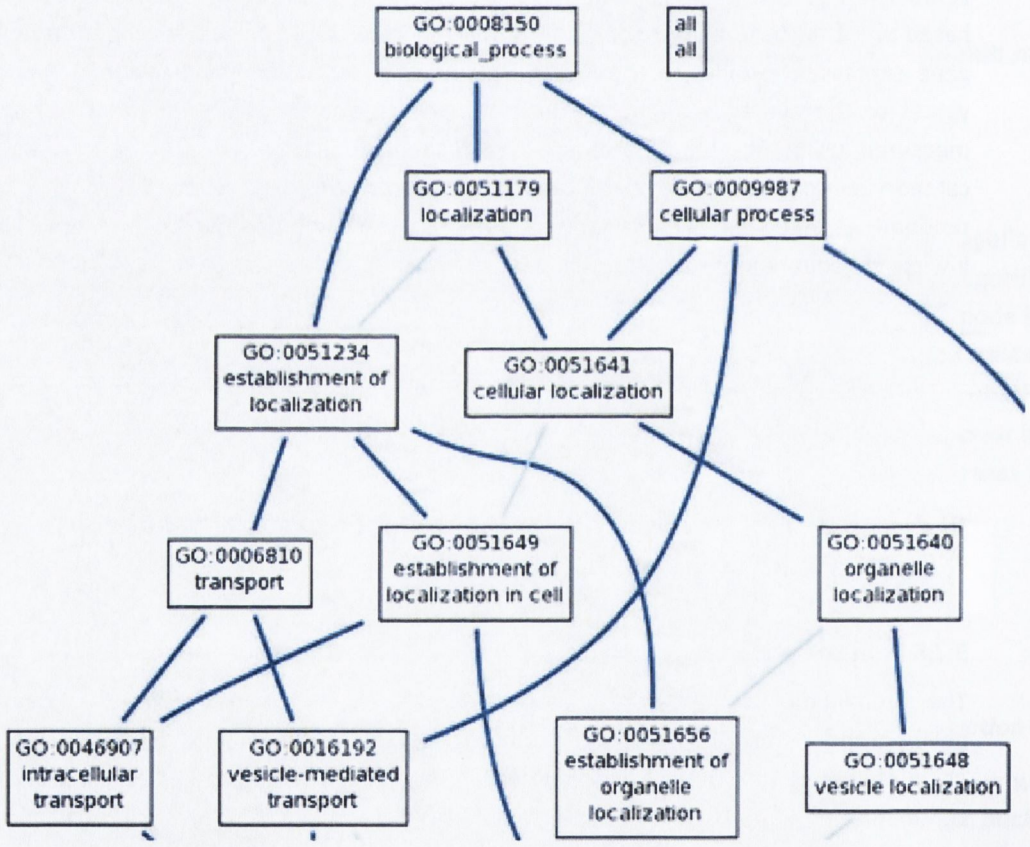


Figure 3.3: Comparing non-matching terms. The hierarchical structure of ontologies like the gene ontology allows for comparison of non-identical terms through consideration of less specific terms. For example the biological processes "intracellular transport" and "vesicle-mediated transport" could be viewed as being more similar to one another than to "establishment of organelle localization" as while all 3 terms represent special cases of "establishment of localization"; "intracellular transport" and "vesicle-mediated transport" also represent special cases of the more specific biological process "transport".3.3 Results

3.3.1 Network neighbourhood of ALS proteins

Ninety thousand nine hundred and thirty nine experimentally validated and non-redundant human protein-protein interactions were retrieved from the biomedical literature and dedicated molecular interaction databases (section 3.2.1). These interactions involved a total of 11850 proteins and after the removal of self-interactions each protein participated in 1-804 interactions (Figure 3.4). Proteins encoded by all 9 training (tALS) genes were represented, as were proteins encoded by 72 of the 93 validation (ALS) genes and 4 of 5 genes included in a reduced validation (rvALS) set (section 3.3.3).

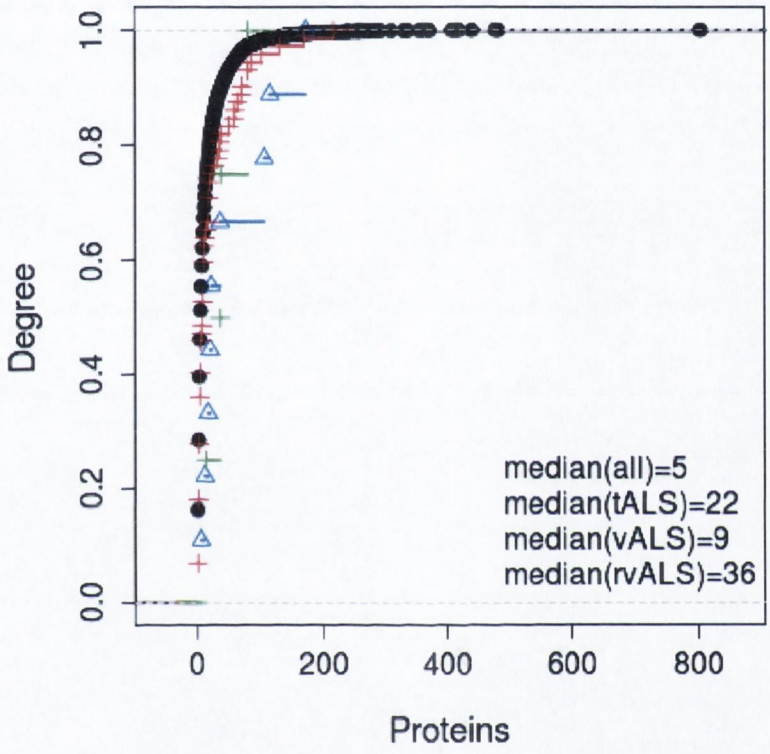


Figure 3.4: Degree distribution. Empirical cumulative density plots denoting the number of interactions partners (degree) of proteins within the assembled interaction network. Data is shown for all proteins in the network (black dots), the tALS set of training genes (blue triangles), the vALS set of validation genes (red crosses) and the rvALS reduced validation set (green crosses).

The immediate network neighbourhoods of proteins encoded by tALS genes included direct interactions with proteins encoded by 407 non-tALS (tALSp1) genes. A direct interaction was also observed between the proteins encoded by 2 tALS genes (*FUS* and *TARDBP*). The tALSp1 proteins also interacted with proteins encoded by a further 5070 non-tALS (tALSp2) genes. Approximately 47% of the total network could therefore be reached from the tALS nodes in a shortest path ≤ 2. The immediate network neighbourhoods of the *DCTN1, FUS, SOD1, SETX* and *TARDBP* nodes formed a single interconnected component whereas the neighbourhoods of the *ALS2, ANG, OPTN* and *VAPB* nodes were isolated from those of other tALS genes (Figure 3.5). The 407 tALSp1 proteins included 362 that interacted with a single tALS protein, 40 that interacted with 2 tALS proteins and 5 that interacted with 3 tALS proteins. Functional enrichment analyses of the tALSp1 genes revealed a multitude of over-represented processes (Table S3.1), with many of the most

significantly enriched annotations relating to RNA processing, the cell cycle, viral infection, intracellular transport and protein trafficking. Independent analyses of the interactors of each tALS gene suggested that the global enrichment for cell cycle terms was primarily driven by *DCTN1* interactors, while RNA processing, viral infection, protein trafficking and intracellular transport terms were enriched among the interactors of multiple tALS genes. Unsurprisingly, it was also observed that some of the annotations which exhibited significant over-representation among interactors of the more isolated tALS genes, like *ALS2* or *ANG,* had not appeared as over-represented in the combined analysis of all tALSp1 genes. Over-represented terms among the interactors shared by multiple tALS proteins primarily related to RNA processing (Table S3.2). Significant overrepresentations of vALS proteins were observed within the 1st and 2nd order interaction neighbourhoods of the tALS proteins (Table 3.1). Such enrichments were not observed within larger neighbourhoods, and restriction of the analyses to the rvALS validation set suggested that these observations did not simply stem from discovery biases.

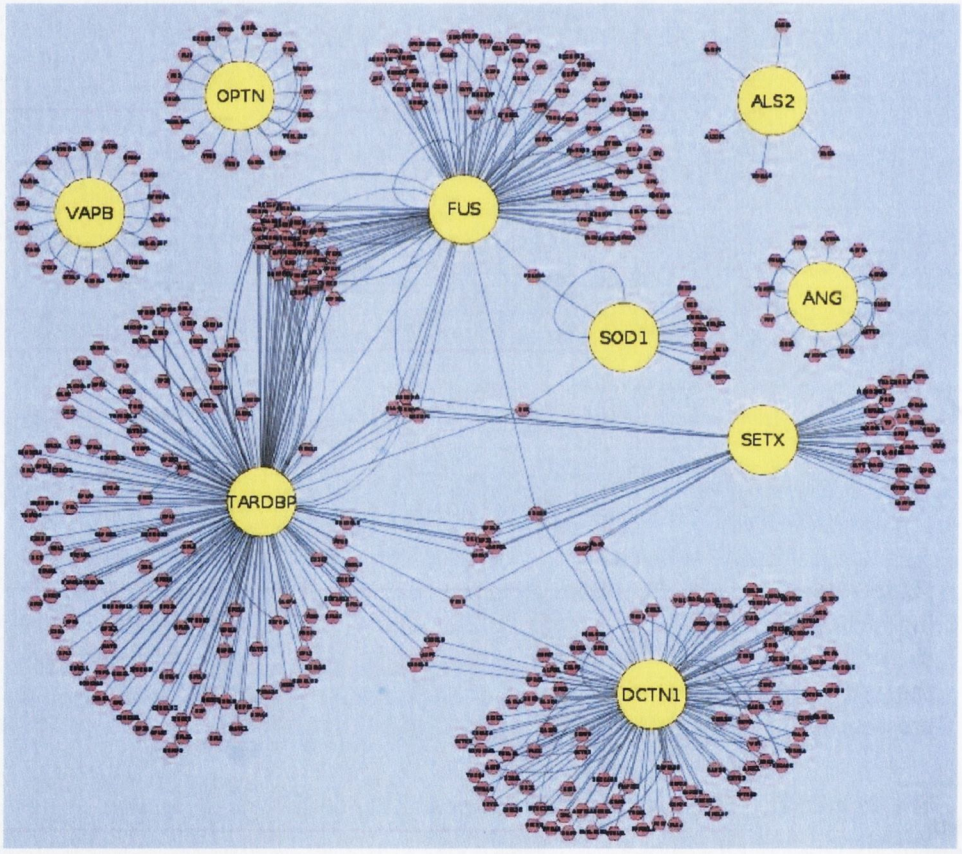


Figure 3.5: The local network neighbourhoods of proteins encoded by the tALS training genes..

Table 3.1: ALS proteins exhibit close network proximity

D	n	vALS	Obs/Exp	p	rvALS	Obs/Exp	p
1	407	9	3.6	8.2×10^{-4}	1	7.2	0.13
2	5477	41	1.2	0.05	4	2.1	0.05
3	10894	68	1.0	0.4	4	1.1	0.76

D Network distance from training ALS genes; n Total number of proteins; vALS Number of proteins from ALS gene validation set; rvALS Number of proteins from the reduced validation set. Obs/ Exp. Observed fold enrichment for vALS and rvALS proteins; p significance of enrichments

3.3.2 Sequence similarity

Analysis of the RefSeq database of protein sequences revealed that proteins encoded by 908 non-tALS genes could be aligned with a tALS protein at an e-value <10 (Figure 3.6). This meant that ~10 of the observed alignments would be expected to have arisen by chance as opposed to reflecting true sequence homology.

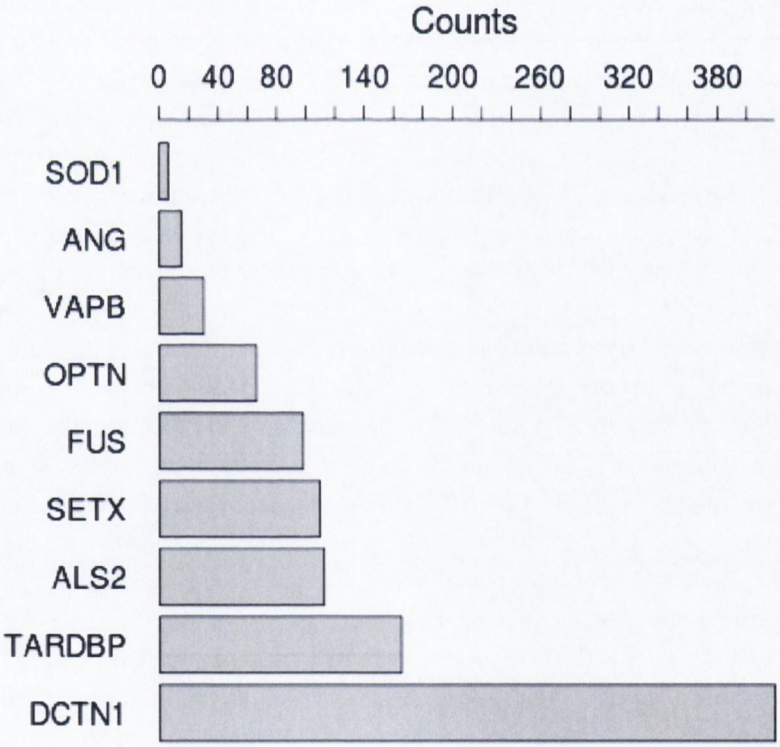


Figure 3.6: Sequence similarity hits. Counts of the number of genes returned as sequence similarity hits for each tALS protein at an e-value threshold <10.

Four genes (*CDC42BPG, KMT2C, RAD50, SPECC1*) were returned as hits for 3 tALS proteins, and a further 104 genes were returned as hits for 2 tALS proteins. Functional enrichment analyses of these proteins suggested a significant enrichment for genes with functions in RNA processing (Table S3.3). The remaining 800 genes were returned as a hit for a single tALS protein only. These

genes appeared to be more diverse in terms of associated molecular functions which enriched annotation categories including RNA processing, the cytoskeleton, transport and GTPase activity (Table S3.4).

OPTN was returned as a sequence similarity hit for *SETX* (e=7.7) but no other tALS-tALS protein alignments were observed. The returned hits also included 7 vALS proteins (Obs/Exp=2.2, p=0.04) and 1 rvALS protein (Obs/Exp=5.9, p=0.16).

3.3.3 Candidate gene prioritization

Given the potential enrichments for ALS related genes described in sections 3.3.1 and 3.3.2, all genes encoding proteins that interacted with a tALS protein, shared an interaction partner with a tALS protein or exhibited potential sequence homology with a tALS protein were included in an initial "test set" of potential candidate genes. This amounted to 5980 genes with the overlap between genes selected on the basis of interaction and sequence similarity data being 403. This overlap represented 63% of the sequence similarity hits occurring within the network (obs/exp=1.3, $p=5.8x10^{-18}$), with 70 of these hits interacting directly with a tALS protein (obs/exp=3.2, $p=7.7x10^{-19}$).

To aid in the selection of the most promising candidates for genetic follow up, a multi-step prioritization strategy was implemented. Firstly 103 genes were automatically selected for screening. These included 40 genes occurring within partially mapped ALS linkage intervals (Table 3.2) and 63 genes associated with conditions regarded as ALS related on the basis of evidence for shared susceptibility loci (Figure 3.7). The remainder of test set genes were then ranked for follow up firstly by their relative importance to the set of tALS proteins in the global protein interaction network described in section 3.3.1, secondly by their extent of protein sequence similarity with the set of tALS proteins and finally by their similarity to a weighted tALS gene functional profile that incorporated gene ontology annotation, pathway database annotation, drug binding annotation, human and mouse phenotype annotation, functional domain annotation, micro RNA binding site annotation, transcription factor binding site annotation, gene family annotation, literature citation and an assortment of associated gene expression values. Ranking by relative network importance was conducted using a custom implementation of the PageRank-priors algorithm (section 3.2.6, section 3.3.4), while ranking by protein sequence similarity was achieved using blastp and ranking by general gene annotation was achieved using the ToppGene webtool (section 3.2.7, section 3.3.5). Blastp and ToppGene based rankings were then combined into an overall "similarity ranking" by ordering genes by the product of the rank ratios achieved in the respective analyses (achieved ranking/ number of genes ranked). Blastp rankings were reset to 5980 (corresponds to a rank ratio of 1) in cases where observed alignments were estimated to have an >50% probability of reflecting chance similarity. An overall prioritization score was then generated by combining the overall similarity rankings with the PageRank-priors rankings, once again by ordering genes in terms of the product of the rank ratios achieved within the respective analyses.

Table 3.2: Linkage interval overlap

Interval	Genes	Test set	Reference
chr9:26181443-33859887	29	16	128
chr9:73802658-85784342	31	8	129
chr18:49989153-57456603	26	10	130
chr20:1-1094252	22	6	131

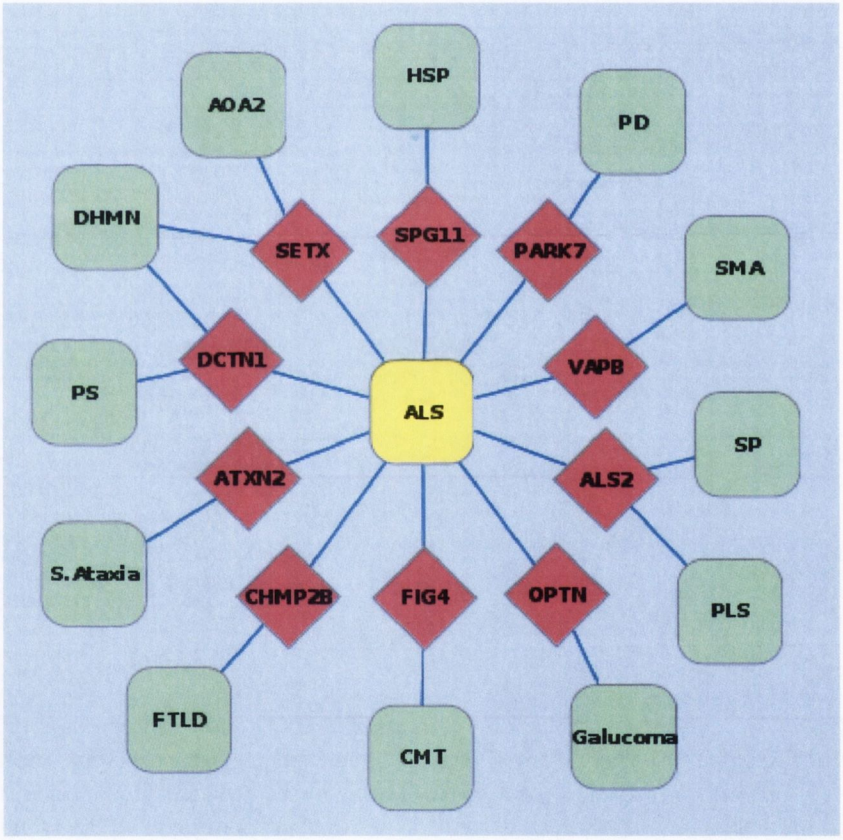
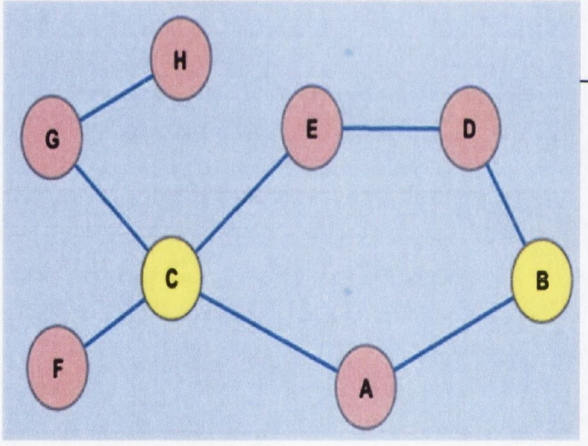


Figure 3.7: Pleiotropy network of ALS related disorders. Outlines conditions which can be linked to ALS by a common susceptibility gene. HSP Hereditary spastic paraplegia; PD Parkinson's disease; SMA Spinal muscular atrophy; SP Spastic paralysis; PLS Primary lateral sclerosis; CMT Charcot-Marie-Tooth disease; FTLD Frontotemporal dementia; S. Ataxia Spinocerebellar ataxia; PS Perry syndrome; DHMN Distal hereditary motor neuropathy, AOA2 Ataxia-oculomotor apraxia type 2. *ATXN2* was not known to represent an ALS gene at the time of the original analysis and is only included for illustrative purposes.

3.3.4 Evaluation of network based prioritization

Ranking by relative network importance was performed using a custom implementation of the PageRank-Priors algorithm (section 3.2.6). Figures 3.8 and 3.9 outline selected results from a series of validation analyses conducted using simple test networks. The mean ranking of vALS genes was 1428 positions lower (better) that of non-vALS genes (fold enrichment=1.3, $p=1.8x10^{-4}$). The median number of interaction partners among vALS proteins was slightly higher than that of the non-vALS proteins (Figure 3.4). The preferential ranking of vALS genes could therefore simply have reflected high global network importance as opposed to use of the tALS genes as seed nodes. To investigate this further, null rankings were computed by assessing importance with respect to 20,000 sets of 9 genes randomly selected from the network. This revealed that mean vALS gene ranking was significantly lower when tALS genes were used to train the analysis (mean fold enrichment=9.8, $p<5x10^{-5}$), confirming that the preferential ranking of vALS genes was at least partially a reflection of seed node selection. Mean ranking of the rvALS genes with respect to the tALS genes was 4843 positions lower than that of non-vALS genes (fold enrichment=5.7,$p=2x10^{-3}$) and once again permutation procedures confirmed that this preferable ranking reflected the use of the tALS gens as seed nodes as opposed to just high global network importance among the rvALS genes (mean fold enrichment=3.8,$p=1.7x10^{-2}$).



	Relative Importance	Rank
A	0.132902	1
D	0.114332	2
E	0.090132	3
G	0.066378	4
F	0.050116	5
H	0.023232	6

Figure 3.8: PageRank-Priors. Demonstrates the application of PageRank-Priors to a simple test network. Nodes "B" and "C" represent the seed nodes to which relative importance was assessed and the remaining nodes represent test nodes for prioritization. A random walk originating at a randomly selected seed node is more likely to traverse node "A" than any other node in the network. Node "A" is therefore assigned the highest relative importance and top ranking.

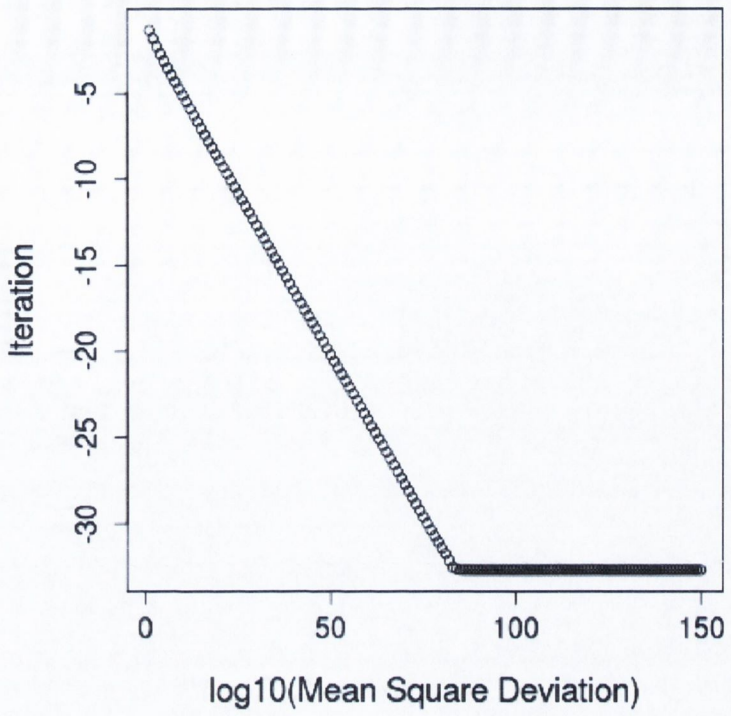


Figure 3.9: PageRank-Priors - Convergence. PageRank-Priors is an iterative algorithm. Plots of the mean square deviation in successive estimates of node importance were used to monitor algorithm performance and assess convergence. The above figure was generated during application of the algorithm to the network described in Figure 3.8. Stable importance estimates were reached in <100 iterations and the assigned rankings remained consistent from the 3rd iteration onwards.

3.3.5 Evaluation of gene annotation based prioritization

Gene annotation based prioritization was conducted using the ToppGene webtool (section 3.2.7). This represented a 2 step process which involved the generation of a weighted functional profile to represent the provided set of training genes and then test gene prioritization based on similarity to this profile. As anticipated, many of the most significantly over-represented terms within the tALS gene profile (Table S3.5) were ALS related. These included biological processes like "cell death", human phenotype annotations like "fasciculations" and "muscle weakness and atrophy", mouse phenotype terms like "neurodegeneration" and disease annotations like "amyotrophic lateral sclerosis". A large proportion of the most significantly over-represented literature citations were also ALS orientated. Other significantly enriched terms included numerous expected annotations like "superoxide dismutase activity" or "tubulin binding" and terms observed to have been over-presented among the tALSp1 genes, including RNA processing, viral infection, GTPase activity, protein localization and cellular transport. Numerous, additional and less expected enrichments were also observed including various processes related to ovulation and the development of primary female sex characteristics, and an assortment of drug interactions, shared micro RNA bindings sites and expression correlations.

The mean ranking of vALS genes was 8254 positions lower (better) than that of non-vALS genes when the whole genome was considered (n=26783 genes, fold enrichment=2.6, $p=2.3x10^{-25}$) and 1646 positions lower when only the test set of potential candidate genes was considered (n=5980

genes, fold enrichment=2.2, $p=1.8x10^{-10}$). The rvALS genes also tended to achieve lower rankings than non-vALS genes, though the observed differences did not reach conventional thresholds for statistical significance (genome: fold enrichment=1.5, p=0.08; test set: fold enrichment=1.5, p=0.14).

3.4 Discussion

In this chapter extensive *in silico* functional profiling of 9 known ALS genes was performed in an effort to identify features that might account for their capacity to influence disease susceptibility. Observations that the proteins encoded by disease associated genes often occur within disease related ubiquitinated neuronal inclusions raises the potential that the underlying biology of ALS genes may not actually be important and what is more important is whether the gene product can be destabilized to become aggregation prone. Regardless of whether or not this is the case, our results strongly suggest that ALS related genes can still be distinguished by their normal biological properties. Specifically we observed a tendency for ALS genes to occur within close proximity in molecular interaction networks and to exhibit significant degrees of sequence homology. We also observed numerous distinguishing features in terms of general gene annotation. As anticipated, the ALS genes themselves, exhibited significant enrichments for terms relating to RNA processing, intracellular transport and protein homeostasis, themes which were also overrepresented among immediate interaction partners and genes with potential protein sequence homology. Numerous other potentially informative correlations were also observed including recurrent enrichments for terms relating to viral infection, gene expression patterns and shared micro RNA binding sites (Tables S3.1-3.14).

One of the primary objectives of this section was to use potentially disease related gene features for the selection of promising candidate genes. To this end we first identified genes that encoded proteins which had been experimentally demonstrated to either interact or share an interaction partner with proteins encoded by known ALS genes. We then also identified genes exhibiting potential protein sequence homology with ALS genes and attempted to further enrich the combined set for ALS related genes through analyses of disease annotation, interaction network topology, sequence homology and general gene annotation. It is probable that performance of the analysis could have been improved by the incorporation of additional data, such as gene coexpression modules, or by refinement of the analyses used to interrogate this data. With regards the latter, one potentially major improvement which could be made would be to weight the importance of the gene features based on their predictive value in identifying training set genes. This was to an extent performed within each component of the analysis (network analysis, sequence analysis, annotation analysis) but was not performed when pooling information from across the analyses.

Despite the room for improvement the results of various evaluations analyses performed using sets of validation disease genes were highly encouraging and supported that each component of the candidate prioritization pipeline provided an enrichment for ALS related genes. However, these analyses have multiple limitations which could prove to be highly important. Firstly, the fact that a pattern has held true for previously discovered disease genes is no guarantee that it will hold true for undiscovered disease genes. Secondly the quality of the results is at best as good as the quality of the underlying data and the current knowledgebases of gene annotation and

molecular interactions are far from complete or error free. Finally, most of the enrichment analyses described in this section are subject to varying types of study bias and the potential impact of these can be extremely difficult to control. For instance, it would be anticipated that known disease genes are generally subjected to more extensive study than non-disease genes, providing increased opportunity for the observation of seemingly relevant gene features and thus favourably biasing validation analyses that use known disease genes. Furthermore, the discovery of a gene's involvement in a given disease is likely to bias it's apparent relationship with any other gene that has also been implicated in that disease, as such genes would be anticipated to exhibit a higher probability of being studied or discussed together. In an effort to assess the potential impact of such biases, the results from the original analyses were used to assess the expected disease relevance of 5 ALS genes not known to have been disease causing at the time. This gene set was small, limiting power to detect significant enrichments but nonetheless the observed results were in line with those obtained using the larger validation set. Furthermore it is worth noting that consistencies in various disease gene features, including molecular interaction network neighbourhood and protein sequence composition, have been reported across various global analyses of disease phenotypes [132-134]. Also, a recently published analysis of the ability of various candidate gene prioritization tools to select newly discovered disease genes, before the introduction of biases into the biomedical literature, showed some very promising results[135].

3.5 Conclusion

Through *in silico* functional profiling we have identified numerous features of known ALS genes that may be disease related and used these features to guide the selection of biologically plausible candidate disease genes. The results from initial validation analyses are encouraging, suggesting that the enrichment strategies do preferentially select for genes that are ALS related. However, these validation analyses are subject to certain limitations and more definitive confirmation stands to be gained from a genetic follow up.

4. TARGETED RESEQUENCING

4.1 Introduction

The development of second generation DNA sequencers and customizable high-throughput target enrichment systems has made it feasible to screen large sample cohorts for variation across large amounts of discontinuous genomic sequence.

In this section I describe the design and implementation of a multiplexed targeted resequencing strategy[136-138], for the analysis of 33 known and 718 candidate ALS genes. Issues explored include the final selection of candidate genes prioritized in section 3, target enrichment kit design, target enrichment efficiency, off-target sequencing, sequence alignment, alignment quality control and sequencing-SNP array genotype concordance rates. Extended variant calling and all subsequent analyses related to clinical parameters are described in sections 5-7.

4.2 Materials and Methods

4.2.1 Target Enrichment Kit Design

Genomic coordinates for the protein-coding exons of all RefSeq transcripts pertaining to 33 known (section 4.3.1) and 718 candidate ALS genes (Table S4.1) were downloaded for the hg19 build of the human genome from the UCSC genome browser[139]. Intervals were extended by 2bp so as to include consensus splice sites and processed to a format recognizable by Agilent's (Santa Clara, CA) e-array tool (https://earray.chem.agilent.com/earray/). E-array was used to design 120bp RNA bait probes complementary to the sequence of target intervals. Design parameters included a tiling frequency of 2X and a maximal overlap of 20bp with sequences masked by RepeatMasker(http://www.repeatmasker.org). Small target intervals were extended prior to upload such that all target bases would be covered by at least 2 probes that overlapped by no more than 80%. Two sequential rounds of probe design were performed, with intervals missed during the first round being extended to facilitate the design of probes which would not excessively overlap repetitive sequences. Finalized bait sequences were submitted to Agilent for incorporation into a customized SureSelect target enrichment kit.

4.2.2 Study participants

Two sample cohorts were considered in this study. The first included 248 Irish ALS cases and 167 age and geographically matched controls. DNA from these individuals was analyzed using the enrichment kit described in section 4.2.1, hereafter referred to as "capture 1", and this cohort will be referred to as the "candidate genes" cohort. The second cohort represented an extension of the first and included 444 cases and 311 controls. DNA from these individuals was analyzed either using capture1, or a distinct SureSelect system, referred to hereafter as "capture2", designed to target the same set of known ALS genes but a distinct set of 1577 candidate disease genes. The candidate genes targeted using capture2 had been selected through homozygosity mapping, association mapping and CNV analyses but will not be discussed further in this work and the extended cohort will be referred to as the "known genes" cohort. All DNA samples were extracted from venous blood (KBioscience, Herts, England). All participating patients were recruited between 1999 and 2011 through the ALS Register of the Republic of Ireland or the ALS Register of

Northern Ireland[7]. All patients were of Irish ancestry and met the revised El Escorial criteria for possible, probable or definite ALS. Patients with an identifiable family history of ALS among 1st, 2nd or 3rd degree relatives were classified as fALS [140], otherwise patients were classified as sALS. Controls were neurologically normal at the time of blood donation and included the spouses of attending patients and volunteers recruited at primary care offices across the country. Informed written consent was obtained from all participants and the study was approved by the Research and Ethics committee in Beaumont Hospital, Dublin.

4.2.3 Targeted resequencing

Preparation of genomic sequencing libraries (Figure 4.1) and subsequent target enrichment was performed as per the "SureSelect Target Enrichment System for Illumina Paired-End Sequencing Library protocol version 2.3" with the following modifications. 1) Only 1 µg of DNA was used in library preparations. 2) DNA fragmentation was performed either using a Covaris system (Covaris, Woburn, MA) as per the manufacturer's instructions, enzymatically or through sonication. Where enzymatic fragmentation was performed, 15.8 µl of DNA (63.3 ng/µl) was incubated with 2 µl NEBNext dsDNA fragmentase (New England Biolabs, Ipswich, MA) in NEBNext fragmentase reaction buffer @ 30 °C for 20 min. 5 µl of 0.5 M EDTA was then used to stop the reaction. Where fragmentation was performed through sonication, 100 µl of DNA (10 ng/µl) was subjected to 60 min of sonication (30 sec on, 30 sec off) using a Bioruptor (Diagenode, Liège, Belgium) at low power. 3) A s described by [137], extended Illumina adapters including 6 bp indexes were used in place of standard adapters (Metabion, Martinsried, Germany). All indexes were designed to be separable by an edit distance $\geq$2. 4) Following adapter ligation, libraries were subjected to electrophoresis through a 2% Low Range Ultra agarose gel (BioRad, Hercules, CA) stained with SYBR Green 1 nucleic acid gel stain (Sigma, Arklow, Ireland) for 90 mins @ 140 V. Fragments of 300 – 400 bp were manually excised from gel, using a Safe Imager™ 2.0 Blue Light Transilluminator (Invitrogen, Carlsbad, CA) for visualization. DNA was extracted from gel slices using QIAquick Gel Extraction Kits (Qiagen, West Sussex, UK) as per the manufacturer's instructions. 5) 5 cycles of PCR were performed prior to target enrichment, and Phusion (New England Biolabs) was used for all PCR reactions.6) Libraries were subjected to target enrichment in pools as opposed to individually[136]. Pools included libraries for 10-48 samples, all of which had been labelled with distinct adapter indexes. Libraries for 97 cases from the candidate genes cohort and 203 cases from the known genes cohort were resequenced by 80 bp paired-end resequencing on an Illumina Genome Analyzer II at TrinSeq, Institute of Molecular Medicine, Trinity College Dublin, Ireland. All other libraries were resequenced by 101 bp single-end sequencing on a HiSeq 2000 at GATC Biotech AG, Konstanz, Switzerland.

4.2.4 Sequence alignment

Raw sequencing reads were received in fastq format (Figure 4.3). Custom python scripts were used to annotate reads with sample of origin based on detected adapter indexes and to convert base quality (BQ) scores to Sanger encoding[141]. Processed reads were aligned to an extended version of the GRCh37 build of the human genome ("http://www.broadinstitute.org/gsa/wiki/index.php/GSA_FTP_Server") using the Burrows-Wheeler aligner (BWA) version 0.6.1[142]. SAMtools version 0.1.18[143] was used to sort and index alignment files. Picard version 1.60 (http://picard.sourceforge.net/) was used to merge alignment

files and remove potential PCR duplicates based on alignment start and end positions. The genome analysis toolkit (GATK) version 2.1-2[144] was used to perform local sequence realignment around known and putative indel sites.

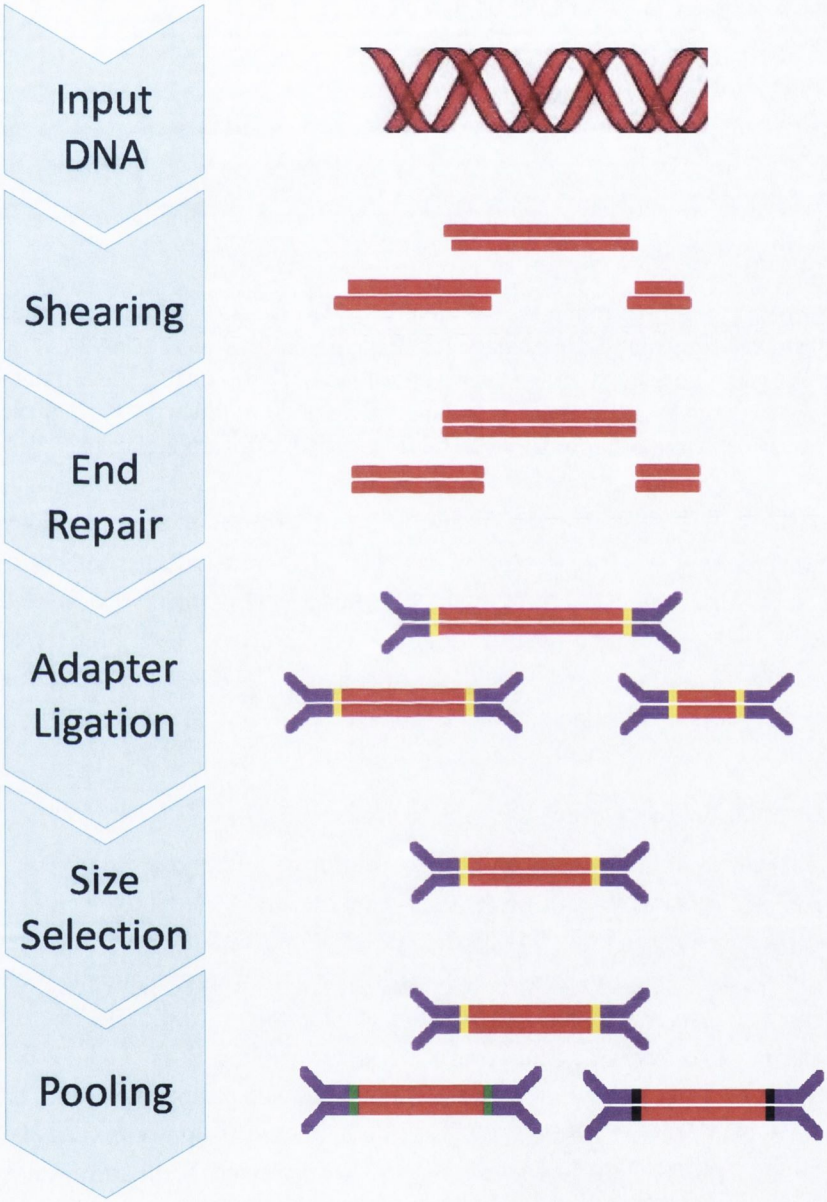


Figure 4.1: Library Preparation. Genomic DNA is sheared and the ends of the resulting fragments blunted in an end repair reaction. Following an a-tailing step, indexed sequencing adapters (yellow denotes the custom 6 bp index and purple the Illumina adapter sequence) are ligated to both ends of each fragment. A size selection step is then performed to retrieve fragments in the desired size range and remove residual adapter sequences. Libraries for multiple samples (labelled with different index sequences) are then pooled for target enrichment as described in Figure 1.3.

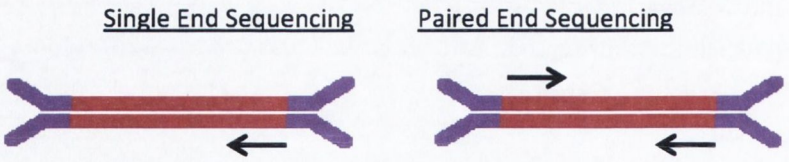


Figure 4.2: Single Vs Paired end sequencing. Under single end mode, only one strand of each read is sequenced. Under paired end mode both strands are sequenced. Black arrows indicate direction of sequencing.

@PSI179204_0014:6:1:1090:9001#0/1
AGAGATAAATTATTCTACAAAACTATAATATCCAAAACAGCCT
+PSI179204_0014:6:1:1090:9001#0/1
a\^c\bY`^bTbb^bccY\ac^cccc^T^cL\\MTb`^bb\\YY\ccb`a\LUMZT`V

Figure 4.3: Fastq format. Each sequencing read is represented by 4 lines. Lines 1 and 3 (blue) denote the read name. Line 2 denotes the assigned base calls. Line 3 denotes the associated base quality (BQ) scores. BQ scores represent Phred scaled (-log10[p]) estimates of the probabilities of associated base calls being erroneous and are encoded as ASCII characters. Adapter indexes utilized in this project represented the first 6 base calls (orange) of each read.

The GATK was also used to recalibrate BQ scores based on empirical sequencing error rates estimated from the frequencies of non-reference base calls at sites not associated with genetic variants represented in dbSNP build 132. Read group, quality score, dinucleotide and machine cycle were incorporated as covariates during the process. Recommended analysis parameters and reference files were used for all steps of sequence alignment and alignment processing and the entire process was implemented and managed using a single python script (Script S2).

4.2.5 Additional data processing and analysis

Quality control of lane index profiles was performed using custom python and R scripts. Depth of coverage calculations were performed using the GATK. Variant calling and genotype assignment was performed using the GATK's UnifiedGenotyper walker. Base calls were only considered for depth of coverage analyses, variant calling or genotype assignment if associated with BQ and alignment mapping quality (MQ) scores >=20, this corresponds to an expected sequencing/ alignment error rate of 0.01. Merging of SNP array datasets and conversion of genotypes encoded within binary PED files to alternate allele counts was performed using PLINK version 1.07[145]. For all analyses the alternate allele was considered to be the non-reference allele observed on the "+" strand. Custom R scripts were used to estimate sequencing-SNParray genotype concordance rates and sample identities. Variants and samples were excluded from concordance analyses if associated with missing data rates ≥80%. Sequence genotypes were reset to missing if associated with genotype quality (GQ) scores <15. Unless otherwise indicated, all data processing and analysis was conducted using python version 2.6.5 (http://www.python.org/), R version 2.14.1 (http://www.r-project.org/) or bash version 4.2.24 (http://www.gnu.org/software/bash/).

4.3 Results

4.3.1 Target enrichment kit design

Thirty three of the more well established ALS genes discussed in section 1 and 718 candidate ALS genes were selected for targeted resequencing. The known ALS genes included 10 implicated in Mendelian forms of the condition (*ALS2, ANG, C9orf72, FUS, OPTN, SETX, SOD1, TARDBP, VAPB, VCP*) and 23 associated with non-Mendelian or uncharacterized influences on disease risk (*ATXN2, CHMP2B, DCTN1, DPP6, ELP3, FGGY, FIG4, GRN, HFE, IFNK, ITPR2, MAPT, MOB3B, NEFH, NIPA1, PARK7, PON1, PON2, PON3, PRPH, SIGMAR1, SPG11, UNC13A*). *ATXN2* was not known to represent an ALS gene at the time of enrichment kit design and had actually been selected as a candidate gene. It was therefore only analyzed among the candidate genes cohort as opposed to the full known genes cohort. Nonetheless, for the purposes of all subsequent analyses, *ATXN2* was regarded as part of the known genes set. The selected set of candidate genes (Table S4.1) included those achieving the highest prioritization scores in section 3, following the removal of genes which had already been included among the known ALS genes set. The genes *TTN* and *SYNE1* were originally selected for analysis but subsequently dropped due to their size. The absolute number of candidate genes was determined by minimum requirements for the size of the cohort to be investigated and the allocated project budget. In all, targeted intervals amounted to 2.02Mb of unique sequence, of which 1.77Mb was annotated as protein-coding.

4.3.2 Pilot study

Before proceeding to the main project, a pilot study was conducted to validate the performance of the target enrichment system and determine an optimal multiplexing strategy. Based on equation 4.1 and results reported by a related study [136], 10 samples were pooled for the initial analysis. Approximately 58 million sequencing reads were generated in the pilot analysis. Ninety nine percent of these began with one of the 10 expected index sequences and sample representation was reasonably uniform (Figure 4.3). The first 6bp of reads that did not contain an expected index generally differed from an expected index by a single nucleotide.

$$n = \frac{(R)(RL)}{(T)(X)}$$

Equation 4.1

n denotes the number of samples to pool, R denotes the expected number of reads mapping to target intervals following the removal of PCR duplicates, RL denotes read length, T denotes the amount of target sequence and X denotes the intended mean sample coverage.

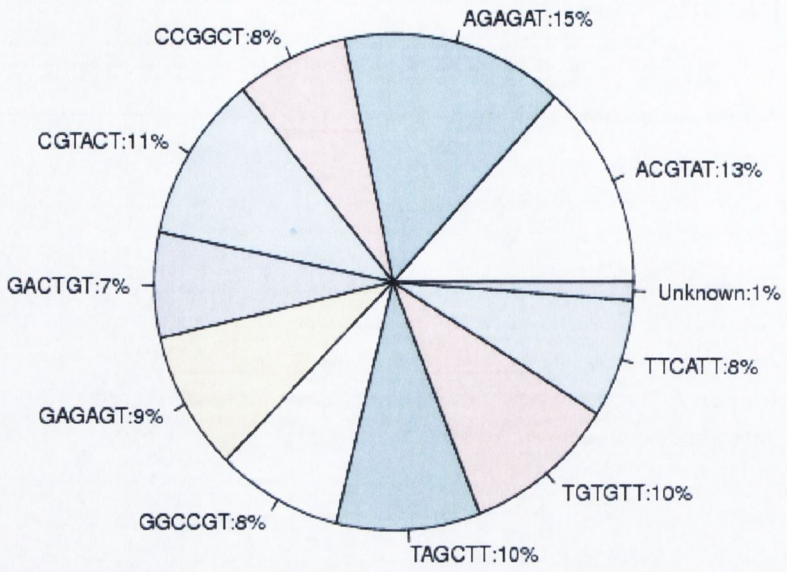


Figure 4.3: Relative barcode representation. Highlights the percentage of reads labelled with each or none of the expected index sequences.

Ninety seven percent of reads were alignable to the reference genome, with 91% of these remaining after the removal of probable PCR duplicates. Only, 16% of resequenced bases mapped to target exons while ~31% mapped to target exons ± 250bp (Figure 4.4). Analysis of "off-target" alignments revealed an over-representation of reads mapping to the mitochondrial genome and that ~13% of off-target nuclear alignments related to regions achieving relatively high coverage. The remaining reads appeared to be distributed relatively uniformly. Further analysis of the off-target intervals associated with high coverage revealed that they generally corresponded to segmental duplications or other forms of repetitive elements. Median sample coverage across target exons was 26X (IQR 4-50X). 72% of sample bases were covered to ≥7X (Figure 4.5), meaning that in 72% of cases, the probability of having sampled both of an individual's chromosomes during resequencing was >0.99. Significant deviations were noted between empirical estimates of sequencing error rates and those which would have been predicted from reported BQ scores. However, the issue appeared to be adequately resolved through the recalibration of BQ scores with respect to various error related covariates (section 4.2.4, Figure 4.6).

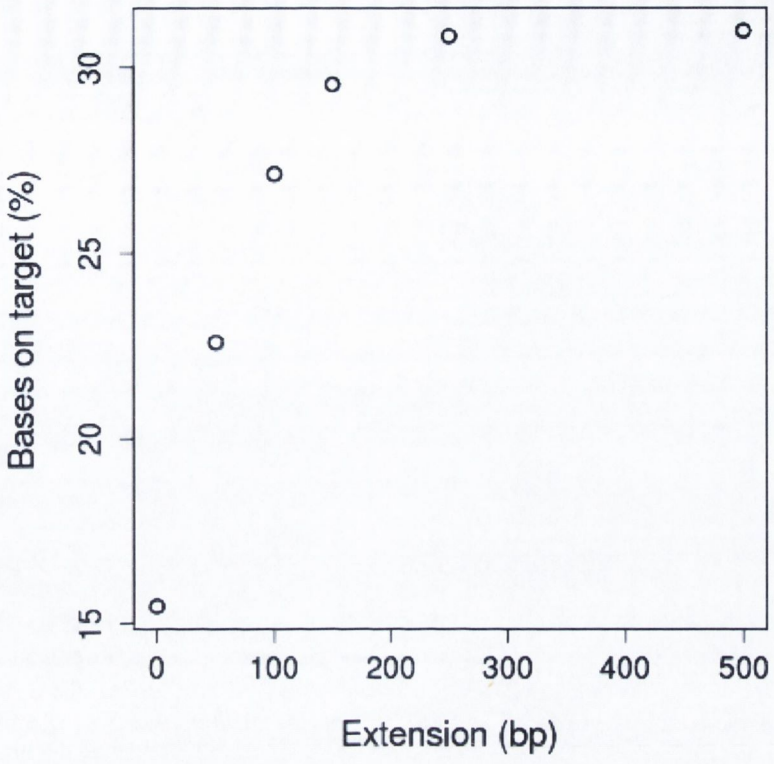


Figure 4.4: Efficiency of target enrichment. Outlines the percentage of aligned bases regarded as on-target with various extensions of target exons.

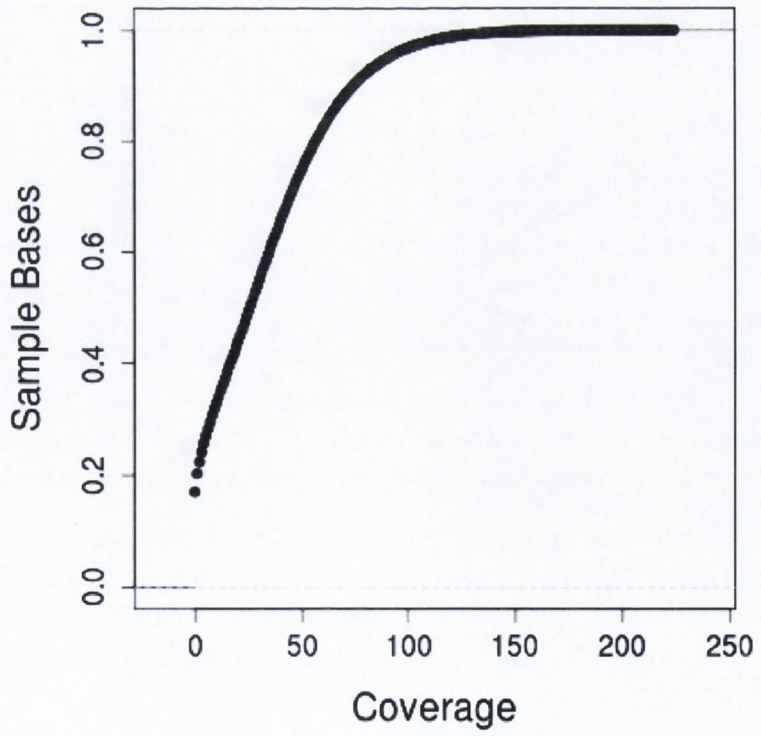


Figure 4.5: Sequence coverage across target intervals. Empirical cumulative density plot outlining the proportion of sample bases covered by a given number of sequencing reads.

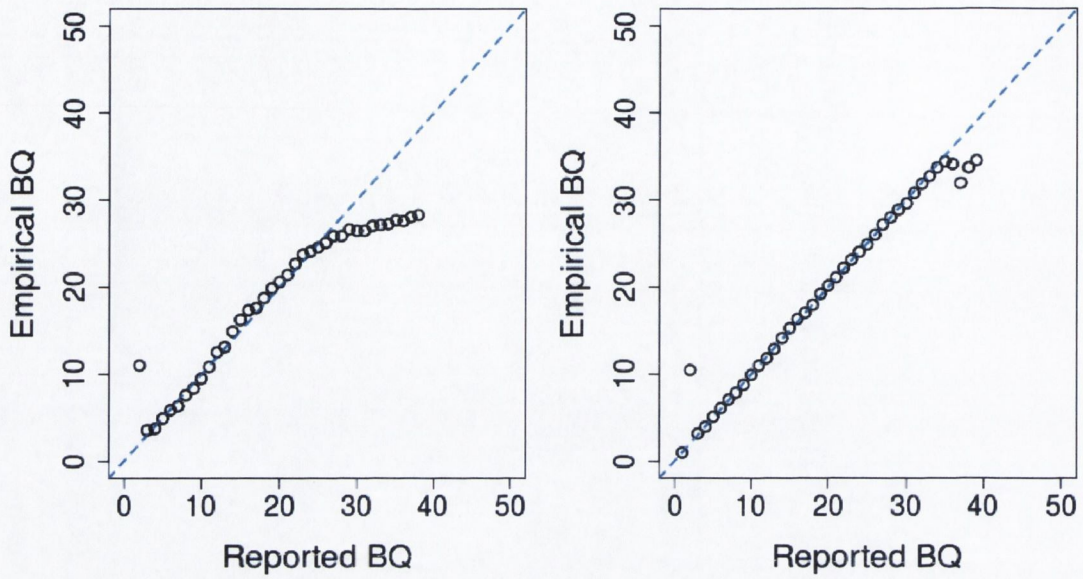


Figure 4.6: Base Quality Score Recalibration. Quantile-quantile plots (QQplots) outlining the relative distributions of empirical BQ score estimates vs. those generated during sequencing. The top and bottom panels describe the data before and after a BQ recalibration step. The underestimation of sequence quality at sites associated with a BQ value of 2 represents an expected artefact.

4.3.3 Extended study

Following the pilot study, the known ALS genes were resequenced across a total of 444 cases and 311 controls. Fifty eight percent of cases were male, 33% presented with bulbar onset disease, 62% presented with spinal onset, mean age of disease onset was 61±12 yrs and 11% were known to have a first, second or third degree relative that had also been afflicted with ALS and were hence classified as fALS. Cognitive evaluation [5] had been performed for 157 cases (35%), 78 (50%) of whom exhibited signs of cognitive impairment. These individuals included 34 fulfilling criteria for executive dysfunction, 21 fulfilling criteria for a diagnosis of ALS FTD, 21 fulfilling criteria for non-executive cognitive impairment and 2 with a diagnosis of Alzheimer's disease. Forty seven percent of the control cohort were male and mean age of sampling was 61±14yrs. The candidate genes were only resequenced in a subset of the known genes cohort that included 248 cases and 167 controls. The demographics of this reduced case cohort were as follows; 59% male, 36% bulbar onset, 57% spinal onset, mean age of onset = 62 ± 11.6 yrs, 15% fALS. Cognitive evaluation was performed for 115 cases (46%), 48 of whom (42%) exhibited signs of cognitive impairment. Forty five percent of the reduced control cohort were male and the mean age of sampling was 61±13.5yrs.

Based on the sequence alignment, PCR duplicate and on-target rates estimated in the pilot study, ~30 samples were pooled for each capture reaction for sequencing conducted with TrinSeq, while 48 samples were pooled per capture reaction for sequencing conducted with GATC. As a result, median target coverage was 29X per sample with 80% of sample bases being covered to >=7X. Analysis of the index sequences detected within each sequencing run, revealed a mix up of 2 lanes ("6_1","6_2") during a second round of sequencing (Figure 4.7). The error was corrected and no further anomalies regarding index composition were observed.

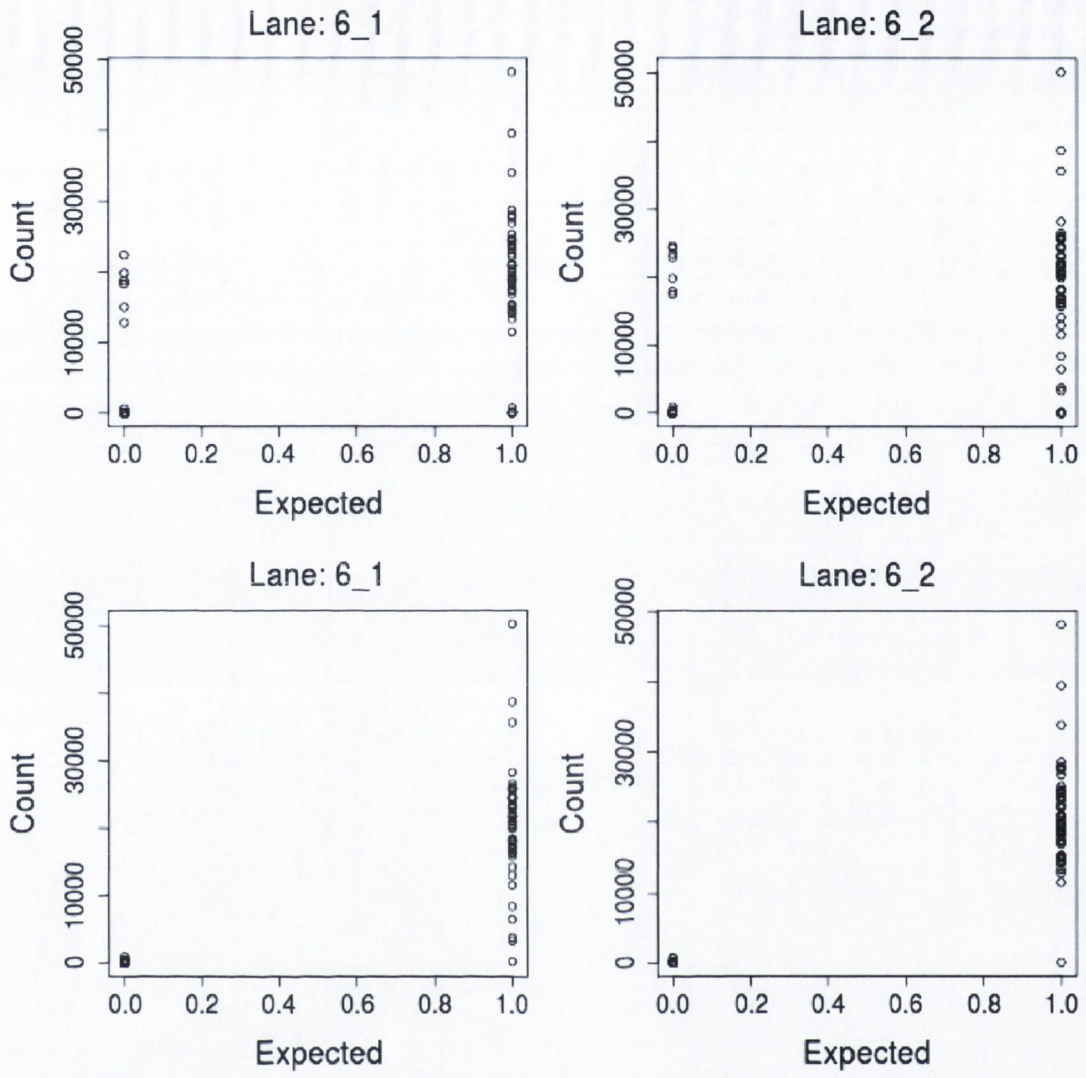


Figure 4.7: Index profiles. Plots outline the counts of various sequences found to constitute the first 6bp of reads generated within 2 distinct sequencer lanes (6_1 and 6_2). Sequences corresponding to expected indexes are assigned an "Expected" value of 1, while unexpected sequences are assigned a value of 0. High frequencies of unexpected sequences were originally observed in lanes 6_1 and 6_2 (top panel). Analysis of the index profiles revealed that the 2 lanes had been inadvertently switched by the sequence provider and reassignment of lane labels resolved the issue (lower panel).

4.3.4 SNP array genotype concordance

Five hundred and sixty seven samples from the known genes cohort (section 4.2.2) had been genotyped using either Illumina HumanHap 550[42], Illumina Human610-Quad (deCODE Genetics, Reykjavik, Iceland) or Illumina OmniExpressExome-8v1 (Atlas Biolabs, Berlin, Germany) SNP BeadChip assays. The overlap between a combined SNP array datasets and the resequenced exons of known ALS genes (±250bp) was 167 variants. To screen for potential sample mix ups, the sequencing genotypes across these sites were compared with the corresponding genotypes in the entirety of the combined SNP array dataset. As expected, each individual exhibited high concordance with themselves but low-moderate concordance with other individuals (Figure 4.8, mean=55%±8.9%). This facilitated the identification of a clear mix up in the barcodes assigned to 2

individuals analyzed in the same capture reaction. Following correction of this error, the median concordance rate between an individual's sequencing and SNParray genotype calls was 1 (IQR=1-1, Figure 4.9). Two individuals were observed to exhibit low concordance rates, suggesting that further sample mix ups were likely to have occurred within either the sequencing or SNParray datasets but that such mix ups were relatively rare. Analysis of concordance rates across SNPs, revealed that one locus exhibited a notably poorer concordance rate than the others. Tests for deviations from HWE indicated that the problem was an excess in homozygous calls within the sequencing genotypes (observed genotype counts (AA,Aa,aa)=169,198,178, expected=132,272, 141, X^2=40.7, p=1.8x10^{-10}), and further analysis revealed this error related to the single but not paired end data (Figure 4.9).

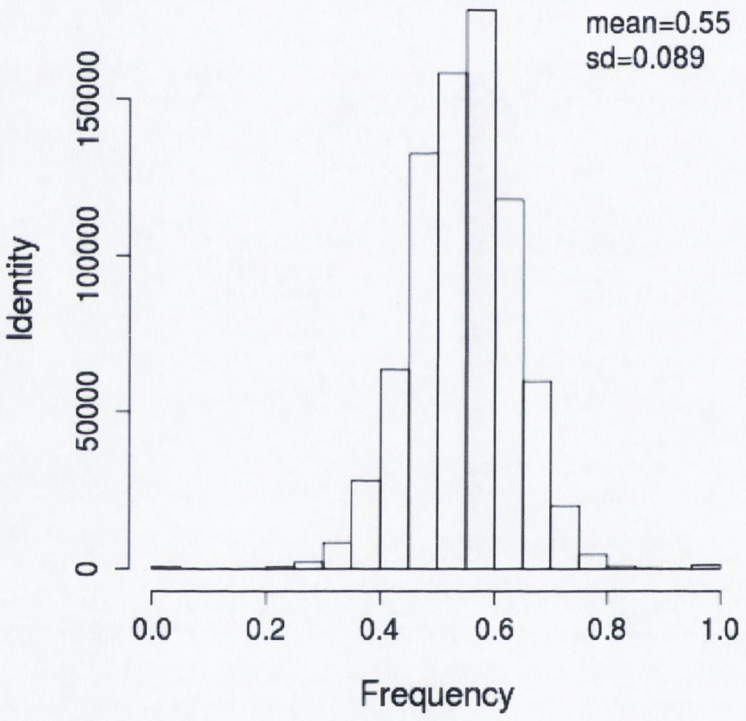


Figure 4.8: Sample identity. Shows the distribution of identity between the sequencing genotype profiles of each sample and the SNParray genotype profiles of all other samples. Instances where identity is ~=1 correspond to situations where the sequencing and SNParray genotype profiles being compared related to the same individual.

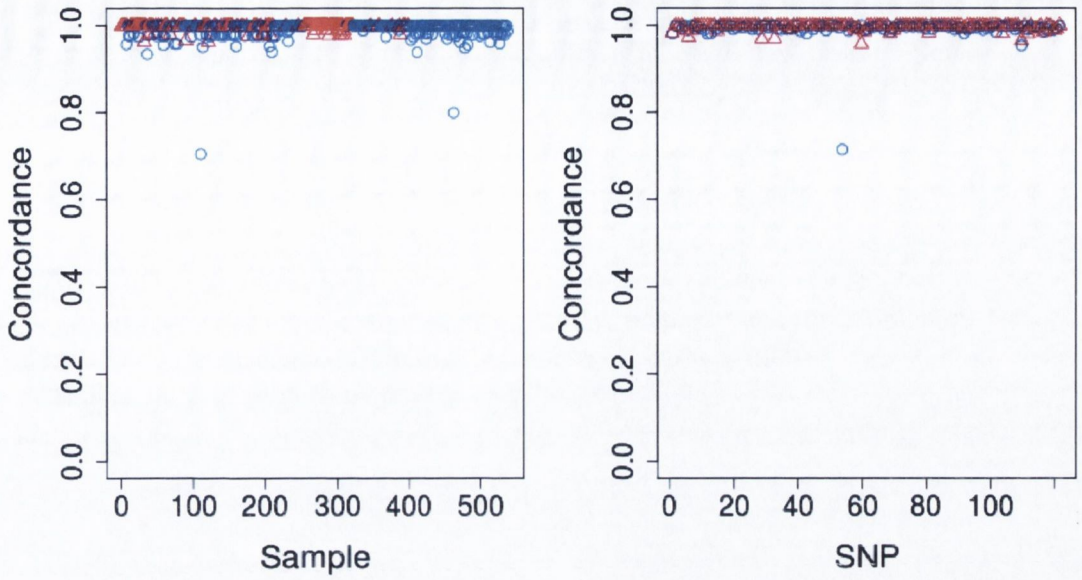


Figure 4.9: Sequencing – SNParray genotype concordance rates. Outlines the rate of concordance between sequencing and SNParray genotype calls for the same individual. The figure on the left shows concordance per sample while the figure on the right shows concordance per site. Red triangles denote concordance rates for individuals resequenced at TrinSeq (paired-end) while blue circles denote rates for individuals resequenced at GATC (single-end).

4.4 Discussion

Targeted resequencing of 33 known ALS genes was performed for a population based cohort of 444 Irish ALS cases and 311 age and geographically matched controls. Two hundred and sixty eight of these cases and 167 controls were also resequenced across a panel of 718 candidate disease genes. The efficiency of target enrichment was relatively poor but in-line with that reported by similar studies[136]. Considerably higher rates would be expected with commercially available exome enrichment kits despite that these utilize the same underlying technology[138]. This most likely reflects the ongoing optimization of the exome enrichment kits and suggest that improved enrichment for target intervals may be achievable through revised probe design. For the most part, off-target sequences appeared to be distributed across the genome relatively uniformly. The observation of an over-representation of mitochondrial reads, presumably reflects the high cellular copy number of mitochondrial DNA with respect to the nuclear genome[146] and apparent peaks of high coverage within the nuclear genome are anticipated to generally have reflected alignment artefacts relating to repetitive elements. There were no unexpected issues with the sample multiplexing strategy employed and the vast majority of sequenced reads could be mapped back to the sample of origin in all cases.

Analyses of empirical estimates of sequencing error rates suggested some systematic inaccuracies in the BQ scores assigned during sequencing. However, this appeared to be largely correctable through recalibration of BQ scores with respect to various sequencing related covariates. Analyses of lane index sequences and sequence-SNParray genotype concordance rates revealed some

further anomalies pertaining to specific samples or genomic sites that were once again predominantly correctable or at least identifiable without the use of reference data.

4.5 Conclusion

Extensive targeted resequencing was conducted for 444 ALS cases and 311 controls using customized enrichment systems from Agilent, the sample multiplexing strategy described by Kenny *et al*[137] and Craig *et al*[136] and Illumina high-throughput sequencing. Detailed analysis and quality control of the resulting dataset revealed numerous surmountable anomalies but by in large suggested the data was of high quality. Further exploration of the generated data in context of clinically relevant parameters is described in sections 5-7.

5. ANALYSIS OF KNOWN ALS GENES

Kenna *et al*, Journal of Medical Genetics, 2013

5.1 Introduction

A detailed understanding of ALS heterogeneity is important to facilitate appropriate stratification of subcohorts for various clinical research purposes. However, the relative importance of identified ALS genes is largely unknown and to date the most comprehensive studies have involved analyses of only 6 or 7 disease genes[147,148]. Accordingly, there is also relatively little known about how frequently patients might develop disease due to the presence of mutations across multiple genes, though a recent study from the Netherlands has highlighted that this is likely to be quite important[67]. There is also little known as to how the genetic aetiology of ALS varies across populations. Previous studies have suggested significant variability in the importance of specific disease genes[149,150] but these studies have involved comparisons of selected patient cohorts and no comparative study of population based cohorts has yet been performed.

In this section, I discuss results from analyses of the targeted resequencing data generated for known ALS genes (section 4). Theses analyses involve characterization of the relative and cumulative frequencies of known and candidate ALS variants, exploration of the significance of disease heterogeneity across populations, and screens for potential signatures of oligogenic disease inheritance.

5.2 Material and methods

5.2.1 Study participants and targeted resequencing

Details of the study cohort, targeted resequencing, sequence alignment and alignment quality control are described in section 4. Base calls were only considered in downstream analyses if the associated base quality (BQ) score was >= 20 and the alignment mapping quality (MQ) score was >=20.

5.2.2 Estimation of variant detection power

The depth of coverage walker provided in the GATK was used to calculate sequence coverage for each individual at each target position. This was then used to estimate the expected total of resequenced chromosomes at each position. The probability of including a mutated chromosome among the expected total was then calculated for a range of minor allele frequencies. Both the sampling of patients from the Irish ALS population and the sampling of chromosomes during resequencing were assumed to follow binomial distributions.

5.2.3 Variant calling, annotation and quality control

Variant calling was performed using the Unified Genotyper walker provided in the GATK. The effects of variants at the protein level were predicted using Variant Effect Predictor version 2.7 (http://www.ensembl.org/info/docs/variation/vep/index.html) and Ensembl 69 (http://www.ensembl.org/). Predictions were based only on transcripts annotated as protein coding and included PolyPhen and SIFT pathogenicity classifications. Genotype frequencies within external control cohorts were imported from the March 2012 release of the 1000 genomes

project (http://www.1000genomes.org/) and the ESP6500 release of the ESP. Variants were assigned a status of either "pass" or "fail" based on whether they met a series of quality control criteria. These criteria included a minimum variant quality score of 30, a genotype call rate ≥10%, and "QD", "FS", "ReadPosRankSum" and "HaplotypeScore" values within the range observed for the subset of variants noted also to have been reported by the 1000 genomes or ESP. Additionally, novel case variants observed to occur only *in cis* with other novel case variants were interpreted as possible misalignments and assigned a status of "fail". Variant quality score, "QD", "FS", "ReadPosRankSum" and "HaplotypeScore" values were calculated using the GATK. Genotype calls associated with genotype quality scores <15 were reset to missing prior to variant quality control.

5.2.4 Single variant analyses

Additive, dominant and recessive tests of variant association with case-control status were performed using Fisher's exact method. Multiple testing correction was performed by the Westfall and Young permutation method. Synonymous variants were not considered in the correction of p-values obtained at nonsynonymous or splice site variants. To avoid issues of population/ analysis stratification, tests were limited to the internal Irish control cohort. Tests were performed with and without sample filtering based on status for established disease mutations and a family history of ALS. As per section 2.2.4, one-tailed binomial tests were used to assess whether the frequencies of variant carriers within Irish (internal controls), European (internal controls, the ESP6500 – European American cohort, the 1000 genomes – European cohort) and global (internal controls, the full ESP6500 cohort, the full 1000 genomes cohort) control cohorts exceeded expectation under high penetrance disease models. The expected carrier frequencies were taken as the product of patient carrier frequencies and the respective population risks for ALS (Irish 1/290; European 1/397; Global 1/397)[6]. High penetrance disease models were rejected when the p-value associated with any 1 of the 3 cohorts was <0.05.

5.2.5 Gene burden analyses

Aggregate tests of association between nonsynonmous gene variants and case-control status were conducted through logistic regression of case-control status with respect to minor allele counts and using the "sequence kernel association test" (SKAT) described by Wu *et al*[151] *(detailed discussion in section 6.2.6 and 6.3.4)*.Tests were implemented using the R package "SKAT" version 0.91 (http://www.hsph.harvard.edu/skat/download/). For all analyses, variants were excluded in the event of a call rate<0.8 and both sample sequencing centre (TrinSeq Vs GATK, section 4) and sample capture (capture1 Vs capture2, section 4) were included as analysis covariates. The burden of putative disease variants across *ANG, C9orf72, FUS, OPTN, SOD1* and *TARDBP* among Irish and Italian ALS populations was compared firstly using the c-alpha tests where singleton variants were collapsed into a single per locus count[152] and secondly using allele count based Fisher exact tests. To avoid artificial inflation of population differences due to biases in missing genotype rates, c-alpha test permutations were performed at each variant site independently.

5.2.6 Analysis of variant co-occurrence

One tailed binomial tests were used to explore whether the frequencies of cases carrying multiple splice site/ nonsynonymous ALS gene variants exceeded chance expectation. These analyses were performed as described in van Blitterswijk *et al.*[67]. Briefly the expected frequency of mutation co-occurrences among cases was taken as per *Equation 5.1*.

$$exp = (x_{ALS}/n_{als}) * (x_{ctrl}/n_{ctrl})$$

Equation 5.1

exp denotes the expected frequency of cases carrying multiple mutations, x denotes the number of cases (ALS) or controls (ctrl) carrying a mutation and n denotes the number of cases (ALS) or controls (ctrl) sequenced.

This model assumes that cases have an elevated probability of carrying one ALS gene mutation (x_{ALS}/n_{als}) but that additional ALS gene variants should occur at the same background rate as observed in controls (x_{ctrl}/n_{ctrl}). The probability of the observed count of cases with multiple mutations was then calculated based on this expected frequency and assuming a binomial distribution. Variants were excluded from these analyses, if the frequency of carriers among European/ Global controls from the ESP6500 and 1000 genomes projects exceeded one of several potential critical values (Table 5.3 and 5.4).

The frequency with which carriers of previously reported ALS variants also carried variants of additional Mendelian disease genes (3/43) was used to estimate the probability of observing multiple Mendelian disease gene variants among all identified carriers of a given ALS variant (i.e. 33/33 *C9orf72* hexanucleotide expansion carriers OR 2/2 *TARDBP*:c.859G>A(p.[G287S]) carriers OR 2/2 *FUS*:c.1574C>T(p.[P525L]) carriers) assuming binomial sampling.

5.2.7 Statistical analyses

Unless otherwise indicated, all statistical analyses were conducted using custom scripts written for R version 2.14.1.

5.3 Results

5.3.1 Targeted resequencing

Study participants were screened for variants within the coding exons of 10 genes previously associated with Mendelian forms of ALS (*ALS2, ANG, C9orf72, FUS, OPTN, SETX, SOD1, TARDBP, VAPB, VCP*) and 23 low penetrance/ tentative ALS genes (*ATXN2, CHMP2B, DCTN1, DPP6, ELP3, FGGY, FIG4, GRN, HFE, IFNK, ITPR2, MAPT, MOB3B, NEFH, NIPA1, PARK7, PON1, PON2, PON3, PRPH, SIGMAR1, SPG11, UNC13A*). At the time of target gene selection the causative variant within the chromosome 9p linkage region had not been resolved and accordingly *MOB3B* and *IFNK* were included as tentative ALS genes[128]. In total, 26.3 x 10^6 sequencing reads mapping to target positions were generated. 99% of target bases were covered by at least 1 sequencing read and on average each target position was covered by a mean of 27.3 sequencing reads per sample (Table 5.1). Power to observe any variant with a patient allele frequency $\geq$ 0.5% was estimated to

range from 0–98.8% across target positions with a median of 98.7% (IQR: 98.3-98.8%, Figure 5.1). A more detailed account of the distribution of variant detection power across target genes can be found in Figure S5.1. 477 potential sequence variants were identified across target intervals (Table S5.1). 95 of these were designated as possible machine errors and are not considered further. To evaluate the accuracy of genotypes inferred across the remainder of sites, genotypes called for 567 samples across 85 variants (target intervals +/- 50 bp) were compared with genotypes previously ascertained using Illumina HumanHap 550[42], Illumina Human610-Quad (deCODE Genetics, Reykjavik, Iceland) and Illumina OmniExpressExome-8v1 (Atlas Biolabs, Berlin, Germany) SNP BeadChip assays. This comparison revealed a genotype concordance rate of 98.9% among sequencing and BeadChip calls following genotype quality control (Figure 5.1).

Table 5.1: Target gene coverage

Gene	Designation	% Sequenced (>=1 X coverage)	Mean Case Coverage
ALS2	Mendelian	100	35.6 + 25.4
ANG	Mendelian	100	30.4 + 22.2
ATXN2	Low penetrance/ tenative	87.1	21.4 + 26.7
C9orf72	Mendelian	100	36.7 + 27.1
CHMP2B	Low penetrance/ tenative	100	25.8 + 19.7
DCTN1	Low penetrance/ tenative	100	16.6 + 17.1
DPP6	Low penetrance/ tenative	98.8	20.6 + 22.2
ELP3	Low penetrance/ tenative	100	32.3 + 21.9
FGGY	Low penetrance/ tenative	100	30.8 + 20.1
FIG4	Low penetrance/ tenative	100	33.4 + 24.3
FUS	Mendelian	100	17.9 + 18.1
GRN	Low penetrance/ tenative	100	10.1 + 12.3
HFE	Low penetrance/ tenative	100	28.7 + 25.4
IFNK	Low penetrance/ tenative	100	38.8 + 21.6
ITPR2	Low penetrance/ tenative	100	31.6 + 21.3
MAPT	Low penetrance/ tenative	100	12.4 + 18.8
MOB3B	Low penetrance/ tenative	100	38.9 + 23.9
NEFH	Low penetrance/ tenative	97.7	11.4 + 13.5
NIPA1	Low penetrance/ tenative	96.1	22 + 20.7
OPTN	Mendelian	100	21.3 + 15.9
PARK7	Low penetrance/ tenative	100	26.7 + 18.1
PON1	Low penetrance/ tenative	100	37.1 + 24.7
PON2	Low penetrance/ tenative	99.9	34.7 + 25.3
PON3	Low penetrance/ tenative	100	32.7 + 22.2
PRPH	Low penetrance/ tenative	99.4	4.2 + 8.7
SETX	Mendelian	100	41.2 + 25.7
SIGMAR1	Low penetrance/ tenative	100	2.9 + 7.5
SOD1	Mendelian	100	29.7 + 24.6
SPG11	Low penetrance/ tenative	100	32.9 + 23.7
TARDBP	Mendelian	100	30.3 + 22.1
UNC13A	Low penetrance/ tenative	98.6	15.5 + 17.9
VAPB	Mendelian	100	31.8 + 22.3
VCP	Mendelian	99.2	34.1 + 23

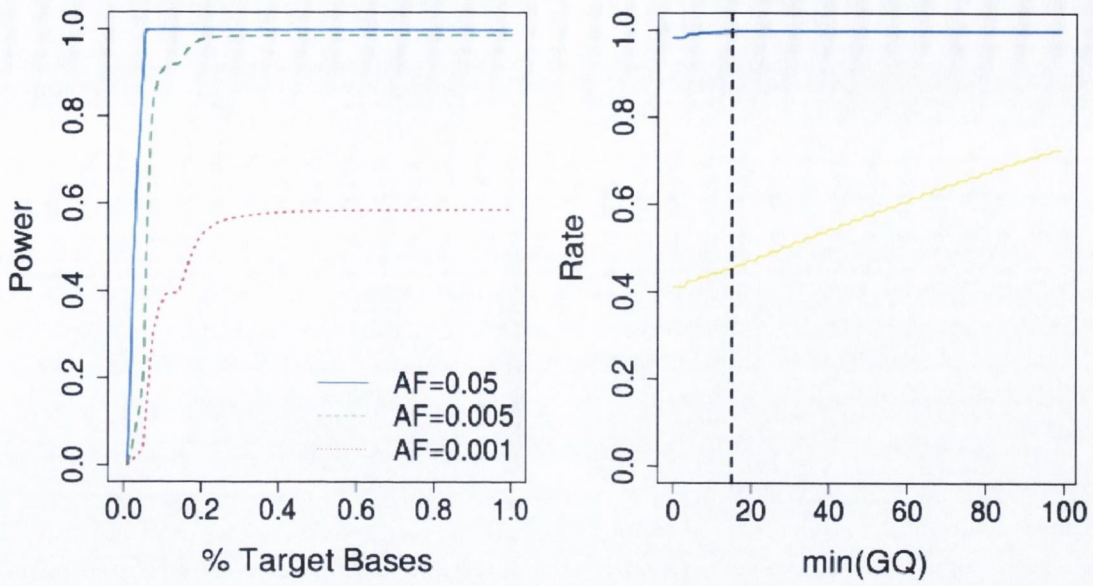


Figure 5.1: Target gene resequencing. The plot to the left represents a cumulative density plot of the distribution of power to observe variants with patient allele frequencies of 0.05, 0.005 and 0.001. Power estimates were derived from the distribution of sequence coverage across target positions (supplementary materials and methods). In the plot to the right, the X-axis denotes the minimum sequencing genotype quality score (GQ) accepted for inference of sample genotypes across 85 variant sites. The blue line denotes the relationship between minimum GQ and the rate of concordance among sample genotypes ascertained by resequencing and sample genotypes ascertained using BeadChip assays. The yellow line denotes the relationship between minimum GQ and missing genotype rate among the sequencing callset (n=567 samples). The vertical dotted line indicates the GQ threshold employed during this study.

5.3.2 ALS related variants

Following the removal of suspected sequencing artefacts and multiple testing correction, no variant was observed to exhibit a statistically significant association with ALS risk (Table S5.1). Surprisingly, aggregate analyses revealed strong evidence for protective effects among nonsyonymous variants of the gene *IFNK* (OR=0.15,=$3x10^{-4}$). As outlined above the original association between ALS and a locus including the *IFNK* gene is believed to be accounted for by the *C9orf72* hexanucleotide repeat expansion[23,24]. It could therefore simply be that the 2 *IFNK* variants driving the observed signal simply occur on a haplotypic backgrounds distinct from that of the expansion and do not actually exert a protective effect on ALS risk. Aggregate analyses did not reveal any other significant gene associations and by in large the frequencies of nonsynonymous gene variants was similar among cases and controls.

To estimate the potential contributions of these genes to Mendelian ALS, the binomial tests were used to screen for nonsynonymous/ splice site variants that could not be excluded as high penetrance ALS mutations. Thirty seven such variants were identified, 2 of which represented previously described ALS variants; *TARDBP*:c.859G>A(p.[G287S])[103,104,153]; *FUS*:c.1574C>T(p.[P525L])[17]. The remaining 35 included 1 splice site variant (*DCTN1*:c.2887-2A>G), 20 missense variants classified as "deleterious" by SIFT [154] or "possibly/ probably damaging" by PolyPhen[155] and 14 missense variants classified as "tolerated" by SIFT and "benign" by PolyPhen.. 76 patients (38.0% of fALS, 14.5% of sALS, 17.1% of combined) carried either one of

these candidate disease variantsor else a previously described ALS variant, with 57 (34.0% of fALS, 10.2% of sALS, 12.8% of combined) carrying those variants within fALS associated genes genes (Table 5.3. *C9orf72* 39;*SETX* 11; *ALS2* 7; *FUS* 2; *TARDBP* 2; *OPTN* 1; *VCP* 1. *ANG, SOD1, VAPB* 0), 21 (6.0% of fALS, 4.6% of sALS, 4.7% of combined) carrying those variants within low penetrance/ tentative ALS genes (Table 5.4. *SPG11* 7; *ELP3* 3; *CHMP2B* 2; *DCTN1* 2; *MAPT* 2; *DPP6* 1; *FGGY* 1; *HFE* 1; *ITPR2* 1; *PON2* 1; *UNC13A* 1. *ATXN2, FIG4, GRN, IFNK, MOB3B, NEFH, NIPA1, PARK7, PON1, PON3, PRPH, SIGMAR1* 0) and 43 (30.0% of fALS, 7.1% of sALS, 9.7% of combined) carrying a known ALS variant (Table 5.3. *C9orf72* 39; *FUS* 2; *TARDBP* 2).

Table5.2: Aggregate Analyses

Gene	Variants	CaseAC	ControlAC	OR	p.Burden	p.SKAT
ALS2	18	125	118	0.74	5.39E-01	3.69E-01
ANG	2	6	7	0.60	8.85E-01	7.62E-01
ATXN2	7	13	9	1.01	9.86E-02	1.00E+00
**C9orf72*	1	92	57	1.13	4.64E-01	4.64E-01
CHMP2B	2	2	0	Inf	1.94E-01	1.94E-01
DCTN1	10	34	18	1.32	2.96E-01	3.40E-01
DPP6	4	236	172	0.96	7.37E-01	6.99E-01
ELP3	5	8	3	1.87	6.19E-01	1.58E-01
FGGY	9	374	304	0.86	1.50E-01	1.08E-01
FIG4	5	156	122	0.90	7.94E-01	1.00E+00
FUS	2	2	1	1.40	1.93E-01	5.86E-01
GRN	2	3	5	0.42	5.45E-01	2.43E-01
HFE	8	266	200	0.93	4.45E-01	2.73E-01
IFNK	2	15	71	0.15	3.00E-04	3.00E-04
ITPR2	15	56	48	0.82	4.74E-01	1.88E-01
MAPT	12	749	596	0.88	6.88E-01	7.47E-01
MOB3B	1	0	1	0.00	1.00E+00	1.00E+00
NEFH	8	373	281	0.93	8.27E-01	2.09E-01
OPTN	4	238	250	0.67	1.33E-01	1.84E-01
PARK7	2	3	7	0.30	4.29E-02	2.17E-02
PON1	5	564	375	1.05	9.67E-01	7.86E-01
PON2	6	446	271	1.15	1.51E-01	1.38E-01
PON3	4	5	4	0.88	1.21E-01	1.00E+00
PRPH	2	4	3	0.93	1.00E+00	6.42E-01
SETX	35	972	770	0.88	3.70E-03	8.47E-01
SIGMAR1	2	1	3	0.23	1.00E+00	6.46E-01
SPG11	27	436	322	0.95	2.03E-01	1.85E-01
TARDBP	2	2	1	1.40	1.00E+00	1.00E+00
UNC13A	8	271	186	1.02	4.81E-02	1.57E-01
VAPB	1	4	1	2.80	3.28E-01	3.28E-01
VCP	1	1	0	Inf	1.00E+00	1.00E+00

*Variants refers to the number of nonsynonymous sites observed per gene following the removal of sites with low call rate (<0.8 of combined case control cohort). Case and control refer to total counts of alternate alleles observed across these sites among cases and controls respectively. OR refers to the ratio of the case mutation rate (combined alternate allele count/number of samples) to the control mutation rate. P.SKAT and p.Burden represent p-values from aggregate tests of association with case-control status. *C9orf72 analysis does not include the repeat expansion.*

Table 5.3: Candidate disease variants of Mendelian ALS genes

Gene	Variant	SIFT	PolyPhen	Model	fALS	sALS	IRL	EUR	GLO
*ALS2	c.3094C>T(p.[Arg1032Cys])	Deleterious	probably damaging	D	0/49	1/375	0/274	NR	NR
*ALS2	c.2606A>C(p.[Gln869Pro])	Deleterious	probably damaging	D	0/50	1/390	0/290	NR	NR
*ALS2	c.2566A>G(p.[Thr856Ala])	Tolerated	possibly damaging	D	0/47	1/373	0/253	NR	NR
*ALS2	c.2408A>G(p.[Lys803Arg])	Tolerated	possibly damaging	D	0/50	1/391	0/289	NR	NR
*ALS2	c.2098A>G(p.[Thr700Ala])	Tolerated	benign	D	0/50	2/386	0/279	NR	NR
C9orf72	**Repeat Expansion**	**-**	**-**	**D**	**15/45**	**24/367**	**0/188**	**NR**	**NR**
FUS	**c.1574C>T(p.[Pro525Leu])**	**Deleterious**	**possibly damaging**	**D**	**0/43**	**2/319**	**0/250**	**NR**	**NR**
OPTN	c.1192C>G(p.[Gln398Glu])	Tolerated	benign	D	0/50	1/368	0/243	NR	NR
SETX	c.7682C>T(p.[Ser2561Leu])	Tolerated	benign	D	0/50	2/358	0/253	1/4553	1/6756
SETX	c.7645G>A(p.[Val2549Ile])	Deleterious	benign	D	0/50	2/389	0/305	0/4605	1/6808
SETX	c.5842A>G(p.[Met1948Val])	Deleterious	probably damaging	D	0/50	1/388	0/279	NR	NR
SETX	c.5587A>G(p.[Thr1863Ala])	Deleterious	probably damaging	D	1/50	0/391	0/278	NR	NR
SETX	c.2975A>G(p.[Lys992Arg])	Deleterious	benign	R	1/50	0/388	0/280	NR	NR
SETX	c.2842C>A(p.[Pro948Thr])	Tolerated	benign	D	1/50	0/387	0/282	NR	NR
SETX	c.2755G>C(p.[Val919Leu])	Tolerated	benign	D	0/50	1/384	0/266	NR	NR
SETX	c.814C>G(p.[His272Asp])	Deleterious	probably damaging	D	0/50	1/387	0/280	NR	NR
TARDBP	**c.859G>A(p.[Gly287Ser])**	**Tolerated**	**benign**	**D**	**0/50**	**2/390**	**0/300**	**NR**	**NR**
VCP	c.2249A>G(p.[Asn750Ser])	Tolerated	benign	D	0/50	1/389	0/306	NR	NR

D Variant may cause disease in a dominant fashion and carrier frequencies relate to heterozygote carriers; R Variants may cause disease in a recessive fashion and carrier frequencies relate to homozygous carriers; NR Variant not reported within either the 1000 genomes or ESP6500 datasets (Samples of European ancestry=4679, Total number of samples =7595). Variants previously associated with ALS are shown in bold. Gene conventionally associated with recessively inherited ALS are marked with an asterix.

Table 5.4: Candidate disease variants of low penetrance/ tentative ALS genes

Gene	Variant	SIFT	PolyPhen	Model	fALS	sALS	IRL	EUR	GLO
CHMP2B	c.118A>G(p.[Lys40Glu])	deleterious	possibly damaging	D	0/50	1/373	0/261	NR	NR
CHMP2B	c.123G>T(p.[Gln41His])	deleterious	possibly damaging	D	0/50	1/373	0/263	NR	NR
DCTN1	c.2887-2A>G(p.?)	-	-	D	0/39	2/239	0/246	NR	NR
DPP6	c.883G>A(p.[Glu295Lys])	deleterious	probably damaging	D	0/50	1/390	0/296	NR	NR
ELP3	c.206G>T(p.[Arg69Leu])	deleterious	benign	D	0/50	1/381	0/294	NR	NR
ELP3	c.326G>A(p.[Cys109Tyr])	deleterious	probably damaging	D	1/50	1/390	0/278	1/4578	1/6781
FGGY	c.1716G>A(p.[Met572Ile])	tolerated	possibly damaging	D	0/48	1/376	0/285	NR	NR
HFE	c.766G>A(p.[Val256Ile])	tolerated	benign	D	0/50	1/388	0/295	NR	NR
ITPR2	c.3614G>A(p.[Arg1205Gln])	tolerated	probably damaging	D	0/50	1/371	0/266	NR	NR
MAPT	c.284C>T(p.[Thr95Met])	tolerated	benign	D	0/10	1/50	0/50	1/4349	1/6549
MAPT	c.698C>T(p.[Pro233Leu])	tolerated	probably damaging	D	0/14	1/79	0/80	NR	NR
PON2	c.661T>G(p.[Ser221Ala])	tolerated	benign	D	0/49	1/381	0/257	NR	NR
*SPG11	c.7324G>C(p.[Ala2442Pro])	deleterious	probably damaging	D	0/50	1/391	0/282	NR	NR
*SPG11	c.4343G>A(p.[Cys1448Tyr])	deleterious	possibly damaging	D	0/50	1/388	0/300	NR	NR
*SPG11	c.3680A>G(p.[Lys1227Arg])	tolerated	benign	D	1/50	0/390	0/302	NR	NR
*SPG11	c.2577A>C(p.[Gln859His])	tolerated	benign	D	0/50	1/380	0/267	NR	NR
*SPG11	c.1930A>T(p.[Thr644Ser])	tolerated	benign	D	1/49	0/382	0/265	NR	NR
*SPG11	c.1529G>A(p.[Ser510Asn])	tolerated	probably damaging	D	0/50	1/389	0/282	NR	NR
*SPG11	c.394A>G(p.[Ser132Gly])	tolerated	benign	D	0/50	1/391	0/279	NR	NR
UNC13A	c.3098T>A(p.[Val1033Asp])	tolerated	benign	D	0/49	1/372	0/286	NR	NR

5.3.3 Patient Phenotypes

Both patients carrying the *TARDBP*:c.859G>A(p.[G287S]) substitution were sporadic and presented with bulbar onset disease at 66/67 yrs of age. One patient remains alive at 51 months from disease onset, while the other died 49 months from disease onset. Both patients were cognitively intact. Both carriers of the *FUS*:c.1574C>T(p.[P525L]) substitution were also sporadic. They exhibited an exceptionally young age of onset (13/21 yrs) and rapid disease progression (disease duration:11-17 months). One patient experienced spinal onset disease while the other experienced bulbar onset. Both were cognitively intact. A detailed account of the phenotype exhibited by *C9orf72* repeat expansion carriers has been provided previously[156]. Further details on the phenotypes of all patients determined carrying known or candidate disease variants are listed inTable S5.2.

5.3.4 Co-occurrences of ALS gene variants

Analysis of the number cases carrying multiple rare or low frequency variants revealed no detectable excesses across either fALS associated genes alone or the entire dataset (Table 5.5, 5.6). 7 patients (4.0% of fALS, 1.3% of sALS, 1.6% of combined) carried 2 variants classified as known or candidate ALS variants in section 5.3.3 (Table S5.2). In the case of 4 of these individuals, both variants fell within fALS associated genes. Notably, these 4 individuals included both identified carriers of the *TARDBP*:c.859G>A(p.[G287S]))variant, who were observed to also carry either an *ALS2*:c.2566A>G(p.Thr856Ala) or *SETX*:c.814C>G(p.His272Asp) substitution. The probability of such an observation for all carriers of any previously reported ALS variant was estimated to be less than 1%, suggesting that these co-occurrences may be pathologically meaningful. Other observed co-occurrences of putative disease variants included *ALS2*:c.2098A>G(p.Thr700Ala) and *SETX*:c.7682C>T(p.Ser2561Leu), the *C9orf72* repeat expansion and *CHMP2B*:c.123G>T(p.Gln41His), the *C9orf72* repeat expansion and *SETX*:c.2842C>A(p.Pro948Thr), the *C9orf72* repeat expansion and *SPG11*:c.3680A>G(p.Lys1227Arg), the *DCTN1*:c.2887-2A>G splice acceptor site variant and *SPG11*:c.1529G>A(p.Ser510Asn). Further details on these patients can be found in Table S5.2.

Table 5.5: Co-occurrence of fALS associated gene variants

Frequency Threshold	Cases with 1 mutation	Cases with 2 mutations	Controls with 1 mutation	Controls with 2 mutations	p
0	55	3	12	0	0.36
0.0005	63	4	17	0	0.45
0.005	79	7	34	0	0.76
0.05	114	16	49	7	0.72

Frequency threshold denotes the cut-off carrier frequency for variants among European and Global reference cohorts. Variants exceeding a given threshold were excluded from analysis. Reference cohorts included individuals resequenced by NHLBI exome sequencing and 1000 genomes project only. Only nonsynonymous and splice variants were included in analyses. P-values denote probabilities that the

observed frequency of cases carrying multiple Mendelian ALS gene variants is in excess of chance expectation (section 5.2.5)

Table 5.6: Co-occurrence of ALS associated gene variant

Frequency Threshold	Cases with 1 variant	Cases with 2 variants	Controls with 1 variant	Controls with 2 variants	p
0	77	6	34	4	0.85
0.0005	103	9	50	7	0.99
0.005	151	30	95	13	1.00
0.05	247	90	160	72	1.00

Frequency threshold, denotes the cut-off carrier frequency for variants among European and Global reference cohorts. Variants exceeding a given threshold were excluded from analysis. Reference cohorts included individuals resequenced by NHLBI exome sequencing and 1000 genomes project only. Only nonsynonymous and splice variants were included in analyses. P-values denote probability that the observed frequency of cases carrying multiple ALS gene variants is in excess of chance expectation (section 5.2.5)

5.3.5 Variation in the frequency of ALS variants across populations

To assess the potential importance of genetic heterogeneity across populations, we compared the estimated frequencies of *ANG, C9orf72, FUS, OPTN, SOD1* and *TARDBP* disease variants among Irish patients with those reported by population based studies of Italian cohorts[147,148]. The difference was statistically significant (combined $p=1.7x10^{-4}$, Table 5.7), supporting a correlation between genetic susceptibility and population of origin. Of the 32 variants analyzed, only the *C9orf72* repeat expansion and the *FUS*:c.1574C>T(p.[P525L]) substitution were observed among both Irish and Italian patients. The *C9orf72* expansion was significantly more common among Irish patients (8.78% vs. 4.39%, $p=3.95x10^{-4}$) while *SOD1* and *TARDBP* variants were significantly more common among Italian patients (*SOD1*: 2.00% vs. 0.00%, $p=3.8x10^{-3}$. *TARDBP*: 2.00% vs. 0.45%, p=0.035). The overall frequencies of *FUS* and *OPTN* variants were similar (FUS: 0.30% vs. 0.45%, p=0.61. *OPTN*: 0.20% vs. 0.23%, p=1). *ANG* variants were identified only among Italian patients but the frequency difference was not significant (0.30% vs. 0.00%, p=0.56).

Table 5.7: The frequencies of ALS variants are population specific

Gene	Variant	% Italian (n=1003)	% Irish (n=444)	p
ANG	c.232A>G(p.K54E)	0.1	0	0.56**
	c.338G>A(p.W89X)	0.1	0	
	c.433C>T(p.R121C)	0.1	0	
C9orf72	Repeat Expansion	4.39	8.78	$3.95x10^{-4}$**
FUS	c.1542G>C(p.R514S)	0.1	0	0.18*, 0.61**
	c.1562G>T(p.R521L)	0.1	0	
	c.1574C>T(p.P525L)	0.1	0.45	
OPTN	c.1192C>G(p.GLn398Glu)	0	0.23	1*, 1**
	c.1499T>C(p.L500P)	0.1	0	
	c.1703T>C(p.L568S)	0.1	0	
SOD1	c.34G>T(p.D11Y)	0.3	0	$3.7x10^{-3}$**
	c.59A>G(p.N19S)	0.1	0	
	c.63C>G(p.F20L)	0.1	0	
	c.115C>G(p.L38V)	0.1	0	
	c.142G>T(p.V47F)	0.1	0	
	c.203T>C(p.L67P)	0.1	0	
	c.256G>A(p.G85S)	0.1	0	
	c.271G>A(p.D90N)	0.1	0	
	c.272A>C(p.D90A)	0.3	0	
	c.281G>A(p.G93D)	0.3	0	
	c.328G>T(p.D109Y)	0.2	0	
	c.400_402delGAA(p.E133del)	0.1	0	
	c.435G>C(p.L144F)	0.1	0	
TARDBP	c.800A>G(p.N267S)	0.2	0	0.026*, 0.035**
	c.859G>A(p.Gly287Ser)	0	0.45	
	c.881G>T(p.G294V)	0.2	0	
	c.909A>C(p.Q303H)	0.1	0	
	c.1009A>G(p.M337V)	0.1	0	
	c.1102G>A(p.G368S)	0.1	0	
	c.1144G>A(p.A382T)	0.9	0	
	c.1147A>G(p.I383V)	0.1	0	

	c.1169A>G(p.N390S)	0.3	0	
Combined	-	9.07	9.91	$1.7x10^{-4}$*, 0.25**

** Comparison of independent variant frequencies using a c-alpha test. ** Comparison of combined variant frequencies using a Fisher's exact test. Analyzed variants include the C9orf72 hexanucleotide repeat expansion and nonsynonymous variants reported as patient specific. Analysis of TARDBP was restricted to exon 6 while analysis of FUS was restricted to exons 14 and 15 and aanalysis of OPTN was restricted to exons 5, 9, 12 and 14* [147]

5.4 Discussion

The results described in this section represent the first reported application of multiplexed high-throughput sequencing in ALS research and more importantly the most extensive survey of relative ALS gene variant frequencies to date. The generated dataset was found to exhibit high sensitivity in terms of predicted power to identify rare patient variants and high accuracy in terms of genotype assignment.

With the exception of the *C9orf72* hexanucleotide expansion, no individual variant was observed to occur significantly more often among cases than controls. Furthermore, aggregate analyses revealed that, by in large, the overall frequencies of nonsynonymous gene variants were relatively similar across cohorts. The study was therefore not sufficiently powered to generate detailed estimates of relative gene/ variant effect sizes but does further support the argument made in section 1 that functional variants of these genes should not be assumed to have pathogenic relevance. Just over 17% of the patient cohort (38.0% of fALS, 14.5% of sALS) were found to carry nonsynonymous/ splice site variants with MAF compatible with high penetrance disease models. However, only 10 of the 33 genes analysed represented well established Mendelian ALS genes and it is anticipated that many if not most of thethese candidate disease variants will not be ALS related. Additionally, variants of ALS2 and *SPG11* were observed solely in heterozygous configurations, but these genes have only been associated with recessively inherited ALS[21,59].

In all, 9.7% of patients (30% of fALS, 7.1% of sALS) carried previously reported ALS variants, with 8.78% carrying the *C9orf72* repeat expansion, 0.45% carrying the *FUS*:c.1574C>T(p.[P525L]) substitution and 0.45% carrying the *TARDBP*:c.859G>A(p.[G287S]) substitution. The phenotype of *FUS*:c.1574C>T(p.[P525L]) carriers was consistent with that reported previously[17], with both patients exhibiting an exceptionally young age of onset and rapid disease progression. Conversely, carriers of the *TARDBP*:c.859G>A(p.[G287S]) variant exhibited a disease of comparatively late onset and slow progression. Review of *TARDBP*:c.859G>A(p.[G287S]) carriers reported across this study and previous publications [103,104,153], revealed that disease onset ranged from 52-70 yrs of age while disease duration ranged from 49 to $\geq$93 months.

Despite prior evidence to suggest at least reasonably high disease penetrance, 62% of *C9orf72* expansion carriers and all *FUS*:c.1574C>T(p.[P525L]) and

TARDBP:c.859G>A(p.[G287S]) carriers were classified as sporadic. Modelling of the effects of penetrance and family size on the rate of familial disease has previously shown that inheritance of high penetrance variants can occur in the absence of a detectable family history[18]. Conversely, the presence of a family history does not necessarily infer the presence of a common disease aetiology[140] and it is not always the case that disease variants are inherited. For example, the *FUS*:c.1574C>T(p.[P525L]) substitution has been reported to occur as a *de novo* event in multiple cases[17]. Taken together our results therefore support the contention that the distinction between familial and sporadic disease is of limited utility from both a clinical and research perspective.

Substantial differences have been described in the general spectrum of rare and low frequency genetic variation across populations[79,91]. A degree of population differentiation would therefore also be anticipated in the genetics of disease pathogenesis. Formal comparison of our results with those reported by studies of major disease genes in Italy, confirmed that this is the case with ALS. There were significant differences in the nature of genetic susceptibility across Irish and Italian patients ($p=1.7x10^{-4}$), with only 2 of the 32 variants identified in either population being identifiable in both. One of these shared variants was the *FUS*:c.1574C>T(p.[P525L]) substitution, which is notable as limited population differentiation may have been predicted *a priori*, given the associated age of mortality and frequency e of *de novo* occurrence. The other was the *C9orf72* repeat expansion, which has been shown to occur with high frequency among various European populations[67,147,157]. However, the expanded allele occurred with a significantly higher frequency in the Irish population than the Italian(Irish=8.78% , Italian=4.39%, $p=3.95x10^{-4}$). Conversely variants of *SOD1*and *TARDBP* were significantly more common among Italian patients than Irish (*SOD1*:Irish=0.00%, Italian=2.00%, $p=3.8x10^{-3}$. *TARDBP*:Irish=0.45%, Italian=2.00%, p=0.035). The absence of *SOD1* variants in the Irish population is particularly striking, as the gene is believed to make significant contributions to disease burden across Scandinavia (9.6%)[102], the USA (7.5%)[158], Germany (12% - familial cases only)[159], Italy (2.1%)[147], France (56% - familial cases only)[160] and Korea (3.9%)[150].

A recent analysis of ALS patients from the Netherlands, revealed that mutations of multiple ALS-associated genes could be identified in 9/57 families analyzed ($p=1.57x10^{-7}$)[67]. This indicated that oligogenic susceptibility may play a highly important role in ALS aetiology. Despite analysis of a larger panel of genes, no such signal could be observed in the Irish population and this remained the case with restriction of the analysis to genes included in the Dutch study (data not shown). However, it was observed that 1.6% of patients (4.0% of fALS, 1.3% of sALS) carried multiple variants known or potential high penetrance ALS variants and that this included all identified carriers of one of the established ALS variants ($p<0.01$). This variant occurred within the *TARDBP* gene (*TARDBP*:c.859G>A(p.[G287S])), which is noteworthy as the strongest evidence for oligogenic based disease in the Dutch analysis also related to a *TARDBP* variant (c.1055A>G(p.[N352S])). *TARDBP*:c.859G>A(p.[G287S]) has previously been identified among patients from Italy[103], France[104], and the United Kingdom[153] but has not yet to our knowledge been reported in healthy controls. As studies which have previously reported the variant analysed *TARDBP* in isolation, it is not known whether carriers identified elsewhere also carry additional disease variants. The potential relevance of oligogenic susceptibility in ALS aetiology means that the exploration of oligogenic models may become increasingly important in disease gene mapping and studies of genotype-phenotype correlations. It also means that known disease genes should

be studied together rather than in isolation and that patients testing positive for disease variants at one locus should not automatically be excluded from subsequent analysis of additional disease loci.

Strong correlations have been observed between patient phenotypes and individual disease variants [17,156] and it may be the case that other clinical features such as the frequency of cognitive impairment, the burden of disability and drug response also vary across populations. While the clinical phenotype within European populations is broadly similar, our data would suggest that general extrapolation of findings from individually characterized ancestral populations must be undertaken with caution. A deeper understanding of disease heterogeneity across populations will require further analysis of representative patient cohorts, and is likely to be of benefit in cohort stratification for future clinical trials.

5.5 Conclusion

The results reveal that 17.1% of Irish cases may carry high penetrance disease variants within the investigated genes but that the overall frequencies of nonsynonymous variants across these genes do not differ dramatically in patients versus controls.Previously established disease variants were identified in9.7% of cases with the *C9orf72* hexanucleotide repeat expansion representing by far the most common. Although the described work represents the most comprehensive comparison of ALS genes to date, the study was nonetheless limited by the exclusion of more recently reported disease genes like *SQSTM1*[60] and *UBQLN2*[35], as well as by the absence of any functional analyses of putative disease variants. Significant differences in the frequencies of disease variants between Irish and other European populations were observed, demonstrating that distinct patient populations cannot always be treated as homogenous. Finally, analyses of variant co-occurrences revealed some indication to support the relevance of oligogenic susceptibility in ALS aetiology and suggested that the *TARDBP*:c.859G>A(p.[G287S]) variant may not cause disease in isolation.

6. CANDIDATE GENE ANALYSIS

6.1 Introduction

Sections 3 and 4 included a description of the prioritization and targeted resequencing of 718 candidate ALS genes. In this section I describe the results from a series of analyses used to screen the resulting dataset for signatures of ALS related variation. These analyses involved distinguishing reliable variant calls from sequencing artefacts through semi-supervised machine learning, estimation of power to detect the independent effects of variants on disease risk, adaptive selection of variants for association testing based on power simulations, controlling for sequencing related biases in association testing and analyses of the independent and aggregate effects of variants on disease susceptibility and patient phenotype. For the purposes of validating analytical methods and extending the work described in section 5, sequencing data for the analyzed panel of known ALS genes is also considered, although only for individuals included in the candidate genes cohort.

6.2 Materials and Methods

6.2.1 Variant calling and annotation

Variant calling, prediction of functional effects and retrieval of external minor allele frequency data was performed as described in section 5.2.3. Alternate minor alleles at multiallelic sites were treated as distinct variants in all analyses. Variants were classified as functional if they occurred within a consensus splice site of a protein-coding transcript, or if they were predicted to alter the amino acid composition of an encoded protein. Variants which could not be assigned PolyPhen scores during prediction analyses were arbitrarily assigned a score of 1 if they had been classified as functional and 0 otherwise. Variants were classified as novel if they did not occur within the dbSNP, 1000 genomes or ESP datasets described in section 2. Variants were considered to represent high confidence calls if they occurred within a series of reference datasets downloaded from the Genome Sequence and Analysis group (GSA) at the broad institute (ftp.broadinstitute.org). These reference datasets include the HapMap[161] dataset version 3.3, indels encountered during an analysis of 98 million publicly available Sanger sequencing traces[162] and a subset of high confidence SNP and indel calls from the 1000 genomes project. One-tailed binomial tests, were used to evaluate if the MAF of variants among various reference populations (Irish, European, Global) exceeded expectation under null models of high disease penetrance. These tests were performed as described in section 2, except that all analyses considered allele as opposed to genotype counts and assumed HWE within the reference populations. Variants were annotated as "possible ALS mutations" if they obtained p-values ≥ 0.05 in all populations analysed and had also been classified as functional.

6.2.2 Variant quality control

Eight quality control statistics were calculated for each variant call. Seven of these were generated by the GATK's UnifiedGenotyper during variant calling. These included measures of the bias in reference vs. alternate allele BQ scores (BaseQRankSum), reference vs. alternate allele strand alignment (FS), reference vs. alternate allele MQ score (MQRankSum), reference vs.

alternate allele read position (ReadPositionRankSum), consistency of the site with 2 and only 2 segregating haplotypes (HaplotypeScore), the number of reads aligning to the site with an MQ of 0 (MQ0) and confidence variant exists normalized by depth of coverage (QD). The 8th statistic (HWE) was calculated using a custom R function and represented the Phred scaled probability of the observed deviation in control genotype frequencies from those expected under HWE (*Equation 6.1*).

$$X^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i}$$

Equation 6.1

O_i and E_i denote the observed and expected frequencies for the i^{th} of the 3 possible genotypic configurations (AA, Aa, aa). The resulting test statistic was assumed to follow a X^2 distribution with 1 degree of freedom. Genotypes relating to samples of male gender were excluded for variants occurring on the X chromosome unless variants fell within the-pseudoautosomal regions.

A multivariate Gaussian mixture model for the expected distributions of these statistics was trained using the values observed across the set of high confidence variant calls. This model was then used to assign each variant a probability score based on the observed test statistic values (*Equation 6.2*, Supp. Script S6.1). Variants were assigned a status of "FAIL", if this score was less than that obtained by 97% of high confidence calls. This methodology is similar to the variant quality score recalibration procedure described by De Pristo *et al.*[163]. Variants were also assigned a status of "FAIL" where the call rate among either the case or control cohorts was <20%. Variants were assigned a status of "SUSPICIOUS" if they did not meet the criteria for a status of "FAIL" but achieved probability scores lower than those observed for 90% of the high confidence callset.

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |\mathrm{E}|}} \exp\left(\frac{1}{2}(x-\mu)' E^{-1}(x-\mu)\right)$$

Equation 6.2

x denotes a vector of the 8 quality control statistic values observed for a given variant. n denotes the length of the vector x. μ denotes a vector of the mean statistic values observed among high confidence calls. E denotes a matrix of the covariance of each test statistic within the high confidence callset. $|E|$ and E^{-1} denote the determinant and inverse of the covariance matrix E.

6.2.3 Genotype QC

As per section 5, sample genotypes were reset to missing where the associated GQ score was less than 15. Heterozygous genotype calls for variants on the X chromosome were reset to missing if the variant occurred outside of the pseudo-autosomal regions (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/) and the carrier was of male gender.

6.2.4 Power calculations

All power simulations were performed using customized R functions. The effective number of independent alleles within a given callset was determined using the Genetic type 1 error calculator (http://statgenpro.psychiatry.hku.hk/gec/)[144]. These effective numbers were then used in place of true callset sizes for the purposes of establishing experiment-wise significance thresholds

6.2.5 Single variant association testing

All single variant association testing was performed using R version 3.0.1. Associations with dichotomous variables, including case-control status, site of onset and status for various cognitive phenotypes, were tested using logistic regression. The limited size of the sample cohort and the low frequency of many observed alleles meant that the commonly used maximum likelihood methods, such as those implemented in PLINK, could not always be applied reliably. This was particularly problematic in cases of linear separation, where a variant was either case or control specific. Logistic regression was therefore performed using Firth's penalized likelihood approach as implemented in the R "logistf" package (http://cran.r-project.org/web/packages/logistf/index.html)[164]. Associations with age of disease onset were tested using linear regression and the R "lm" function. Associations with patient survival (time from diagnosis until death or the most check for survival status) were tested using cox proportional-hazards regression and the "coxph" function implemented in the R "survival" package (http://cran.r-project.org/web/packages/survival/). For analyses of case-control status, experiment-wise significance thresholds were determined by Bonferroni correction for the effective number of independent tests. This allowed for incorporation of sample gender, status for known ALS mutations (section 5), sequencing centre of analysis (TrinSeq vs. GATC, section 4), age of onset (cases) and age of sampling (controls) as sample covariates. For all other analyses multiple testing correction was performed using permutation procedures. This facilitated correction for the analysis of multiple correlated phenotypes as well as just the analysis of multiple alleles per phenotype but prohibited the use of sample as opposed to sequencing related covariates. All covariates except sample sequencing centre were therefore dropped for these analyses and sequence centre remained linked to the sequence data for a given individual as opposed to the phenotypic assignment. However, for exploratory purposes and the Manhattan plots shown in FigureS6.3-S6.9, uncorrected analyses were also performed with the inclusion of sample gender and status for known ALS mutations.

6.2.6 Aggregate association analyses

The combined influence of variants on phenotypic variables was assessed using a battery of 7 statistical tests (BTw1, BTw2, SKATw2, BTw3, BTw4, SKATw4 and metaSKAT). These tests represented alternate implementations of 3 basic analyses (BT, SKAT and metaSKAT) with one of 5 variant weighting strategies (Table 6.1). The BT tests were based upon regression of phenotypic variables with respect to either the number of variants (BTw1, BTw3) or the number of minor alleles (BTw2, BTw4) per individual and represented special cases of the weighted sum test described by Madsen and Browning[165]. The SKAT and metaSKAT analyses were based upon regression with respect to minor allele counts and represented the sequence kernel association and combined sum tests proposed by Wu *et al.*[151] and Ionita-Laza *et al.*[166]. All tests were

conducted using functions provided in the R package "SKAT" version 0.91 (http://www.hsph.harvard.edu/skat/download/). Permutation procedures were used to perform all multiple testing correction, with correction parameters including the number of statistical tests performed per aggregate unit, the number of units analysed and where relevant the number of phenotypic parameters analyzed. All permutations were implemented using customized R functions. Sequencing centre was included as a sample covariate in all analyses but once again the use of permutation procedures prohibited the use of covariates. Unlike the single variant analyses, patient survival was modelled as a continuous variable in all aggregate analyses with patients believed to still being excluded from analysis.

6.2.7 Data analyses

Unless otherwise indicated all data analyses were performed using custom scripts written for python version 2.6.5 (http://www.python.org/), R version 3.0.1 (http://www.r-project.org/) or bash version 4.2.24 (http://www.gnu.org/software/bash/).

Table 6.1: Variant weighting strategies

Weight	Tests	Description	Assumption
w1	BTw1	Variants are assigned a weighting of 1 if classified as "possible ALS mutations" (section 6.2.1) and 0 otherwise	Disease related variants act with high penetrance
w2	BTw2, SKATw 2	Variants are weighted in accordance with the density of a beta distribution with shape paramaters 1 and 25 at the point of the variant MAF[151]	The probability of a variant being disease related is inversely proportional to it's MAF
w3	BTw3	Variants are weighted by the product of w1, a scaling factor (100) and assigned PolyPhen scores	Disease related variants act with high penetrance and are more likely to be classified as damaging in *in silico* predtictions
w4	BTw4, SKATw 4	Variants are weighted in accordance with the product of w2 and assigned PolyPhen scores	The probability of a variant being disease related is inversely proportional to it's MAF but directly proportional to the score assigned by PolyPhen
w5	metaS KAT	Variants are classified as rare or common based upon whether or not the MAF exceeds $\frac{1}{\sqrt{2n}}$, with n representing the number of samples genotyped[166]. Rare variants are weighted as per w2 while common variants are weighted based on density under a beta distribution has shape paramaters 0.5 and 0.5[166]	The probability of a variant being disease related is inversly proportional to MAF but disease is determined by both rare and common variation

6.3 Results

6.3.1 Variant quality control

As stated in section 4.3.3, median coverage of target exons was 29X per sample with 80% of target bases being covered to >=7X per sample. Thirty five thousand nine hundred and ninety three putative variants were called within target exons ± 250bp. Twenty four thousand and seventy one (67%) of these were regarded as known genetic variants having been reported by dbSNP, the 1000 genomes or the NHLBI ESP. The remaining 11922 (33%) calls, were considered as novel variants. Sixteen thousand nine hundred and sixty eight variants (47%), had been previously reported by the HapMap project[161], an analysis of publicly available Sanger sequencing traces[162] or within a high confidence subset of SNP and indel calls from the 1000 genomes project(section 6.2.1). As the least likely to represent high-throughput sequencing artefacts, these latter variants were regarded as high confidence calls.

Although minimal confidence criteria were imposed during variant calling (section 5.2.3), the generated callset would still be anticipated to include numerous sequencing artefacts. These can result in both the calling of variation where it does not exist and introduce unreliable genotype assignment where variation does exist. Eight quality control statistics (section 6.2.1) designed to capture signatures of such events were calculated for each variant. Seven of these were based on the properties of sequencing reads, while the eight was based on deviations in control genotype frequencies from those expected under HWE. Each variant was then assigned a score based on the probability of observing the set of obtained statistic values under a multivariate Gaussian mixture model that had been trained using the set of 16968 high confidence calls (section 6.2.1). An important advantage of this approach over analyzing each statistic independently is that there is the potential to detect unusual correlations between statistics as opposed to just unusual values for a given statistic.

To provide a desirable balance between sensitivity and selectivity, variants were only removed from the dataset (assigned a status of "FAIL") if the associated probability score was lower than that achieved by the upper 97% of the training set. However, as an extra layer of quality of control, retained variants which achieved scores lower than those achieved by 90% of the training set were assigned a status of "SUSPICIOUS". As expected, these filtering strategies resulted in a greater loss of novel than known variants (Figure6.1), which was desirable given that the set of novel variants would be anticipated to include a higher proportion of artefacts. Review of the pair-wise distributions of the QC statistics confirmed that as expected, filtered variants exhibited either highly deviant values for one or multiple test statistics, or else moderately unusual values across multiple statistics (Figure6.2).

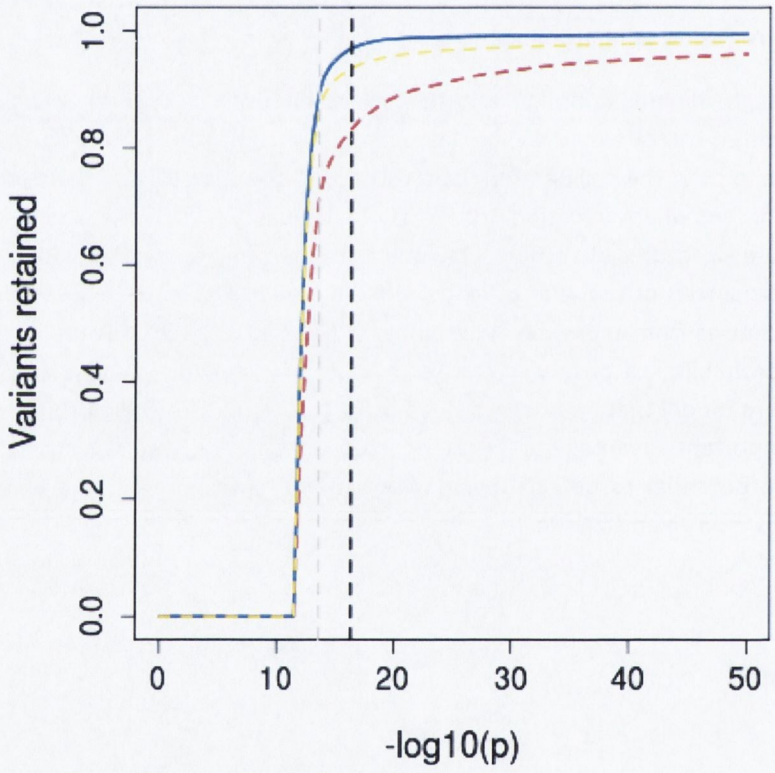


Figure 6.1: Variant filtering. Each variant was assigned a probability score based upon the relative values of 8 quality control test statistics. The x-axis denotes a maximum log scaled probability score accepted for variant retention while the y-axis denotes the proportion of variants retained. The blue line denotes the retention rate among high confidence (training) variants, while the yellow and red lines denote the retention rates among known and novel variants. The black and grey vertical line denote the employed thresholds for assignment of a status of "FAIL" and "SUSPICIOUS". These thresholds were selected such that 90% of the training set would be assigned a status of "PASS" while 97% would be assigned a status of "PASS" or "SUSPICIOUS".

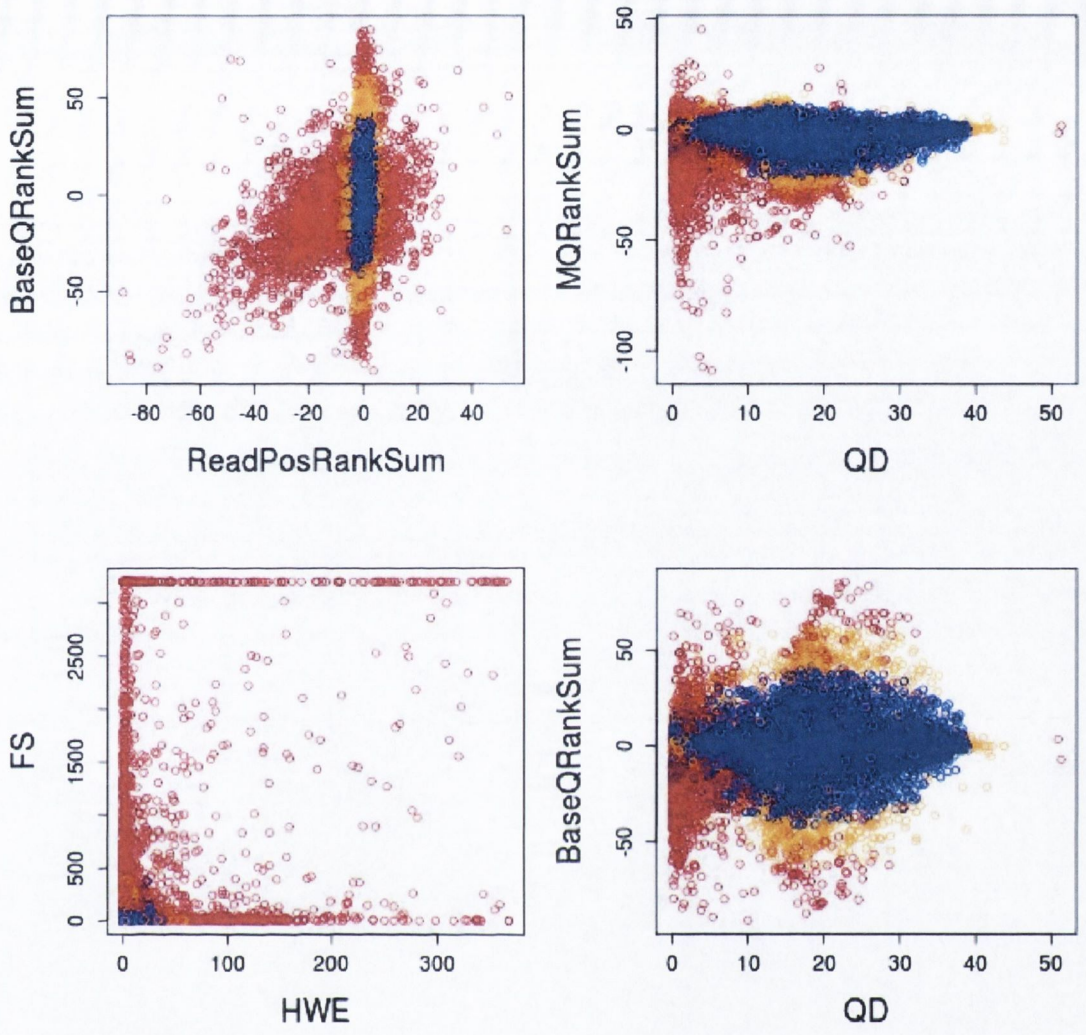


Figure 6.2: Pair-wise distributions of quality control statistics. Each plot shows the relative distribution of 2 quality control statistics. Variants assigned a status of "PASS" before the analysis of missing genotype rates are denoted by blue circles, while those assigned a status of "SUSPICIOUS" or "FAIL" are denoted by orange and red circles respectively. Clear correlations between statistics like BaseQRankSum and ReadPosRankSum, can be observed among calls anticipated to represent sequencing artefacts while such correlations are not observed among variants assigned a status of "PASS". The full set of pair-wise plots can be found in FigureS6.2.

6.3.2 Patient variants

Following the quality control described in section 6.3.1 and subsequent filtering based on missing genotype rates (Section 6.2.2), 22694 variants remained within the callset. These included 2994 variants assigned a status of "SUSPICIOUS", 15902 previously reported variants, 6792 novel variants, 3066 variants classified as functional, 593 variants classified as possible ALS mutations under dominant disease models and 3 variants classified as possible ALS mutations under recessive models. Individually, the possible ALS mutations occurred in 1-28 cases. However, the higher frequency variants exhibited significant sequencing centre dependent biases (TrinSeq vs. GATC, section 4). Following the implementation of an extremely stringent quality control measure, where variant calls were not considered unless observed in at least one individual

analyzed at sequencing centre 2, the number of possible ALS mutations fell to 367. Individually, these variants were observed within 1-3 ALS cases, with 346 being observed in a single case, 20 being observed in 2 cases and only a c.2849A>C[p.Gln950Pro] missense variant of the gene *PCNT* being observed in 3 cases. Exploratory analyses revealed that some of the variants carried by multiple patients exhibited notable phenotypic correlations (Figure6.2). For example, both patients carrying a c.853C>T[p.Arg285X] nonsense mutation within the *DDX50* gene exhibited an aggressive disease with bulbar onset, while all observed carriers of *CREBBP*:c.6624A>C(p.Gln2208His) or *NCOR1*:c.4472C>T(p.Thr1491Ile) missense variants exhibited aggressive disease with spinal onset. It was also observed that both carriers of a *PITPNM3*:c.548G>C(p.Cys183Ser) missense variant also carried additional known and/or candidate disease variants and that counter to expectation, possible ALS mutations observed in multiple cases tended to achieve higher PolyPhen scores than those observed in a single case (Figure6.3).

Table 6.2: Recurrent patient variants

Mutation	PolyPhen	Known ALS Mutation	fALS	Gender	Age (Years)	Site	Cognitive Phenotype	Survival (Months)	Deceased
DDX50:c.853C>T(p.Arg285X)	1	N	N	F	76	bulbar	NA	14	Y
DDX50:c.853C>T(p.Arg285X)	1	N	N	F	65	bulbar	FTD	20	Y
DYNC1H1:c.8200G>A(p.Val2734Met)	0.544	N	Y	M	51	spinal	NA	41	Y
DYNC1H1:c.8200G>A(p.Val2734Met)	0.544	N	N	F	69	spinal	NA	49	Y
RNASE2:c.106C>G(p.Gln36Glu)	0.908	N	N	M	49	NA	NA	NA	N
RNASE2:c.106C>G(p.Gln36Glu)	0.908	N	Y	M	49	spinal	NA	34	Y
RIN3:c.598C>T(p.Arg200Trp)	0	N	N	M	84	spinal	NA	99	Y
RIN3:c.598C>T(p.Arg200Trp)	0	N	N	M	74	spinal	NA	96	Y
PRKCB:c.1183G>A(p.Ala395Thr)	0.983	N	Y	M	61	spinal	NA	30	Y
PRKCB:c.1183G>A(p.Ala395Thr)	0.983	N	N	F	69	bulbar	NA	20	Y
CREBBP:c.6624A>C(p.Gln2208His)	0	N	N	F	68	spinal	NA	12	Y
CREBBP:c.6624A>C(p.Gln2208His)	0	N	N	M	57	spinal	NA	15	Y
NCOR1:c.4472C>T(p.Thr1491Ile)	0.821	N	N	M	73	spinal	NA	14	Y
NCOR1:c.4472C>T(p.Thr1491Ile)	0.821	N	N	M	57	spinal	NA	15	Y
MED25:c.1508C>T(p.Thr503Met)	0.485	N	N	M	47	spinal	NA	32	Y
MED25:c.1508C>T(p.Thr503Met)	0.485	N	N	F	50	spinal	NA	32	Y
MTOR:c.5863C>T(p.Arg1955Cys)	0.124	N	N	M	59	bulbar	NA	14	Y
MTOR:c.5863C>T(p.Arg1955Cys)	0.124	N	N	M	30	spinal	NA	46	Y
SRXN1:c.302G>A(p.Arg101His)	1	N	N	M	77	spinal	NA	65	Y
SRXN1:c.302G>A(p.Arg101His)	1	N	N	F	57	spinal	NA	49	Y

PCNT:c.2849A>C(p.Gln950Pro)	0.996	N	N	M	61	spinal	NA	115	N
PCNT:c.2849A>C(p.Gln950Pro)	0.996	C9orf72	N	M	NA	spinal	NA	NA	N
PCNT:c.2849A>C(p.Gln950Pro)	0.996	N	N	M	58	spinal	N.Exec	33	Y
PCNT:c.4930A>T(p.Thr1644Ser)	0.052	N	N	F	NA	NA	NA	NA	N
PCNT:c.4930A>T(p.Thr1644Ser)	0.052	N	N	F	28	spinal	NA	94	Y
PCNT:c.8116G>A(p.Val2706Met)	0.995	N	N	F	61	spinal	NA	28	Y
PCNT:c.8116G>A(p.Val2706Met)	0.995	N	N	M	63	spinal	NA	33	Y
DYNC1I2:c.824G>A(p.Arg275His)	0.109	N	N	M	68	spinal	NA	32	Y
DYNC1I2:c.824G>A(p.Arg275His)	0.109	N	Y	F	46	bulbar	NA	NA	N
CASP8:c.74C>T(p.Pro25Leu)	0.915	N	N	M	42	spinal	NA	148	N
CASP8:c.74C>T(p.Pro25Leu), PITPNM3:c.548G>C(p.Cys183Ser)	0.915, 1	N	N	M	80	spinal	NA	29	Y
MAP1B:c.1855G>A(p.Glu619Lys)	0.851	N	N	M	77	bulbar	NA	31	Y
MAP1B:c.1855G>A(p.Glu619Lys)	0.851	N	N	M	49	spinal	NA	37	Y
CTTNBP2:c.194G>A(p.Arg65His)	0.999	C9orf72	Y	F	82	bulbar	NA	12	Y
CTTNBP2:c.194G>A(p.Arg65His)	0.999	N	N	M	59	bulbar	NA	NA	N
PCM1:c.32G>T(p.Gly11Val)	0.98	N	N	F	47	spinal	NA	247	N
PCM1:c.32G>T(p.Gly11Val)	0.98	N	N	F	55	spinal	NA	110	Y
ELP3:c.326G>A(p.Cys109Tyr)	0.998	N	N	M	73	spinal	NA	42	Y
ELP3:c.326G>A(p.Cys109Tyr), PITPNM3:c.548G>C(p.Cys183Ser)	0.998, 1	C9orf72	Y	M	51	bulbar	NA	15	Y
PRKDC:c.5524A>G(p.Ile1842Val)	0.013	N	N	M	54	spinal	NA	14	Y
PRKDC:c.5524A>G(p.Ile1842Val)	0.013	N	Y	M	47	bulbar	NA	138	Y

Tables provide a description of patients carrying one or more of the 21 possible ALS variants observed across multiple individuals after highly stringent quality control filtering. Age and site refer to age and site of disease onset respectively. N.Exec denotes non-executive cognitive impairment.

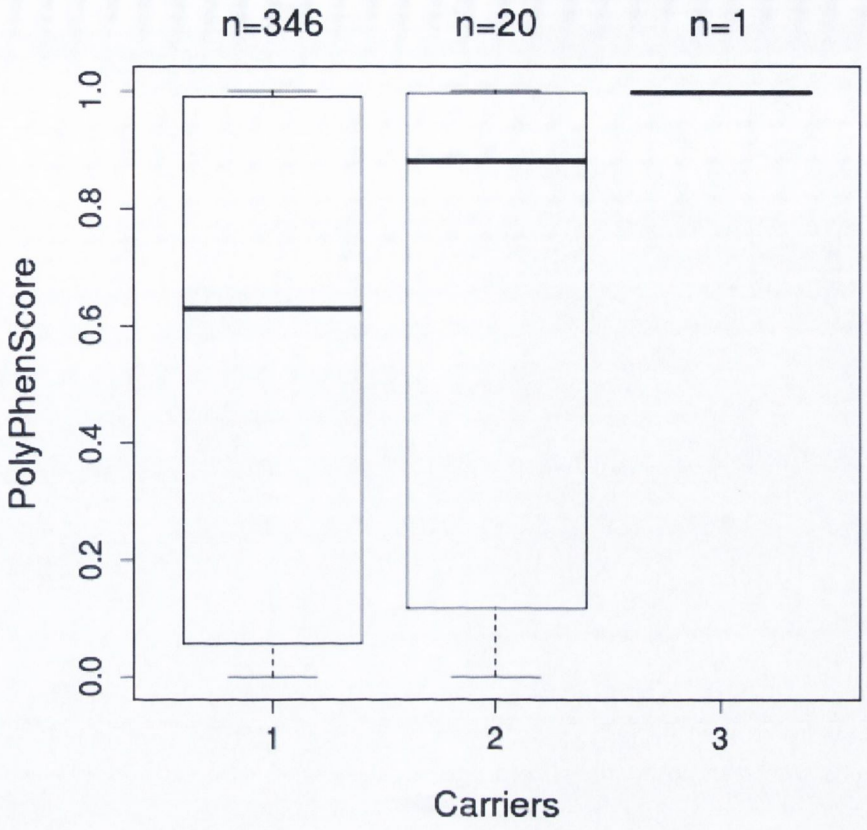


Figure6.3: Distribution of PolyPhen scores. Boxplots outline the distribution of PolyPhen scores among patient variants classified as possible ALS mutations with respect to the number of identified carriers. To reduce the impact of potential sequencing errors or Irish polymorphisms at sites that are difficult to genotype, only variants passing highly stringent quality control measures are considered.

6.3.3 Analysis of independent variant effects

Following quality control, variant call rates ranged from 20-100% while variant MAF ranged from 0.12-50% (Figure6.4). This inferred substantially variability across loci regarding power to detect disease related effects. Before proceeding to analyses of independent effects on disease risk, it was therefore decided to reduce the callset to variants for which it was actually possible to detect a significant association.

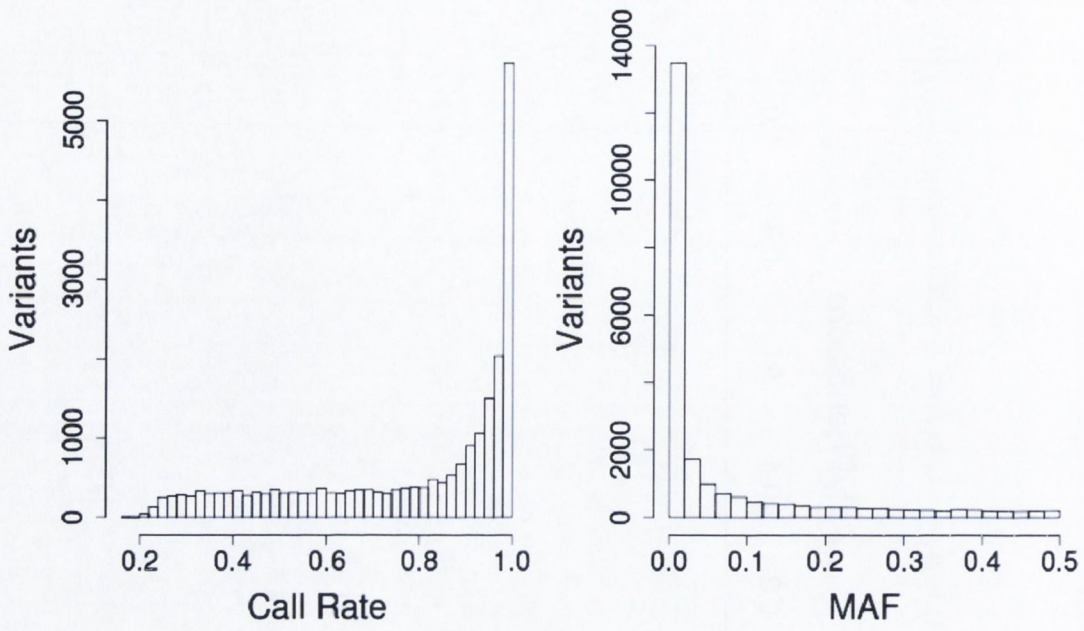


Figure6.4: Variant call rate and MAF

To achieve this reduction, the minimum possible p-value (minP) observable in an allele count based test was calculated for each site. This represented the p-value obtained in a Fisher's exact test, where all observations of the minor allele were assumed to represent occurrences among cases. An appropriate minP threshold was then obtained by progressively filtering variants, until the lowest minP value corresponded to the Bonferroni corrected threshold for experiment-wise significance (experiment-wise type I error rate = 5%, Figure6.5). A threshold of 10^{-5} was determined as the optimum for an analysis intended to incorporate as much variation as possible while a threshold of 10^{-4} was selected for an analysis restricted to functional variants. Power simulations revealed that following minP filtering, the probabilities of detecting significant associations with case-control status within the minP filtered callset were in excess of 80% where variants had MAF $\geq$10% and OR were $\geq$3 or where variants had MAF $\geq$1-5% and OR were $\geq$ 12-5 (Figure6.6). Restriction of the analysis to functional variants provided some increase in the minimum effect size sensitivity but this was not substantial.

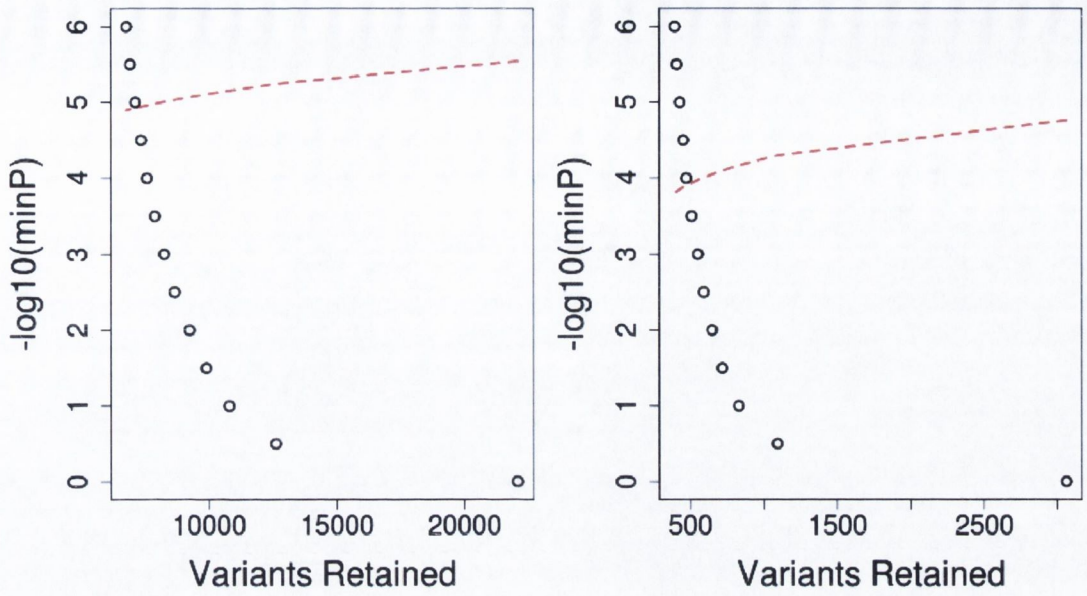


Figure6.5: Adaptive filtering of callset for single variant association testing. The minimum possible p-value (minP) observable in an allelic test of association with case-control status was calculated for each variant. This represented the result from a Fisher exact test that incorporated the relative missing genotype rates among cases and controls as well as the combined variant MAF. For each test it was assumed that all observations of the minor allele represented occurrences among cases. The x-axes of both plots denotes the number of variants retained after filtering based on a given minP threshold while the y-axes denotes the log transform of those minP thresholds. At each threshold, the effective number of independent tests is calculated by correcting for LD and used to determine a Bonferroni corrected threshold for experiment-wise significance (denoted by red line, experiment-wise type I error rate = 5%). The plot to the left relates to an analysis where as much variation as possible is considered while the plot to the right relates to an analysis where only functional variants are considered.

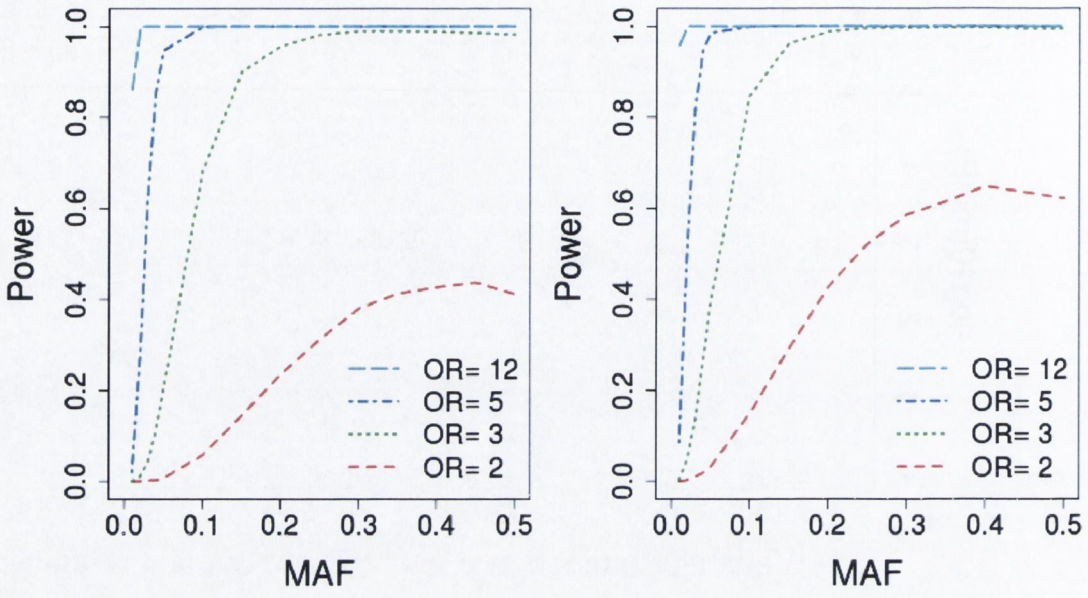


Figure6.6 Power Simulations. Plots outline power to detect allelic associations with case-control status across various effect sizes and variant MAF. Each data-point represents the result from 10,000 simulations at a diploid locus. The numbers of cases and controls included in each simulation represent the average of that observed across sites passing variant QC. The first plot outlines the results where the criteria for statistical significance represents the Bonferroni corrected threshold selected for an analysis of as much variation as possible ($p \leq 10^{-5}$) while the second plot outlines the results where the criteria for statistical significance represents the Bonferroni corrected threshold selected for an analysis restricted to functional variants ($p \leq 10^{-4}$).

Initial analyses of associations with case-control status revealed a significant inflation of the observed p-values with respect to null expectation (Figure6.7). This was primarily driven by variants which had been assigned a status of "SUSPICIOUS", most of which occurred in the HLA locus on chromosome 6. Further analyses revealed that this inflation was primarily driven by sequencing centre biases. Following correction for this through incorporation of sequencing centre as a sample covariate (section 6.2.5), one association result met the criteria for statistical significance (Figure6.8). However, this related to a variant that had been assigned a status of "SUSPICIOUS" and review of available dbSNP annotation suggested that it was likely to represent a recurrent sequencing artefact.

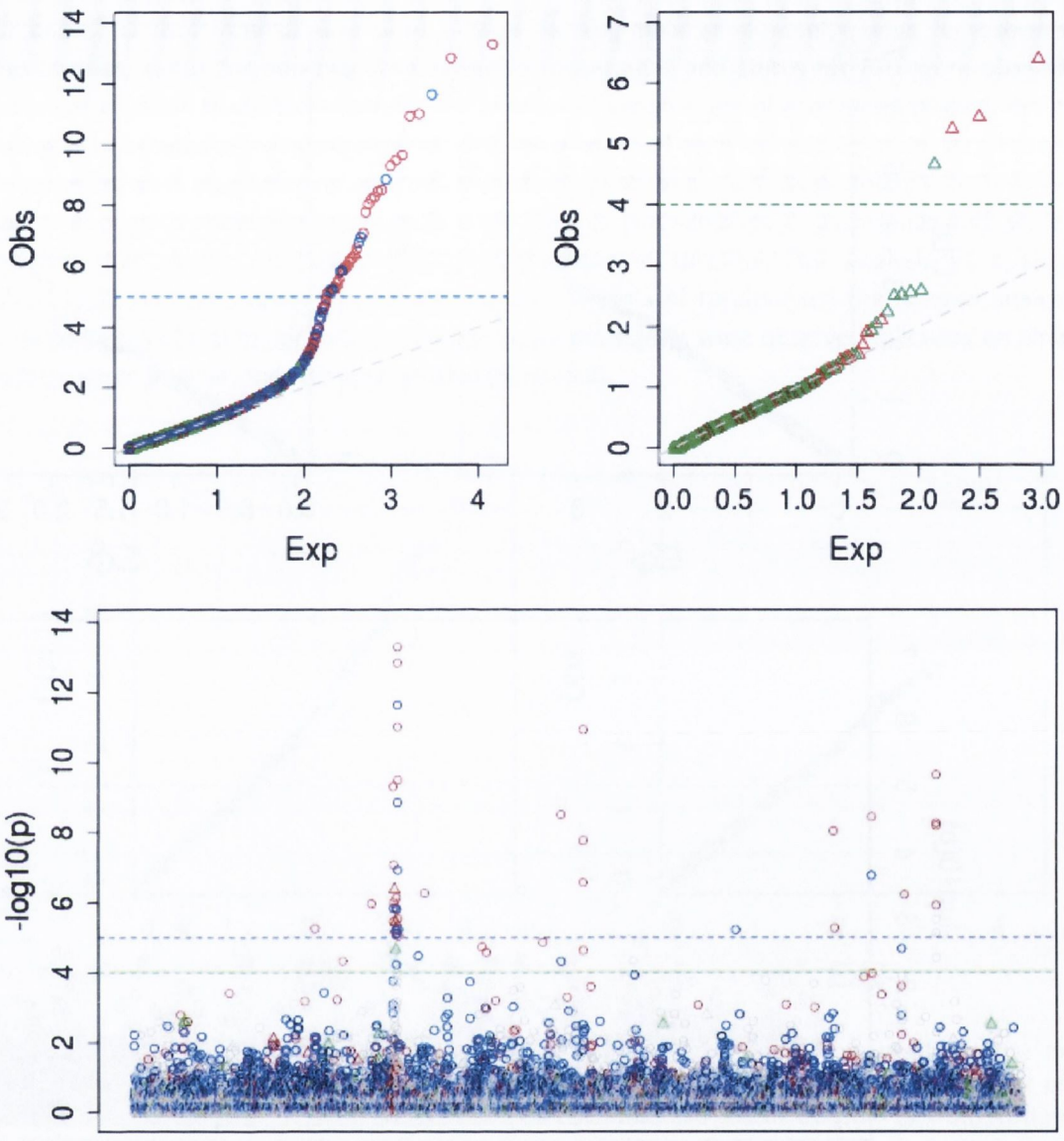


Figure 6.7: Uncorrected association analysis. Plots outline the probabilities of observed associations between variant minor allele counts and case-control status, as calculated using Fisher exact method. The top row includes QQ-plots where either the full dataset or only functional variants were considered. The bottom plot represents a Manhattan plot where variants were ordered by genomic position. Blue circles and green triangles denote non-functional and functional variants respectively. Variants have been re-coloured red if assigned a status of "SUSPICIOUS" during variant QC and grey if they failed to meet minP thresholds. Horizontal blue and green lines denote the thresholds selected for a statistically significant associations among non-functional and functional variants respectively.

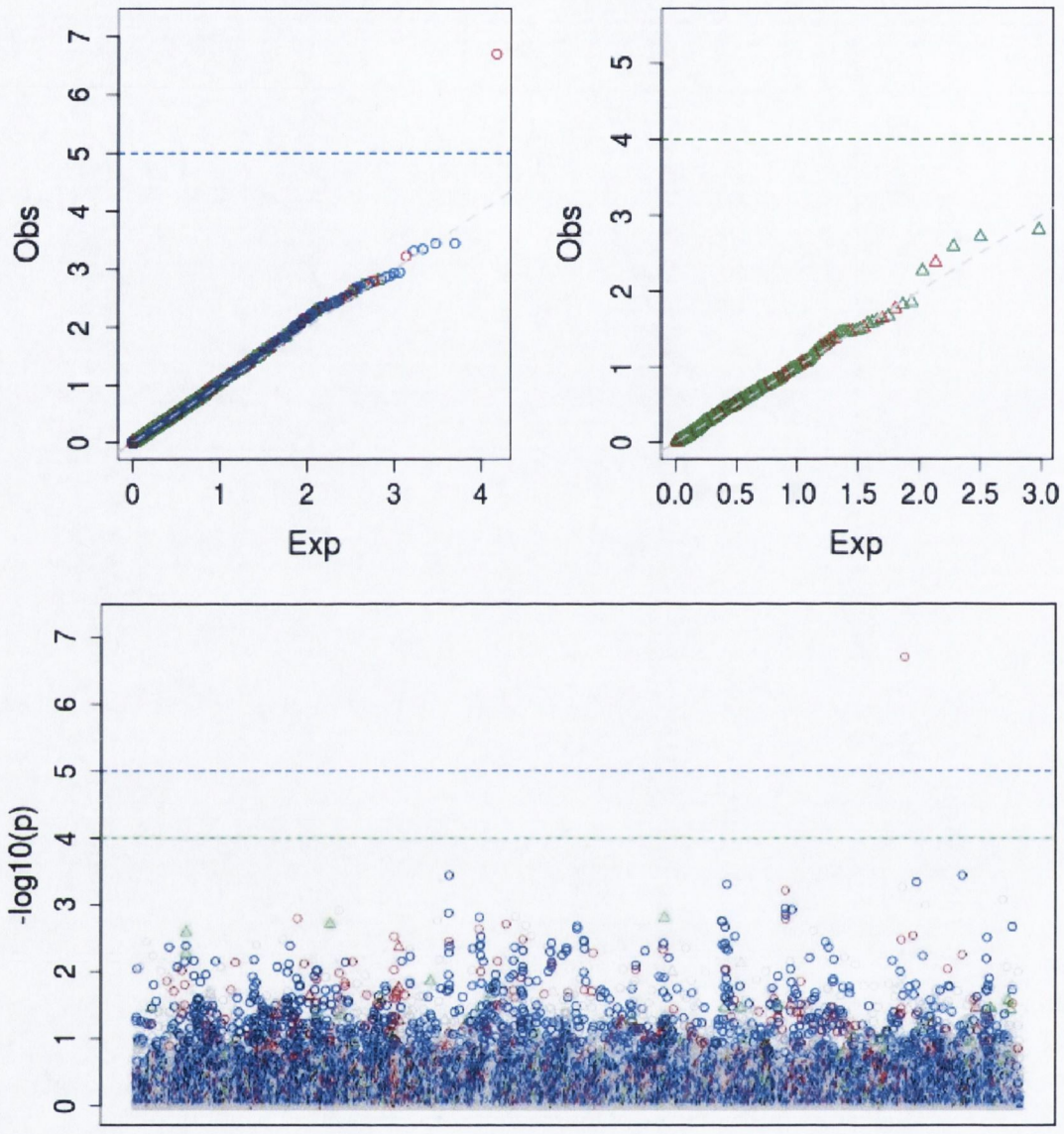


Figure 6.7: Corrected association analysis. Plots denote the same data described in Figure6.6, except where association testing has been performed using a logistic regression framework that incorporated sequencing centre as well as sample gender, age of disease onset (cases), age of sampling (controls) and status for known ALS mutations as covariates. Blue circles and green triangles denote non-functional and functional variants respectively. Variants have been re-coloured red where the status assigned during variant QC was "SUSPICIOUS" and grey where they did not meet required minP thresholds. Horizontal blue and green lines denote the thresholds for statistically significant associations among non-functional and functional variants respectively

No statistically significant associations with other clinical parameters, including age of onset, site of onset, survival from disease onset, status for cognitive impairment, status for executive dysfunction, status for non-executive cognitive impairment and status for FTD were observed after permutation based correction for the number of variants and phenotypes analyzed. Review of the distributions of p-values observed with the analysis of each individual parameter, revealed that results were largely in line with null expectation, with some deflation among tests involving the rarer cognitive phenotypes (FigureS6.3-6.9). One variant exhibited an association with patient survival that almost reached the Bonferroni corrected threshold for analysis of a single phenotype. However, review of QQ and Manhattan plots of results from the survival analyses revealed signs of systematic inflation and no stand-out signals were observed following empirical estimation of p-values through permutation (Figure6.8).

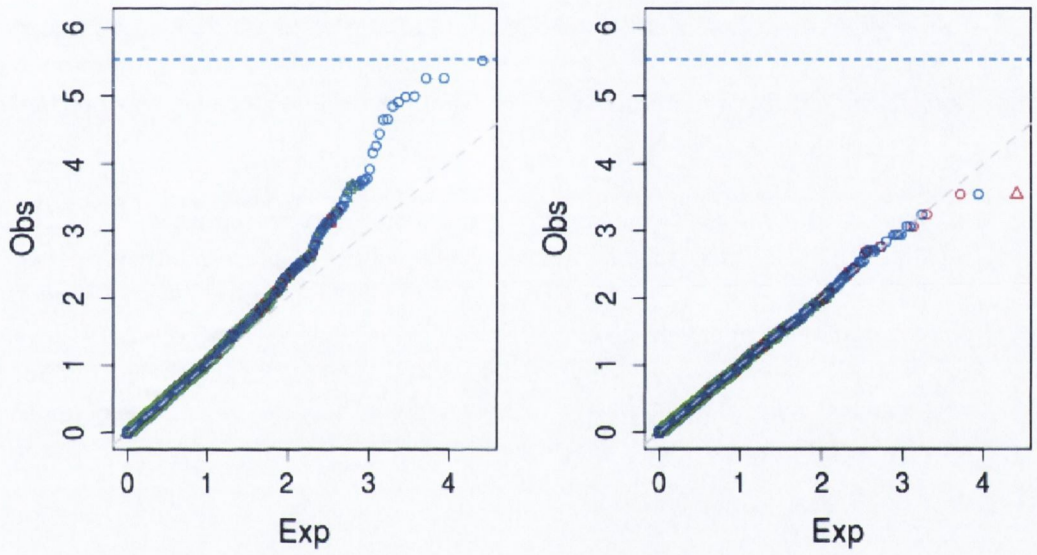


Figure6.8 Analysis of independent effects on patient survival. QQ plots outlining the null probabilities observed in tests for association between patient survival from disease onset and variant minor allele counts. The plot to the left outlines p-values estimated from Cox-proportional hazards regression with sample gender, status for known ALS mutations and sequencing centre being incorporated as covariates. The plot to the right outlines the results, where p-values are recalculated empirically through permutation.

6.3.4 Analysis of aggregated variant effects

As discussed in sections 1, 2 and 5, most ALS variants have very low MAF even among patients. This can make it extremely difficult, if not impossible to demonstrate independent effects through association testing. However, most established ALS genes have been associated with multiple disease variants and an assortment of burden and variance component analyses (section 6.2.6) were therefore used to test aggregate variant groupings for combined influences on disease risk or patient phenotype. Each of these tests catered for slightly different disease scenarios (Table 6.1), with the burden (BT) tests being expected to achieve higher power in situations where effect sizes are unidirectional and the variance component analyses (SKAT, metaSKAT) being expected to achieve higher power in the presence of neutral variation and mixed directions of effect[151]. The aggregate units tested included each of the 718 candidate ALS genes, the 36 known ALS genes described in section 5, the subset of candidate genes reported to exhibit sequence similarity to ALS proteins in section 3 and various gene sets representing local neighbourhoods within the molecular interaction networks described in section 3. Permutation procedures were used to correct for the number of statistical tests performed per grouping, the number of grouping analyzed, and where relevant, the number of patient phenotypes analysed. Only 3 groupings achieved fully corrected p-values < 0.95. These included the genes *TAFIL* and *NTRK1,* which achieved the lowest p-values in tests of association with case-control status (Table 6.3) and the *PRLR* gene, which achieved the lowest p-value in tests of association with patient phenotype (Table 6.4). Encouragingly, the top associations with disease risk included the gene *ANG* while the top associations with patient phenotype included the genes *FUS* and *PON2*. The p-value achieved by *ANG* following correction for the number of statistical tests per grouping but not the number of groupings analyzed (EMP1) was 0.023 while the corresponding p-values obtained by *FUS* and *PON2* were 0.021 and 0.0067 respectively.

Table 6.3: Top hits from aggregate tests of association with disease risk

Locus	Risk (EMP1)	Risk (EMP2)	Phenotype (EMP1)	Phenotype (EMP2)
TAF1L	0.001667	0.313333	0.958889	1
NTRK1	0.003333	0.368333	0.911111	1
TCF4	0.006667	1	0.695556	1
STAT3	0.008333	1	0.642222	1
CALCOCO1	0.01	0.968333	0.695556	1
PABPC1	0.013333	1	1	1
PLK4	0.021667	1	0.368889	1
ANG	0.023333	1	1	1
LRPPRC	0.033333	1	0.524444	1
FLNA	0.033333	1	1	1

The EMP1 columns denote the minimum p-value observed per grouping following correction for the number of statistical tests performed per grouping. In the case of tests of association with patient phenotype, this includes correction for the number of phenotypic categories analyzed. The EMP2 columns denote the fully corrected p-values, where correction for the number of groupings analyzed has also been performed.

Table 6.4: Top 10 hits from aggregate tests of association with phenotypic paramaters

Locus	Risk (EMP1)	Risk (EMP2)	Phenotype (EMP1)	Phenotype (EMP2)
PRLR	0.705	1	0.001111	0.261111
FECH	0.935	1	0.002222	0.975556
EPS15	0.303333	1	0.005556	0.998889
PON2	1	1	0.006667	1
RALBP1	0.286667	1	0.011111	1
DDX5	0.463333	1	0.013333	1
MSI1	0.578333	1	0.018889	0.998889
CCNE1	0.713333	1	0.018889	1
CDH1	0.998333	1	0.02	1
FUS	0.728333	1	0.021111	1

The EMP1 columns denote the minimum p-value observed per grouping following correction for the number of statistical tests performed per grouping. In the case of tests of association with phenotypic parameters this includes correction for the number of phenotypes analyzed. The EMP2 columns denote the fully corrected p-values, where correction for the number of groupings analyzed has also been performed.

6.4 Discussion

Differing classes of disease variation would be anticipated to produce disease signatures best detected using differing types of statistical analysis. In this section, an extensive battery of tests was employed to screen for disease related effects but after catering for the influence of sequencing and analysis artefacts, no statistically significant effects were observed, either on disease risk or patient phenotype. However, the results do not infer the absence of such associations and there were many encouraging signs that if taken further, the analyses performed in this section could still lead to the discovery of new and important insights into ALS pathogenesis.

Firstly, it is worth noting that following corrections for multiple testing, statistically significant effects could not be detected across the known ALS genes, despite the presence of known ALS mutations. It can therefore safely be concluded that the analyses suffered from a lack of power. Secondly, filtering based on MAF and predicted functional effects was found to represent a reliable means of selecting out disease variants like *TARDBP*:c.859G>A(p.[G287S]) and *FUS*:c.1574C>T(p.[P525L]) and revealed over 300 patient variants deemed to represent possible ALS mutations. Twenty one of these were observed among multiple patients and, contrary to expectation, these variants tended to achieve higher PolyPhen scores than those observed in a single patient. Although anecdotal, this observation could be an indication for the presence of pathogenic variants with MAF comparable to *FUS*:c.1574C>T(p.[P525L]) and *TARDBP*:c.859G>A(p.[G287S]). It was also encouraging to note that some of these variants exhibited unusual phenotypic correlations which held across the, albeit limited, set of identified carriers. Sample sizes are not sufficient to draw definitive conclusions but these observations could represent signatures of variants that cause atypical disease phenotypes and further analysis of *DDX50:*c.853C>T[p.Arg285X] *CREBBP*:c.6624A>C(p.Gln2208His), *NCOR1*:c.4472C>T(p.Thr1491Ile) in particular, would seem warranted.

Another highly encouraging observation was that the intervals that exhibited the most significant aggregate associations with case-control status included the *ANG* gene(EMP1=0.023), while those that exhibited the most significant associations with patient phenotype included the *FUS* (EMP1=0.021) and *PON2* (EMP1=0.0067) genes. These observations seem unlikely to represent coincidence as *ANG* has previously been demonstrated to associate with case-control status in aggregate analyses[167] while *FUS* variants are well known to associate with atypical patient phenotypes[17]. It would therefore seem that despite the very limited sample sizes, the aggregate analyses were still sufficiently powered to achieve favourable prioritization of genes with ALS related effects. Further follow up on the other top hits from these analyses, most particularly *TAFIL*, *NTRK1* and *PRLR*, may therefore yield demonstrable effects on ALS risk and/or manifestation.

Finally, it is worth highlighting that certain aspects of the analyses described in this section may benefit from further refinement. Alternate variant filtering strategies or reductions in the number of statistical tests used to screen for disease effects could potentially provide meaningful gains in power. Conversely the implementation of additional statistical tests may also prove useful. For example, a combined analysis of clinical parameters may allow for the detection of effects that were not observable with independent consideration of each parameter and may increase sensitivity to detect variants associated with atypical disease phenotypes.

6.5 Conclusion

No statistically significant associations between disease risk or patient phenotype could be detected, either with single variant or aggregate analyses. However, numerous patient variants were found to represent possible ALS mutations and there were encouraging signs that at least some of these may indeed be ALS related. Additionally, despite being underpowered, there was evidence to indicate that the aggregate analyses may still have provided an efficient enrichment for loci with disease related effects. It may therefore yet be the case that new disease genes can be discovered through a more comprehensive follow up of selected candidate disease variants and the top hits from aggregate analyses.

7. WHOLE MITOCHONDRIAL GENOME ANALYSIS

7.1 Introduction

Several lines of neuropathological and biochemical evidence suggest that mitochondrial dysfunction may play an important role in ALS pathobiology. This includes reported associations between disease status and abnormal mitochondrial morphology, the accumulation of mitochondrial DNA damage and decreased respiratory chain function[168,169]. Such anomalies could represent features of the disease without playing meaningful roles in disease onset or progression, however it has also been reported that the toxicity of at least certain ALS variants may be mediated through effects on mitochondrial homeostasis[105,170]. Furthermore, various rare ALS–like disorders have been reported to be the result of mitochondrial DNA mutations[171-173]. Though potentially explicable by phenomena such as non-paternity or genomic imprinting, it has also been reported that disease susceptibility is more frequently inherited through the maternal than paternal lineage[9]. Despite these observations, a comprehensive investigation into the extent to which mitochondrial variation may determine ALS susceptibility or clinical phenotype has never been undertaken.

As discussed in section 4, "off target" reads generated during targeted resequencing experiments include an over-representation of sequences aligning to the MT genome. This presumably results from the comparatively high copy number of mitochondrial DNA with respect to nuclear DNA. In this section, I discuss the analysis of off-target MT alignments generated over the course of the works described in sections 4-6. This analysis includes an investigation of MT coverage, sequencing centre and target enrichment related biases and a reimplementation of the disease variant screening strategies discussed in section 6.

7.2 Materials and Methods

7.2.1 Study participants

Study participants included the known genes cohort described in section 4.

7.2.2 Data analysis

Variant calling was performed as described in section 5. All genotypes were re-coded to haploid configurations (1 – carries alternate allele, 0 – does not), with individuals displaying signs of heteroplasmy (assigned a heterozygous genotype) being treated as homoplasmic variant carriers. Variant quality control and annotation were performed as per section 6 with the following modifications 1) the HWE quality control statistic was dropped from the analysis, 2) minor allele counts for external reference cohorts were imported only from the 'ALL.chrMT.phase1_samtools_si.20101123.snps.low_coverage.genotypes.vcf.gz' file (downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/), 3) due to the potentially confounding effects of heteroplasmy and the limited number of individuals within the control and 1000 genomes reference panels, the one tailed binomial tests described in section 2 were not performed. Depth of coverage calculations were performed as per section 4. Power and association analyses were performed as per section 6, except that all individuals were treated as haploid. Unless otherwise indicated, all analyses were performed

using custom scripts written for python version 2.6.5 (http://www.python.org/), R version 2.14.1/ 3.0.1 (http://www.r-project.org/) or bash version 4.2.24 (http://www.gnu.org/software/bash/).

7.3 Results

7.3.1 Sequence coverage

In total, 7112728 (0.03%) of the ~2.5 billion previously generated sequencing reads aligned to the mitochondrial genome. Following sample QC, the removal of low quality alignments and the masking of low quality base calls, median sample coverage was 3X (IQR=2-6X) and total coverage of mitochondrial bases ranged from 70-7453 X (Fig 7.1). Samples tended to achieve higher coverage if capture 1 or sequencing centre 2 was used in data preparation (Fig 7.1, section 4). Some regional inconsistencies to this pattern were noted and are most likely explained by off-target probe affinity and differences in power to map single vs. paired end sequencing reads. Fourteen samples were dropped from further analysis due to missing genotype rates in excess of 80%, reducing the dataset to 440 cases and 301 controls.

7.3.2 Variant callset

One thousand and thirty seven variant calls passed minimal confidence and quality control criteria (Fig S7.1). Of these, 112 were patient specific (absent from controls and the 1000 genomes project: 34 missense; 1 loss of stop site; 4 tRNA; 51 synonymous; 22 intergenic), while 40 were both patient specific and detectable across multiple cases (15 missense; 1 tRNA; 15 synonymous; 9 intergenic). The most frequently recurring case only variant was a c.8A>G(p.His3Arg) substitution within the *CO3* gene, detected in 6 cases. The remainder of patient specific variants were individually observed within 1-3 cases, with 56 patients (13%) carrying a case only variant expected to alter protein structure and 5 patients (1%) carrying a case only variant within a tRNA gene (Table 7.1). In 99.8% of cases, the observed sequencing data were compatible with homoplasmic genotype configurations. However, 1625 potential instances of heteroplasmy were encountered, including 3 of the 6 *CO3*:c.8A>G(p.His3Arg) occurrences (Table 7.1). Sanger sequencing confirmed the *CO3*:c.8A>G(p.His3Arg) status of the 3 carriers assigned homoplasmic genotypes but not the 3 assigned heteroplasmic genotypes. The only variant encountered which had previously been implicated in an ALS-like disorder was the *ND2*:p.Ala331Thr substitution[172], and this was observed at a higher frequency among controls (17/260; 6.5%) than cases (23/406; 5.7%).

7.3.3 Power analyses

Median power to detect ALS mutations was estimated as $\geq$87% (IQR$\geq$86-88%), assuming a patient MAF$\geq$0.5% and a homoplasmic genotype configuration (Fig 7.1). Power simulations, based on Bonferroni correction for multiple testing and assuming average missing genotype rates, indicated that the cohort size was not sufficient to test individual variants for disease association in the event of MAF<2% (Fig 7.2). Following removal of these variants and correction of the remaining callset for linkage disequilibrium, an experiment-wide significance threshold of 3.7×10^{-4} was selected for single variant analyses of disease risk. Further simulations suggested that power to

detect associations at this threshold was 93% for variants with MAF of 10% and OR of 3 and 87% for variants with MAF of 2% and OR of 6 (Fig 7.2). Repetition of the analyses with the prior exclusion of variants not expected to alter protein structure yielded similar results, including a minimum testable MAF of 2%, an experiment-wide significance threshold of 2.4×10^{-3}, an estimate of 84%

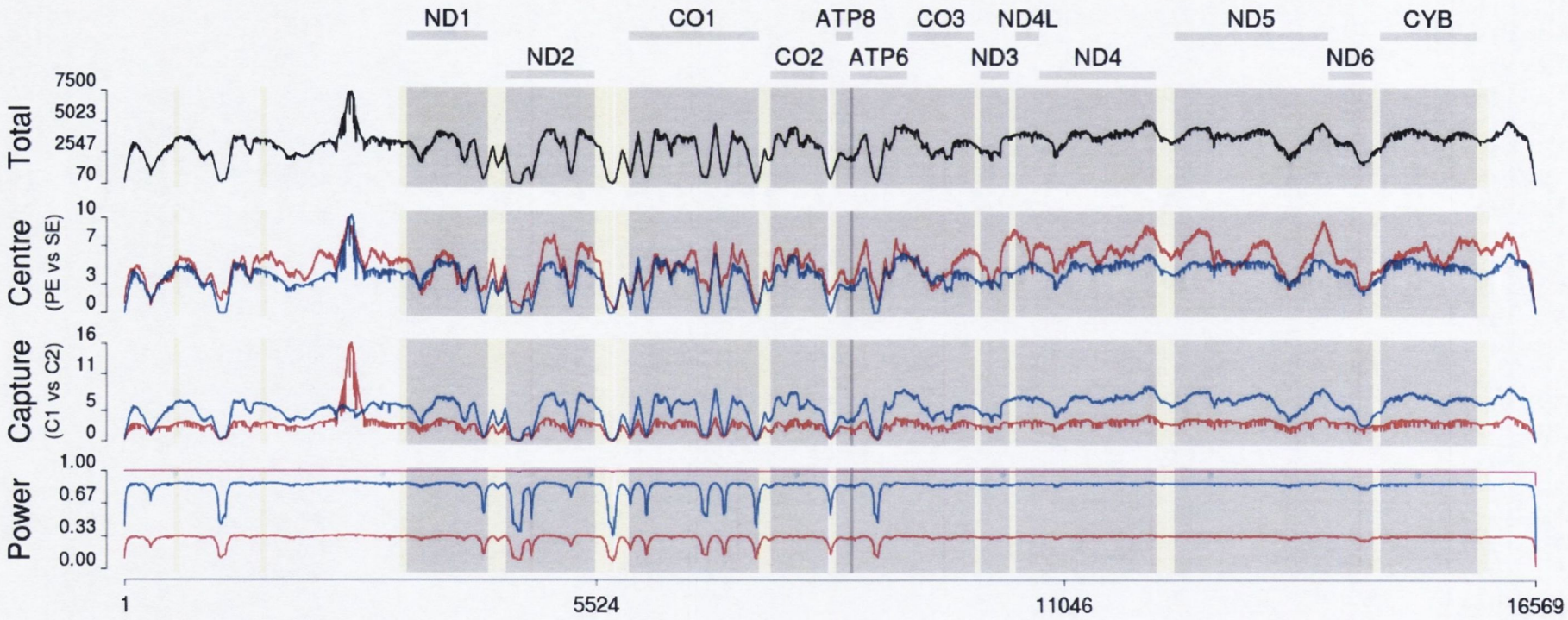


Figure7.1 Mitochondrial coverage. Panel 1 (Total): Total coverage across all samples. Panel 2 (Centre): Mean coverage per sample where the cohort is divided based on sequencing centre (blue: centre 1; red: centre 2). Panel 3 (Capture): Mean coverage per sample where the cohort is divided based on target enrichment system (blue: capture 1, red: capture 2). Panel 4 (Power): Power to detect ALS mutations of given patient MAF (purple - 5%, blue - 0.5%, red - 0.1%), assuming homoplasmic genotype configurations. For all panels, sequence coverage and power estimates relate to that observed following sample, alignment and base quality QC. For all panels, grey shaded regions denote protein-coding sequence while yellow shaded regions denote tRNA coding sequence.

power to detect disease associations in the event of MAF=10% and OR =2.5 and 95% power in the event of MAF=2% and OR=6 (Fig. 2).

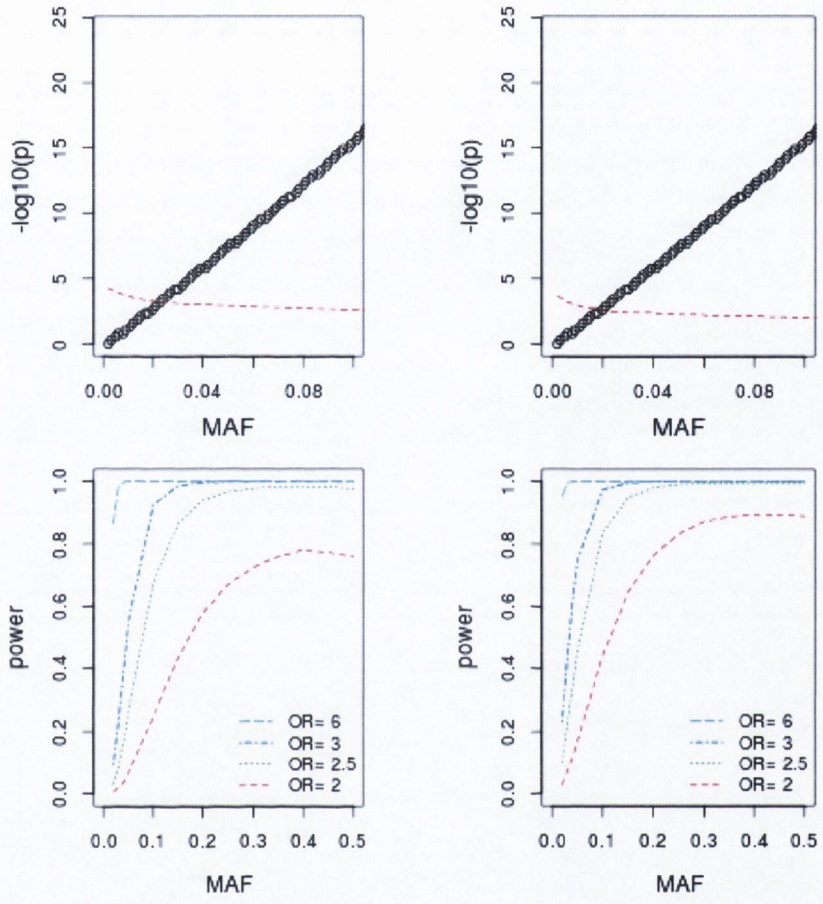


Fig 7.2: Power simulations. Top panel: Black dots denote the minimum p-value observable for a variant of given MAF in a test of association with case-control status. Tests assume the average missing genotype rates within the observed sample cohort and are based on Fisher's exact method. Dotted red lines denote the Bonferroni corrected thresholds for experiment-wide significance (fwe<0.05), where variants below a given MAF are excluded from analysis. Bottom panel: Plots denote power to detect variant associations with case-control status, assuming various MAF and effect sizes. Criteria for experiment-wide significance correspond to Bonferroni corrected thresholds within the observed datasets, following the removal of variants with MAF<2%. For both upper and lower panels, the plots to the left describe an analysis which considers all observed variation while the plots to the right describe an analysis which considers only nonsynonymous variation.

7.3.4 Association testing

Burden analyses revealed highly suggestive evidence for an association between disease risk and the occurrence of rare nonsynonymous *ND4* gene variants, with 11 patients having been found to carry case specific variants and 0 controls having been found to carry control specific variants (Table 7.3, burden.fwe=6.6×10^{-2}). Subsequent implementation of aggregate analyses intended to

cater for additional disease models, did not increase the statistical evidence for the *ND4* association (all.fwe=0.12) or provide any strong indications for other protein coding/ tRNA gene disease associations (Table 7.3, Table 7.4). Nonsynonymous variants within the gene *ND5* achieved the top single variant associations with both disease risk (*ND5*:c.1294A>G[p.Thr432Ala]: case MAF=6/393, control MAF= 0/263, p=0.01; *ND5*:c.1372G>A[p.Ala458Thr]: case MAF=50/382, control MAF=48/255, p=0.02, *ND5*:c.1301A>G[p.Gln434Arg]: case MAF=12/397, control MAF=2/262, p=0.02) and patient phenotype (*ND5*:c.1598C>T[p.Thr533Met]: obs=4/8 tested carriers with frontotemporal dementia [FTD OMIM=600274], exp=1/8 tested carriers with FTD, $p=1.8x10^{-3}$, fwe=0.6)(Fig 7.3). *ND5* also exhibited the top association with patient phenotype under aggregate analyses, a signal which appeared to reflect a comparatively high frequency of patients with executive dysfunction among carriers of rare nonsynonymous mutations (Table 7.3, obs=2/2, exp=0/2, burden.fwe=0.4). No global signatures of aggregate associations were observed across the combined sets of nonsynonymous (risk: burden.fwe=0.34,all.fwe=0.45; phenotype: burden.fwe=0.96, all.fwe=0.89) or tRNA gene (risk: burden.fwe=0.98, all.fwe=1.00; phenotype: burden.fwe=0.99, all.fwe=0.50) variants.

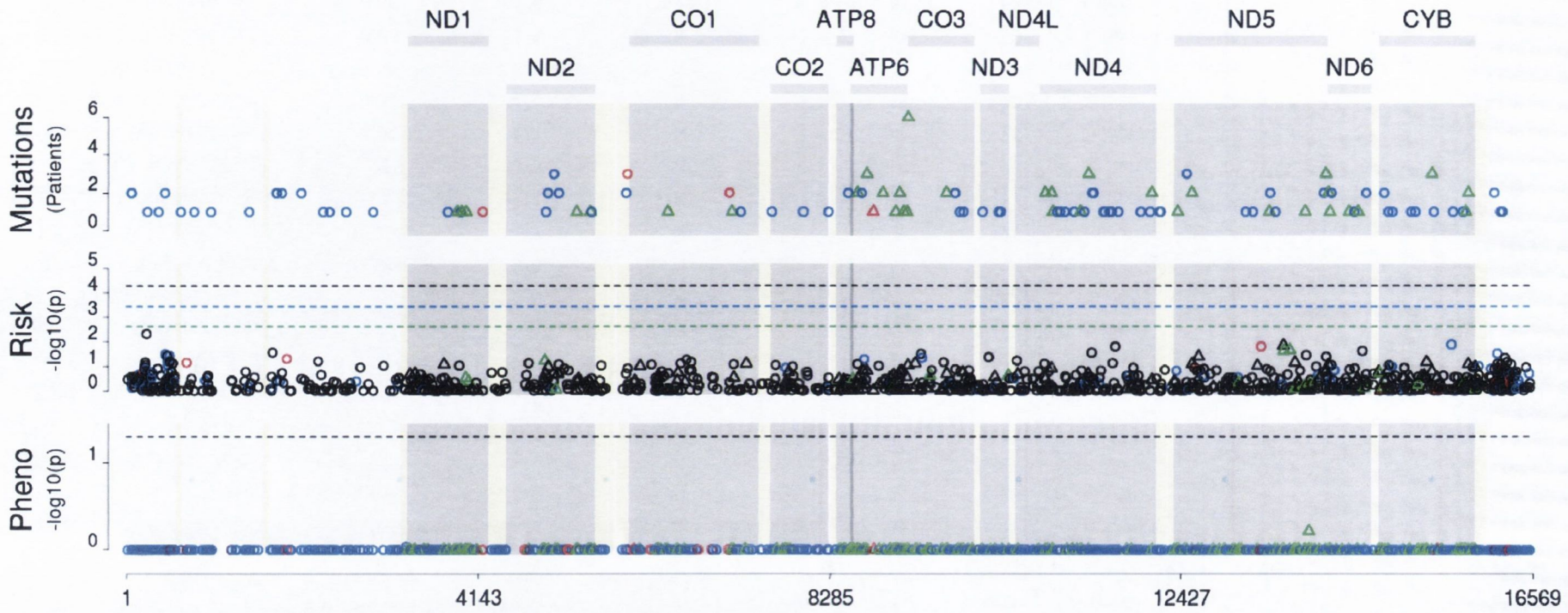


Fig 7.3: Single variant analyses. Panel 1 (Mutations): Observed carrier frequencies for variants detected exclusively within the case cohort (absent from controls and 1000 genomes). Panel 2 (Risk): P-values from single variant tests of association with disease risk. Dotted lines denote Bonferroni corrected thresholds for experiment-wide significance at a type I error rare <5% (black - all variants including those falling below the testable MAF threshold (2%), blue - variants with MAF > 2%, green nonsynonymous variants with MAF > 2%). Panel 3 (Pheno): Minimum p-value from single variant tests of association with clinically relevant patient parameters (age of onset, site of onset, status for cognitive impairment, status for FTD, status for executive cognitive dysfunction, status for non-executive cognitive dysfunction, survival, meta-analysis). P-values have been fully corrected for the total number of variants and clinical features analyzed. The dotted black line denotes the corrected threshold for experiment-wide significance at a type I error rare <5%. In panels 2-4, nonsynonymous variants are represented as green triangles; all other variants are represented as blue circles. Variants have been re-coloured red if assigned a status of "SUSPICIOUS" during QC and black (applies to panel 2 only) where the MAF fell below the testable threshold determined through power simulation. For all panels, grey shaded regions denote protein-coding sequence while yellow shaded regions denote tRNA coding sequence.

Table 7.1: Patient specific variants of protein-coding genes

Mutation(s)	PolyPhen	Mutated Reads	Known ALS Mutation	fALS	Gender	Age (yrs)	Site	Cognitive Phenotype	Survival (mnths)	Deceased
ATP6:c.199A>G(p.Thr67Ala)	0.012	2/8	N	N	M	48	bulbar	none	73	N
ATP6:c.199A>G(p.Thr67Ala)	0.012	7/7	N	N	F	66	spinal	FTD	35	Y
ATP6:c.199A>G(p.Thr67Ala), ND4:c.1325C>T(p.Ser442Phe)	0.012, 0.003	1/3; 3/3	N	N	F	47	spinal	none	38	Y
ATP6:c.277A>T(p.Thr93Ser)	0.987	7/7	N	N	M	78	spinal	NA	133	N
ATP6:c.394G>A(p.Gly132Ser)	0.999	4/4	N	N	M	68	bulbar	NA	24	Y
ATP6:c.394G>A(p.Gly132Ser)	0.999	8/8	N	N	M	55	spinal	NA	NA	N
ATP6:c.544T>G(p.Ser182Ala), ND4:c.148T>C(p.Phe50Leu)	0.195, 0.003	4/4; 3/3	N	N	F	52	bulbar	NA	29	Y
ATP6:c.593T>C(p.Leu198Pro)	0.998	1/5	N	N	F	72	spinal	NA	74	Y
ATP6:c.593T>C(p.Leu198Pro)	0.998	8/8	N	N	F	61	spinal	NA	18	Y
ATP6:c.640T>C(p.Phe214Leu)	0.98	14/14	N	N	M	58	bulbar	FTD	52	Y
ATP6:c.679T>A(p.X227Lys)	1	9/9	C9orf72	Y	F	82	bulbar	NA	12	Y
ATP6:c.93C>A(p.Ile31Met)	0.384	6/6	N	N	M	68	spinal	NA	32	Y
ATP6:c.93C>A(p.Ile31Met)	0.384	1/1	N	N	F	74	spinal	NA	35	Y
CO1:c.1226A>G(p.Tyr409Cys)	0.012	11/11	N	N	M	81	spinal	NA	13	Y
CO1:c.464T>C(p.Val155Ala)	0.265	3/3	N	N	M	36	spinal	NA	98	N
CO3:c.464A>G(p.Asn155Ser)	0.01	1/1	N	N	F	28	spinal	NA	94	Y
CO3:c.464A>G(p.Asn155Ser)	0.01	4/4	N	N	M	70	bulbar	NA	27	Y
CO3:c.8A>G(p.His3Arg)	0.66	5/5	N	N	F	69	spinal	NA	49	Y
CO3:c.8A>G(p.His3Arg)	0.66	1/1	N	N	M	59	bulbar	NA	NA	N
CO3:c.8A>G(p.His3Arg)	0.66	1/6	C9orf72	N	M	62	spinal	FTD	52	Y
CO3:c.8A>G(p.His3Arg)	0.66	1/10	N	N	F	69	bulbar	none	26	Y
CO3:c.8A>G(p.His3Arg)	0.66	1/6	C9orf72	N	F	44	spinal	none	38	Y

CO3:c.8A>G(p.His3Arg)	0.66	7/7	N	N	M	60	spinal	NA	21	Y
CYB:c.1028T>C(p.Val343Ala)	0.986	7/7	N	N	F	77	spinal	NA	43	Y
CYB:c.1057G>A(p.Val353Met)	0.145	1/2	N	N	M	57	spinal	NA	15	Y
CYB:c.1057G>A(p.Val353Met)	0.145	2/2	N	N	M	NA	spinal	NA	NA	N
CYB:c.631A>G(p.Ile211Val)	0.965	1/3	N	N	F	64	spinal	NA	25	Y
CYB:c.631A>G(p.Ile211Val)	0.965	1/1	N	N	M	43	spinal	NA	106	Y
CYB:c.631A>G(p.Ile211Val)	0.965	7/7	N	N	F	66	bulbar	NA	30	Y
ND1:c.586A>G(p.Thr196Ala)	0.012	5/5	N	N	M	57	spinal	none	55	N
ND1:c.637A>G(p.Ile213Val)	0.003	4/4	N	N	F	73	spinal	NA	20	Y
ND1:c.689A>G(p.Asn230Ser)	0.265	4/4	N	N	F	70	spinal	none	44	N
ND2:c.824G>A(p.Ser275Asn)	0.002	8/8	N	N	M	63	spinal	NA	19	Y
ND2:c.983C>T(p.Thr328Met)	0.057	12/12	N	N	M	58	spinal	none	44	N
ND4:c.1325C>T(p.Ser442Phe)	0.003	1/1	N	N	F	61	spinal	NA	28	Y
ND4:c.148T>C(p.Phe50Leu)	0.003	1/4	N	Y	M	67	spinal	NA	79	Y
ND4:c.163A>G(p.Thr55Ala)	0.014	4/4	N	N	M	NA	spinal	NA	NA	N
ND4:c.487G>T(p.Ala163Ser)	0.268	3/7	N	N	F	76	spinal	NA	26	Y
ND4:c.578A>G(p.Asn193Ser)	0.028	9/9	N	N	M	65	spinal	none	20	Y
ND4:c.578A>G(p.Asn193Ser)	0.028	18/18	N	Y	F	66	spinal	none	55	Y
ND4:c.578A>G(p.Asn193Ser)	0.028	12/12	N	N	M	49	spinal	NA	28	Y
ND4:c.73A>G(p.Ile25Val)	0.727	4/5	N	N	F	82	spinal	alzheimers	30	Y
ND4:c.73A>G(p.Ile25Val)	0.727	1/12	N	N	F	51	bulbar	none	43	Y
ND5:c.1130G>C(p.Ser377Thr)	0.639	5/5	N	N	M	57	spinal	executive dysfunction	39	N
ND5:c.1552T>C(p.Cys518Arg)	0	4/4	N	Y	M	78	spinal	NA	32	Y
ND5:c.1793C>T(p.Thr598Ile)	0.536	1/3	N	N	M	64	spinal	NA	40	Y
ND5:c.1793C>T(p.Thr598Ile)	0.536	1/6	N	N	M	59	spinal	executive dysfunction	27	Y
ND5:c.1793C>T(p.Thr598Ile)	0.536	5/5	N	N	M	50	spinal	NA	21	Y
ND5:c.221C>T(p.Thr74Ile)	0.005	1/5	N	Y	M	47	bulbar	NA	138	Y

ND5:c.221C>T(p.Thr74Ile)	0.005	9/9	N	Y	F	56	spinal	NA	94	N
ND5:c.53C>T(p.Pro18Leu)	0	7/7	C9orf72	N	M	59	bulbar	NA	25	Y
ND6:c.162A>C(p.Met54Ile)	0.008	4/4	N	N	F	80	bulbar	NA	18	Y
ND6:c.269T>C(p.Val90Ala)	0.002	4/4	N	N	M	47	spinal	NA	19	Y
ND6:c.494A>C(p.Tyr165Ser)	0.702	9/9	N	N	F	52	spinal	NA	13	Y
ND6:c.511G>A(p.Ala171Thr)	0.006	2/2	N	N	M	83	spinal	NA	38	Y
ND6:c.511G>A(p.Ala171Thr)	0.006	21/22	N	N	M	43	spinal	none	35	Y

Mutated reads Number of sequencing reads supporting mutant allele/ total number of sequencing reads observed

Table 7.2: Patient specific variants of tRNA genes

Mutations	Mutated Reads	Known ALS Mutation	fALS	Gender	Age (yrs)	Site	Cognitive Phenotype	Survival (mnths)	Deceased
TF:n.68A>G	4/4	N	N	M	42	spinal	NA	148	N
TH:n.36T>A	3/3	N	N	F	67	bulbar	NA	121	Y
TH:n.54C>T	4/4	N	N	F	NA	NA	NA	NA	N
TY:n.17G>A	1/5	N	N	M	76	spinal	NA	28	Y
TY:n.17G>A	6/6	N	N	F	76	bulbar	executive dysfunction	26	Y

Mutated reads Number of sequencing reads supporting mutant allele/ total number of sequencing reads observed

Table 7.3 Aggregate analyses of protein coding genes

Gene	Patient					Control	Disease Risk		Phenotype	
	Carriers	Age (yrs)	Site of onset	Cognitive status (impaired/tested)	Survival (months)	Carriers	Burden.fwe	All.fwe	Burden.fwe	All.fwe
ATP6	13	64+11	0.38	2/4	35+18	1	0.84	0.69	0.83	1.00
ATP8	0	-	-	0/0	-	0	1.00	1.00	1.00	0.90
CO1	2	58+32	0	0/0	13	3	1.00	1.00	1.00	0.99
CO2	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
CO3	8	58+15	0.38	1/3	44+25	4	1.00	1.00	1.00	0.99
CYB	6	61+13	0.17	0/0	44+36	4	1.00	0.95	1.00	1.00
ND1	3	67+9	0	0/2	20	0	1.00	1.00	1.00	1.00
ND2	2	60+4	0	0/1	19	1	1.00	1.00	1.00	1.00
ND3	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
ND4	11	62+12	0.18	1/5	38+18	0	6.6x10-2	0.12	1.00	1.00
ND4L	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
ND5	8	59+9	0.25	2/2	47+45	6	1.00	0.99	0.40	0.99
ND6	5	61+19	0.2	0/1	25+11	1	1.00	1.00	1.00	1.00

Patient Describes carriers of variants observed exclusively within the case cohort; Control Describes carriers of variants observed exclusively within the control cohort.. Age/ Survival Mean ± SD; Site of onset Proportion of cases with bulbar onset; Burden.fwe Family wise error rate for association observed in rare variant burden analysis; All.fwe Family wise error rate for most significant association observed across extended set of aggregate analyses.

Table 7.4: Aggregate Analyses of tRNA genes

Gene	Patient					Control	Disease Risk		Phenotype	
	Carriers	Age (yrs)	Site of onset	Cognitive status (impaired/tested)	Survival (months)	Carriers	Burden.fwe	All.fwe	Burden.fwe	All.fwe
TA	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TC	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TD	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TE	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TF	1	42	0	0/0	-	0	1.00	1.00	1.00	1.00
TG	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TH	2	67	1	0/0	121	2	1.00	1.00	1.00	1.00
TI	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TK	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TL1	0	-	-	0/0	-	0	1.00	1.00	1.00	0.88
TL2	0	-	-	0/0	-	1	1.00	1.00	1.00	1.00
TM	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TN	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TP	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TQ	0	-	-	0/0	-	1	1.00	1.00	1.00	1.00
TR	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TS1	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TS2	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TT	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TV	0	-	-	0/0	-	0	1.00	1.00	1.00	1.00
TW	0	-	-	0/0	-	1	1.00	1.00	1.00	1.00
TY	2	76+0	0.5	1/1	27+1	0	1.00	1.00	1.00	1.00

Patient Describes carriers of variants observed exclusively within the case cohort; Control Describes carriers of variants observed exclusively within the control cohort.. Age/ Survival Mean $\pm$ SD; Site of onset Proportion of cases with bulbar onset; Burden.fwe Family wise error rate for association observed in rare variant burden analysis; All.fwe Family wise error rate for most significant association observed across extended set of aggregate analyses.

7.4 Discussion

The results of our analyses demonstrate that off-target data generated during targeted high-throughput resequencing projects can be used to conduct comprehensive mitochondrial DNA screening. The depths of sample coverage obtained would not generally have been sufficient for reliably assaying variants that occur at low heteroplasmy but would have been suitably adequate for the detection of variants occurring at high copy number. Our study involved analyses of datasets generated using customized sequence enrichment kits which had been found to exhibit lower efficiency than might be expected with commercially available exome enrichment kits(section 4). The relative amount of off-target and hence mitochondrial DNA data generated in exome studies is therefore likely to be notably less, however this may be somewhat offset by the fact that substantially higher numbers of sequencing reads are generated per sample.

One of the primary objectives pursued in this section was to assess whether mitochondrial DNA variants implicated in ALS-like disorders also contribute to ALS aetiology. We observed a variant previously associated with ALS by Lin *et al.* to occur with higher frequency among controls than cases suggesting that it is unlikely to actually be disease causing. We did not observe the variants reported by Borthwick *et al* or Comi *et al*. This could reflect that they represent very rare causes of ALS or that they are unlikely to be encountered within a more strictly defined ALS case-control cohort. However, it should also be noted that the evidence supporting the pathogenicity of these variants is limited. The pathogenic relevance of the 35 nonsynonymous and 4 tRNA variants observed exclusively within the case cohort is also uncertain, as MAF were too low to reliably assess independent effects. Nonetheless, our results outline a maximum potential contribution for high penetrance mitochondrial DNA variants to ALS burden, namely 13% of patients for rare nonsynonymous mutations and 1% of patients for rare tRNA gene variants.

Of the 39 case-only nonsynonymous or tRNA gene variants, 16 were notable for having been observed across multiple patients. The most frequent of these appeared to be a *CO3:*c.8A>G(p.His3Arg) mutation that was detected within 6 patients. These patients appeared to include 3 homoplasmic and 3 heteroplasmic (variant allele observed in 1/6-1/10 sequencing reads per sample; Table 1) carriers, but Sanger sequencing could only validate the presence of the mutation within the cases assigned homoplasmic genotypes. It could be that the remaining individuals were erroneously called as heteroplasmic due to a recurrent sequencing artefact at a true variant site, however an alternative and perhaps more likely explanation is that the variant was simply carried at too low a copy number for detection by standard PCR-Sanger sequencing protocols. This in turn raises the question of how biologically relevant carrying a variant at such low heteroplasmy is likely to be? Typically, the phenotypic effects of mitochondrial variants are observed only at heteroplasmies >60-90%[174], but certain variants have been reported to cause disease at much lower levels[175]. Furthermore, the potential for heteroplasmy across tissues should also be noted, meaning that low copy number in blood does not necessarily infer low copy number across other cell types. It is also worth noting that if the *CO3:*c.8A>G(p.His3Arg) mutation does indeed occur at low copy number, it's presence within additional carriers could easily have been missed. As indicated a limitation of our study is that it was not powered to meaningful factor heteroplasmy into most analyses. This could prove to be an important issue and one which argues for the use of deeper sample resequencing in future analyses.

Although the independent effects of rare variants (MAF <2%) could not be reliably quantified, the study was powered to characterize those of more common variation. Furthermore, gene based aggregate analyses allowed for the detection of a potential association between disease risk and the overall burden of rare nonsynonymous *ND4* variants. After multiple testing correction, the null probability of this association was estimated to be $6.6x10^{-2}$, with 11 of the 440 cases (2.5%) having been observed to carry case-specific variants and 0 of the 301 controls (0%) having being observed to carry control specific variants. No individual variant or other aggregate grouping exhibited an association as suggestive as that observed across *ND4*. However, the fact that variants of *ND5* achieved the top associations in single variant analyses of disease risk, single variant analyses of patient phenotype and aggregate analyses of patient phenotype is notable, particularly since these signals concerned different *ND5* variants. It is also notable that for both aggregate and single variant analyses, the observed phenotypic associations concerned higher than expected burdens of patients with cognitive impairment. However, analyses of further patient cohorts will be required to definitively determine whether or not these trends may represent genuine disease associations.

7.5 Conclusion

Our results demonstrate the utility of harnessing off-target data from exon-resequencing studies for the purposes of mitochondrial DNA screening, outline the maximum potential contribution of high penetrance mitochondrial DNA variants to ALS burden and reveal the MAF and potential effect sizes of mitochondrial variants observable among ALS patients. In particular, our analyses highlight the genes *ND5* and more notably *ND4,* as candidate disease loci for further investigation.

8. DISCUSSION AND FUTURE WORK

8.1 Extending statistical analyses of known and candidate ALS genes

The greatest challenge faced over the course of the works described in this thesis was attempting to overcome limited statistical power in case-control association analyses. The difficulty was driven primarily by the heterogeneity of the patient population, the low MAF and potential for low effect sizes among many variants of potential interest, non-negligible background gene mutation rates, multiple testing burdens and above all the limited size of available sample cohorts. Sections 2, 5, 6 and 7 involved considerable investigation into the clinical relevance of known and candidate ALS genes through a combination of single variant and/ or gene aggregate analyses. In each section power to detect single variant associations with disease risk ranged from modest to negligible, and no significant independent effects could be confirmed. Overall aggregate analyses appeared to prove more useful for the purposes of highlighting candidate disease loci (sections 6,7) and studying population based disease heterogeneity (section 5). However, the lack of positive associations with known ALS genes in section 5, despite the presence of known as well as simply promising candidate disease mutations, highlights that even with aggregation strategies; association analyses were still considerably underpowered. Extending the analyses conducted in this work to larger samples cohorts will therefore be required to obtain a more complete understanding of the relevance of investigated loci. A promising start to this might be to mine genotype data from public repositories, however as highlighted in section 6, this will require very careful control of analysis as well as population based stratification issues. Deeper patient phenotyping through more detailed family history collection, molecular biomarker analyses and patient imaging could also serve to boost detection power at loci that exhibit specific or atypical genotype-phenotype correlations.

One aspect in which single variant analyses did prove somewhat informative was in exploring the applicability of monogenic disease models, though as discussed in section 2, the employed methodology had several important limitations. This was a specific objective in section 2 where their use was therefore justifiable but their use in highlighting candidate disease variants in sections 5-7 could be criticised, particularly since in sections 1-3, it was suggested that high penetrance is likely to apply to only a very small percentage of ALS related variants. This is a fair point and conventional association analyses that are free from prior assumptions about variant pathogenicity would constitute a preferred method for variant prioritization in an ideal setting. However, it should be noted that for a given MAF, high impact variants would be anticipated to be the most amenable to replication in subsequent association analyses or in functional validation assays (section 8.2). There is therefore a certain practical merit to highlighting potential instances of this mutation class.

8.2 Biochemical characterization of candidate disease variants

Given the extensive heterogeneity in ALS and the importance of both low penetrance and low frequency variation, it may prove that analyses of relative variant frequencies or gene mutation rates alone, will not be sufficient to confirm the relevance of certain disease loci. In such instances, progress might still be achievable through biochemical characterization of candidate

variants. Much is known about the pathological and molecular hallmarks of ALS. Well characterized disease features include progressive paralysis, motor neuron loss, the accumulation of ubiquitinated neuronal inclusions containing TDP-43 and/or other disease related proteins, more generalized abnormalities in the localization and post-translational modification of certain disease related proteins and several less well defined features such as the accumulation of RAN translation products in *C9orf72* based models. Several quantifiable endpoints are therefore available for use in both *in vitro* and *in vivo* mutational screening and as research into established disease variants progresses; further such endpoints may become available. Assessing variant pathogenicity through functional screening involves many obstacles which may not always be easily surmounted. Firstly there is a trade off between model throughput and complexity. For example, transfection of mutant constructs into yeast or human cell lines is in itself relatively scalable. Likewise advances in microscopy staging and automation systems, image analysis software, RNA sequencing technologies and mass spectroscopy are allowing for increasingly comprehensive and high throughput analyses of protein localization, protein state and gene expression profiles. However, how well a complex multisystem disorder like ALS can be modelled using such model systems is uncertain and many phenotypic endpoints of interest, such as paralysis, cannot be assessed. More complex cultures that include for example primary motor neurons and supporting glial cell could prove more relevant but may still not be sufficient to recapitulate relevant disease features and are also less scalable. *In vivo* models such as flies, zebra fish or mice are more complex again but suffer from even lower throughput, as well as higher costs. Furthermore, if ALS is typically a late onset disorder and mutations generally do not act with high penetrance, it may be that the expression of disease related features will require the presence of multiple genetic/ environmental stimuli. A common strategy to attempt to overcome this in the past has included transgenic over-expression of mutant constructs at extremely high levels, so as to exacerbate toxic effects, but the physiological relevance of this is widely disputed. In addition to testing the effects of candidate variants identified through mapping analyses, functional assays could also serve as important precursors to the mapping process itself. For example the association between *ATXN2* and human ALS was observed only after the yeast orthologue (*Pbp1*) was discovered to represent a modifier of TDP-43 toxicity in yeast and Drosophila during an unbiased over/under expression screening assay[19].

8.3 Candidate gene studies – Pros and cons

Candidate gene resequencing represented one of the core components of the work described in this thesis. Some discussion on the pros and cons of candidate gene approaches would therefore seem warranted, particularly given the availability of seemingly more comprehensive mapping strategies like linkage analyses, GWAS and exome resequencing.

In the vast majority of cases, the lack of a sufficient cohort of informative family members means that it is not possible to map ALS causing mutations through linkage or other family-orientated analyses. This reduces available statistical genetics methodologies to population based approaches like GWAS, exome wide association studies and candidate gene analyses. GWAS have helped provide significant insights into ALS aetiology, with notable examples including the replicated associations of variants in *DPP6* and *UNC13A* with disease susceptibility[41-43,63,176]. However, in the absence of appropriate methods for patient stratification, the successful

application of GWAS in conditions like ALS requires very large sample cohorts to overcome the effects of genetic heterogeneity. Achieving these sample numbers is extremely difficult given the frequency of the disease and it is therefore no surprise that most success stories in ALS GWAS have been the fruit of international collaborations. Another potentially important limitation with GWAS, is that where based on genotyping assays as opposed to whole genome sequencing, it is generally not anticipated to assay causal variants directly. The probability of success therefore depends upon the extent to which the effects of disease variants can be captured through linkage disequilibrium. Improvements in the density of SNP arrays and the development of reliable imputation procedures is allowing for increasingly comprehensive screening but it may prove that in certain situations, the only way to the effects of disease variants is through direct genotyping. An example of this might be where mutations tend to occur *de novo,* such as *FUS*:c.1574C>T(p.[P525L]) [17], as such variants would not be anticipated to exhibit consistent LD profiles. Whole genome sequencing has the power to eliminate this concern, but is highly expensive. Furthermore, a full GWAS that involved testing of every observed variant would incur an extremely high multiple testing burden, necessitating the analysis of potentially unachievable sample cohorts. Downgrading to an exome wide association analysis has the disadvantage of losing power to detect effects at non-exonic variants but the major advantages of substantial reductions in multiple testing burdens and per sample analysis costs. Even so the size of the exome would still necessitate analysis of very large sample cohorts and the cost of exome sequencing remains relatively high.

Focused candidate gene resequencing incurs an even lower multiple testing burden and per sample costs than exome sequencing. In this study, 1.77Mb of sequence was analyzed across 248 cases and 167 controls. The amount of sequence targeted by most commercially available exome enrichment kits is approximately 30Mb, meaning that exome sequencing would have resulted in a 17 fold increase in the genomic search space. To maintain equivalent power with the same case-control ratio and under simple Bonferroni extrapolation of significance thresholds, the analysis would have had to have been extended to approximately 4216 cases and 2839 controls. This might be achievable through international collaboration but represents an unrealistic objective for an analysis of the Irish population alone. Furthermore, it would have resulted in a 288 fold increase in sequencing requirements. There are therefore considerable advantages to pursuing candidate gene approaches where appropriate methods for candidate gene selection can be devised.

As a final point, it is also worth addressing the fact that a high proportion of reported candidate gene associations fail to replicate in subsequent analyses[177]. This could be interpreted as reflecting some kind of intrinsic limitation in the candidate gene approach but is more likely to reflect how candidate gene studies are commonly executed and/or interpreted. Often, reported associations have been based on the observations of weak statistical evidence and in such cases reasonable probabilities of error should be anticipated. Additionally, there are numerous instances where reported statistical modelling may not have been entirely appropriate, a reasonably common example among earlier studies having been to fail to account for multiple testing. Population structure may also have represented an important confounder in many studies. Correcting for such structure in candidate gene studies can be difficult but should still be achievable if the set of analyzed genes is large enough. Otherwise the use of strict sample ascertainment procedures or additional genotype data may be required. Finally, it should also be highlighted that the inability to replicate many associations may at times simply reflects power within replication study or genuine population based differences in disease aetiology and/or LD

patterns. This could for instance account for why the rs12608932 variant in *UNC13A* has been observed to associate with ALS in a large collaborative GWAS[63] and an independent analysis of the Dutch population[176] but not within analyses of French and Italian populations[178,179].

8.4 Screening for ALS related structural variation

Despite that the most common ALS variants identified to date have represented repeat expansions[19,23,24], research into the role of SV in ALS aetiology has been comparatively limited[41,180,181]. This primarily reflects that SV has been inherently difficult assay with most available technologies. In an effort to overcome the limitations of conventional approaches, we are currently conducting a pilot study for the use of mate pair sequencing in genome-wide SV screening. This approach is most commonly used in genome completion and the assembly of highly repetitive regions[182] but has also been used to explore genomic instability in cancer[183] and the basis of certain developmental disorders_ENREF_168. The approach constitutes a specialized type of paired end sequencing where insert sizes of several kilobases, as opposed to ~400bp, are achieved through an extended library preparation protocol (http://www.illumina.com/technology/mate_pair_sequencing_assay.ilmn). The advantage of the strategy over conventional paired end sequencing is that a given genomic spanning coverage can be achieved with substantially less sequencing (Figure 8.1). Furthermore the approach can allow for more detailed characterization of larger SV classes. The major disadvantage is that the library preparation protocol is substantially more involved and expensive. Furthermore relatively large amounts ($\geq$10 ug) of high quality input DNA are required. Commercially available mate pair library preparation kits, cost ~€300 per sample and so a "home-made" alternative is currently being developed. The results from an initial pilot run are anticipated at the beginning of next year and will involve efforts to characterize the SV across a *C9orf72* repeat expansion carrier and a patient that has tested negative for all disease mutations across the ALS genes described in section 5.

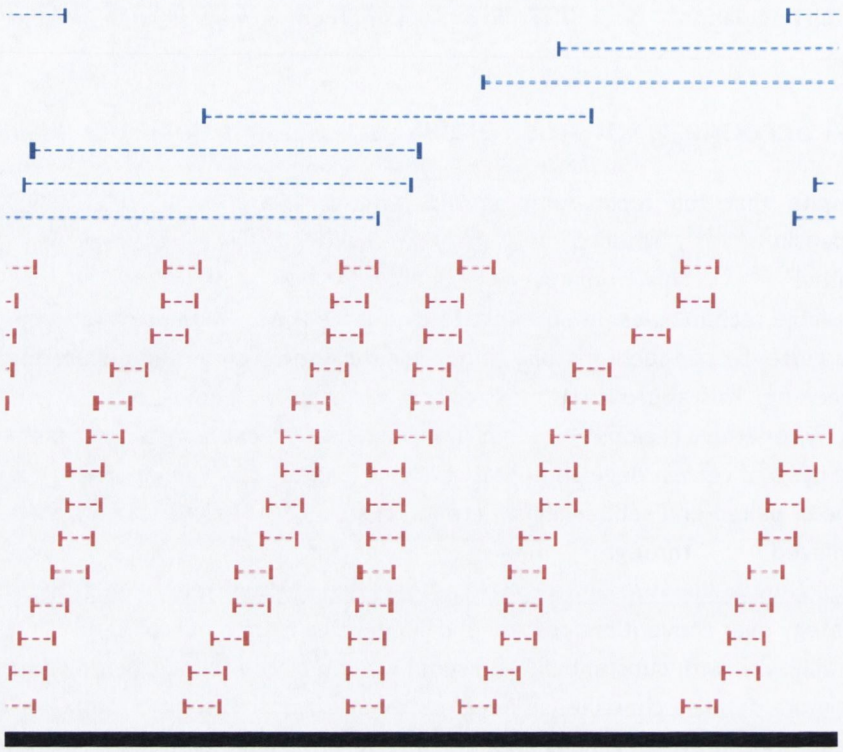


Figure 8.1: Spanning coverage. Outlines the number of sequencing reads required to achieve a mean of 3X coverage across a simulated 10Kb interval (black). With standard paired end sequencing (insert size=400bp) 60 read pairs are required (red). With mate pair sequencing (insert size=5Kb), only 6 read pairs are required (blue), corresponding to an ~10 fold reduction in sequencing.

8.5 Design of future gene mapping efforts

As it stands, identified ALS variants account for a considerable though loosely quantified proportion of disease risk. Obtaining an accurate and complete understanding of this semi-explained component is currently an area of vital importance and will require an in depth analysis of relevant loci in larger cohorts than have been studied so far. Given the allelic heterogeneity associated with these loci and the importance of rare variation, such analyses would be better served through gene resequencing as opposed to simple genotyping of previously identified variants. A full understanding may also require the identification of relevant environmental factors as well as a deeper exploration of the relevance of oligogenic disease inheritance. Determining the unidentified component of ALS genetic risk is likely to require a combination of different methodologies. As discussed in section 1.1.7, common variation appears to account for ~12% of phenotypic variance[66]. Further delineation of this component is likely to be achievable by extending traditional SNP array based GWAS to larger sample cohorts, and it is worth noting that this has been proving with various other common-relatively common disease phenotypes[184,185]. The major hurdle to will be obtaining the required sample sizes but this is likely to simply prove a matter of time and collaboration. The remainder of unidentified genetic risk is accounted for by

rare variants and/ or genetic interaction effects. As discussed above demonstrating independent effects for rare variants will often not be possible with achievable sample cohorts. Where such mutations concentrate in specific genes, domains, pathways or other definable aggregate units, burden and variance component based analyses may facilitate mapping. However, as highlighted by the results from section 5, it should still be expected that large sample cohorts will be required. Once again given the MAF involved and the probable relevance of allelic heterogeneity, SNP array based approaches are not likely to be sufficient. Instead whole genome or exome sequencing will likely be required. Currently sequencing costs allow for considerable more scalability with exome than genome resequencing. However, those costs are falling rapidly and as there is little reason to assume that disease related variation must be confined to coding sequence, whole genome sequencing should shortly represent the preferred option for at least a variant discovery component of future studies.

Where simple case-control analyses of variant frequencies are insufficient, incorporation of additional clinical data may allow for important gains in power. Likewise, it may also prove beneficial to recognize the overlap between the genetics of ALS and other disease phenotypes and attempt to exploit this through pooled sample analyses. Functional validation of candidate disease variants by assaying for effects on disease related phenotypic endpoints in cell or animal models may also play an increasing important role in future studies. This could prove particularly important for the purposes of identifying disease variants among sets of tightly linked SNPs or within loci found to exhibit associations with disease risk under aggregate analyses. As discussed above, the design and interpretation of suitably informative assays and the application of these assays in sufficient throughput are likely to represent considerable hurdles but may be surmountable. In certain situations it may also prove possible to empower mapping by taking into account prior knowledge or assumptions regarding disease models/ inheritance. For example, as with *FUS*:P525L carriers, *de novo* mutational occurrence could very well play an important role in severe early onset ALS phenotypes. In such cases there is also an increased probability that both parents will be alive at the time of diagnosis. This raises the potential to identify disease variants by conducting focused screens for *de novo* variation using genome-wide analyses of both patients and their unaffected parents. However, the low frequency of such phenotypes will likely mean that obtaining adequate statistical power still proves a considerable challenge. Furthermore, it should not be expected that early onset ALS phenotypes will necessarily involve *de novo* mutational events. For the purposes of exploring the relevance of genetic interaction effects, simplistic approaches such as pairwise tests of all variants assayed in GWAS, are likely to prove inadequate. However, alternative methodologies based on more advanced machine learning algorithms such recursive partitioning[186], which incorporate prior biological or which involve functional validation of candidate interactions may prove effective.

BIBLIOGRAPHY

1 Miller, R. G., Mitchell, J. D. & Moore, D. H. Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). *Cochrane database of systematic reviews* **3**, CD001447, doi:10.1002/14651858.CD001447.pub3 (2012).

2 Traynor, B. J. *et al.* Incidence and prevalence of ALS in Ireland, 1995-1997: a population-based study. *Neurology* **52**, 504-509 (1999).

3 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* **250**, 473-479 (2003).

4 Byrne, S. & Hardiman, O. Familial aggregation in amyotrophic lateral sclerosis. *Ann Neurol* **67**, 554, doi:10.1002/ana.21883 (2010).

5 Phukan, J. *et al.* The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study. *J Neurol Neurosurg Psychiatry* **83**, 102-108, doi:10.1136/jnnp-2011-300188 (2012).

6 Johnston, C. A. *et al.* Amyotrophic lateral sclerosis in an urban setting: a population based study of inner city London. *J Neurol* **253**, 1642-1643, doi:10.1007/s00415-006-0195-y (2006).

7 O'Toole, O. *et al.* Epidemiology and clinical features of amyotrophic lateral sclerosis in Ireland between 1995 and 2004. *J Neurol Neurosurg Psychiatry* **79**, 30-32 (2008).

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

9 Fang, F. *et al.* Familial aggregation of amyotrophic lateral sclerosis. *Ann Neurol* **66**, 94-99, doi:10.1002/ana.21580 (2009).

10 Byrne, S. *et al.* Aggregation of neurologic and neuropsychiatric disease in ALS kindreds: A population based case controlled cohort study of Familial and Sporadic ALS. *Ann Neurol*, doi:10.1002/ana.23969 (2013).

11 Farrer, M. J. *et al.* DCTN1 mutations in Perry syndrome. *Nat Genet* **41**, 163-165, doi:10.1038/ng.293 (2009).

12 Moreira, M. C. *et al.* Senataxin, the ortholog of a yeast RNA helicase, is mutant in ataxia-ocular apraxia 2. *Nat Genet* **36**, 225-227, doi:10.1038/ng1303 (2004).

13 Sarfarazi, M. *et al.* Localization of the fourth locus (GLC1E) for adult-onset primary open-angle glaucoma to the 10p15-p14 region. *Am J Hum Genet* **62**, 641-652, doi:10.1086/301767 (1998).

14 Beck, J. *et al.* Large C9orf72 hexanucleotide repeat expansions are seen in multiple neurodegenerative syndromes and are more frequent than expected in the UK population. *Am J Hum Genet* **92**, 345-353, doi:10.1016/j.ajhg.2013.01.011 (2013).

15 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* **36**, 377-381 (2004).

16 Sacconi, S. *et al.* Patients with a phenotype consistent with facioscapulohumeral muscular dystrophy display genetic and epigenetic heterogeneity. *J Med Genet* **49**, 41-46 (2012).

17 Conte, A. *et al.* P525L FUS mutation is consistently associated with a severe form of juvenile amyotrophic lateral sclerosis. *Neuromuscul Disord* **22**, 73-75 (2012).

18 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).

19 Elden, A. C. *et al.* Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature* **466**, 1069-1075 (2010).

20 Hadano, S. *et al.* A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nat Genet* **29**, 166-173, doi:10.1038/ng1001-166 (2001).

21 Yang, Y. *et al.* The gene encoding alsin, a protein with three guanine-nucleotide exchange factor domains, is mutated in a form of recessive amyotrophic lateral sclerosis. *Nat Genet* **29**, 160-165 (2001).

22 Greenway, M. J. *et al.* ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nat Genet* **38**, 411-413 (2006).

23 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).

24 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).

25 Mitchell, J. *et al.* Familial amyotrophic lateral sclerosis is associated with a mutation in D-amino acid oxidase. *Proc Natl Acad Sci U S A* **107**, 7556-7561 (2010).

26 Puls, I. *et al.* Mutant dynactin in motor neuron disease. *Nat Genet* **33**, 455-456, doi:10.1038/ng1123 (2003).

27 Kwiatkowski, T. J. J. *et al.* Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* **323**, 1205-1208 (2009).

28 Vance, C. *et al.* Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science* **323**, 1208-1211 (2009).

29 Maruyama, H. *et al.* Mutations of optineurin in amyotrophic lateral sclerosis. *Nature* **465**, 223-226 (2010).

30 Wu, C.-H. *et al.* Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. *Nature* **488**, 499-503 (2012).

31 Chen, Y. Z. *et al.* DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *Am J Hum Genet* **74**, 1128-1135, doi:10.1086/421054 (2004).

32 Al-Saif, A., Al-Mohanna, F. & Bohlega, S. A mutation in sigma-1 receptor causes juvenile amyotrophic lateral sclerosis. *Ann Neurol* **70**, 913-919, doi:10.1002/ana.22534 (2011).

33 Rosen, D. R. *et al.* Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362**, 59-62, doi:10.1038/362059a0 (1993).

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

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

36 Nishimura, A. L. *et al.* A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. *Am J Hum Genet* **75**, 822-831 (2004).

37 Johnson, J. O. *et al.* Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron* **68**, 857-864 (2010).

38 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* **33**, 1345-1351, doi:10.1002/humu.22157 (2012).

39 Cox, L. E. *et al.* Mutations in CHMP2B in lower motor neuron predominant amyotrophic lateral sclerosis (ALS). *PLoS One* **5**, e9872 (2010).

40 Parkinson, N. *et al.* ALS phenotypes with mutations in CHMP2B (charged multivesicular body protein 2B). *Neurology* **67**, 1074-1077, doi:10.1212/01.wnl.0000231510.89311.8b (2006).

41 Blauw, H. M. *et al.* A large genome scan for rare CNVs in amyotrophic lateral sclerosis. *Hum Mol Genet* **19**, 4091-4099, doi:10.1093/hmg/ddq323 (2010).

42 Cronin, S. *et al.* A genome-wide association study of sporadic ALS in a homogenous Irish population. *Hum Mol Genet* **17**, 768-774 (2008).

43 van Es, M. A. *et al.* Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat Genet* **40**, 29-31, doi:10.1038/ng.2007.52 (2008).

44 Simpson, C. L. *et al.* Variants of the elongator protein 3 (ELP3) gene are associated with motor neuron degeneration. *Hum Mol Genet* **18**, 472-481, doi:10.1093/hmg/ddn375 (2009).

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

46 Dunckley, T. *et al.* Whole-genome analysis of sporadic amyotrophic lateral sclerosis. *The New England journal of medicine* **357**, 775-788, doi:10.1056/NEJMoa070174 (2007).

47 van Es, M. A. *et al.* ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study. *Lancet Neurol* **6**, 869-877, doi:10.1016/S1474-4422(07)70222-3 (2007).

48 Landers, J. E. *et al.* Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. *Proc Natl Acad Sci U S A* **106**, 9004-9009, doi:10.1073/pnas.0812937106 (2009).

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

50 Blauw, H. M. *et al.* NIPA1 polyalanine repeat expansions are associated with amyotrophic lateral sclerosis. *Hum Mol Genet* **21**, 2497-2502, doi:10.1093/hmg/dds064 (2012).

51 Al-Chalabi, A. *et al.* Deletions of the heavy neurofilament subunit tail in amyotrophic lateral sclerosis. *Hum Mol Genet* **8**, 157-164 (1999).

52 Annesi, G. *et al.* DJ-1 mutations and parkinsonism-dementia-amyotrophic lateral sclerosis complex. *Ann Neurol* **58**, 803-807, doi:10.1002/ana.20666 (2005).

53 Cronin, S., Greenway, M. J., Prehn, J. H. & Hardiman, O. Paraoxonase promoter and intronic variants modify risk of sporadic amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* **78**, 984-986, doi:10.1136/jnnp.2006.112581 (2007).

54 Saeed, M. *et al.* Paraoxonase cluster polymorphisms are associated with sporadic ALS. *Neurology* **67**, 771-776, doi:10.1212/01.wnl.0000227187.52002.88 (2006).

55 Slowik, A. *et al.* Paraoxonase gene polymorphisms and sporadic ALS. *Neurology* **67**, 766-770, doi:10.1212/01.wnl.0000219565.32247.11 (2006).

56 van Blitterswijk, M. *et al.* Rare and common paraoxonase gene variants in amyotrophic lateral sclerosis patients. *Neurobiol Aging* **33**, 1845.e1841-1843 (2012).

57 Wills, A. M. *et al.* A large-scale international meta-analysis of paraoxonase gene polymorphisms in sporadic ALS. *Neurology* **73**, 16-24, doi:10.1212/WNL.0b013e3181a18674 (2009).

58 Meyer, T. *et al.* Early-onset ALS with long-term survival associated with spastin gene mutation. *Neurology* **65**, 141-143, doi:10.1212/01.wnl.0000167130.31618.0a (2005).

59 Orlacchio, A. *et al.* SPATACSIN mutations cause autosomal recessive juvenile amyotrophic lateral sclerosis. *Brain* **133**, 591-598 (2010).

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

61 Rubino, E. *et al.* SQSTM1 mutations in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Neurology* **79**, 1556-1562, doi:10.1212/WNL.0b013e31826e25df (2012).

62 Ticozzi, N. *et al.* Mutational analysis reveals the FUS homolog TAF15 as a candidate gene for familial amyotrophic lateral sclerosis. *Am J Med Genet B Neuropsychiatr Genet* **156B**, 285-290 (2011).

63 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* **41**, 1083-1087, doi:10.1038/ng.442 (2009).

64 Awadalla, P. *et al.* Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am J Hum Genet* **87**, 316-324, doi:10.1016/j.ajhg.2010.07.019 (2010).

65 Chio, A. *et al.* A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. *Hum Mol Genet* **18**, 1524-1532, doi:10.1093/hmg/ddp059 (2009).

66 Fogh, I. *et al.* A genome-wide association meta-analysis identifies a novel locus at 17q11.2 associated with sporadic amyotrophic lateral sclerosis. *Hum Mol Genet*, doi:10.1093/hmg/ddt587 (2014).

67 van Blitterswijk, M. *et al.* Evidence for an oligogenic basis of amyotrophic lateral sclerosis. *Hum Mol Genet* **21**, 3776-3784 (2012).

68 Spencer, P. S. *et al.* Guam amyotrophic lateral sclerosis-parkinsonism-dementia linked to a plant excitant neurotoxin. *Science* **237**, 517-522 (1987).

69 Chio, A., Benzi, G., Dossena, M., Mutani, R. & Mora, G. Severely increased risk of amyotrophic lateral sclerosis among Italian professional football players. *Brain* **128**, 472-476, doi:10.1093/brain/awh373 (2005).

70 Horner, R. D. *et al.* Occurrence of amyotrophic lateral sclerosis among Gulf War veterans. *Neurology* **61**, 742-749 (2003).

71 Shibata, N. Transgenic mouse model for familial amyotrophic lateral sclerosis with superoxide dismutase-1 mutation. *Neuropathology : official journal of the Japanese Society of Neuropathology* **21**, 82-92 (2001).

72 Donnelly, C. J. *et al.* RNA toxicity from the ALS/FTD C9ORF72 expansion is mitigated by antisense intervention. *Neuron* **80**, 415-428, doi:10.1016/j.neuron.2013.10.015 (2013).

73 Al-Chalabi, A. *et al.* The genetics and neuropathology of amyotrophic lateral sclerosis. *Acta Neuropathol* **124**, 339-352, doi:10.1007/s00401-012-1022-4 (2012).

74 Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94**, 441-448 (1975).

75 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467 (1977).

76 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

77 Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum Mol Genet* **19**, R227-240, doi:10.1093/hmg/ddq416 (2010).

78 Liu, L. *et al.* Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology* **2012**, 251364, doi:10.1155/2012/251364 (2012).

79 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

80 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).

81 Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745-755 (2011).

82 Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-35 (2010).

83 Williams, K. *et al.* UBQLN2/ubiquilin 2 mutation and pathology in familial amyotrophic lateral sclerosis. *Neurobiol Aging* (2012).

84 Giardine, B. *et al.* PhenCode: connecting ENCODE data with mutations and phenotype. *Hum Mutat* **28**, 554-562, doi:10.1002/humu.20484 (2007).

85 Lill, C. M., Abel, O., Bertram, L. & Al-Chalabi, A. Keeping up with genetic discoveries in amyotrophic lateral sclerosis: The ALSoD and ALSGene databases. *Amyotroph Lateral Scler* **12**, 238-249 (2011).

86 den Dunnen, J. T. & Antonarakis, S. E. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* **15**, 7-12, doi:10.1002/(SICI)1098-1004(200001)15:1<7::AID-HUMU4>3.0.CO;2-N (2000).

87 Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res* **40**, D84-90 (2012).

88 McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-2070 (2010).

89 Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718-719 (2011).

90 Armon, C. Sports and trauma in amyotrophic lateral sclerosis revisited. *J Neurol Sci* **262**, 45-53 (2007).

91 Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69 (2012).

92 Gellera, C. *et al.* Identification of new ANG gene mutations in a large cohort of Italian patients with amyotrophic lateral sclerosis. *Neurogenetics* **9**, 33-40 (2008).

93 Fernandez-Santiago, R. *et al.* Identification of novel Angiogenin (ANG) gene missense variants in German patients with amyotrophic lateral sclerosis. *J Neurol* **256**, 1337-1342 (2009).

94 Wu, D. *et al.* Angiogenin loss-of-function mutations in amyotrophic lateral sclerosis. *Ann Neurol* **62**, 609-617 (2007).

95 Daoud, H. *et al.* Resequencing of 29 candidate genes in patients with familial and sporadic amyotrophic lateral sclerosis. *Arch Neurol* **68**, 587-593 (2011).

96 Millecamps, S. *et al.* Questioning on the role of D amino acid oxidase in familial amyotrophic lateral sclerosis. *Proc Natl Acad Sci U S A* **107**, E107; author reply E108 (2010).

97 Munch, C. *et al.* Point mutations of the p150 subunit of dynactin (DCTN1) gene in ALS. *Neurology* **63**, 724-726 (2004).

98 Corrado, L. *et al.* Mutations of FUS gene in sporadic amyotrophic lateral sclerosis. *J Med Genet* **47**, 190-194 (2010).

99 Del Bo, R. *et al.* Novel optineurin mutations in patients with familial and sporadic amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* **82**, 1239-1243 (2011).

100 Belzil, V. V. *et al.* Analysis of OPTN as a causative gene for amyotrophic lateral sclerosis. *Neurobiol Aging* **32**, 555.e513-554 (2011).

101 Hirano, M. *et al.* Senataxin mutations and amyotrophic lateral sclerosis. *Amyotroph Lateral Scler* **12**, 223-227 (2011).

102 Andersen, P. M. *et al.* Phenotypic heterogeneity in motor neuron disease patients with CuZn-superoxide dismutase mutations in Scandinavia. *Brain* **120 (Pt 10)**, 1723-1737 (1997).

103 Corrado, L. *et al.* High frequency of TARDBP gene mutations in Italian patients with amyotrophic lateral sclerosis. *Hum Mutat* **30**, 688-694 (2009).

104 Kabashi, E. *et al.* TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nat Genet* **40**, 572-574 (2008).

105 Ferraiuolo, L., Kirby, J., Grierson, A. J., Sendtner, M. & Shaw, P. J. Molecular pathways of motor neuron injury in amyotrophic lateral sclerosis. *Nature reviews. Neurology* **7**, 616-630, doi:10.1038/nrneurol.2011.152 (2011).

106 Turner, F. S., Clutterbuck, D. R. & Semple, C. A. M. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* **4**, R75 (2003).

107 Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J. & Pickard, B. S. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* **6**, 55, doi:10.1186/1471-2105-6-55 (2005).

108 Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nat Biotechnol* **24**, 537-544, doi:10.1038/nbt1203 (2006).

109 Kohler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* **82**, 949-958, doi:10.1016/j.ajhg.2008.02.013 (2008).

110 Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J. & Pickard, B. S. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* **22**, 773-774 (2006).

111 Chen, J., Xu, H., Aronow, B. J. & Jegga, A. G. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics* **8**, 392 (2007).

112 Chen, J., Aronow, B. J. & Jegga, A. G. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* **10**, 73 (2009).

113 Walsh, D. M. & Selkoe, D. J. A beta oligomers - a decade of discovery. *Journal of neurochemistry* **101**, 1172-1184, doi:10.1111/j.1471-4159.2006.04426.x (2007).

114 Bader, G. D., Betel, D. & Hogue, C. W. V. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-250 (2003).

115 Stark, C. *et al.* The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* **39**, D698-704 (2011).

116 Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-772 (2009).

117 Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **39**, D52-57, doi:10.1093/nar/gkq1237 (2011).

118 Aranda, B. *et al.* The IntAct molecular interaction database in 2010. *Nucleic Acids Res* **38**, D525-531 (2010).

119 Ceol, A. *et al.* MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res* **38**, D532-539 (2010).

120 Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, D449-451 (2004).

121 Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* **39**, D691-697 (2011).

122 Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res* **37**, D674-679 (2009).

123 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

124 Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**, D130-135, doi:10.1093/nar/gkr1079 (2012).

125 Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**, W305-311 (2009).

126 White, S. & Smyth, P. Algorithms for Estimating Relative Importance in Networks. *In KDD'03:Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining ACM Press*, 266-275 (2003).

127 Popescu, M., Keller, J. M. & Mitchell, J. A. Fuzzy measures on the Gene Ontology for gene product similarity. *IEEE/ACM Trans Comput Biol Bioinform* **3**, 263-274 (2006).

128 Le Ber, I. *et al.* Chromosome 9p-linked families with frontotemporal dementia associated with motor neuron disease. *Neurology* **72**, 1669-1676 (2009).

129 Hosler, B. A. *et al.* Linkage of familial amyotrophic lateral sclerosis with frontotemporal dementia to chromosome 9q21-q22. *JAMA* **284**, 1664-1669 (2000).

130 Hand, C. K. *et al.* A novel locus for familial amyotrophic lateral sclerosis, on chromosome 18q. *Am J Hum Genet* **70**, 251-256 (2002).

131 Sapp, P. C. *et al.* Identification of two novel loci for dominantly inherited familial amyotrophic lateral sclerosis. *Am J Hum Genet* **73**, 397-403 (2003).

132 Goh, K.-I. *et al.* The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-8690 (2007).

133 Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci U S A* **105**, 4323-4328 (2008).

134 Yandell, M. *et al.* Genome-wide analysis of human disease alleles reveals that their locations are correlated in paralogous proteins. *PLoS Comput Biol* **4**, e1000218 (2008).

135 Bornigen, D. *et al.* An unbiased evaluation of gene prioritization tools. *Bioinformatics* **28**, 3081-3088, doi:10.1093/bioinformatics/bts581 (2012).

136 Kenny, E. M. *et al.* Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Res* **18**, 31-38 (2011).

137 Craig, D. W. *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* **5**, 887-893 (2008).

138 Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-189 (2009).

139 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102. Article published online before print in May 2002 (2002).

140 Byrne, S. *et al.* Proposed criteria for familial amyotrophic lateral sclerosis. *Amyotroph Lateral Scler* **12**, 157-159, doi:10.3109/17482968.2010.545420 (2011).

141 Cock, P. J., 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, doi:10.1093/nar/gkp1137 (2010).

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

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

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

145 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).

146 Picardi, E. & Pesole, G. Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat Methods* **9**, 523-524, doi:10.1038/nmeth.2029 (2012).

147 Chio, A. *et al.* Extensive genetics of ALS: A population-based study in Italy. *Neurology* **79**, 1983-1989 (2012).

148 Lattante, S. *et al.* Contribution of major amyotrophic lateral sclerosis genes to the etiology of sporadic disease. *Neurology* **79**, 66-72, doi:10.1212/WNL.0b013e31825dceca (2012).

149 van Es, M. A. *et al.* Large-scale SOD1 mutation screening provides evidence for genetic heterogeneity in amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* **81**, 562-566 (2010).

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

151 Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93, doi:10.1016/j.ajhg.2011.05.029 (2011).

152 Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet* **7**, e1001322 (2011).

153 Kirby, J. *et al.* Broad clinical phenotypes associated with TAR-DNA binding protein (TARDBP) mutations in amyotrophic lateral sclerosis. *Neurogenetics* **11**, 217-225 (2010).

154 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081 (2009).

155 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249 (2010).

156 Byrne, S. *et al.* Cognitive and clinical characteristics of patients with amyotrophic lateral sclerosis carrying a C9orf72 repeat expansion: a population-based cohort study. *Lancet Neurol* (2012).

157 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).

158 Brown, J. A. *et al.* SOD1, ANG, TARDBP and FUS mutations in amyotrophic lateral sclerosis: a United States clinical testing lab experience. *Amyotroph Lateral Scler* **13**, 217-222 (2012).

159 Niemann, S. *et al.* Familial ALS in Germany: origin of the R115G SOD1 mutation by a founder effect. *J Neurol Neurosurg Psychiatry* **75**, 1186-1188 (2004).

160 Millecamps, S. *et al.* SOD1, ANG, VAPB, TARDBP, and FUS mutations in familial amyotrophic lateral sclerosis: genotype-phenotype correlations. *J Med Genet* **47**, 554-560 (2010).

161 International HapMap, C. The International HapMap Project. *Nature* **426**, 789-796, doi:10.1038/nature02168 (2003).

162 Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res* **21**, 830-839, doi:10.1101/gr.115907.110 (2011).

163 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

164 Heinze, G. & Schemper, M. A solution to the problem of separation in logistic regression. *Statistics in medicine* **21**, 2409-2419, doi:10.1002/sim.1047 (2002).

165 Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**, e1000384 (2009).

166 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*, doi:10.1016/j.ajhg.2013.04.015 (2013).

167 van Es, M. A. *et al.* Angiogenin variants in Parkinson disease and amyotrophic lateral sclerosis. *Ann Neurol* **70**, 964-973 (2011).

168 Sasaki, S. & Iwata, M. Mitochondrial alterations in the spinal cord of patients with sporadic amyotrophic lateral sclerosis. *Journal of neuropathology and experimental neurology* **66**, 10-16, doi:10.1097/nen.0b013e31802c396b (2007).

169 Wiedemann, F. R., Manfredi, G., Mawrin, C., Beal, M. F. & Schon, E. A. Mitochondrial DNA and respiratory chain function in spinal cords of ALS patients. *Journal of neurochemistry* **80**, 616-625 (2002).

170 Duffy, L. M., Chapman, A. L., Shaw, P. J. & Grierson, A. J. Review: The role of mitochondria in the pathogenesis of amyotrophic lateral sclerosis. *Neuropathology and applied neurobiology* **37**, 336-352, doi:10.1111/j.1365-2990.2011.01166.x (2011).

171 Comi, G. P. *et al.* Cytochrome c oxidase subunit I microdeletion in a patient with motor neuron disease. *Ann Neurol* **43**, 110-116, doi:10.1002/ana.410430119 (1998).

172 Lin, F. H. *et al.* Detection of point mutations in codon 331 of mitochondrial NADH dehydrogenase subunit 2 in Alzheimer's brains. *Biochemical and biophysical research communications* **182**, 238-246 (1992).

173 Borthwick, G. M. *et al.* Motor neuron disease in a patient with a mitochondrial tRNAIle mutation. *Ann Neurol* **59**, 570-574, doi:10.1002/ana.20758 (2006).

174 Rossignol, R. *et al.* Mitochondrial threshold effects. *The Biochemical journal* **370**, 751-762, doi:10.1042/BJ20021594 (2003).

175 Sacconi, S. *et al.* A functionally dominant mitochondrial DNA mutation. *Hum Mol Genet* **17**, 1814-1820, doi:10.1093/hmg/ddn073 (2008).

176 Diekstra, F. P. *et al.* UNC13A is a modifier of survival in amyotrophic lateral sclerosis. *Neurobiol Aging* **33**, 630 e633-638, doi:10.1016/j.neurobiolaging.2011.10.029 (2012).

177 Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nat Genet* **29**, 306-309, doi:10.1038/ng749 (2001).

178 Daoud, H. *et al.* Analysis of the UNC13A gene as a risk factor for sporadic amyotrophic lateral sclerosis. *Arch Neurol* **67**, 516-517, doi:10.1001/archneurol.2010.46 (2010).

179 Chio, A. *et al.* UNC13A influences survival in Italian amyotrophic lateral sclerosis patients: a population-based study. *Neurobiol Aging* **34**, 357 e351-355, doi:10.1016/j.neurobiolaging.2012.07.016 (2013).

180 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* **4**, e8175, doi:10.1371/journal.pone.0008175 (2009).

181 Cronin, S. *et al.* Analysis of genome-wide copy number variation in Irish and Dutch ALS populations. *Hum Mol Genet* **17**, 3392-3398, doi:10.1093/hmg/ddn233 (2008).

182 van Heesch, S. *et al.* Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. *BMC Genomics* **14**, 257, doi:10.1186/1471-2164-14-257 (2013).

183 Hillmer, A. M. *et al.* Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. *Genome Res* **21**, 665-675, doi:10.1101/gr.113555.110 (2011).

184 Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**, 1150-1159, doi:10.1038/ng.2742 (2013).

185 European Alzheimer's Disease, I. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-1458, doi:10.1038/ng.2802 (2013).

186 Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* **10**, 392-404, doi:10.1038/nrg2579 (2009).

Appendix 1 – Supplemental results

Contents:

Table S2.1: ALSoD mutations.
Table S3.1: Functional enrichment results – tALS interactors
Table S3.2: Functional enrichment results –Shared interactors
Table S3.3: Functional profile – tALS sequence similarity hits
Table S3.4: Functional profile – shared tALS sequence similarity hits
Table S3.5: Functional profile – tALS genes
Table S4.1: Candidate genes
Figure S5.1 Variant detection power
Table S5.1: ALS gene variants
Table S5.2: Carriers of known and candidate ALS variants
Figure S6.1: Histograms outlining distribution of variant quality control statistics
Figure S6.2: Pair-wise distributions of variant quality control statistics.
Figure S6.3: Single variant analyses: age of disease onset.
Figure S6.4: Single variant analyses: site of disease onset
Figure S6.5: Single variant analyses: status for cognitive impairment.
Figure S6.6: Single variant analyses: status for executive dysfunction.
Figure S6.7: Single variant analyses: status for non-executive cognitive impairment.
Figure S6.8: Single variant analyses: status for FTD
Figure S6.9 Single variant analyses: patient Survival.
Table S6.1 Aggregate Analyses

Table S2.1: ALSoD Mutations

Mutation	dbSNP	EUR	AFR	AMR	ASN	US
ALS2:c.280A>G(p.Ile94Val) ALS2:c.305A>G(p.His102Arg)	rs3219154 .	4311, 218, 4	2015, 204, 3	172, 9, 0	286, 0, 0	5726, 399, 6
ALS2:c.1102G>A(p.Val368Met) ALS2:c.1619G>A(p.Gly540Glu)	rs3219156 .	43, 823, 3654	99, 764, 1300	1, 27, 153	0, 0, 286	120, 1439, 4500
ALS2:c.2143C>T(p.Gln715*) ALS2:c.2464G>A(p.Val822Met)	rs121908139 .					
ALS2:c.2992C>T(p.Arg998*) ALS2:c.4217G>A(p.Arg1406Lys)	rs121908137 .					
ANG:c.3G>A(p.Met1Ile) ANG:c.35T>C(p.Phe12Ser)	rs201068740 .					
ANG:c.61C>T(p.Pro21Ser)	rs149672657	4298, 2, 0	2203, 0, 0			6501, 2, 0
ANG:c.107A>T(p.Gln36Leu) ANG:c.112T>C(p.Tyr38His)	rs121909535 .					
ANG:c.122A>T(p.Lys41Ile)	rs121909536	4655, 23, 0	2445, 4, 0	181, 0, 0	286, 0, 0	6477, 26, 0
ANG:c.121A>G(p.Lys41Glu) ANG:c.137A>G(p.Asp46Gly)	rs121909537 .					

ANG:c.155G>A(p.Ser52Asn)	rs121909542					
ANG:c.164G>A(p.Arg55Lys)	rs121909538					
ANG:c.189C>G(p.Cys63Trp)	rs121909539					
ANG:c.191A>T(p.Lys64Ile)	rs121909540					
ANG:c.208A>G(p.Ile70Val)	rs121909541	4295, 5, 0	2201, 2, 0			6496, 7, 0
ANG:c.232A>G(p.Lys78Glu)	rs141055235	4299, 1, 0	2203, 0, 0			6502, 1, 0
ANG:c.176T>C(p.Leu59Pro)	rs11541242					
ANG:c.250A>G(p.Lys84Glu)	rs17560	4678, 0, 0	2381, 67, 1	180, 1, 0	286, 0, 0	6452, 51, 0
ANG:c.407C>T(p.Pro136Leu)	rs121909543	4299, 1, 0	2203, 0, 0			6502, 1, 0
ANG:c.409G>A(p.Val137Ile)	rs121909544					
ANG:c.413A>G(p.His138Arg)	.					
ANG:c.434G>A(p.Arg145His)	.					
ANG:c.433C>T(p.Arg145Cys)	.					
BCL11B:c.95A>T(p.Glu32Val)	.					
BCL11B:c.685C>T(p.Pro229Ser)	.					
BCL6:c.218A>G(p.Asn73Ser)	.					
CDH13:c.193C>T(p.Arg65Cys)	.	4182, 1, 0	2026, 0, 0			6208, 1, 0
CDH13:c.308C>T(p.Ala103Val)	.					
CDH13:c.337G>C(p.Gly113Arg)	rs183971768	4490, 17, 0	2178, 2, 0	181, 0, 0	286, 0, 0	6046, 17, 0

CDH13:c.736C>T(p.Arg246Trp)	.	4128, 1, 0	1920, 0, 0			6048, 1, 0
CDH13:c.1099G>C(p.Glu367Gln)	rs200000145	4155, 13, 0	1957, 0, 0			6112, 13, 0
CDH22:c.274G>A(p.Glu92Lys)	rs76517618	4677, 1, 0	2449, 0, 0	181, 0, 0	250, 35, 1	6502, 1, 0
CDH22:c.401C>T(p.Thr134Met)	.					
CDH22:c.1598G>A(p.Arg533His)	rs117065621	4647, 31, 0	2445, 4, 0	180, 1, 0	286, 0, 0	6471, 32, 0
CNTN6:c.940T>G(p.Phe314Val)	rs144474590					
CNTN6:c.2861A>T(p.Glu954Val)	.					
CRIM1:c.1217A>G(p.Asn406Ser)	rs139895660	4674, 4, 0	2449, 0, 0	181, 0, 0	285, 1, 0	6499, 4, 0
CRYM:c.505C>T(p.Arg169Cys)	rs148349291	4299, 1, 0	2199, 0, 0			6498, 1, 0
DAO:c.113G>A(p.Arg38His)	.	4299, 1, 0	2203, 0, 0			6502, 1, 0
DAO:c.595C>T(p.Arg199Trp)	rs139166976	4295, 5, 0	2202, 1, 0			6497, 6, 0
DCTN1:c.1712T>C(p.Met571Thr)	rs121909343					
DCTN1:c.2353C>T(p.Arg785Trp)	rs121909344	4293, 7, 0	2201, 2, 0			6494, 9, 0
DCTN1:c.3302G>A(p.Arg1101Lys)	rs121909345					
DCTN1:c.3746C>T(p.Thr1249Ile)	rs72466496	4639, 39, 0	2445, 4, 0	180, 1, 0	286, 0, 0	6460, 43, 0
DIAPH3:c.701T>C(p.Ile234Thr)	rs200839105	4059, 5, 0	1805, 0, 0			5864, 5, 0
DIAPH3:c.1733C>T(p.Pro578Leu)	rs76366906	4501, 7, 0	2176, 1, 0	180, 1, 0	268, 18, 0	6055, 6, 0

DIAPH3:c.1763C>T(p.Pro588Leu)	rs111260336	4145, 362, 7	2158, 28, 0	169, 11, 1	286, 0, 0	5710, 359, 7
DIAPH3:c.1787C>T(p.Pro596Leu)	rs150023947	4483, 47, 1	2182, 3, 0	180, 1, 0	286, 0, 0	6044, 47, 1
DIAPH3:c.3125G>A(p.Arg1042His)	rs200189161	4432, 5, 0	2045, 0, 0	181, 0, 0	286, 0, 0	5854, 4, 0
DIAPH3:c.3571C>T(p.Arg1191*)	.					
DOC2B:c.626G>T(p.Arg209Leu)	.					
FEZF2:c.563G>A(p.Gly188Asp)	rs199850439	4288, 5, 0	2199, 1, 0			6487, 6, 0
FIG4:c.143A>G(p.Asp48Gly)	.					
FIG4:c.157G>T(p.Asp53Tyr)	rs121908290					
FIG4:c.547C>T(p.Arg183*)	rs121908288					
FIG4:c.1162A>G(p.Arg388Gly)	.					
FIG4:c.1207C>T(p.Gln403*)	.					
FIG4:c.1231A>G(p.Ile411Val)	.					
FIG4:c.1940A>G(p.Tyr647Cys)	rs150301327	4297, 2, 0	2203, 0, 0			6500, 2, 0
FIG4:c.2705T>C(p.Ile902Thr)	.					
FUS:c.467G>A(p.Gly156Glu)	.					
FUS:c.571G>A(p.Gly191Ser)	rs148758737	4299, 1, 0	2197, 0, 0			6496, 1, 0
FUS:c.646C>T(p.Arg216Cys)	.					
FUS:c.674G>T(p.Gly225Val)	.					
FUS:c.688G>T(p.Gly230Cys)	.					
FUS:c.700C>T(p.Arg234Cys)	.					
FUS:c.701G>T(p.Arg234Leu)	.					
FUS:c.730C>T(p.Arg244Cys)	.					

FUS:c.760A>G(p.Met254Val)	.					
FUS:c.1483C>T(p.Arg495*)	.					
FUS:c.1520G>A(p.Gly507Asp)	.					
FUS:c.1529A>G(p.Lys510Arg)	.					
FUS:c.1542G>T(p.Arg514Ser)	.					
FUS:c.1540A>G(p.Arg514Gly)	.					
FUS:c.1543G>T(p.Gly515Cys)	.					
FUS:c.1551C>G(p.His517Gln)	rs121909667					
FUS:c.1549C>G(p.His517Asp)	.					
FUS:c.1553G>A(p.Arg518Lys)	rs121909669					
FUS:c.1561C>G(p.Arg521Gly)	rs121909668; rs121909670					
FUS:c.1562G>A(p.Arg521His)	rs121909671					
FUS:c.1561C>T(p.Arg521Cys)	rs121909668; rs121909670					
FUS:c.1561C>A(p.Arg521Ser)	.					
FUS:c.1562G>T(p.Arg521Leu)	.					
FUS:c.1564A>G(p.Arg522Gly)	.					
FUS:c.1571G>C(p.Arg524Thr)	.					
FUS:c.1572G>C(p.Arg524Ser)	.					
FUS:c.1574C>T(p.Pro525Leu)	.					
GRB14:c.1519C>T(p.His507Tyr)	rs144301087	4283, 17, 0	2203, 0, 0			6486, 17, 0
LUM:c.596T>C(p.Leu199Pro)	rs147975710	4636, 42, 0	2445, 4, 0	180, 1, 0	286, 0, 0	6460, 43, 0

NEFH:c.119C>T(p.Ala40Val)	.					
NEFH:c.1037G>A(p.Arg346His)	.					
NEFH:c.2601G>T(p.Lys867Asn)	rs138156220	4300, 0, 0	2175, 28, 0			6475, 28, 0
NEFH:c.2753A>G(p.Glu918Gly)	rs189881592	378, 0, 0	246, 0, 0	181, 0, 0	283, 3, 0	
NETO1:c.1201G>A(p.Gly401Arg)	rs149193005	4652, 26, 0	2444, 5, 0	181, 0, 0	286, 0, 0	6473, 30, 0
OMA1:c.205C>T(p.His69Tyr)	rs75220198	4658, 20, 0	2353, 94, 2	175, 6, 0	286, 0, 0	6397, 104, 2
OMA1:c.815A>G(p.Glu272Gly)	rs139938730	4285, 15, 0	2198, 5, 0			6483, 20, 0
OMA1:c.1093G>T(p.Asp365Tyr)	rs77980955	4634, 44, 0	2442, 7, 0	181, 0, 0	286, 0, 0	6457, 46, 0
OPTN:c.67G>T(p.Gly23*)	.					
OPTN:c.177G>C(p.Lys59Asn)	.					
OPTN:c.277G>C(p.Ala93Pro)	.					
OPTN:c.287G>T(p.Arg96Leu)	.					
OPTN:c.476G>T(p.Gly159Val)	.					
OPTN:c.493C>T(p.Gln165*)	.					
OPTN:c.811C>T(p.Arg271Cys)	.					
OPTN:c.844A>C(p.Thr282Pro)	.					
OPTN:c.941A>T(p.Gln314Leu)	rs142812715	4297, 3, 0	2203, 0, 0			6500, 3, 0
OPTN:c.1192C>T(p.Gln398*)	.					
OPTN:c.1360C>G(p.Gln454Glu)	.					
OPTN:c.1433A>G(p.Glu478Gly)	.					
OPTN:c.1442C>T(p.Ala481Val)	.	4299, 1, 0	2203, 0, 0			6502, 1, 0
OPTN:c.1670A>C(p.Lys557Thr)	.					
PCP4:c.172G>T(p.Ala58Ser)	.					

SOD1:c.137T>G(p.Phe46Cys)	rs121912457
SOD1:c.140A>G(p.His47Arg)	rs121912443
SOD1:c.139C>G(p.His47Asp)	.
SOD1:c.142G>T(p.Val48Phe)	.
SOD1:c.143T>C(p.Val48Ala)	.
SOD1:c.146A>G(p.His49Arg)	.
SOD1:c.147T>G(p.His49Gln)	.
SOD1:c.148G>A(p.Glu50Lys)	.
SOD1:c.164C>G(p.Thr55Arg)	.
SOD1:c.179G>T(p.Ser60Ile)	.
SOD1:c.184G>C(p.Gly62Arg)	.
SOD1:c.197A>G(p.Asn66Ser)	.
SOD1:c.199C>G(p.Pro67Ala)	.
SOD1:c.199C>T(p.Pro67Ser)	.
SOD1:c.200C>G(p.Pro67Arg)	.
SOD1:c.203T>C(p.Leu68Pro)	.
SOD1:c.203T>G(p.Leu68Arg)	.
SOD1:c.217G>A(p.Gly73Ser)	rs121912455
SOD1:c.217G>T(p.Gly73Cys)	.
SOD1:c.229G>T(p.Asp77Tyr)	.
SOD1:c.230A>T(p.Asp77Val)	.
SOD1:c.242A>G(p.His81Arg)	rs121912458
SOD1:c.251A>G(p.Asp84Gly)	.
SOD1:c.255G>C(p.Leu85Phe)	.
SOD1:c.253T>G(p.Leu85Val)	rs121912452

SOD1:c.256G>C(p.Gly86Arg)	rs121912436					
SOD1:c.256G>A(p.Gly86Ser)	.					
SOD1:c.260A>T(p.Asn87Ile)	.					
SOD1:c.260A>G(p.Asn87Ser)	rs11556620					
SOD1:c.259A>G(p.Asn87Asp)	.					
SOD1:c.261T>A(p.Asn87Lys)	.					
SOD1:c.263T>C(p.Val88Ala)	.					
SOD1:c.262G>A(p.Val88Met)	.					
SOD1:c.268G>A(p.Ala90Thr)	.					
SOD1:c.269C>T(p.Ala90Val)	.					
SOD1:c.272A>T(p.Asp91Val)	.					
SOD1:c.272A>C(p.Asp91Ala)	rs80265967	4669, 9, 0	2449, 0, 0	181, 0, 0	286, 0, 0	6496, 7, 0
SOD1:c.281G>T(p.Gly94Val)	.					
SOD1:c.281G>C(p.Gly94Ala)	rs121912438					
SOD1:c.280G>T(p.Gly94Cys)	rs121912437; rs121912460					
SOD1:c.280G>C(p.Gly94Arg)	rs121912437; rs121912460					
SOD1:c.281G>A(p.Gly94Asp)	.					
SOD1:c.280G>A(p.Gly94Ser)	.					
SOD1:c.284T>C(p.Val95Ala)	rs202198235					
SOD1:c.287C>T(p.Ala96Val)	.					
SOD1:c.286G>A(p.Ala96Thr)	.					

SOD1:c.287C>G(p.Ala96Gly)	.
SOD1:c.290A>T(p.Asp97Val)	.
SOD1:c.289G>A(p.Asp97Asn)	rs121912459
SOD1:c.292G>A(p.Val98Met)	.
SOD1:c.292G>C(p.Val98Leu)	.
SOD1:c.298A>G(p.Ile100Val)	.
SOD1:c.302A>G(p.Glu101Gly)	rs121912439
SOD1:c.301G>A(p.Glu101Lys)	.
SOD1:c.305A>G(p.Asp102Gly)	.
SOD1:c.304G>A(p.Asp102Asn)	.
SOD1:c.304G>T(p.Asp102Tyr)	.
SOD1:c.304G>C(p.Asp102His)	.
SOD1:c.313A>T(p.Ile105Phe)	rs121912445
SOD1:c.317C>T(p.Ser106Leu)	.
SOD1:c.319C>G(p.Leu107Val)	rs121912440
SOD1:c.319C>T(p.Leu107Phe)	.
SOD1:c.326G>T(p.Gly109Val)	.
SOD1:c.328G>T(p.Asp110Tyr)	.
SOD1:c.335G>A(p.Cys112Tyr)	.
SOD1:c.338T>C(p.Ile113Thr)	rs74315452
SOD1:c.339C>G(p.Ile113Met)	.
SOD1:c.341T>C(p.Ile114Thr)	rs121912441
SOD1:c.340A>T(p.Ile114Phe)	.
SOD1:c.344G>C(p.Gly115Ala)	.

SOD1:c.346C>G(p.Arg116Gly)	.
SOD1:c.352C>G(p.Leu118Val)	rs199474723
SOD1:c.355G>T(p.Val119Leu)	.
SOD1:c.355G>C(p.Val119Leu)	.
SOD1:c.355G>A(p.Val119Met)	.
SOD1:c.362A>T(p.His121Leu)	.
SOD1:c.374A>T(p.Asp125Val)	.
SOD1:c.374A>G(p.Asp125Gly)	.
SOD1:c.376G>C(p.Asp126His)	.
SOD1:c.380T>C(p.Leu127Ser)	.
SOD1:c.380T>A(p.Leu127*)	rs121912454
SOD1:c.382G>C(p.Gly128Arg)	.
SOD1:c.401A>T(p.Glu134Val)	.
SOD1:c.404G>A(p.Ser135Asn)	rs121912451
SOD1:c.413C>G(p.Thr138Arg)	.
SOD1:c.412A>G(p.Thr138Ala)	.
SOD1:c.416G>A(p.Gly139Glu)	.
SOD1:c.418A>C(p.Asn140His)	.
SOD1:c.420C>A(p.Asn140Lys)	.
SOD1:c.418A>G(p.Asn140Asp)	.
SOD1:c.422C>G(p.Ala141Gly)	.
SOD1:c.424G>T(p.Gly142*)	.
SOD1:c.425G>A(p.Gly142Glu)	.
SOD1:c.425G>C(p.Gly142Ala)	.
SOD1:c.435G>T(p.Leu145Phe)	.
SOD1:c.435G>C(p.Leu145Phe)	.

SOD1:c.434T>C(p.Leu145Ser)	rs121912446					
SOD1:c.436G>A(p.Ala146Thr)	rs121912447					
SOD1:c.437C>G(p.Ala146Gly)	.					
SOD1:c.439T>C(p.Cys147Arg)	.					
SOD1:c.441T>A(p.Cys147*)	.					
SOD1:c.443G>A(p.Gly148Asp)	.					
SOD1:c.442G>C(p.Gly148Arg)	.					
SOD1:c.442G>A(p.Gly148Ser)	.					
SOD1:c.446T>G(p.Val149Gly)	.					
SOD1:c.445G>A(p.Val149Ile)	.					
SOD1:c.449T>C(p.Ile150Thr)	.					
SOD1:c.455T>C(p.Ile152Thr)	rs121912449					
SOD1:c.455T>G(p.Ile152Ser)	.					
SOX5:c.1085A>C(p.Gln362Pro)	rs144670919	4298, 2, 0	2203, 0, 0			6501, 2, 0
SPG11:c.118C>T(p.Gln40*)	.					
SPG11:c.266G>A(p.Trp89*)	.					
SPG11:c.267G>A(p.Trp89*)	.					
SPG11:c.2198T>G(p.Leu733*)	.					
SPG11:c.5970C>G(p.Tyr1990*)	.					
SPG11:c.5974C>T(p.Arg1992*)	rs200793464	378, 0, 0	246, 0, 0	181, 0, 0	285, 1, 0	
SPG11:c.6157G>A(p.Val2053Met)	rs149003934	4296, 2, 0	2197, 1, 0			6493, 3, 0
SQSTM1:c.98C>T(p.Ala33Val)	rs200396166	3688, 6, 0	1787, 0, 0			5475, 6, 0

SQSTM1:c.457G>A(p.Val153Ile)	rs145056421	4291, 9, 0	2203, 0, 0			6494, 9, 0
SQSTM1:c.683C>T(p.Pro228Leu)	rs151191977	4296, 4, 0	2202, 1, 0			6498, 5, 0
SQSTM1:c.952T>C(p.Ser318Pro)	.					
SQSTM1:c.961C>T(p.Arg321Cys)	rs140226523	4663, 5, 0	2288, 145, 4	180, 1, 0	286, 0, 0	6343, 134, 4
SQSTM1:c.1108T>C(p.Ser370Pro)	rs143956614	4300, 0, 0	2175, 27, 1			6475, 27, 1
SQSTM1:c.1175C>T(p.Pro392Leu)	rs104893941	4656, 22, 0	2448, 1, 0	175, 6, 0	286, 0, 0	6484, 19, 0
SQSTM1:c.1231G>A(p.Gly411Ser)	rs143511494	4300, 0, 0	2202, 1, 0			6502, 1, 0
SQSTM1:c.1273G>A(p.Gly425Arg)	.					
SYT9:c.460G>A(p.Val154Met)	rs78477754	4385, 283, 6	2369, 76, 2	169, 11, 1	286, 0, 0	6154, 335, 8
SYT9:c.712A>G(p.Ile238Val)	.					
SYT9:c.1057C>G(p.Leu353Val)	rs117876446	4574, 100, 0	2438, 9, 0	180, 1, 0	286, 0, 0	6393, 104, 0
TAF15:c.91G>A(p.Ala31Thr)	.					
TAF15:c.1156G>A(p.Asp386Asn)	rs147145070	4299, 1, 0	2203, 0, 0			6502, 1, 0
TAF15:c.1163G>A(p.Arg388His)	rs140268553	4293, 7, 0	2203, 0, 0			6496, 7, 0
TAF15:c.1184G>A(p.Arg395Gln)	rs71381481	4677, 0, 0	2446, 2, 0	181, 0, 0	286, 0, 0	6499, 2, 0
TARDBP:c.269C>T(p.Ala90Val)	rs80356715	4298, 2, 0	2202, 1, 0			6500, 3, 0
TARDBP:c.506A>G(p.Asp169Gly)	rs80356717					

TARDBP:c.800A>G(p.Asn267Ser)	rs80356718	377, 1, 0	246, 0, 0	181, 0, 0	286, 0, 0
TARDBP:c.859G>A(p.Gly287Ser)	rs80356719				
TARDBP:c.869G>C(p.Gly290Ala)	rs121908395				
TARDBP:c.875G>A(p.Ser292Asn)	.				
TARDBP:c.881G>C(p.Gly294Ala)	rs80356721				
TARDBP:c.881G>T(p.Gly294Val)	rs80356721				
TARDBP:c.883G>C(p.Gly295Arg)	rs80356723				
TARDBP:c.883G>A(p.Gly295Ser)	rs80356723				
TARDBP:c.892G>A(p.Gly298Ser)	rs4884357				
TARDBP:c.931A>G(p.Met311Val)	rs80356725				
TARDBP:c.943G>A(p.Ala315Thr)	rs80356726				
TARDBP:c.962C>G(p.Ala321Gly)	.				
TARDBP:c.991C>A(p.Gln331Lys)	rs80356727				
TARDBP:c.995G>A(p.Ser332Asn)	rs80356728				
TARDBP:c.1004G>A(p.Gly335Asp)	rs80356729				
TARDBP:c.1009A>G(p.Met337Val)	rs80356730				

TARDBP:c.1028A>G(p.Gln343Arg)	rs80356731			
TARDBP:c.1035C>A(p.Asn345Lys)	rs80356732			
TARDBP:c.1042G>T(p.Gly348Cys)	rs80356733			
TARDBP:c.1055A>G(p.Asn352Ser)	rs80356734			
TARDBP:c.1083G>T(p.Arg361Ser)	rs80356735			
TARDBP:c.1087C>G(p.Pro363Ala)	.			
TARDBP:c.1122T>A(p.Tyr374*)	.			
TARDBP:c.1132A>G(p.Asn378Asp)	.			
TARDBP:c.1135T>C(p.Ser379Pro)	rs80356738			
TARDBP:c.1136C>G(p.Ser379Cys)	rs80356739			
TARDBP:c.1144G>A(p.Ala382Thr)	rs11689432			
TARDBP:c.1144G>C(p.Ala382Pro)	rs11689432			
TARDBP:c.1147A>G(p.Ile383Val)	rs80356740			
TARDBP:c.1150G>C(p.Gly384Arg)	.			
TARDBP:c.1153T>G(p.Trp385Gly)	.			
TARDBP:c.1168A>G(p.Asn390Asp)	rs80356741			
TARDBP:c.1169A>G(p.Asn390Ser)	rs80356742	4299, 1, 0	2203, 0, 0	6502, 1, 0
TARDBP:c.1178C>T(p.Ser393Leu)	rs80356743			

microtubule cytoskeleton	1.599e-29	87	869
protein localization to endoplasmic reticulum	6.621e-29	39	119
Genes involved in Formation and Maturation of mRNA Transcript	2.809e-28	49	152
mRNA processing	3.573e-28	64	410
translational elongation	6.311e-28	37	109
RNA catabolic process	6.312e-28	48	212
multi-organism reproductive process	6.676e-28	83	719
cytoskeletal part	3.004e-27	103	1292
viral reproduction	1.356e-26	80	696
multi-organism cellular process	1.831e-26	80	699
mRNA catabolic process	3.229e-26	44	186
nucleoplasm	4.661e-26	109	1483
ribosomal subunit	9.84e-26	36	134
Genes involved in Cell Cycle, Mitotic	1.239e-25	64	306
translational initiation	1.585e-25	40	152
Genes involved in Peptide chain elongation	3.22e-25	36	84
protein targeting to membrane	7.628e-25	39	148
viral genome expression	3.023e-24	39	153
viral transcription	3.023e-24	39	153
nuclear-transcribed mRNA catabolic process	3.651e-24	41	174
Genes involved in Viral mRNA Translation	6.099e-24	35	84
viral infectious cycle	8.414e-24	45	222
cellular protein complex disassembly	8.735e-24	39	157
spliceosomal complex	3.115e-23	35	145
protein complex disassembly	3.141e-23	39	162
Genes involved in Influenza Viral RNA Transcription and Replication	3.29e-23	37	100
ribosome	4.321e-23	40	204
Genes involved in Insulin Synthesis and Secretion	5.802e-23	41	129
cellular macromolecule localization	6.44e-23	88	945
Cytoplasmic Ribosomal Proteins	1.642e-22	35	91
mitotic cell cycle	2.605e-22	79	785
Genes involved in Translation	3.874e-22	39	120
microtubule-based process	5.81e-22	58	427
establishment of protein localization to organelle	5.903e-22	42	208
Genes involved in Influenza Life Cycle	8.161e-22	41	137
Genes involved in Formation of a pool of free 40S subunits	9.245e-22	35	95
cellular protein localization	1.017e-21	86	941
cytosolic part	1.857e-21	37	186

cellular component disassembly	3.435e-21	48	294
macromolecular complex disassembly	4.135e-21	39	183
Genes involved in GTP hydrolysis and joining of the 60S ribosomal subunit	5.119e-21	36	106
microtubule cytoskeleton organization	6.985e-21	47	285
Genes involved in Regulation of gene expression in beta cells	1.032e-20	35	101
cell cycle process	2.358e-20	90	1070
structural constituent of ribosome	2.397e-20	35	157
intracellular protein transport	6.442e-20	71	698
protein localization to organelle	5.652e-19	59	505
cell cycle phase	5.838e-19	78	864
macromolecular complex subunit organization	6.724e-19	92	1167
Genes involved in Metabolism of proteins	6.747e-19	47	215
microtubule organizing center organization	7.14e-19	25	69
Genes involved in Regulation of beta-cell development	1.046e-18	35	114
translation	1.164e-18	58	495
multi-organism process	1.203e-18	101	1385
cell cycle	2.563e-18	101	1399
protein targeting	4.42e-18	56	474
catalytic step 2 spliceosome	6.927e-18	24	80
microtubule organizing center part	2.897e-17	25	94
centrosome organization	6.719e-17	23	65
structural molecule activity	2.981e-16	60	621
cytosolic small ribosomal subunit	3.216e-16	17	36
cellular process involved in reproduction	9.759e-16	57	549
cytosolic large ribosomal subunit	1.041e-15	19	52
ribonucleoprotein complex biogenesis	2.363e-15	37	232
RNA localization	5.775e-15	29	137
cellular macromolecule catabolic process	1.305e-14	63	698
nucleobase-containing compound catabolic process	1.361e-14	66	760
protein complex subunit organization	2.792e-14	72	899
centrosome cycle	4.661e-14	18	44
RNA export from nucleus	1.122e-13	21	69
protein transport	1.439e-13	86	1252
heterocycle catabolic process	2.133e-13	66	802
cellular nitrogen compound catabolic process	2.419e-13	66	804
aromatic compound catabolic process	2.575e-13	66	805
cytoskeleton organization	3.107e-13	66	808
centriole	3.37e-13	18	59

nuclear export	5.815e-13	25	114
mRNA binding	9.063e-13	22	93
establishment of protein localization	9.303e-13	86	1291
organic cyclic compound catabolic process	1.428e-12	66	833
nucleic acid transport	1.568e-12	26	130
RNA transport	1.568e-12	26	130
establishment of RNA localization	1.568e-12	26	130
mRNA export from nucleus	1.617e-12	19	60
large ribosomal subunit	2.402e-12	19	75
Genes involved in mRNA 3-end processing	2.607e-12	17	34
Ribosomal proteins	5.405e-12	15	113
small ribosomal subunit	8.718e-12	17	60
macromolecule catabolic process	8.922e-12	67	887
enzyme binding	1.477e-11	73	1086
termination of RNA polymerase II transcription	2.848e-11	16	44
mRNA transport	4.433e-11	23	112
microtubule	6.771e-11	35	340
nucleobase-containing compound transport	9.581e-11	26	153
single organism reproductive process	3.307e-10	69	1003
Genes involved in Transport of Mature mRNA derived from an Intron-Containing Transcript	6.7e-10	18	51
nucleoside-triphosphatase activity	6.701e-10	56	761
centrosome duplication	1.108e-09	13	31
virus-host interaction	1.983e-09	38	370
Genes involved in Diabetes pathways	2.864e-09	48	383
mRNA 3'-end processing	3.353e-09	18	77
ribonucleoprotein complex assembly	3.399e-09	20	99
Genes involved in Translation initiation complex formation	4.216e-09	18	56
nuclear body	4.593e-09	29	274
pyrophosphatase activity	5.231e-09	56	801
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	5.773e-09	56	803
hydrolase activity, acting on acid anhydrides	7.375e-09	56	808
ribonucleoprotein complex subunit organization	8.98e-09	20	104
RNA 3'-end processing	1.029e-08	19	93
single-stranded RNA binding	1.061e-08	13	41
interaction with host	1.267e-08	39	412
nucleocytoplasmic transport	1.909e-08	35	342
cellular macromolecular complex assembly	2.11e-08	45	540
purine NTP-dependent helicase activity	2.185e-08	19	108

ATP-dependent helicase activity	2.185e-08	19	108
nuclear transport	2.452e-08	35	345
centriole replication	3.456e-08	9	14
Huntington disease	4.101e-08	25	126
ribosome biogenesis	4.296e-08	23	153
interspecies interaction between organisms	4.459e-08	40	449
symbiosis, encompassing mutualism through parasitism	4.459e-08	40	449
Genes involved in Formation of the ternary complex, and subsequently, the 43S complex	5.141e-08	16	49
prenatal lethality	5.835e-08	82	1466
DNA-dependent transcription, termination	9.315e-08	17	81
negative regulation of mRNA splicing, via spliceosome	1.162e-07	8	11
helicase activity	1.536e-07	21	149
negative regulation of RNA splicing	1.885e-07	9	16
DEAH (Asp-Glu-Ala-His) box genes	2.381e-07	6	17
negative regulation of mRNA processing	3.407e-07	8	12
Tubulins	4.252e-07	7	31
FoxO family signaling	5.867e-07	15	49
nuclear speck	7.544e-07	19	148
poly(A) RNA binding	1.367e-06	7	11
regulation of mRNA splicing, via spliceosome	1.464e-06	11	33
posttranscriptional regulation of gene expression	1.507e-06	34	380
spindle organization	1.675e-06	16	84
nucleoplasm part	1.853e-06	48	824
regulation of RNA splicing	1.936e-06	14	62
embryonic lethality between implantation and placentation	2.207e-06	34	372
embryonic lethality	2.7e-06	59	945
Genes involved in RNA Polymerase II Transcription	4.725e-06	19	92
spliceosomal complex assembly	5.884e-06	11	37
microtubule-based movement	7.089e-06	21	164
RNA helicase activity	1.299e-05	9	28
poly-purine tract binding	1.329e-05	7	14
mRNA splice site selection	1.491e-05	8	17
spindle	1.508e-05	22	235
Role of Calcineurin-dependent NFAT signaling in lymphocytes	2.42e-05	14	54
perinuclear region of cytoplasm	2.573e-05	32	469
macromolecular complex assembly	3.204e-05	58	1002
tubulin binding	3.269e-05	19	165
structural constituent of cytoskeleton	4.057e-05	13	76

regulation of mRNA processing	5.785e-05	11	45
cytoskeleton-dependent intracellular transport	5.792e-05	13	67
vesicle	6.538e-05	51	1012
regulation of mRNA stability	6.867e-05	10	36
Caspases	7.188e-05	4	11
regulation of microtubule-based process	8.636e-05	15	95
extracellular vesicular exosome	9.312e-05	11	63
complete embryonic lethality between implantation and somite formation	0.0001038	25	252
kinase binding	0.0001164	30	401
embryonic lethality between implantation and somite formation	0.0001248	26	273
extracellular membrane-bounded organelle	0.0001298	11	65
extracellular organelle	0.0001298	11	65
ATPase activity	0.0001446	28	361
protein kinase binding	0.0001619	28	363
chromosome	0.0001628	37	641
ATPase activity, coupled	0.0001656	24	279
organelle transport along microtubule	0.0001702	9	30
organelle assembly	0.0001759	15	100
regulation of RNA stability	0.0002085	10	40
cytoplasmic stress granule	0.0003009	7	23
early cellular replicative senescence	0.0003188	11	50
poly-pyrimidine tract binding	0.0003672	5	8
protein domain specific binding	0.000381	35	542
abnormal apoptosis	0.0004782	51	870
Parkinson disease	0.0005059	16	88
regulation of microtubule cytoskeleton organization	0.0005284	13	80
RNA-dependent ATPase activity	0.0005524	7	22
chromosomal part	0.000678	32	543
CRD-mediated mRNA stability complex	0.0006953	4	5
microtubule binding	0.0006992	14	112
nitrogen compound transport	0.0007155	37	556
abnormal cell death	0.0007298	53	934
mRNA stabilization	0.0007786	8	26
RNA stabilization	0.0007786	8	26
La ribonucleoprotein domain family	0.000799	3	7
RAB, member RAS oncogene family	0.0008241	6	62
pigment granule	0.0008656	12	94
melanosome	0.0008656	12	94

double-stranded RNA binding	0.0009249	9	44
membrane-bounded vesicle	0.001027	45	926
ATP binding	0.001116	68	1482
small conjugating protein ligase binding	0.001145	16	151
ubiquitin protein ligase binding	0.001145	16	151
negative regulation of macromolecule metabolic process	0.001207	67	1365
dynactin binding	0.00134	4	5
microtubule associated complex	0.001387	14	134
structure-specific DNA binding	0.001406	20	230
spindle pole	0.00151	12	99
negative regulation of metabolic process	0.001514	70	1460
abnormal cellular replicative senescence	0.001595	11	58
Genes involved in Proteolytic cleavage of SNARE complex proteins	0.001599	7	16
regulation of centrosome cycle	0.001689	7	20
dynactin complex	0.002048	4	6
regulation of centriole replication	0.002296	5	8
ribonucleoprotein granule	0.002589	10	71
cytoplasmic microtubule	0.002751	7	31
nuclear chromosome	0.00279	21	295
kinetochore	0.002815	12	105
antigen processing and presentation of exogenous peptide antigen via MHC class II	0.002817	13	92
increased apoptosis	0.002877	37	566
Hedgehog signaling events mediated by Gli proteins	0.002964	11	48
p53 pathway	0.003337	12	58
'de novo' protein folding	0.003539	10	53
transcription from RNA polymerase II promoter	0.003685	68	1437
antigen processing and presentation of peptide antigen via MHC class II	0.0041	13	95
Genes involved in Botulinum neurotoxicity	0.004139	7	18
rRNA metabolic process	0.004359	14	111
negative regulation of nucleobase-containing compound metabolic process	0.0046	50	938
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	0.004631	13	96
regulation of centrosome duplication	0.00464	6	15
spindle assembly	0.00496	9	43
GTPase activity	0.00531	19	230
abnormal prenatal growth/weight/body size	0.00653	47	839

negative regulation of nitrogen compound metabolic process	0.006639	50	950
ATP-dependent RNA helicase activity	0.006957	6	21
rRNA binding	0.006968	7	31
polysome	0.007009	6	24
CRD-mediated mRNA stabilization	0.008251	4	5
negative regulation of cellular metabolic process	0.008264	64	1353
M phase	0.008311	34	541
poly(U) RNA binding	0.009002	4	7
RS domain binding	0.009002	4	7
transcription factor binding	0.009249	28	447
alternative mRNA splicing, via spliceosome	0.00935	7	25
RNA polymerase II transcription factor binding	0.00943	10	72
microtubule anchoring	0.009537	8	35
Cytoplasmic dyneins	0.009794	3	15

Query denotes the number of query genes annotated with a given term. Total denotes the total number of genes annotated with a term. P denotes the significant of the enrichment for query genes after Bonferroni correction. Only results associated with Bonferroni corrected p-values <0.01 and relating to annotations from GO, Pathway, Gene Family, Human Phenotype and Mouse Phenotype databases are shown.

Table S3.2: Functional enrichment results –Shared interactors

Annotation	p	Query	Total
ribonucleoprotein complex	1.118e-28	28	556
RNA binding	2.365e-28	32	857
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	3.324e-28	23	209
mRNA splicing, via spliceosome	3.324e-28	23	209
RNA splicing, via transesterification reactions	5.859e-28	23	214
RNA splicing	6.186e-27	25	331
RNA processing	1.339e-25	29	655
mRNA metabolic process	9.666e-25	28	617
mRNA processing	1.408e-24	25	410
Genes involved in mRNA Splicing	1.975e-21	18	107
Genes involved in Elongation and Processing of Capped Transcripts	1.448e-19	18	134
Genes involved in Processing of Capped Intron-Containing Pre-mRNA	2.521e-19	18	138
Genes involved in Formation and Maturation of mRNA Transcript	1.542e-18	18	152
mRNA processing	2.781e-18	17	127
spliceosomal complex	1.033e-16	14	145
catalytic step 2 spliceosome	1.74e-16	12	80
Genes involved in Gene Expression	3.197e-14	21	425
nucleoplasm	3.052e-11	23	1483
single-stranded RNA binding	3.16e-09	7	41
RNA helicase activity	2.132e-08	6	28
regulation of mRNA splicing, via spliceosome	6.562e-07	6	33
regulation of RNA splicing	7.271e-07	7	62
poly(A) RNA binding	2.755e-06	4	11
regulation of mRNA processing	4.695e-06	6	45
poly-purine tract binding	8.304e-06	4	14
alternative mRNA splicing, via spliceosome	1.173e-05	5	25
double-stranded RNA binding	2.265e-05	5	44
negative regulation of mRNA splicing, via spliceosome	2.686e-05	4	11
mRNA binding	3.741e-05	6	93
negative regulation of mRNA processing	4.02e-05	4	12
RNA-dependent ATPase activity	5.965e-05	4	22
purine NTP-dependent helicase activity	9.088e-05	6	108

ATP-dependent helicase activity	9.088e-05	6	108
negative regulation of RNA splicing	0.0001465	4	16
helicase activity	0.0005959	6	149
mRNA stabilization	0.001177	4	26
RNA stabilization	0.001177	4	26
DEAH (Asp-Glu-Ala-His) box genes	0.001489	2	17
ATPase activity, coupled	0.002023	7	279
ATP-dependent RNA helicase activity	0.003957	3	21
cytoplasmic stress granule	0.004211	3	23
Deregulation of CDK5 in Alzheimers Disease	0.004439	3	11
regulation of mRNA stability	0.004534	4	36
viral reproductive process	0.004631	10	542
single-stranded DNA binding	0.005035	4	65
regulation of RNA stability	0.006972	4	40
CRD-mediated mRNA stability complex	0.009741	2	5

Query denotes the number of query genes annotated with a given term. Total denotes the total number of genes annotated with a term. P denotes the significant of the enrichment for query genes after Bonferroni correction. Only results associated with Bonferroni corrected p-values <0.01 and relating to annotations from GO, Pathway, Gene Family, Human Phenotype and Mouse Phenotype databases are shown.

Table S3.3: Functional profile – tALS sequence similarity hits

Annotation	p	Query	Total
RNA binding	5.776e-58	170	857
mRNA processing	1.285e-34	92	410
cytoskeletal part	1.139e-29	158	1292
RNA splicing	1.518e-29	77	331
RNA processing	9.276e-28	107	655
mRNA metabolic process	4.695e-26	101	617
mRNA splicing, via spliceosome	2.303e-21	53	209
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	2.303e-21	53	209
RNA splicing, via transesterification reactions	7.806e-21	53	214
mRNA processing	1.078e-18	37	127
Keratins	4.622e-18	20	61
regulation of RNA splicing	8.42e-18	28	62
microtubule cytoskeleton	3.479e-16	101	869
mRNA binding	1.756e-15	31	93
Genes involved in mRNA Splicing	2.922e-15	31	107
microtubule-based process	1.368e-14	66	427
Genes involved in Processing of Capped Intron-Containing Pre-mRNA	1.693e-14	34	138
regulation of mRNA splicing, via spliceosome	6.26e-14	19	33
Genes involved in Elongation and Processing of Capped Transcripts	4.174e-13	32	134
motor activity	1.146e-11	32	131
Kinesins	1.337e-11	13	39
Genes involved in Formation and Maturation of mRNA Transcript	1.971e-11	32	152
microtubule organizing center	5.692e-11	62	487
alternative mRNA splicing, via spliceosome	6.996e-11	15	25
regulation of mRNA processing	1.1e-10	19	45
GTPase regulator activity	1.11e-10	61	446
cytoskeleton organization	1.532e-10	88	808
cytoskeletal protein binding	2.399e-10	76	643
intermediate filament cytoskeleton	3.078e-10	39	229
nucleoside-triphosphatase regulator activity	3.647e-10	61	458
centrosome	2.987e-09	50	376
laminin complex	3.599e-09	9	10

intermediate filament	9.436e-09	33	189
microtubule	1.182e-08	46	340
pancreatic ribonuclease activity	1.254e-08	10	13
regulation of alternative mRNA splicing, via spliceosome	1.293e-08	12	19
Genes involved in mRNA 3-end processing	2.64e-08	14	34
Ras guanyl-nucleotide exchange factor activity	7.31e-08	25	112
spliceosomal complex	1.456e-07	27	145
regulation of small GTPase mediated signal transduction	2.443e-07	48	365
Zinc fingers, FYVE type	2.959e-07	9	31
phospholipid binding	2.99e-07	58	497
endoribonuclease activity, producing 3'-phosphomonoesters	3.073e-07	10	16
microtubule-based movement	3.618e-07	30	164
guanyl-nucleotide exchange factor activity	6.991e-07	31	183
RNA export from nucleus	7.277e-07	19	69
mRNA 3'-end processing	7.694e-07	20	77
cell leading edge	8.683e-07	36	259
Structural maintenance of chromosomes proteins	9.286e-07	5	6
small GTPase regulator activity	1.143e-06	41	299
Genes involved in Transport of Mature mRNA derived from an Intron-Containing Transcript	1.325e-06	15	51
Rho guanyl-nucleotide exchange factor activity	1.358e-06	19	75
structural molecule activity	2.341e-06	65	621
RNA localization	2.826e-06	26	137
mRNA export from nucleus	4.045e-06	17	60
RNA transport	4.49e-06	25	130
establishment of RNA localization	4.49e-06	25	130
nucleic acid transport	4.49e-06	25	130
RNA 3'-end processing	4.507e-06	21	93
nucleoside-triphosphatase activity	4.594e-06	74	761
myosin complex	8.092e-06	16	64
negative regulation of mRNA splicing, via spliceosome	1.136e-05	8	11
microtubule motor activity	1.152e-05	18	76
microtubule binding	1.219e-05	22	112
nucleoplasm	1.394e-05	112	1483
microtubule cytoskeleton organization	1.692e-05	38	285
ribonucleoprotein complex	1.806e-05	55	556
termination of RNA polymerase II transcription	2.267e-05	14	44
laminin-1 complex	2.843e-05	6	7
mRNA transport	2.941e-05	22	112

endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 3'-phosphomonoesters	3.222e-05	10	23
negative regulation of mRNA processing	3.268e-05	8	12
pyrophosphatase activity	3.943e-05	74	801
basal lamina	4.02e-05	9	20
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	4.366e-05	74	803
regulation of Ras protein signal transduction	4.779e-05	36	271
hydrolase activity, acting on acid anhydrides	5.62e-05	74	808
microtubule associated complex	6.871e-05	22	134
amb2 Integrin signaling	9.544e-05	14	59
Ras protein signal transduction	0.0001255	43	372
spindle	0.0001344	30	235
nucleobase-containing compound transport	0.0001418	25	153
M phase	0.0001419	55	541
intermediate filament cytoskeleton organization	0.0001834	11	30
nuclear export	0.0002086	21	114
Genes involved in RNA Polymerase II Transcription	0.0002118	17	92
single-stranded RNA binding	0.0002455	12	41
intermediate filament-based process	0.0002723	11	31
tubulin binding	0.0002759	25	165
GTPase binding	0.0003092	23	144
ruffle	0.0003405	20	124
actin binding	0.0004934	40	357
paraspeckles	0.0005454	5	6
posttranscriptional regulation of gene expression	0.0006396	42	380
regulation of Rho protein signal transduction	0.0006475	25	165
cell cycle process	0.0006617	87	1070
mRNA 3'-UTR binding	0.0007097	9	24
negative regulation of RNA splicing	0.0007192	8	16
actin cytoskeleton	0.0007634	38	366
contractile fiber	0.0008044	23	166
small GTPase mediated signal transduction	0.0009592	58	618
myofibril	0.001002	22	156
hemidesmosome assembly	0.001091	7	12
Sparse or absent hair	0.001109	25	207
GTPase activator activity	0.001177	31	251
catalytic step 2 spliceosome	0.001228	15	80
mRNA splice site selection	0.001303	8	17

intracellular transport	0.001591	90	1142
contractile fiber part	0.001723	21	149
AU-rich element binding	0.002089	7	15
Tripartite motif-containing family	0.002699	8	67
cell cycle	0.003035	104	1399
Genes involved in Gene Expression	0.00324	39	425
cell cycle phase	0.003423	72	864
regulation of translation	0.003425	29	229
poly(A) RNA binding	0.003568	6	11
microtubule plus-end binding	0.003568	6	11
Genes involved in Rho GTPase cycle	0.003725	18	124
lamellipodium	0.004471	18	122
small GTPase binding	0.00452	20	133
cell projection assembly	0.004543	27	207
M phase of mitotic cell cycle	0.004767	40	381
cell projection	0.005223	88	1234
nucleocytoplasmic transport	0.005848	37	342
Rho protein signal transduction	0.006279	26	198
actin filament-based process	0.006607	46	473
Nail dystrophy	0.006732	11	49
translation regulator activity, nucleic acid binding	0.006844	6	12
Centaurins	0.006877	4	14
cytoskeleton-dependent intracellular transport	0.007138	14	67
nuclear transport	0.007207	37	345
cell-substrate junction assembly	0.008819	13	59
cell projection organization	0.009504	75	938
microtubule organizing center part	0.009576	15	94

Query denotes the number of query genes annotated with a given term. Total denotes the total number of genes annotated with a term. P denotes the significant of the enrichment for query genes after Bonferroni correction. Only results associated with Bonferroni corrected p-values <0.01 and relating to annotations from GO, Pathway, Gene Family, Human Phenotype and Mouse Phenotype databases are shown.

Table S3.4: Functional profile – shared tALS sequence similarity hits

Annotation	p	Query	Total
RNA binding	1.342e-58	67	857
mRNA processing	2.969e-29	36	410
RNA processing	3.77e-28	41	655
mRNA metabolic process	2.115e-25	38	617
RNA splicing	4.7e-24	30	331
mRNA binding	2.512e-23	20	93
mRNA processing	2.522e-20	19	127
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	1.242e-19	23	209
mRNA splicing, via spliceosome	1.242e-19	23	209
RNA splicing, via transesterification reactions	2.156e-19	23	214
ribonucleoprotein complex	2.065e-14	25	556
Genes involved in mRNA Splicing	4.698e-12	13	107
Genes involved in Elongation and Processing of Capped Transcripts	9.15e-11	13	134
Genes involved in Processing of Capped Intron-Containing Pre-mRNA	1.343e-10	13	138
spliceosomal complex	1.612e-10	13	145
Genes involved in Formation and Maturation of mRNA Transcript	4.696e-10	13	152
single-stranded RNA binding	7.321e-10	9	41
catalytic step 2 spliceosome	2.791e-09	10	80
regulation of RNA splicing	4.879e-09	10	62
poly(A) RNA binding	5.288e-09	6	11
poly-purine tract binding	3.384e-08	6	14
alternative mRNA splicing, via spliceosome	1.382e-07	7	25
regulation of mRNA splicing, via spliceosome	1.179e-06	7	33
positive regulation of RNA splicing	2.227e-06	5	10
Genes involved in Gene Expression	2.262e-06	15	425
regulation of mRNA stability	2.269e-06	7	36
regulation of RNA stability	4.965e-06	7	40
AU-rich element binding	5.516e-06	5	15
nucleoplasm	7.747e-06	26	1483
mRNA stabilization	1.118e-05	6	26
RNA stabilization	1.118e-05	6	26
regulation of mRNA processing	1.178e-05	7	45

positive regulation of mRNA splicing, via spliceosome	2.244e-05	4	6
mRNA 3'-UTR binding	7.458e-05	5	24
translation regulator activity	7.458e-05	5	24
regulation of alternative mRNA splicing, via spliceosome	9.828e-05	5	19
posttranscriptional regulation of gene expression	0.000101	14	380
translation regulator activity, nucleic acid binding	0.0001497	4	12
Genes involved in mRNA 3-end processing	0.0002141	5	34
mRNA 3'-end processing	0.0005308	7	77
laminin-1 complex	0.0009612	3	7
positive regulation of mRNA processing	0.001034	4	13
cytoplasmic stress granule	0.001191	4	23
RNA 3'-end processing	0.001928	7	93
laminin complex	0.003257	3	10
translation repressor activity	0.005885	3	10
BRE binding	0.008076	2	2
termination of RNA polymerase II transcription	0.008112	5	44

Query denotes the number of query genes annotated with a given term. Total denotes the total number of genes annotated with a term. P denotes the significant of the enrichment for query genes after Bonferroni correction. Only results associated with Bonferroni corrected p-values <0.01 and relating to annotations from GO, Pathway, Gene Family, Human Phenotype and Mouse Phenotype databases are shown

Table S3.5: Functional profile – tALS genes

Category	Annotation	p	Query	Total
GO-BP	cell death	0	9	1385
GO-BP	death	0	9	1390
HP	Pseudobulbar signs	0	4	11
HP	Fasciculations	0	4	14
HP	Involuntary muscle contractions	0	4	18
HP	Pathologic changes in anterior horn cells and lateral corticospinal tracts	0	3	5
HP	Lower motor neuron signs	0	3	6
HP	Muscle weakness and atrophy	0	3	7
Drug	celecoxib	0	7	273
Drug	peroxynitrite	0	7	376
Drug	thiourea	0	4	128
Disease	Amyotrophic Lateral Sclerosis	0	5	15
Disease	AMYOTROPHIC LATERAL SCLEROSIS 1; ALS1	0	3	5
HP	Abnormality of the corticospinal tracts	1e-06	5	90
HP	Pseudobulbar palsy	2e-06	3	10
HP	Pseudobulbar paralysis	2e-06	3	10
HP	Dysarthria	1e-05	5	141
GO-BP	positive regulation of viral reproduction	1.1e-05	2	11
HP	Involuntary movements	3.1e-05	5	179
HP	Amyotrophy involving the musculature of the hand	3.9e-05	2	4
HP	Saccadic smooth pursuit	3.9e-05	2	4
HP	Amyotrophy involving the upper limbs	3.9e-05	2	4
HP	Neurological speech impairment	4.5e-05	5	193
GO-BP	regulation of viral reproduction	4.8e-05	2	22
HP	Muscle cramps	5.1e-05	3	30
HP	Abnormality of the peripheral nervous system	6.1e-05	5	205
Drug	polyglutamine	7.8e-05	4	461
HP	Amyotrophic lateral sclerosis	0.000137	2	7
HP	Muscle weakness	0.000137	5	242

CoExp	UVC_HIGH_D5_DN	0.000164	2	32
HP	Pallor	0.000182	2	8
GO-BP	positive regulation of protein kinase activity	0.000213	3	260
Comp	MORF_FBL	0.000217	3	134
GO-CC	cell soma	0.000221	3	271
HP	Amyotrophy involving the extremities	0.000234	2	9
GO-BP	positive regulation of kinase activity	0.000241	3	271
GO-BP	ovarian follicle development	0.000252	2	50
GO-BP	positive regulation of transferase activity	0.000265	3	280
HP	Juvenile onset	0.000292	2	10
HP	Abnormality of the motor neurons	0.000302	3	54
Drug	neon	0.000304	2	47
Drug	tetrahydrocortisol	0.00033	2	49
Drug	thymine glycol	0.000344	2	50
HP	Atrophy/Degeneration involving motor neurons	0.000356	2	11
GO-BP	viral infectious cycle	0.000388	2	62
GO-BP	double-strand break repair	0.000414	2	64
Comp	MORF_HAT1	0.000417	3	167
Drug	berberine	0.000419	3	280
GO-CC	nucleolus	0.000466	4	881
GO-BP	negative regulation by host of viral transcription	0.000484	1	1
GO-BP	myelin maintenance	0.000484	1	1
GO-BP	myelin maintenance in the peripheral nervous system	0.000484	1	1
GO-BP	symbiosis, encompassing mutualism through parasitism	0.000495	2	70
Pathway	Amyotrophic lateral sclerosis (ALS)	0.000549	2	56
Drug	tetraethylenepentamine	0.000566	1	1
Drug	Pyrophosphate 2-	0.000566	1	1
GO-MF	copper ion binding	0.000571	2	76
Drug	helium	0.000581	2	65
GO-BP	reproductive process	0.000602	4	915
GO-BP	viral reproductive process	0.000614	2	78
GO-BP	reproduction	0.000615	4	920
Comp	MORF_ANP32B	0.00062	3	191

Drug	leuprolide acetate	0.000674	2	70
GO-BP	ovulation cycle process	0.000678	2	82
GO-BP	female gonad development	0.000678	2	82
Drug	galantamine	0.000753	2	74
GO-BP	regulation of protein kinase activity	0.000754	3	400
GO-BP	ovulation cycle	0.000763	2	87
GO-BP	development of primary female sexual characteristics	0.000763	2	87
GO-BP	viral reproduction	0.000763	2	87
HP	Onset in adolescence	0.000771	2	16
Comp	GNF2_ELAC2	0.000802	2	43
GO-BP	regulation of kinase activity	0.000851	3	417
Drug	amprenavir	0.000857	2	79
GO-BP	female sex differentiation	0.000871	2	93
MP	increased sensitivity to induced morbidity/mortality	0.000901	3	267
MP	abnormal neuromuscular synapse morphology	0.000932	2	57
HP	Slow progression	0.000934	3	79
GO-CC	angiogenin-PRI complex	0.00094	1	2
GO-BP	regulation of transferase activity	0.000943	3	432
GO-BP	placenta development	0.000967	2	98
GO-BP	protein targeting to Golgi	0.000968	1	2
GO-BP	3'-UTR-mediated mRNA stabilization	0.000968	1	2
GO-BP	regulation of endosome size	0.000968	1	2
GO-BP	modulation of transcription in other organism during symbiotic interaction	0.000968	1	2
GO-BP	modulation by host of symbiont transcription	0.000968	1	2
GO-BP	negative regulation of viral transcription	0.000968	1	2
GO-BP	modulation by host of viral transcription	0.000968	1	2
GO-MF	protein homodimerization activity	0.000991	3	444
Comp	GNF2_DENR	0.001	2	48
MicroRNA	hsa-miR-143:TargetScan	0.001042	3	322
Drug	argon	0.001062	2	88
Drug	aluminum carbonate	0.001131	1	2

Drug	cerous chloride	0.001131	1	2
Drug	2,3-diaminonaphthalene	0.001131	1	2
Drug	Metalloporphyrins	0.001131	1	2
Drug	tocopherols	0.001135	2	91
HP	Abnormality of musculature	0.001138	6	617
HP	Impaired smooth pursuit	0.001215	2	20
Drug	Antimony	0.001236	2	95
GO-CC	neuron projection	0.001241	3	489
MP	neurodegeneration	0.001294	3	302
GO-MF	tubulin binding	0.001324	2	116
HP	Spasticity	0.001341	4	211
HP	Upper motor neuron abnormality	0.001341	2	21
Drug	allopurinol	0.001369	2	100
Comp	MORF_DEK	0.001434	3	254
GO-MF	oxidoreductase activity, acting on superoxide radicals as acceptor	0.001437	1	3
GO-MF	superoxide dismutase activity	0.001437	1	3
GO-BP	negative regulation of cholesterol biosynthetic process	0.001452	1	3
Comp	MORF_RAN	0.001501	3	258
Comp	MORF_SOD1	0.001605	3	264
Comp	MORF_ESPL1	0.001613	2	61
Comp	MORF_FEN1	0.001666	2	62
Domain	EWS_RNA_bd	0.001691	1	3
Drug	talbutal	0.001697	1	3
Drug	Bupropion	0.001697	1	3
Comp	MORF_BUB3	0.001714	3	270
Comp	MORF_HDAC2	0.001733	3	271
MP	increased sensitivity to xenobiotic induced morbidity/mortality	0.001741	2	78
MP	abnormal somatic motor system morphology	0.001741	2	78
Drug	AlCl3	0.001743	2	113
Comp	MORF_CSNK2B	0.001828	3	276
GO-MF	cytoskeletal protein binding	0.001829	3	549
MP	abnormal induced morbidity/mortality	0.001861	3	342
CoExp	HEARTFAILURE_ATRIA_DN	0.001899	2	109
GO-MF	protein serine/threonine kinase activator activity	0.001916	1	4

Drug	malathion	0.00193	2	119
GO-BP	retrograde transport, vesicle recycling within Golgi	0.001935	1	4
GO-BP	activation of phospholipase A2 activity	0.001935	1	4
GO-BP	protein localization in Golgi apparatus	0.001935	1	4
GO-BP	relaxation of vascular smooth muscle	0.001935	1	4
Comp	GNF2_RBBP6	0.001943	2	67
GO-BP	gonad development	0.001958	2	140
MP	abnormal synapse morphology	0.002016	2	84
Disease	SPINAL MUSCULAR ATROPHY, PROXIMAL, ADULT, AUTOSOMAL DOMINANT	0.002048	1	1
Disease	AMYOTROPHIC LATERAL SCLEROSIS 2, JUVENILE; ALS2	0.002048	1	1
Disease	Sleep Apnea Syndromes	0.002048	1	1
Disease	SPASTIC PARALYSIS, INFANTILE-ONSET ASCENDING; IAHSP	0.002048	1	1
Disease	SPINOCEREBELLAR ATAXIA, AUTOSOMAL RECESSIVE 1; SCAR1	0.002048	1	1
Disease	AMYOTROPHIC LATERAL SCLEROSIS 10; ALS10	0.002048	1	1
Disease	PERRY SYNDROME	0.002048	1	1
Disease	Hypoventilation	0.002048	1	1
Disease	Mood Disorders	0.002048	1	1
Disease	AMYOTROPHIC LATERAL SCLEROSIS 9; ALS9	0.002048	1	1
Disease	AMYOTROPHIC LATERAL SCLEROSIS 4, JUVENILE; ALS4	0.002048	1	1
Disease	NEURONOPATHY, DISTAL HEREDITARY MOTOR, TYPE VIIB; HMN7B	0.002048	1	1
Disease	AMYOTROPHIC LATERAL SCLEROSIS 6	0.002048	1	1
Disease	AMYOTROPHIC LATERAL SCLEROSIS 8; ALS8	0.002048	1	1

Disease	PRIMARY LATERAL SCLEROSIS, JUVENILE; PLSJ	0.002048	1	1
HP	Abnormal EMG findings	0.002062	2	26
Drug	superoxide	0.002174	3	494
GO-BP	nervous system development	0.002255	4	1300
Drug	propamocarb	0.002261	1	4
Drug	Fructose	0.002261	1	4
Drug	2-aminophenol	0.002261	1	4
GO-BP	RNA processing	0.002315	3	590
Drug	coenzyme Q10	0.002333	2	131
GO-BP	positive regulation of catalytic activity	0.002372	3	595
GO-MF	Rab guanyl-nucleotide exchange factor activity	0.002394	1	5
GO-BP	positive regulation of Rac protein signal transduction	0.002418	1	5
GO-BP	positive regulation of viral genome replication	0.002418	1	5
GO-BP	relaxation of muscle	0.002418	1	5
GO-BP	positive regulation of phospholipase A2 activity	0.002418	1	5
GO-BP	development of primary sexual characteristics	0.002423	2	156
Comp	MORF_PCPRMT1	0.002431	2	75
Domain	RRM_1	0.002664	2	143
Pathway	hypertension	0.002686	1	4
Pathway	removal of superoxide radicals	0.002686	1	4
Pathway	Cardiac Protection Against ROS	0.002686	1	4
Drug	thioacetamide	0.002696	2	141
MP	abnormal xenobiotic induced morbidity/mortality	0.002736	2	98
HP	Diffuse axonal swelling	0.002764	1	1
HP	Decreased distal vibration sense	0.002764	1	1
HP	Breathing difficulty due to vocal cord paralysis	0.002764	1	1
HP	Pseudobulbar symptoms (uncontrolled laughter, weeping)	0.002764	1	1
HP	Early involvement of the corticospinal pathways	0.002764	1	1
HP	Spastic paraplegia, lower limb	0.002764	1	1

HP	Sural nerve biopsy shows loss of large myelinated fibers	0.002764	1	1
HP	Decreased distal touch sense	0.002764	1	1
HP	Pallor of dorsal columns of the spinal cord	0.002764	1	1
HP	Pontocerebellar atrophy	0.002764	1	1
HP	alveolar hypoventilation	0.002764	1	1
HP	Chewing difficulties	0.002764	1	1
HP	Spasticity of facial muscles	0.002764	1	1
HP	Spastic dysarthria	0.002764	1	1
HP	Spasticity of pharyngeal muscles	0.002764	1	1
Domain	RRM	0.002813	2	147
Domain	CAP_GLY_1	0.002817	1	5
Domain	CAP_GLY	0.002817	1	5
Domain	CAP-Gly	0.002817	1	5
Domain	Motile_Sperm	0.002817	1	5
Domain	VPS9	0.002817	1	5
Domain	MSP	0.002817	1	5
Domain	CAP_GLY_2	0.002817	1	5
Domain	MSP	0.002817	1	5
Domain	MSP	0.002817	1	5
Domain	VPS9	0.002817	1	5
Drug	Pyroglutamic Acid	0.002826	1	5
Drug	Antioxidants	0.002826	1	5
Drug	pyrrolidino-4-iodotamoxifen	0.002826	1	5
GO-MF	Rac guanyl-nucleotide exchange factor activity	0.002872	1	6
GO-MF	Ran guanyl-nucleotide exchange factor activity	0.002872	1	6
GO-BP	modification by host of symbiont morphology or physiology	0.002902	1	6
GO-BP	activation of protein kinase B activity	0.002902	1	6
GO-BP	diacylglycerol biosynthetic process	0.002902	1	6
HP	Sleep disturbances	0.002931	2	31
GO-BP	rhythmic process	0.002969	2	173
Domain	RNP1_RNA_bd	0.003122	2	155
GO-MF	enzyme binding	0.00315	3	664
Domain	RRM	0.003161	2	156
Domain	a_b_plait_nuc_bd	0.003242	2	158

GO-CC	cell leading edge	0.003267	2	187
GO-BP	reproductive structure development	0.00328	2	182
GO-MF	protein phosphatase 2B binding	0.00335	1	7
Comp	MORF_RAD23A	0.003352	3	340
Pathway	Glutathione Metabolism	0.003357	1	5
Domain	VPS9	0.003379	1	6
GO-BP	negative regulation of cellular biosynthetic process	0.003383	3	674
GO-BP	myelination in the peripheral nervous system	0.003384	1	7
GO-BP	auditory receptor cell stereocilium organization	0.003384	1	7
GO-BP	ensheathment of axons in the peripheral nervous system	0.003384	1	7
GO-BP	auditory receptor cell morphogenesis	0.003384	1	7
GO-BP	negative regulation of smooth muscle contraction	0.003384	1	7
GO-BP	Golgi to plasma membrane protein transport	0.003384	1	7
GO-BP	regulation of phospholipase A2 activity	0.003384	1	7
GO-BP	retina homeostasis	0.003384	1	7
GO-BP	regulation of viral transcription	0.003384	1	7
Drug	cobyric acid	0.00339	1	6
Drug	Ghrelin	0.00339	1	6
MicroRNA	TCATCTC,MIR-143:MSigDB	0.003481	2	134
Comp	MORF_UBE2N	0.003486	2	90
Comp	GNF2_APEX1	0.003486	2	90
GO-BP	negative regulation of biosynthetic process	0.003513	3	683
GO-BP	sex differentiation	0.003568	2	190
GO-MF	protein dimerization activity	0.0036	3	696
Domain	a_b_plait_nuc_bd	0.003699	2	169
GO-BP	cellular response to stress	0.003751	3	699
Drug	3-amino-1,2,4-triazole	0.003758	2	167
GO-BP	removal of superoxide radicals	0.003867	1	8
GO-BP	Schwann cell development	0.003867	1	8
GO-BP	viral transcription	0.003867	1	8
GO-BP	auditory receptor cell development	0.003867	1	8

GO-BP	positive regulation of Rac GTPase activity	0.003867	1	8
GO-BP	viral genome expression	0.003867	1	8
GO-BP	hydrogen peroxide biosynthetic process	0.003867	1	8
Drug	AA 861	0.003955	1	7
Drug	Gibberellin A8	0.003955	1	7
Drug	butobarbitone	0.003955	1	7
Drug	cyclobarbital	0.003955	1	7
Drug	Albendazole	0.003955	1	7
GO-BP	response to oxidative stress	0.004061	2	203
Drug	riluzole	0.004072	2	174
Disease	GLAUCOMA, NORMAL TENSION, SUSCEPTIBILITY TO	0.004093	1	2
Disease	Parkinsonian Disorders	0.004093	1	2
Disease	Mental Disorders	0.004093	1	2
Disease	Schistosomiasis mansoni	0.004093	1	2
GO-BP	neurofilament cytoskeleton organization	0.004349	1	9
Domain	RNAse_Pc	0.004503	1	8
Domain	RNASE_PANCREATIC	0.004503	1	8
Drug	Aclarubicin	0.004518	1	8
Drug	deferiprone	0.004518	1	8
Drug	citiolone	0.004518	1	8
Drug	Desoxycorticosterone	0.004518	1	8
MP	abnormal neuron morphology	0.004546	4	1048
Drug	Asbestos, Crocidolite	0.004661	3	646
MP	abnormal motor neuron morphology	0.004701	2	129
GO-MF	identical protein binding	0.004703	3	765
GO-MF	RNA binding	0.00472	3	766
Comp	MORF_PTPN11	0.004721	2	105
GO-BP	inner ear receptor stereocilium organization	0.004832	1	10
Drug	panfuran	0.004835	2	190
HP	Amyotrophy	0.005009	3	141
Domain	MORN	0.005065	1	9
Domain	RnaseA	0.005065	1	9
Drug	3',4',7-trihydroxyisoflavone	0.005082	1	9
Drug	Uric Acid	0.005082	1	9

Drug	Triethylenetetramine	0.005082	1	9
MP	pillar cell degeneration	0.005165	1	6
MP	suppressed circulating follicle stimulating hormone level	0.005165	1	6
Drug	uric acid	0.005291	2	199
GO-BP	interaction with symbiont	0.005314	1	11
GO-BP	receptor recycling	0.005314	1	11
GO-BP	Golgi to plasma membrane transport	0.005314	1	11
GO-BP	negative regulation of steroid biosynthetic process	0.005314	1	11
GO-BP	inner ear receptor cell development	0.005314	1	11
GO-BP	negative regulation of steroid metabolic process	0.005314	1	11
GO-BP	response to superoxide	0.005314	1	11
GO-BP	positive regulation of hydrolase activity	0.005402	2	235
HP	Sural nerve biopsy shows chronic axonal neuropathy	0.00552	1	2
HP	Cerebellar atrophy, progressive	0.00552	1	2
HP	Lower motor neuron disease	0.00552	1	2
HP	Anarthria	0.00552	1	2
Domain	MORN	0.005627	1	10
Domain	MORN	0.005627	1	10
GO-CC	cytoplasmic dynein complex	0.00563	1	12
GO-CC	dendrite cytoplasm	0.00563	1	12
Comp	MORF_DNMT1	0.005644	2	115
Drug	hydroxyl radical	0.005659	2	206
Drug	DIDS	0.005713	2	207
GO-BP	reproductive process in a multicellular organism	0.005719	2	242
GO-BP	multicellular organism reproduction	0.005719	2	242
GO-MF	Rac GTPase activator activity	0.005737	1	12
GO-BP	response to oxygen radical	0.005796	1	12
GO-BP	regulation of cholesterol biosynthetic process	0.005796	1	12
GO-BP	negative regulation of muscle contraction	0.005796	1	12
GO-BP	mRNA transcription	0.005796	1	12
GO-BP	multi-organism process	0.005962	3	824

Pathway	FLUMAZENILPATHWAY	0.006036	1	9
Pathway	Huntington's disease	0.006042	2	188
GO-CC	cell projection cytoplasm	0.006098	1	13
MicroRNA	miR-143:PicTar	0.006128	2	179
Domain	RNaseA	0.006188	1	11
Drug	ferric nitrilotriacetate	0.006208	1	11
Drug	Meda	0.006208	1	11
Drug	cyanoacrylate	0.006208	1	11
GO-MF	Rac GTPase binding	0.006214	1	13
GO-BP	Schwann cell differentiation	0.006277	1	13
GO-BP	diacylglycerol metabolic process	0.006277	1	13
GO-CC	dendrite	0.006307	2	262
GO-BP	posttranscriptional regulation of gene expression	0.006379	2	256
Drug	hypoxanthine	0.00643	2	220
GO-BP	negative regulation of cellular metabolic process	0.00661	3	855
Pathway	Free Radical Induced Apoptosis	0.006705	1	10
Pathway	FREEPATHWAY	0.006705	1	10
Domain	RCC1/BLIP-II	0.006749	1	12
GO-BP	regulation of T cell differentiation in the thymus	0.006759	1	14
GO-BP	regulation of Rac protein signal transduction	0.006759	1	14
Drug	3-isopropylmalate	0.006771	1	12
Drug	tetradecylselenoacetic acid	0.006771	1	12
CoExp	UV_ESR_WS_UNREG	0.006772	1	11
MicroRNA	GGGACCA,MIR-133B:MSigDB	0.006881	2	190
MicroRNA	GGGACCA,MIR-133A:MSigDB	0.006881	2	190
MP	thin stria vascularis	0.006882	1	8
HP	Generalized muscle weakness	0.006951	2	48
MP	abnormal spinal cord grey matter morphology	0.006991	2	158
GO-MF	pancreatic ribonuclease activity	0.007167	1	15
GO-BP	oocyte maturation	0.00724	1	15
GO-BP	regulation of viral genome replication	0.00724	1	15
GO-BP	muscle homeostasis	0.00724	1	15
Domain	RNaseA	0.007309	1	13
Domain	RNaseA	0.007309	1	13

Drug	Nitrogen Dioxide	0.007333	1	13
Drug	terephthalic acid	0.007333	1	13
Drug	Doxepin	0.007333	1	13
Drug	Citric Acid	0.007333	1	13
HP	Weakness	0.007529	2	50
Drug	carbon tetrachloride	0.00761	2	240
GO-CC	cell projection	0.007682	3	929
GO-BP	superoxide anion generation	0.007721	1	16
GO-BP	negative regulation of smooth muscle cell proliferation	0.007721	1	16
Drug	N5,N10-methylenetetrahydromethanopterin	0.007895	1	14
Drug	Kleptose	0.007895	1	14
TFBS	V$CREB_Q4_01	0.007957	2	171
CoExp	HEATSHOCK_YOUNG_UP	0.007999	1	13
Pathway	Lissencephaly gene (LIS1) in neuronal migration and development	0.008042	1	12
Drug	4Fe-4S	0.008171	2	249
Disease	GLAUCOMA, PRIMARY OPEN ANGLE; POAG	0.008171	1	4
GO-BP	mitosis	0.008174	2	291
GO-BP	nuclear division	0.008174	2	291
GO-BP	mRNA stabilization	0.008202	1	17
GO-BP	endosome organization	0.008202	1	17
GO-BP	myeloid cell homeostasis	0.008202	1	17
GO-BP	RNA stabilization	0.008202	1	17
GO-BP	modification of morphology or physiology of other organism during symbiotic interaction	0.008202	1	17
HP	Loss of spinal cord anterior horn cells	0.008271	1	3
HP	Difficulty in tongue movements	0.008271	1	3
HP	Hypergammaglobulinemia	0.008271	1	3
HP	Emg shows evidence of denervation	0.008271	1	3
HP	Shortening of the achilles tendon	0.008271	1	3
HP	Movement abnormalities of the tongue	0.008271	1	3
HP	Onset in early adulthood	0.008271	1	3

GO-BP	negative regulation of metabolic process	0.008278	3	927
GO-BP	M phase of mitotic cell cycle	0.008448	2	296
Drug	cyclobenzaprine	0.008457	1	15
Drug	tosufloxacin	0.008457	1	15
Drug	carbonate radical	0.008457	1	15
Drug	lenalidomide	0.008457	1	15
GO-MF	endoribonuclease activity, producing 3'-phosphomonoesters	0.008595	1	18
GO-MF	beta-tubulin binding	0.008595	1	18
GO-BP	regulation of mitochondrial membrane potential	0.008682	1	18
GO-BP	behavioral fear response	0.008682	1	18
HP	Distal amyotrophy	0.008748	2	54
GO-BP	organelle fission	0.008837	2	303
GO-CC	basal lamina	0.008901	1	19
GO-BP	RNA splicing	0.00895	2	305
Domain	Reg_chr_condens	0.008989	1	16
Drug	propionate	0.009014	2	262
Drug	fenamiphos	0.009019	1	16
Drug	Optison	0.009019	1	16
Drug	flavone	0.009019	1	16
Drug	tempol	0.009019	1	16
Drug	S-formylglutathione	0.009019	1	16
GO-BP	DNA repair	0.009063	2	307
GO-MF	mRNA 3'-UTR binding	0.00907	1	19
GO-BP	DNA fragmentation involved in apoptosis	0.009162	1	19
GO-BP	virus-host interaction	0.009162	1	19
GO-BP	plasma membrane organization	0.009162	1	19
GO-BP	behavioral defense response	0.009162	1	19
GO-BP	positive regulation of Rho GTPase activity	0.009162	1	19
GO-BP	intra-Golgi vesicle-mediated transport	0.009162	1	19
Pathway	LONGEVITYPATHWAY	0.009377	1	14
GO-BP	reproductive developmental process	0.009407	2	313
GO-BP	regulation of catalytic activity	0.009438	3	972
MP	abnormal cochlear nerve	0.009453	1	11

	morphology			
Domain	RCC1_2	0.009549	1	17
Domain	RCC1_3	0.009549	1	17
Domain	RCC1_1	0.009549	1	17
TFBS	TGACGTCA_V$ATF3_Q6	0.009555	2	188
Drug	nitronium	0.00958	1	17
Drug	oxonic acid	0.00958	1	17
Cytoband	chr10p14	0.009606	1	30
CoExp	HSIAO_LIVER_SPECIFIC_GENES	0.009612	2	250
CoExp	PENG_GLUTAMINE_DN	0.009612	2	250
GO-BP	response to copper ion	0.009643	1	20

GO-BP GO biological process; GO-MF GO molecular function; GO-CC GO cellular component; HP human phenotype; MP mouse phenotype; Comp Computational; CoExp Co-expression. Query denotes the number of query genes annotated with a given term. Total denotes the total number of genes annotated with a term. P denotes the significant of the enrichment for query genes. For brevity, only results associated with p-values <0.01are shown and results relating to citations within PubMed articles have been excluded.

Table S4.1: Candidate genes

HGNC	Network Hit	Sequence/ Similarity Hit	Disease Annotation	PageRank Priors	ToppGene	BLASTp	Overall
ALPK2	F	T	Linkage- chr18	NA	NA	NA	NA
POLI	T	F	Linkage- chr18	NA	NA	NA	NA
TCF4	T	F	Linkage- chr18	NA	NA	NA	NA
LMAN1	T	F	Linkage- chr18	NA	NA	NA	NA
FECH	T	F	Linkage- chr18	NA	NA	NA	NA
GRP	T	F	Linkage- chr18	NA	NA	NA	NA
MALT1	T	F	Linkage- chr18	NA	NA	NA	NA
MBD2	T	F	Linkage- chr18	NA	NA	NA	NA
DCC	T	F	Linkage- chr18	NA	NA	NA	NA
NEDD4L	T	F	Linkage- chr18	NA	NA	NA	NA
TRIB3	T	F	Linkage- chr20	NA	NA	NA	NA
FAM110A	T	F	Linkage- chr20	NA	NA	NA	NA
CSNK2A1	T	F	Linkage- chr20	NA	NA	NA	NA
PSMF1	T	F	Linkage- chr20	NA	NA	NA	NA
SRXN1	F	T	Linkage- chr20	NA	NA	NA	NA
RBCK1	T	F	Linkage- chr20	NA	NA	NA	NA
B4GALT1	T	F	Linkage- chr9p	NA	NA	NA	NA
NOL6	T	F	Linkage- chr9p	NA	NA	NA	NA
BAG1	T	F	Linkage- chr9p	NA	NA	NA	NA

DNAJA1	T	F	Linkage- chr9p	NA	NA	NA	NA
CHMP5	T	F	Linkage- chr9p	NA	NA	NA	NA
IFT74	F	T	Linkage- chr9p	NA	NA	NA	NA
NFX1	T	F	Linkage- chr9p	NA	NA	NA	NA
PRSS3	T	F	Linkage- chr9p	NA	NA	NA	NA
UBE2R2	T	F	Linkage- chr9p	NA	NA	NA	NA
TAF1L	T	F	Linkage- chr9p	NA	NA	NA	NA
TOPORS	T	F	Linkage- chr9p	NA	NA	NA	NA
PLAA	T	F	Linkage- chr9p	NA	NA	NA	NA
TEK	T	F	Linkage- chr9p	NA	NA	NA	NA
DDX58	T	F	Linkage- chr9p	NA	NA	NA	NA
SMU1	T	F	Linkage- chr9p	NA	NA	NA	NA
APTX	T	F	Linkage- chr9p	NA	NA	NA	NA
PRUNE2	T	F	Linkage- chr9q	NA	NA	NA	NA
TLE1	T	F	Linkage- chr9q	NA	NA	NA	NA
CEP78	T	F	Linkage- chr9q	NA	NA	NA	NA
OSTF1	T	F	Linkage- chr9q	NA	NA	NA	NA
VPS13A	T	T	Linkage- chr9q	NA	NA	NA	NA
ZFAND5	T	F	Linkage- chr9q	NA	NA	NA	NA
ANXA1	T	F	Linkage- chr9q	NA	NA	NA	NA
RORB	T	F	Linkage- chr9q	NA	NA	NA	NA
APP	T	F	Related disease: Alzheimers	NA	NA	NA	NA
PSEN1	T	F	Related disease: Alzheimers	NA	NA	NA	NA
PSEN2	T	F	Related disease: Alzheimers	NA	NA	NA	NA
APOE	T	F	Related disease: Alzheimers	NA	NA	NA	NA

ITM2B	T	F	Related disease: British/Danish dementia	NA	NA	NA	NA
SNAP29	T	F	Related disease: Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome	NA	NA	NA	NA
NDRG1	T	F	Related disease: CMT	NA	NA	NA	NA
CTDP1	T	F	Related disease: CMT	NA	NA	NA	NA
GAN	T	F	Related disease: CMT	NA	NA	NA	NA
MED25	T	F	Related disease: CMT	NA	NA	NA	NA
HSPB8	T	F	Related disease: CMT	NA	NA	NA	NA
LMNA	T	T	Related disease: CMT	NA	NA	NA	NA
ARHGEF10	F	T	Related disease: CMT	NA	NA	NA	NA
TRPV4	T	F	Related disease: CMT	NA	NA	NA	NA
GARS	T	F	Related disease: CMT	NA	NA	NA	NA
SOX10	T	F	Related disease: CMT	NA	NA	NA	NA
FGD4	T	F	Related disease: CMT	NA	NA	NA	NA
KIF1B	T	F	Related disease: CMT	NA	NA	NA	NA
RAB7A	T	F	Related disease: CMT	NA	NA	NA	NA
HOXD10	T	F	Related disease: CMT	NA	NA	NA	NA
HSPB1	T	F	Related disease: CMT	NA	NA	NA	NA
DNM2	T	F	Related disease: CMT	NA	NA	NA	NA
YARS	T	F	Related disease: CMT	NA	NA	NA	NA
MPZ	T	F	Related disease: CMT1B	NA	NA	NA	NA
GJB1	T	F	Related disease: CMT1X	NA	NA	NA	NA

ATN1	T	F	Related disease: DRPLA	NA	NA	NA	NA
MYOC	T	T	Related disease: Glaucoma 1A, primary open angle, juvenile-onset	NA	NA	NA	NA
WDR36	T	F	Related disease: Glaucoma 1, open angle, G	NA	NA	NA	NA
IGHMBP2	F	T	Related disease: HMN	NA	NA	NA	NA
BSCL2	T	F	Related disease: HMN	NA	NA	NA	NA
IKBKAP	T	F	Related disease: HSAN3	NA	NA	NA	NA
NTRK1	T	T	Related disease: HSAN4	NA	NA	NA	NA
NGF	T	F	Related disease: HSAN5	NA	NA	NA	NA
SCP2	T	F	Related disease: Leukoencephalopathy with dystonia and motor neuropathy	NA	NA	NA	NA
WNK1	T	T	Related disease: Neuropathy, hereditary sensory and autonomic, type II	NA	NA	NA	NA
CCT5	T	F	Related disease: Neuropathy, hereditary sensory, with spastic paraplegia	NA	NA	NA	NA
ELAVL4	T	T	Related disease: Neuropathy, paraneoplastic sensory	NA	NA	NA	NA
PARK2	T	F	Related disease: Parkinson	NA	NA	NA	NA
LRRK2	F	T	Related disease: Parkinson	NA	NA	NA	NA
UCHL1	T	F	Related disease: Parkinson	NA	NA	NA	NA

SNCA	T	F	Related disease: Parkinson	NA	NA	NA	NA
HTRA2	T	F	Related disease: Parkinson	NA	NA	NA	NA
ATXN1	T	F	Related disease: SCA1	NA	NA	NA	NA
ATXN10	T	F	Related disease: SCA10	NA	NA	NA	NA
PPP2R2B	T	F	Related disease: SCA12	NA	NA	NA	NA
PRKCG	T	F	Related disease: SCA14	NA	NA	NA	NA
ITPR1	T	F	Related disease: SCA15	NA	NA	NA	NA
TBP	T	F	Related disease: SCA17	NA	NA	NA	NA
ATXN2	T	F	Related disease: SCA2	NA	NA	NA	NA
ATXN3	T	T	Related disease: SCA3	NA	NA	NA	NA
PLEKHG4	F	T	Related disease: SCA4	NA	NA	NA	NA
SPTBN2	T	T	Related disease: SCA5	NA	NA	NA	NA
CACNA1A	F	T	Related disease: SCA6	NA	NA	NA	NA
C10orf2	T	F	Related disease: Sensory ataxia neuropathy, dysarthria, and ophthalmoparesis	NA	NA	NA	NA
KIF5A	F	T	Related disease: Spastic paraplegia 10	NA	NA	NA	NA
HSPD1	T	F	Related disease: Spastic paraplegia-13	NA	NA	NA	NA
SPAST	T	F	Related disease: Spastic paraplegia-4	NA	NA	NA	NA
GJC2	T	F	Related disease: Spastic paraplegia-44	NA	NA	NA	NA
SPG7	T	T	Related disease: Spastic	NA	NA	NA	NA

			paraplegia-7				
L1CAM	T	F	Related disease: SPG1	NA	NA	NA	NA
SPG20	T	F	Related disease: SPG20	NA	NA	NA	NA
SLC16A2	T	F	Related disease: SPG22	NA	NA	NA	NA
UBA1	T	F	Related disease: Spinal muscular atrophy, X-linked 2, infantile	NA	NA	NA	NA
CCS	T	T	NA	9	271	4	1
EWSR1	T	T	NA	104	120	2	2
RAC1	T	F	NA	3	3	5826	3
HNRNPA1	T	T	NA	25	86	15	4
TAF15	T	T	NA	168	520	1	5
ESR1	T	F	NA	8	5	5826	6
POLR2A	T	T	NA	10	205	199	7
HNRNPD	T	T	NA	155	165	6	8
YWHAB	T	F	NA	2	451	1655	9
BCL2	T	F	NA	12	10	5826	10
RAB5A	T	F	NA	1	378	5826	11
CLIP1	T	T	NA	329	74	12	12
HNRNPA2B1	T	T	NA	118	154	16	13
FOS	T	F	NA	65	22	451	14
IKBKG	T	T	NA	236	395	5	15
PPARGC1A	T	T	NA	794	4	175	16
SFRS1	T	T	NA	161	17	152	17
GRM5	T	T	NA	69	72	251	18
HNRNPM	T	T	NA	35	505	82	19

TP53	T	F	NA	153	1	5826	20
HNRPDL	T	T	NA	187	610	8	21
PAFAH1B1	T	F	NA	141	9	1038	22
UPF1	T	T	NA	414	776	3	23
SFRS3	T	T	NA	165	171	81	24
HNRNPR	T	T	NA	125	369	70	25
RNPS1	T	T	NA	119	358	84	26
NOS1	T	T	NA	1850	6	144	27
HNRNPA0	T	T	NA	124	1221	27	28
CASP3	T	F	NA	196	15	1468	29
KARS	T	F	NA	14	667	1240	30
DYNLL1	T	F	NA	140	35	960	31
AKAP9	T	T	NA	66	1595	65	32
RNASE1	T	T	NA	533	335	19	33
PABPN1	T	T	NA	701	46	102	34
UBC	T	F	NA	7	565	5826	35
CTNNB1	T	F	NA	204	18	1462	36
ALS2CL	T	T	NA	6	809	5826	37
PPP3CA	T	F	NA	13	247	5826	38
RNH1	T	F	NA	37	1173	317	39
CLTC	T	F	NA	94	98	817	40
SFPQ	T	T	NA	307	314	63	41
SFRS13A	T	T	NA	388	227	59	42
DNAJB1	T	F	NA	16	1394	933	43
G3BP2	T	T	NA	576	158	60	44

RAB5C	T	F	NA	4	3040	5826	45
APOA1	T	F	NA	92	172	656	46
U2AF2	T	T	NA	159	518	101	47
SFRS6	T	T	NA	129	981	72	48
ARHGEF2	T	F	NA	1366	12	185	49
HTT	T	F	NA	62	45	5826	50
MSI2	T	T	NA	600	1019	11	51
MAPRE1	T	F	NA	166	182	366	52
ALB	T	F	NA	83	130	1444	53
EPHA4	T	F	NA	52	643	618	54
RAB5B	T	F	NA	5	3093	5826	55
STX4	T	F	NA	24	671	1658	56
BRCA1	T	F	NA	446	31	678	57
HNRNPH1	T	T	NA	89	1540	119	58
HSPA5	T	T	NA	265	336	118	59
STAT3	T	F	NA	622	60	197	60
SFRS9	T	T	NA	189	516	120	61
GRB2	T	F	NA	43	592	962	62
RELA	T	F	NA	128	19	5826	63
ATM	T	T	NA	1200	23	214	64
HSPA8	T	F	NA	78	257	995	65
TF	T	F	NA	101	131	1224	66
MYO6	T	F	NA	71	58	5826	67
HNRNPF	T	T	NA	178	702	107	68
SFRS4	T	T	NA	180	1040	75	69

MSI1	T	T	NA	605	2032	9	70
PLK1	T	F	NA	95	197	1253	71
PABPC4	T	T	NA	487	471	54	72
MAPK14	T	F	NA	340	44	890	73
DCTN2	T	F	NA	179	94	1025	74
RBM5	T	T	NA	205	775	100	75
SSR4	T	F	NA	11	1228	5826	76
RIN2	T	T	NA	146	1944	66	77
VDAC1	T	F	NA	106	134	1822	78
CDK1	T	F	NA	63	554	967	79
PTBP1	T	F	NA	122	593	371	80
YWHAZ	T	F	NA	93	199	1786	81
EP300	T	T	NA	84	71	5826	82
STX1A	T	F	NA	23	431	5826	83
NPM1	T	F	NA	314	59	920	84
CREBBP	T	F	NA	82	77	5826	85
ACTA1	T	F	NA	20	581	5826	86
SAFB	T	T	NA	953	136	94	87
FLII	T	F	NA	73	750	713	88
HMOX1	T	F	NA	3757	2	687	89
RPGR	T	T	NA	404	2023	22	90
GRIN1	T	F	NA	283	37	1862	91
ACTN2	T	F	NA	26	1627	1447	92
GRIN2A	T	F	NA	1880	32	176	93
IGF1R	T	F	NA	722	70	348	94

MOV10	T	T	NA	531	3310	10	95
CYLD	T	T	NA	822	309	67	96
SFRS2	T	T	NA	175	26	5826	97
JUN	T	F	NA	248	16	5826	98
MAPK1	T	F	NA	396	30	1635	99
NDE1	T	T	NA	183	559	255	100
VAPA	T	T	NA	18	892	5826	101
TAX1BP1	T	T	NA	64	2354	322	102
RALY	T	F	NA	190	1190	131	103
MYC	T	F	NA	120	61	5826	104
VAMP2	T	F	NA	22	701	5826	105
COL4A3BP	T	F	NA	50	1995	591	106
GSK3B	T	F	NA	134	200	1690	107
NFKB1	T	F	NA	312	14	5826	108
ACTC1	T	F	NA	28	531	5826	109
CDH2	T	T	NA	1455	63	179	110
HNRNPU	T	F	NA	91	367	1641	111
CYCS	T	F	NA	1255	11	1511	112
HRAS	T	F	NA	645	79	448	113
HNRNPUL1	T	F	NA	97	1713	341	114
YWHAG	T	F	NA	27	644	5826	115
RCC1	T	T	NA	1713	296	34	116
NDEL1	T	T	NA	760	85	375	117
ACTB	T	F	NA	31	533	5826	118
HOXB2	T	F	NA	19	1356	5826	119

HNRNPH2	T	T	NA	195	1534	153	120
PTEN	T	F	NA	30	603	5826	121
HECW1	T	F	NA	15	2508	5826	122
ELANE	T	F	NA	29	665	5826	123
RNF19A	T	F	NA	17	1991	5826	124
HNRNPC	T	F	NA	48	308	5826	125
TRAF3	T	F	NA	58	1551	866	126
XPO1	T	F	NA	96	488	1424	127
XRCC6	T	F	NA	336	91	1231	128
CASP8	T	F	NA	517	106	586	129
DST	T	T	NA	412	399	201	130
TUBB	T	F	NA	36	552	5826	131
YWHAE	T	F	NA	126	92	5826	132
TSC1	T	T	NA	530	245	267	133
SSB	T	T	NA	479	589	124	134
TNFRSF1A	T	F	NA	298	109	1320	135
RPS5	T	F	NA	99	672	1166	136
STAT1	T	F	NA	634	29	1823	137
APBA1	T	T	NA	211	1300	228	138
PABPC1	T	T	NA	81	184	5826	139
OSBPL9	T	F	NA	47	3113	782	140
RNASE4	F	T	NA	5826	157	13	141
TGFBR1	T	F	NA	454	359	239	142
DDX5	T	F	NA	147	346	1466	143
SNAP23	T	F	NA	150	2283	219	144

CDC42	T	F	NA	310	27	5826	145
SIRT1	T	F	NA	929	65	522	146
BCL2L1	T	F	NA	706	55	910	147
PPP2R1A	T	F	NA	148	458	1142	148
GNAI2	T	F	NA	102	891	1039	149
PRKCE	T	F	NA	496	148	605	150
GHR	T	F	NA	2126	24	526	151
AR	T	F	NA	401	21	5826	152
FST	T	F	NA	34	927	5826	153
G3BP1	T	T	NA	433	1571	74	154
CDK5	T	F	NA	805	114	424	155
PRKCA	T	F	NA	262	327	832	156
CELF1	T	T	NA	3432	298	28	157
MAP2K1	T	T	NA	731	170	331	158
NCL	T	T	NA	85	221	5826	159
SFRS12	T	T	NA	435	1323	90	160
ARHGEF7	T	T	NA	1355	144	196	161
NEDD1	T	F	NA	223	1292	296	162
TFRC	T	F	NA	39	837	5826	163
MAPK8	T	F	NA	780	73	748	164
DHX9	T	F	NA	103	189	5826	165
PHB2	T	F	NA	112	1600	645	166
STX1B	T	F	NA	45	3240	1349	168
RUVBL2	T	F	NA	256	722	504	169
TRA2B	T	T	NA	1837	480	50	170

ACTA2	T	F	NA	33	1441	5826	171
CDKN1B	T	F	NA	1009	8	5826	172
RBMX	T	T	NA	132	150	5826	173
AKT1	T	F	NA	302	50	5826	174
PCBP2	T	F	NA	123	744	1339	175
HNRNPH3	T	T	NA	1842	211	114	176
HNRNPK	T	F	NA	145	878	892	177
FBLN1	T	F	NA	38	1133	5826	178
EPB41	T	F	NA	107	688	1749	179
FYN	T	F	NA	243	699	598	180
HSF1	T	T	NA	678	352	225	181
TRAF6	T	F	NA	86	1476	1153	182
OFD1	T	F	NA	208	766	679	183
SFRS11	T	F	NA	194	444	1287	184
YWHAQ	T	F	NA	327	151	1878	185
SP1	T	F	NA	334	47	5826	186
PUF60	T	T	NA	500	2180	61	187
HLA-B	T	F	NA	21	4357	5826	188
UBB	T	F	NA	272	275	1457	189
RBM3	T	T	NA	4500	232	36	190
YBX1	T	F	NA	158	147	5826	191
CSTF2	T	T	NA	1479	888	39	192
THRA	T	F	NA	342	354	899	193
SNAP25	T	F	NA	280	289	1464	194
U2AF1	T	F	NA	170	563	1433	195

HSP90AA1	T	F	NA	111	249	5826	196
RPS3	T	F	NA	79	428	5826	197
ATP6AP1	T	F	NA	44	1233	5826	198
RB1	T	F	NA	415	42	5826	199
CBS	T	F	NA	720	143	706	200
TUBA1A	T	F	NA	152	767	1324	201
SART3	T	T	NA	2037	263	99	202
TDGF1	T	F	NA	40	1672	5826	203
PPARG	T	F	NA	1394	49	860	204
GRM1	T	F	NA	67	618	5826	205
CEP250	T	F	NA	228	1321	461	206
DMD	T	T	NA	915	387	194	207
ITGB1	T	F	NA	637	78	1593	208
KRAS	T	T	NA	1807	139	232	209
RPN1	T	F	NA	110	2019	867	210
TBCE	T	T	NA	4173	229	47	211
TRAF2	T	F	NA	279	1336	381	212
MAP3K7	T	F	NA	242	393	1556	213
CREB1	T	F	NA	1175	56	1054	214
XRCC5	T	F	NA	413	187	1601	215
GTF3A	T	F	NA	77	557	5826	216
RBM39	T	T	NA	1717	998	35	217
NFKBIA	T	F	NA	504	207	1071	218
NF1	T	F	NA	2737	52	427	219
TIAL1	F	T	NA	5374	302	30	220

CTNND1	T	F	NA	1362	51	1009	221
SOD2	T	F	NA	1640	7	5826	222
HDAC1	T	F	NA	407	233	1438	223
VAMP1	T	F	NA	32	3506	5826	224
RBM14	T	T	NA	4160	365	32	225
NXF1	T	F	NA	418	736	456	226
PRKACA	T	F	NA	117	322	5826	227
RBM10	T	T	NA	1581	476	96	228
RAB8A	T	F	NA	72	2408	1744	229
CDK4	T	F	NA	1093	66	1103	230
CDK2	T	F	NA	470	168	1720	231
ATF6	T	F	NA	41	2438	5826	232
STX2	T	F	NA	49	1622	5826	233
SMAD3	T	F	NA	292	355	1572	234
KIF11	T	F	NA	395	576	666	235
PPP1CC	T	F	NA	635	161	1097	236
TNF	T	F	NA	1104	13	5826	237
NCOR1	T	F	NA	641	679	269	238
RHOA	T	F	NA	274	113	5826	239
CIRBP	T	T	NA	2844	655	38	240
RIPK1	T	F	NA	55	1458	5826	241
TUBA4A	T	F	NA	142	1337	1226	242
ACTR1A	T	F	NA	184	1800	640	243
CALR	T	F	NA	841	290	423	244
NEK2	T	F	NA	213	749	1297	245

DCTN3	T	F	NA	232	663	1292	246
IRS1	T	F	NA	785	286	471	247
PRKDC	T	F	NA	638	560	378	248
PPP3R1	T	F	NA	913	491	234	249
SFRS7	T	T	NA	138	313	5826	250
PPP2CA	T	F	NA	648	36	5826	251
TRA2A	T	T	NA	2912	695	37	252
TNFSF8	T	F	NA	42	3070	5826	253
DDX3X	T	F	NA	154	281	5826	254
FN1	T	F	NA	424	351	1230	255
ILF3	T	F	NA	156	279	5826	256
SMAD1	T	F	NA	489	385	938	257
FGFR1	T	F	NA	756	210	831	258
SNRPA	T	T	NA	1572	548	111	259
ATP2A2	T	F	NA	2042	167	258	260
BCL6	T	F	NA	624	318	794	261
ERBB3	T	F	NA	748	33	5826	262
RPL10A	T	F	NA	105	2057	1619	263
RUNX1	T	F	NA	344	507	1183	264
ACTR1B	T	F	NA	426	707	655	265
HMGB1	T	F	NA	1116	105	987	266
APEX1	T	F	NA	374	93	5826	267
DAXX	T	F	NA	940	127	1017	268
SF3B4	T	T	NA	1546	1556	43	269
TNFAIP3	T	F	NA	389	513	1033	270

MAP3K5	T	F	NA	643	389	674	271
CALCOCO2	T	T	NA	406	3850	134	272
CLASP1	T	F	NA	186	248	5826	273
ETS1	T	F	NA	1107	392	282	274
HIF1A	T	F	NA	998	108	1176	275
SMC1A	T	F	NA	172	292	5826	276
TBK1	T	F	NA	57	2314	5826	277
DYNC1H1	T	T	NA	174	293	5826	278
MAP2K3	T	F	NA	383	550	1083	279
EEF1A1	T	F	NA	558	569	644	280
TAB2	T	T	NA	742	1149	210	281
VHL	T	F	NA	109	658	5826	282
SYNCRIP	T	T	NA	362	115	5826	283
HNRNPA3	T	T	NA	137	456	5826	284
RABGEF1	T	T	NA	1264	1400	73	285
PPM1L	T	F	NA	54	2836	5826	286
BUB1	T	F	NA	988	435	314	287
HTATSF1	T	T	NA	1687	1187	57	288
PTPN1	T	F	NA	1067	500	250	289
GAPDH	T	F	NA	557	416	926	290
STX5	T	F	NA	51	3211	5826	291
SMC3	T	T	NA	1194	297	373	292
PTK2B	T	F	NA	88	1018	5826	293
RXRA	T	F	NA	299	817	1052	294
CEP63	T	F	NA	200	1837	823	295

PCBP1	T	F	NA	114	646	5826	296
LRPPRC	T	F	NA	113	664	5826	297
ARHGEF6	T	T	NA	1601	975	76	298
PTBP2	T	F	NA	378	860	779	299
SRC	T	F	NA	249	933	1214	300
RPS16	T	F	NA	80	5826	1296	301
TOP1	T	F	NA	282	1057	934	302
SFRS5	T	T	NA	182	319	5826	303
TOP2A	T	F	NA	1083	527	256	304
KPNA2	T	F	NA	1017	261	608	305
FGFR3	T	F	NA	736	218	1336	306
TUBG1	T	F	NA	127	628	5826	307
FASLG	T	F	NA	1135	180	742	308
ERBB2	T	F	NA	462	96	5826	309
FGFR1OP	T	F	NA	229	886	1626	310
HNRNPL	T	F	NA	171	396	5826	311
MLH1	T	F	NA	1216	198	643	312
MYH9	T	F	NA	1078	206	759	313
CALM3	T	F	NA	259	215	5826	314
XRN2	T	F	NA	537	1262	384	315
CELF4	F	T	NA	5438	772	25	316
RBM8A	T	T	NA	1980	1522	44	317
CEP290	T	T	NA	219	277	5826	318
EEF1G	T	F	NA	90	1327	5826	319
PIK3R1	T	F	NA	296	185	5826	320

MAP3K11	T	F	NA	1099	175	912	321
BAX	T	F	NA	934	39	5826	322
BMPR2	T	F	NA	459	634	989	323
KIF5B	T	T	NA	768	1120	280	324
SOS1	T	F	NA	902	220	1118	325
EFNB2	T	F	NA	498	1572	356	326
HDAC2	T	F	NA	616	75	5826	327
MRE11A	T	F	NA	2692	288	166	328
CTGF	T	F	NA	1359	648	187	329
EGFR	T	F	NA	253	246	5826	330
POLR2B	T	T	NA	1404	703	167	331
DHX38	T	F	NA	203	2198	915	332
AKTIP	T	T	NA	409	2992	252	333
HERC3	T	T	NA	3307	2752	14	334
TXN	T	F	NA	740	62	5826	335
PITPNM1	T	F	NA	60	3665	5826	336
RNF11	T	F	NA	56	4123	5826	337
CCNE1	T	F	NA	1407	219	561	338
VEGFA	T	F	NA	1004	41	5826	339
NOS2	T	F	NA	1203	28	5826	340
ARRB2	T	F	NA	87	1878	5826	341
CSNK1E	T	F	NA	130	2676	1772	342
RAB11B	T	F	NA	74	2729	5826	343
PDGFRA	T	F	NA	3030	84	552	344
MTNR1A	T	F	NA	53	4809	5826	345

ROCK1	T	T	NA	1125	568	362	346
USP20	T	F	NA	46	5826	5826	347
APAF1	T	F	NA	1974	162	488	348
NR3C1	T	F	NA	534	104	5826	349
HAP1	T	F	NA	293	236	5826	350
IGF1	T	F	NA	1590	20	5826	351
ARHGDIA	T	F	NA	291	239	5826	352
CALM1	T	F	NA	192	461	5826	353
HNRNPAB	T	T	NA	477	124	5826	354
RALBP1	T	F	NA	506	459	1474	355
CSNK1D	T	F	NA	139	2659	1782	356
CEP76	T	F	NA	221	2500	871	357
APC	T	T	NA	909	54	5826	358
TSG101	T	T	NA	788	1650	237	359
EIF6	T	F	NA	263	2319	732	360
MEF2C	T	F	NA	2295	202	358	361
TBC1D17	T	F	NA	68	3618	5826	362
MECP2	T	F	NA	2365	102	688	363
CDH1	T	F	NA	743	254	1730	364
PITPNM3	T	F	NA	61	4466	5826	365
RPA1	T	F	NA	1987	361	261	366
PKD2	T	F	NA	1936	356	281	367
MOV10L1	F	T	NA	5826	3527	7	368
SPEN	T	T	NA	2752	864	71	369
STK11	T	F	NA	1120	1158	213	370

RAC2	T	F	NA	116	1310	5826	371
HDAC6	T	F	NA	1207	349	631	372
JAK2	T	F	NA	769	752	580	373
BICD2	T	F	NA	380	1150	937	374
BBS4	T	F	NA	403	174	5826	375
MSH2	T	F	NA	2049	153	606	376
RPS6	T	F	NA	326	1005	1332	377
ZMYM2	T	F	NA	377	2264	491	378
CFTR	T	F	NA	287	283	5826	379
PTPN11	T	F	NA	699	87	5826	380
PRKCB	T	F	NA	483	1079	750	381
SF3B1	T	F	NA	108	1673	5826	382
CALM2	T	F	NA	276	317	5826	383
USF2	T	F	NA	355	2197	576	384
CAV1	T	F	NA	590	588	1130	385
VDR	T	F	NA	1174	166	1488	386
PRKCD	T	F	NA	359	216	5826	387
APBB1	T	F	NA	2171	196	468	388
RASA1	T	F	NA	467	659	1362	389
ALMS1	T	F	NA	238	409	5826	390
SLC25A6	T	F	NA	98	2250	5826	391
PRLR	T	F	NA	739	877	566	392
VIM	T	T	NA	245	390	5826	393
AKAP13	T	F	NA	1548	372	401	394
MAPK9	T	F	NA	1444	228	769	395

KDR	T	F	NA	1198	439	568	396
NDC80	T	T	NA	1144	872	310	397
DVL1	T	F	NA	343	243	5826	398
WDR20	T	F	NA	76	3835	5826	399
SF1	T	F	NA	332	258	5826	400
RNASE6	F	T	NA	5517	729	40	401
ARF6	T	F	NA	131	1325	5826	402
NCOA2	T	F	NA	844	1168	389	403
CASP1	T	F	NA	1111	173	1751	404
PCNA	T	F	NA	625	455	1479	405
CD2BP2	T	F	NA	163	3463	1453	406
MAP2K6	T	F	NA	394	865	1416	407
RAB14	T	F	NA	1356	473	487	408
PITPNM2	T	F	NA	59	5826	5826	409
TUBB2C	T	F	NA	201	611	5826	410
MDM2	T	F	NA	538	141	5826	411
GRIN2B	T	F	NA	945	69	5826	412
RPL5	T	F	NA	366	1506	927	413
RNASE2	F	T	NA	5560	357	85	414
CCDC53	T	F	NA	75	4320	5826	415
NRAS	T	T	NA	867	1842	248	416
CDKN1A	T	F	NA	1012	64	5826	417
PCM1	T	T	NA	169	861	5826	418
XBP1	T	F	NA	1760	80	1768	419
KIF3B	T	F	NA	732	2706	212	420

RBMS1	F	T	NA	5514	651	49	421
AHR	T	F	NA	1539	420	467	422
EIF2C2	T	F	NA	457	186	5826	423
ABL1	T	F	NA	252	460	5826	424
UBE2I	T	F	NA	255	453	5826	425
HGS	T	F	NA	488	1728	588	426
CDK5RAP2	T	T	NA	198	698	5826	427
RUVBL1	T	T	NA	492	3009	340	428
JPH3	F	T	NA	5358	1280	26	429
FAS	T	F	NA	1293	145	1876	430
TUBB4	T	F	NA	151	1191	5826	431
PLK4	T	F	NA	217	604	5826	432
TRIM28	T	F	NA	429	1062	1157	433
CDKN2A	T	F	NA	443	203	5826	434
PARP1	T	F	NA	710	111	5826	435
NTRK2	T	F	NA	1823	89	1627	436
SAFB2	T	T	NA	2102	1352	89	437
RPL6	T	F	NA	370	2399	636	438
CALCOCO1	T	T	NA	1165	3219	106	439
HAUS2	T	F	NA	239	1749	1795	440
HDAC3	T	F	NA	646	426	1725	441
DDX1	T	F	NA	185	842	5826	442
RAN	T	F	NA	996	430	1007	443
KIT	T	F	NA	1428	256	974	444
DDIT3	T	F	NA	1931	146	956	445

EPS15	T	T	NA	1101	4372	87	446
PRDX1	T	F	NA	1524	40	5826	447
GSN	T	F	NA	1055	260	1573	448
IMMT	T	T	NA	100	3245	5826	449
NDN	T	F	NA	2410	250	426	450
CIAO1	T	F	NA	328	4594	430	451
GATA1	T	T	NA	1648	1064	198	452
FLNA	T	F	NA	428	235	5826	453
MMP2	T	F	NA	2100	129	1012	454
KHDRBS1	T	F	NA	994	517	881	455
NCK1	T	F	NA	325	1370	1506	456
ESR2	T	F	NA	893	300	1742	457
RPL7	T	F	NA	346	1563	1221	458
PRKAR2B	T	F	NA	202	802	5826	459
IKBKE	T	F	NA	70	5826	5826	460
ILF2	T	F	NA	167	1189	5826	461
DDX17	T	F	NA	157	1342	5826	462
ETS2	T	F	NA	1709	967	215	463
CDC37	T	F	NA	1100	909	452	464
THOC4	T	F	NA	173	1140	5826	465
MAP1B	T	F	NA	1142	68	5826	466
EZR	T	F	NA	733	978	720	467
NRIP1	T	F	NA	856	100	5826	468
CTCF	T	F	NA	2099	152	922	469
VCL	T	F	NA	1408	450	665	470

BAD	T	F	NA	675	135	5826	471
SREBF1	T	F	NA	2471	961	121	472
CLIP2	T	T	NA	4383	2894	18	473
CDK5R1	T	F	NA	1827	34	5826	474
JPH2	F	T	NA	5545	1682	24	475
NCBP2	T	T	NA	2101	1092	135	476
CASP7	T	F	NA	308	2223	1175	477
NFKB2	T	F	NA	273	541	5826	478
RAD50	T	T	NA	3057	329	268	479
SF3A2	T	F	NA	144	1926	5826	480
SCHIP1	T	F	NA	3279	103	755	481
ANXA2	T	F	NA	1119	408	1086	482
SNRPB2	T	F	NA	1908	684	289	483
PLEKHF2	T	T	NA	1714	5826	41	484
RNASEN	T	F	NA	181	1218	5826	485
CCND1	T	F	NA	1075	81	5826	486
CP	T	F	NA	3419	107	717	487
MPHOSPH9	T	F	NA	136	2393	5826	489
ADRB2	T	F	NA	843	448	1507	490
LDHA	T	F	NA	2346	126	1069	491
TCERG1	T	F	NA	683	879	1001	492
NF2	T	F	NA	1535	326	919	493
MYO1E	T	F	NA	133	2634	5826	494
SMARCA4	T	F	NA	811	1798	398	495
MYCBP2	T	T	NA	1887	38	5826	496

DUSP1	T	F	NA	2693	76	1539	497
RPL13	T	F	NA	364	3476	647	498
SRRM1	T	F	NA	193	1198	5826	499
CD4	T	F	NA	608	183	5826	500
VCAN	T	F	NA	1259	1545	270	501
VAV1	T	F	NA	658	844	1148	502
DTNBP1	T	F	NA	665	587	1636	503
NR2F2	T	F	NA	3634	176	434	504
BMP7	T	F	NA	773	1872	414	505
RAF1	T	F	NA	485	251	5826	506
MAPK10	T	F	NA	2518	269	485	507
ETV6	T	F	NA	2983	347	306	508
MTOR	T	F	NA	1118	405	1223	509
LMNB1	T	F	NA	1536	751	428	510
RIN3	T	T	NA	162	1828	5826	511
USO1	T	T	NA	698	3542	262	512
PICK1	T	F	NA	1316	921	444	513
HOMER1	T	F	NA	639	1128	928	514
WWP1	T	T	NA	1472	1372	254	515
NOS3	T	T	NA	2391	25	5826	516
VAMP8	T	F	NA	281	2147	1716	517
CKAP5	T	F	NA	176	1593	5826	518
AP2M1	T	F	NA	1435	487	752	519
CETN2	T	F	NA	230	910	5826	520
SRPK1	T	F	NA	143	2570	5826	521

BNIP3	T	F	NA	2750	99	1228	522
RBMS2	F	T	NA	5826	723	62	523
NEDD9	T	T	NA	1489	1487	238	524
A2BP1	T	T	NA	2218	2174	83	525
INSR	T	F	NA	944	112	5826	526
PDS5B	T	T	NA	4406	454	147	527
SERGEF	F	T	NA	5556	1464	33	528
TP73	T	F	NA	1016	1170	505	529
CTTNBP2	T	T	NA	3173	694	151	530
LCK	T	F	NA	512	253	5826	531
RBMY1A1	T	T	NA	3347	1764	56	532
RAB11FIP3	T	F	NA	916	1584	437	533
MAF	T	F	NA	1904	169	1461	534
PCNT	T	F	NA	214	1127	5826	535
EEA1	T	F	NA	149	2532	5826	536
NUP98	T	F	NA	1199	713	683	537
BCAR3	T	F	NA	316	3419	953	538
AP2B1	T	F	NA	1644	590	540	539
LYAR	T	F	NA	207	4329	1657	540
DDOST	T	F	NA	115	4025	5826	541
COIL	T	T	NA	939	2783	247	542
RSPH1	F	T	NA	5826	2749	17	543
RPL4	T	F	NA	361	2005	1341	544
MCC	T	F	NA	264	4040	1113	545
CXCR4	T	F	NA	1058	101	5826	546

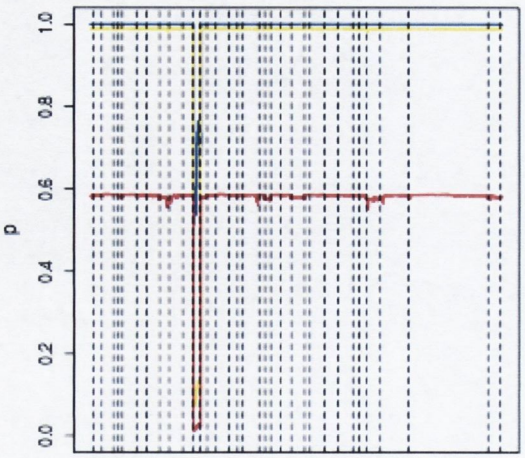
PARD3	T	F	NA	1201	972	517	547
BOLL	F	T	NA	5520	973	53	548
DRD2	T	F	NA	1937	140	1802	549
RACGAP1	T	F	NA	1213	786	642	550
HPRT1	T	F	NA	2173	123	1741	551
OCRL	T	F	NA	480	2312	773	552
RBM4B	T	T	NA	3780	1832	45	553
MYH11	T	T	NA	1073	2056	287	554
BIN1	T	F	NA	719	2076	489	555
CCAR1	T	F	NA	191	1613	5826	556
SPI1	T	F	NA	352	509	5826	557
AKT2	T	F	NA	1721	1135	290	558
KALRN	T	T	NA	1978	2059	123	559
TGFB1	T	F	NA	1871	48	5826	560
CASP4	T	F	NA	295	4285	949	561
CD40	T	F	NA	923	788	945	562
HLA-DRB1	T	F	NA	121	4122	5826	563
WBSCR16	F	T	NA	5518	1276	42	564
TRIP13	F	T	NA	1898	1547	180	565
SAR1A	T	F	NA	432	2765	790	566
CHM	T	F	NA	197	1611	5826	567
CENPJ	T	F	NA	212	1378	5826	568
RCBTB2	F	T	NA	5515	2569	21	569
ELAVL1	T	T	NA	473	331	5826	570
EXOSC10	T	F	NA	481	2117	902	571

HSP90AB1	T	F	NA	437	380	5826	572
APOB	T	F	NA	1777	57	5826	573
TPM3	T	F	NA	680	201	5826	574
DYNC1I2	T	F	NA	225	1284	5826	575
STK24	T	F	NA	1249	1868	286	576
DHX15	T	F	NA	177	2263	5826	577
GH1	T	F	NA	632	226	5826	578
SCNN1A	T	F	NA	387	2368	1217	579
SLC2A4	T	F	NA	802	1366	715	580
STMN1	T	F	NA	2285	240	964	581
ARIH2	T	F	NA	1542	2801	149	582
MCM3	T	F	NA	1091	1243	537	583
ADAP1	T	T	NA	1715	2573	141	584
DLG4	T	F	NA	501	328	5826	585
ZBTB16	T	F	NA	812	912	1065	586
BCL2L11	T	F	NA	1224	457	1229	587
BLM	T	F	NA	2548	268	725	588
PQBP1	T	T	NA	1053	2361	295	589
H2AFX	T	F	NA	268	971	5826	590
RAP1A	T	T	NA	963	2603	299	591
BBS1	T	F	NA	398	490	5826	592
PPP1CA	T	F	NA	493	1913	1044	593
CASP9	T	F	NA	1314	404	1309	594
CEP57	T	T	NA	231	1381	5826	595
TNFRSF1B	T	F	NA	297	831	5826	596

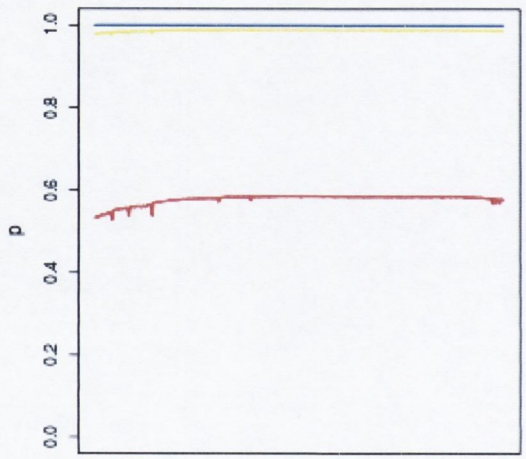
ATF2	T	F	NA	1069	493	1440	597
CP110	T	F	NA	216	1636	5826	598
GRIK2	T	F	NA	2834	195	826	599
EEF2	T	F	NA	505	344	5826	600
TP63	T	F	NA	357	608	5826	601
ILK	T	F	NA	800	1245	849	602
CSTF2T	T	T	NA	5318	1489	46	603
MAP3K10	T	F	NA	954	1796	469	604
SMAD4	T	F	NA	321	764	5826	605
DNMT3A	T	F	NA	2507	209	1079	606
NOX1	T	F	NA	1693	340	1205	607
CPNE4	T	F	NA	349	5826	671	608
BID	T	F	NA	1103	1314	541	609
DDX50	T	F	NA	574	2674	668	610
KIFAP3	T	F	NA	1652	631	681	611
SYK	T	F	NA	881	632	1535	612
CDC5L	T	F	NA	516	3315	625	613
RTN4	T	F	NA	564	316	5826	614
MSN	T	F	NA	1562	287	1609	615
RPL18A	T	F	NA	623	2497	622	616
ERCC2	T	T	NA	3969	273	390	617

ID Entrez Gene ID; Network Candidate Denotes whether gene was included as a potential candidate on the basis of protein interaction data; Sequence Similarity Candidate Denotes whether gene was included as a potential candidate on the basis of protein sequence analyses; Disease Annotation; Denotes whether gene,

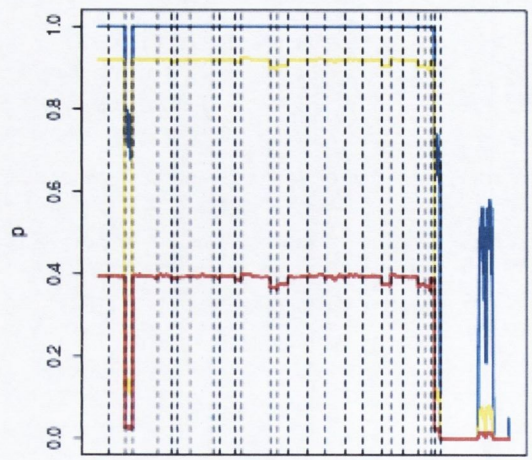
occurs in an ALS linkage interval (Table 3.2) or has been implicated in a condition related to ALS (Figure 3.7); PageRank-Priors, ToppGene, BLASTp, Overall Ranking after exclusion of candidates with disease annotation values other than "NA".



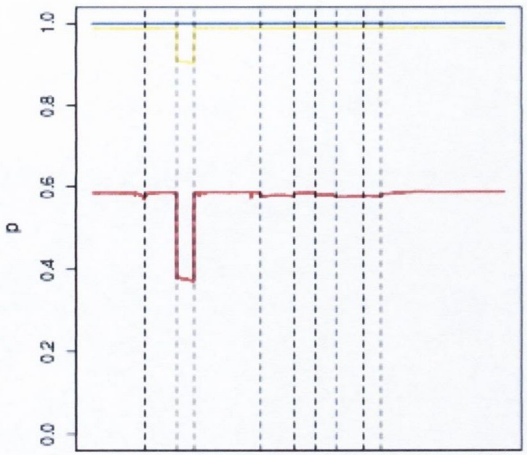
ALS2



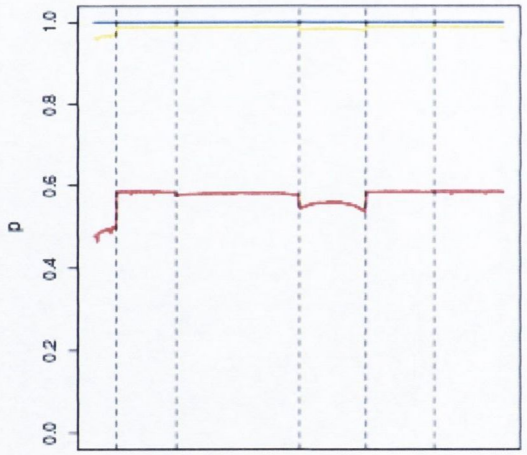
ANG



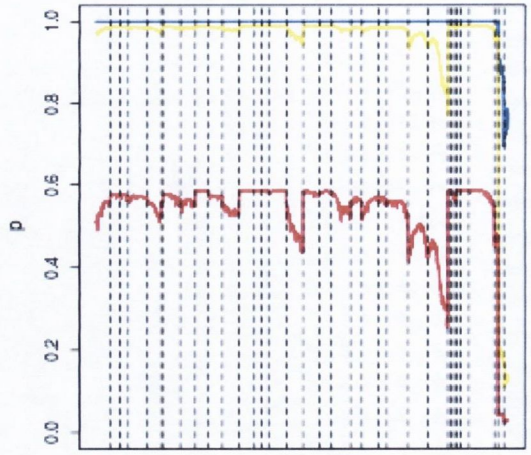
ATXN2



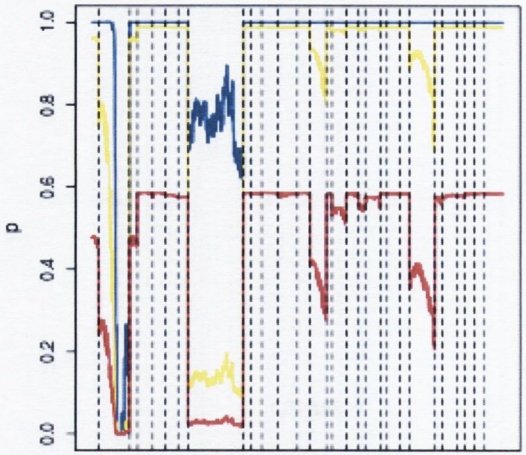
C9orf72



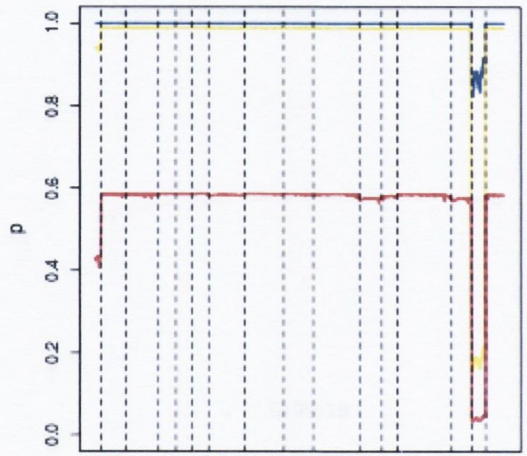
CHMP2B



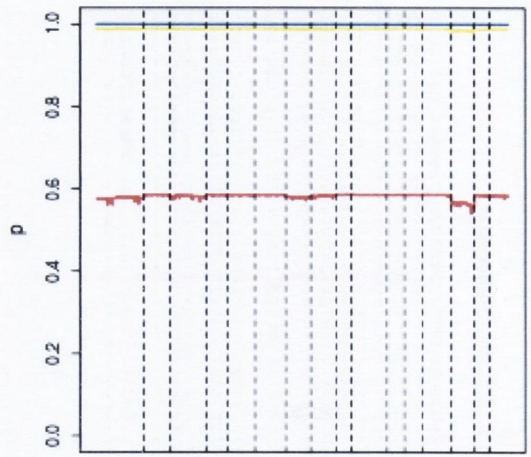
DCTN1



DPP6



ELP3



FGGY

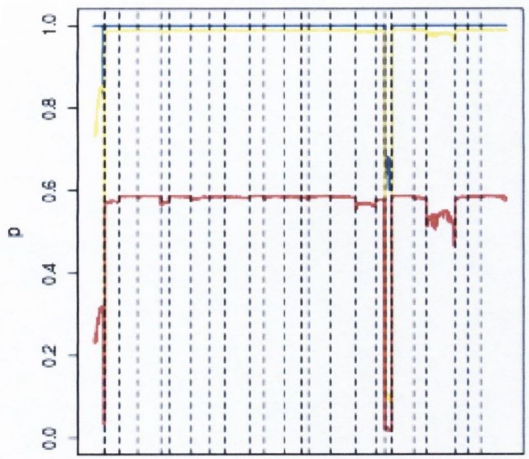
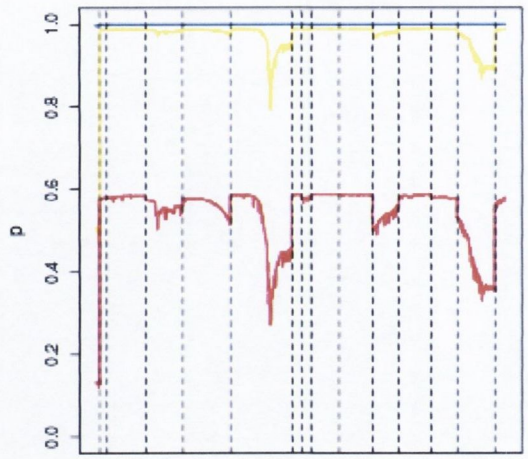
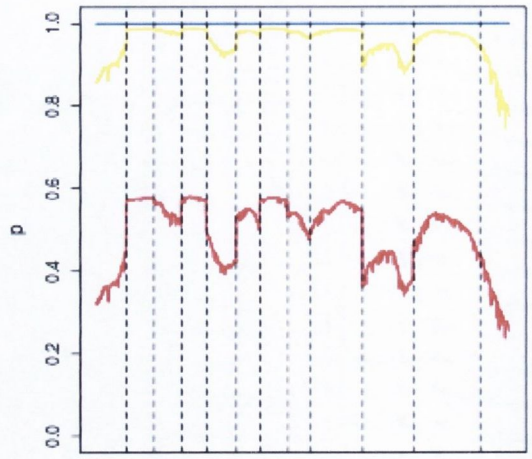


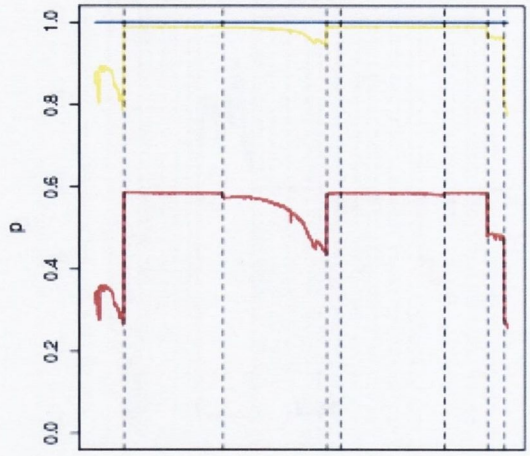
FIG4



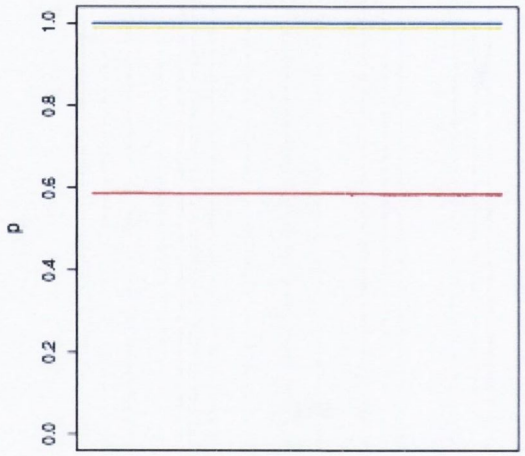
FUS



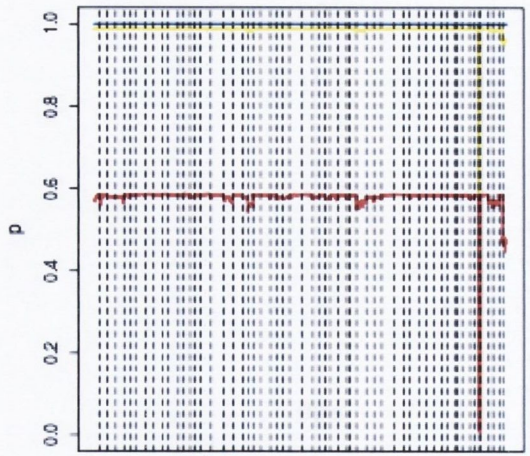
GRN



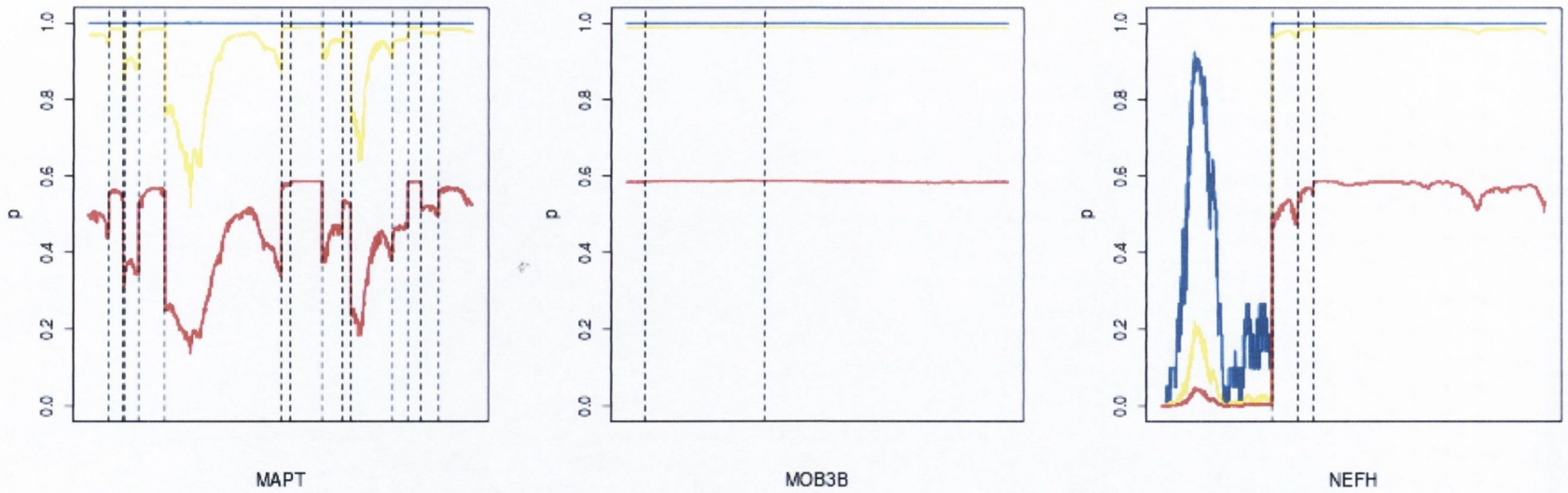
HFE

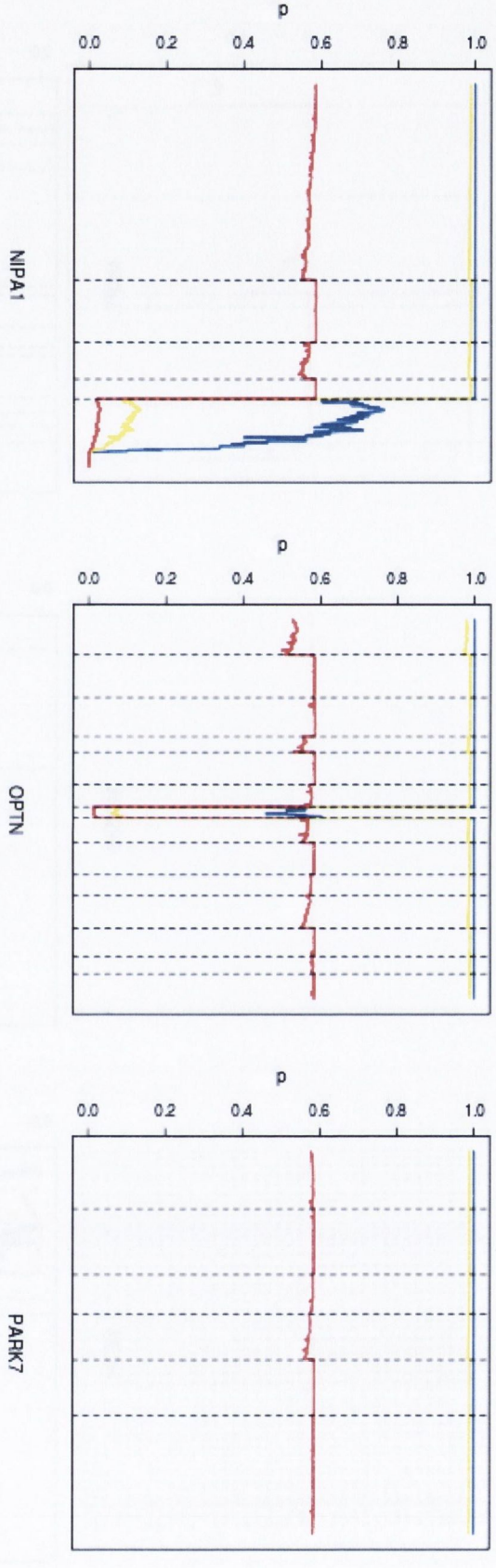


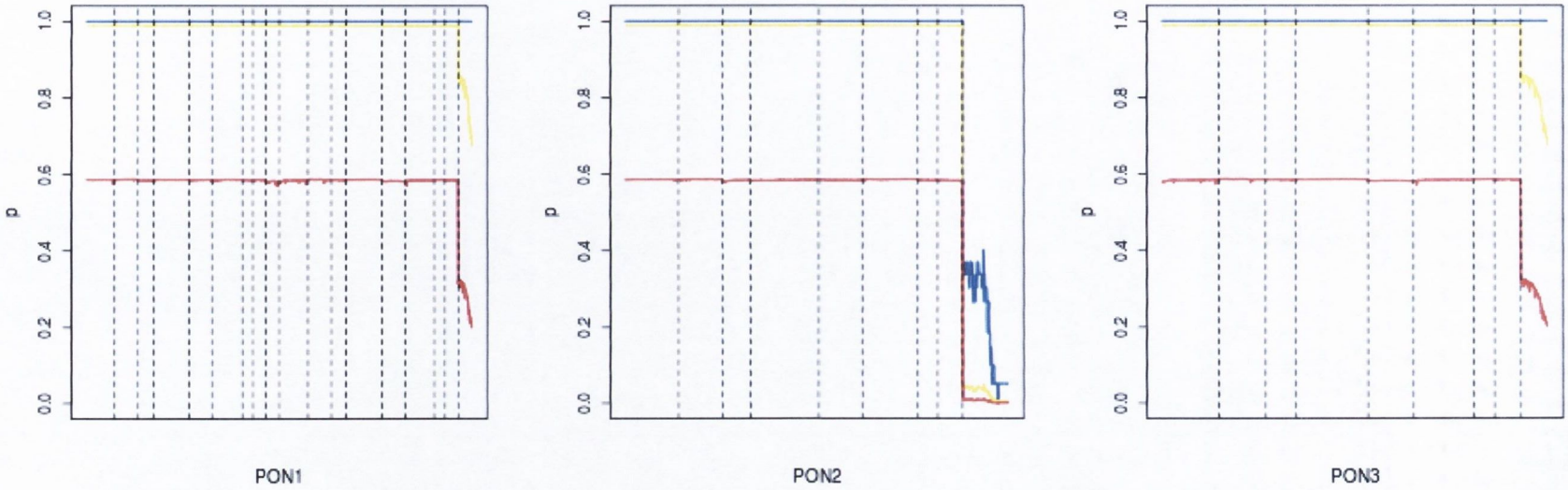
IFNK

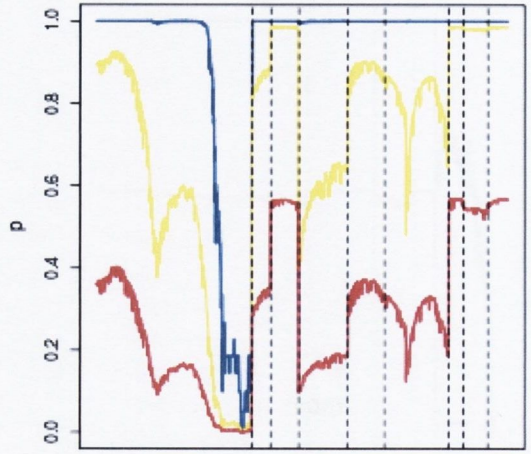


ITPR2

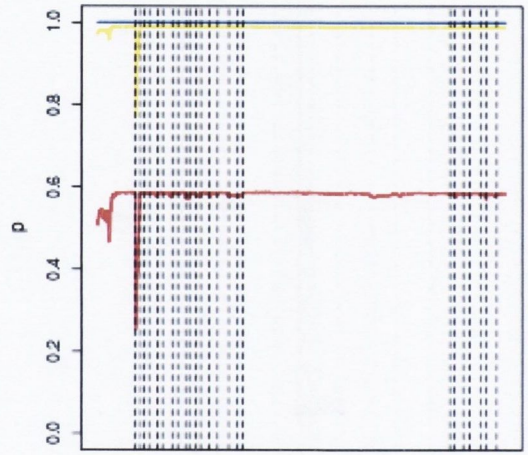

p
0.0
0.2
0.4
0.6
0.8
1.0
MAPT
MOB3B
NEFH


p
0.0
0.2
0.4
0.6
0.8
1.0
NIPA1
p
0.0
0.2
0.4
0.6
0.8
1.0
OPTN
p
0.0
0.2
0.4
0.6
0.8
1.0
PARK7

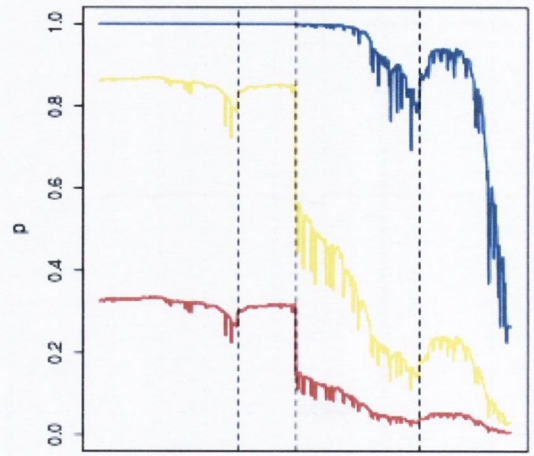

p
0.0
0.2
0.4
0.6
0.8
1.0
PON1
PON2
PON3



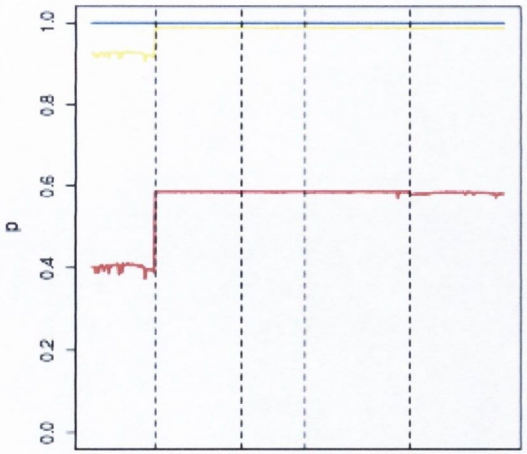
PRPH



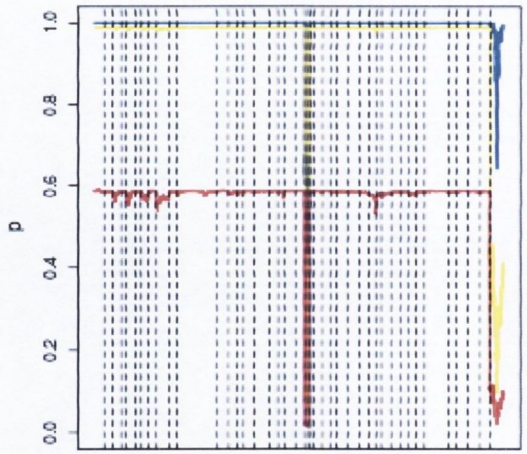
SETX



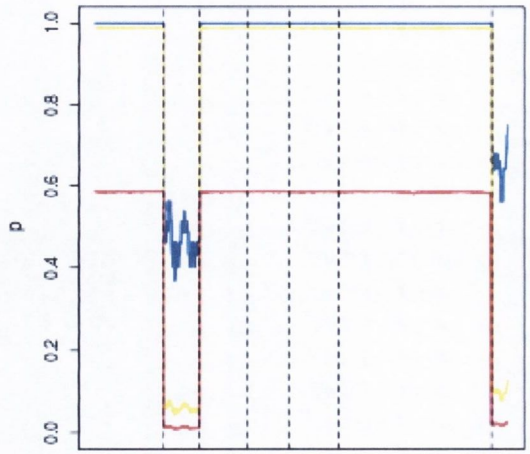
SIGMAR1



SOD1



SPG11



TARDBP

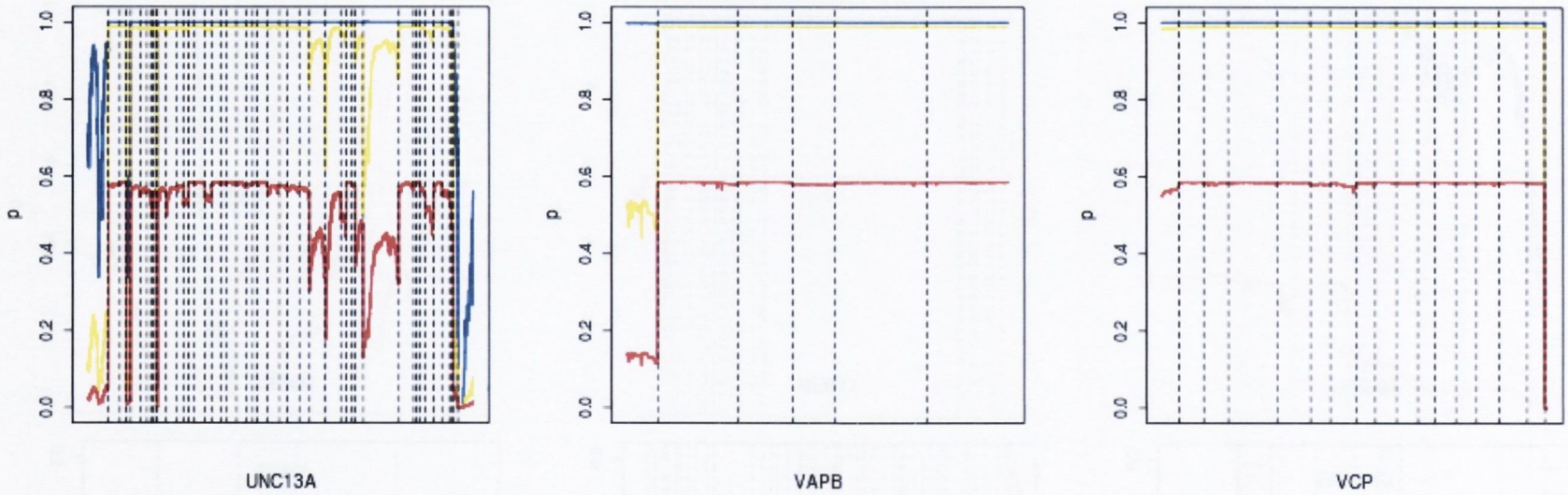


Figure S5.1: Variant detection power . The horizontal axes denote position within the coding sequence of each target gene. The vertical axes denote the probability that a variant with an allele frequency of 5% (blue line), 0.5% (yellow line) or 0.1% (red line) among Irish ALS patients was sampled during resequencing. The coding sequence of a target gene refers to the union of all coding transcripts reported in Ensembl 69. Exon boundaries are demarcated by vertical dotted lines

Table S5.1: ALS gene variants

Gene	DNA	Protein	SIFT; PP	OR	p	fALS	sALS	Irish	European	Global
TARDBP	c.312C>T, c.435C>T	p.(=)		Inf	1.0E+00	50,0,0	382,1,0	270,0,0	270,0,0	270,0,0
TARDBP	c.859G>A, c.511G>A	p.Gly287Ser, p.Gly171Ser	0;0	Inf	5.2E-01	50,0,0	388,2,0	300,0,0	300,0,0	300,0,0
TARDBP	c.975C>T, c.627C>T	p.(=)		0.33	5.7E-01	49,0,0	363,1,0	272,2,0	272,2,0	272,2,0
TARDBP	c.1122T>C, c.774T>C	p.(=)		0.00	4.1E-01	50,0,0	385,0,0	301,1,0	4601,1,0	6803,2,0
TARDBP	c.1129T>A, c.781T>A	p.Ser377Thr, p.Ser261Thr	0;1	0.00	4.1E-01	50,0,0	386,0,0	302,1,0	302,1,0	302,1,0
FGGY	c.49G>A	p.Val17Ile	1;2	1.90	1.0E+00	47,1,0	308,2,0	225,1,0	3800,8,0	5367,9,0
FGGY	c.129T>G	p.Asn43Lys	0;0	0.88	2.7E-01	17,24,8	143,187,31	117,139,46	1848,1892,521	2391,3053,1098
FGGY	c.188G>C	p.Cys63Ser	1;1	Inf	1.0E+00	49,0,0	351,1,0	293,0,0	4251,2,0	6529,5,0
FGGY	**c.411_412insG**	**p.Val140GlyfsX25**		**Inf**	**1.0E+00**	**50,0,0**	**365,1,0**	**299,0,0**	**299,0,0**	**299,0,0**
FGGY	c.516G>A, c.252G>A	p.(=)		0.89	3.0E-01	17,24,9	137,199,49	91,151,44	1899,2327,737	2388,3679,1812
FGGY	c.516G>C, c.252G>C	p.(=)		Inf	1.0E+00	50,0,0	384,1,0	286,0,0	286,0,0	286,0,0
FGGY	c.614G>A, c.350G>A	p.Ser205Asn, p.Ser117Asn	0;0	Inf	1.0E+00	50,0,0	388,1,0	284,0,0	4582,2,0	6785,2,0

FGGY	**c.622A>G, c.358A>G**	**p.Lys208Glu, p.Lys120Glu**	**0;0**	**0.00**	**3.9E-01**	**50,0,0**	**389,0,0**	**280,1,0**	**280,1,0**	**280,1,0**
FGGY	c.630T>A, c.366T>A	p.(=)		Inf	1.0E+00	50,0,0	387,1,0	281,0,0	281,0,0	281,0,0
FGGY	c.800G>T, c.536G>T	p.Gly267Val, p.Gly179Val	1;2	0.18	1.7E-01	49,0,0	359,1,0	294,4,0	4958,18,0	7872,20,0
FGGY	c.857C>T, c.593C>T	p.Thr286Met, p.Thr198Met	1;2	0.82	2.7E-01	39,5,2	292,55,3	238,46,8	4364,579,27	7223,635,28
FGGY	c.903G>T, c.639G>T, c.6G>T	p.(=)		0.74	1.0E+00	50,0,0	357,1,0	302,1,0	4600,3,0	6803,3,0
FGGY	c.979A>C, c.715A>C, c.82A>C	p.Asn327His, p.Asn239His, p.Asn28His	1;2	Inf	1.0E+00	50,0,0	383,1,0	295,0,0	4586,9,0	6789,9,0
FGGY	c.1053A>G, c.156A>G, c.789A>G	p.(=)		0.32	5.6E-01	50,0,0	389,1,0	281,2,0	4580,3,0	6783,3,0
*FGGY	c.1221+2T>C, c.957+2T>C, c.324+2T>C			0.78	7.8E-01	49,1,0	384,6,0	287,6,0	4924,47,0	7835,52,0
FGGY	c.1560C>T, c.1488C>T, c.1224C>T, c.591C>T	p.(=)		0.00	4.5E-01	46,0,0	310,0,0	285,1,0	285,1,0	285,1,0
FGGY	c.1716G>A, c.747G>A, c.1644G>A, c.1380G>A	p.Met572Ile, p.Met249Ile, p.Met548Ile, p.Met460Ile	0;1	Inf	1.0E+00	48,0,0	375,1,0	285,0,0	285,0,0	285,0,0
PARK7	**c.49A>C**	**p.Met17Leu**	**1;0**	**Inf**	**1.0E+00**	**49,0,0**	**375,1,0**	**283,0,0**	**661,0,0**	**1373,1,0**
PARK7	c.234C>T, c.117C>T	p.(=)		Inf	1.0E+00	48,0,0	345,1,0	240,0,0	4906,12,0	7317,490,27
PARK7	c.293G>A, c.233G>A,	p.Arg98Gln,	0;0	0.3	1.8E-01	46,0,0	294,3,0	239,6,0	4809,113,1	7715,123,1

	c.176G>A	p.Arg78Gln, p.Arg59Gln		5						
PARK7	c.501A>G, c.414A>G, c.441A>G	p.(=)		0.35	5.7E-01	50,0,0	383,1,0	305,2,0	4967,18,0	7881,20,0
PARK7	c.535G>A, c.475G>A, c.448G>A	p.Ala179Thr, p.Ala159Thr, p.Ala150Thr	0;0	0.00	4.2E-01	50,0,0	378,0,0	303,1,0	4976,6,0	7890,8,0
OPTN	c.102G>A	p.(=)		0.91	4.8E-01	24,16,3	120,112,12	104,102,14	2564,1980,354	5049,2389,376
OPTN	c.123G>A	p.(=)		0.94	1.0E+00	39,2,0	225,8,0	224,9,0	4806,104,1	7710,116,1
OPTN	**c.187C>A, c.16C>A**	**p.Gln63Lys, p.Gln6Lys**	**1;2**	**0.66**	**1.0E+00**	**49,0,0**	**361,1,0**	**271,1,0**	**271,1,0**	**271,1,0**
OPTN	c.213G>A, c.42G>A	p.(=)		Inf	1.0E+00	49,1,0	384,0,0	284,0,0	284,0,0	284,0,0
OPTN	c.287G>A, c.116G>A	p.Arg96His, p.Arg39His	0;0	0.00	3.9E-01	47,0,0	368,0,0	264,1,0	642,1,0	1354,2,0
OPTN	c.293T>A, c.122T>A	p.Met98Lys, p.Met41Lys	0;0	0.83	5.4E-01	48,2,0	357,21,1	267,18,1	4654,303,7	6956,885,39
OPTN	c.489A>G, c.318A>G	p.(=)		Inf	1.0E+00	50,0,0	388,1,0	306,0,0	4968,16,0	7881,19,0
OPTN	c.513C>T, c.342C>T	p.(=)		0.00	4.1E-01	50,0,0	384,0,0	299,1,0	4598,2,0	6801,2,0
OPTN	c.964A>G, c.946A>G	p.Lys322Glu, p.Lys316Glu	0;0	Inf	3.7E-01	0,0,44	0,0,294	0,1,195	0,12,4862	2,138,7650
OPTN	c.1192C>G, c.1174C>G	p.Gln398Glu, p.Gln392Glu	0;0	Inf	1.0E+00	50,0,0	367,1,0	243,0,0	243,0,0	243,0,0
OPTN	c.1569G>A, c.1551G>A	p.(=)		Inf	1.0E+00	48,0,0	341,1,0	279,0,0	279,0,0	279,0,0

ATXN2	c.3708G>A, c.873G>A, c.483G>A, c.2913G>A, c.2919G>A, c.2787G>A, c.711G>A	p.(=)	Inf	2.8E-01	37,0,0	207,3,0	165,0,0	4805,38,0	7714,45,0
ATXN2	c.3491G>A, c.266G>A, c.656G>A, c.494G>A, c.145G>A, c.2702G>A, c.2570G>A, c.2696G>A	p.Ser1164Asn, p.Ser89Asn, p.Ser219Asn, p.Ser165Asn, p.Val49Ile, p.Ser901Asn, p.Ser857Asn, p.Ser899Asn	Inf	1.0E+00	36,1,0	210,0,0	165,0,0	4464,1,0	6667,1,0
ATXN2	c.3411C>T, c.2490C>T, c.186C>T, c.2616C>T, c.576C>T, c.2622C>T, c.414C>T, c.65C>T	p.(=), p.Thr22Met	Inf	1.0E+00	37,0,0	209,1,0	165,0,0	165,0,0	165,0,0
ATXN2	c.3322C>T, c.97C>T, c.487C>T, c.2527C>T, c.2533C>T	p.Pro1108Ser, p.Pro33Ser, p.Pro163Ser, p.Pro843Ser, p.Pro845Ser	0;2	0.21	3.0E-01	37,0,0	200,1,0	147,3,0	4811,17,0	7724,20,0
ATXN2	c.3000A>G, c.2205A>G, c.2939A>G, c.2211A>G, c.165A>G, c.2133A>G, c.57A>G	p.(=), p.Tyr980Cys	1.34	1.0E+00	37,0,0	208,2,0	164,1,0	4803,40,0	7712,47,0
ATXN2	c.1472A>G, c.605A>G, c.677A>G	p.Asn491Ser, p.Asn202Ser, p.Asn226Ser	0;0	Inf	1.6E-01	36,1,0	206,4,0	163,0,0	4821,20,0	7733,24,0

ATXN2	c.1123G>A, c.256G>A, c.328G>A, c.151G>A, c.352G>A	p.Asp375Asn, p.Asp86Asn, p.Asp110Asn, p.Asp51Asn, p.Asp118Asn	1;0	0.00	3.9E-01	37,0,0	210,0,0	158,1,0	158,1,0	158,1,0
ATXN2	c.743G>A	p.Ser248Asn	0;1	0.47	4.4E-01	36,0,0	197,3,0	145,4,0	4790,37,0	6858,776,108
ATXN2	**c.319C>G**	**p.Leu107Val**	**0;**	**0.00**	**1.0E+00**	**0,0,0**	**0,0,0**	**0,0,0**	**19,122,237**	**416,351,324**
ITPR2	c.8002G>A	p.Ala2668Thr	0;0	1.18	1.0E+00	48,0,0	356,5,0	287,3,0	4884,28,0	7671,33,0
ITPR2	**c.7518G>A**	**p.(=)**		**0.00**	**4.0E-01**	**48,0,0**	**377,0,0**	**286,1,0**	**286,1,0**	**286,1,0**
ITPR2	c.7401T>G	p.(=)		0.00	1.6E-01	50,0,0	376,0,0	277,2,0	4357,3,0	6193,3,0
ITPR2	c.6966T>C	p.(=)		1.03	1.0E+00	50,0,0	388,3,0	302,2,0	4420,5,0	6317,5,0
ITPR2	c.6833T>C	p.Leu2278Pro	1;1	0.65	1.0E+00	49,1,0	383,0,0	281,1,0	4757,2,0	7318,3,0
ITPR2	c.6529A>C	p.Lys2177Gln	1;0	0.00	3.9E-01	50,0,0	390,0,0	278,1,0	4383,7,0	6242,7,0
ITPR2	c.6390G>A	p.(=)		Inf	1.0E+00	50,0,0	378,1,0	277,0,0	4402,1,0	6329,1,0
ITPR2	c.6389C>T	p.Ser2130Leu	0;0	Inf	1.0E+00	50,0,0	386,1,0	282,0,0	4404,1,0	6326,1,0
ITPR2	**c.6346G>C**	**p.Ala2116Pro**	**1;2**	**0.93**	**1.0E+00**	**48,1,0**	**375,5,0**	**263,4,0**	**263,4,0**	**263,4,0**

ITPR2	c.6022G>A	p.Ala2008Thr	1;2	0.10	2.0E-02	46,0,0	350,1,0	203,5,0	4302,6,0	6152,6,0
ITPR2	c.5569C>T	p.(=)		0.20	1.5E-01	50,0,0	384,1,0	255,3,0	4698,18,0	7239,19,0
ITPR2	c.5331C>T	p.(=)		0.00	3.8E-01	49,0,0	379,0,0	263,1,0	4360,2,0	6211,2,0
ITPR2	c.5175G>A	p.(=)		1.09	5.4E-01	27,19,3	227,129,14	166,81,12	2888,1631,230	5279,1829,234
ITPR2	c.5118T>C	p.(=)		Inf	1.0E+00	48,0,0	367,1,0	256,0,0	256,0,0	256,0,0
ITPR2	c.4962G>A	p.(=)		3.25	4.1E-01	50,0,0	383,5,0	283,1,0	4804,34,1	7511,39,1
ITPR2	c.4482C>T	p.(=)		1.05	6.8E-01	18,28,2	151,175,43	106,122,25	2194,2023,490	4106,2590,532
ITPR2	c.4407A>G	p.(=)		1.12	3.3E-01	17,29,3	144,186,50	106,126,27	2044,2124,557	3976,2691,598
ITPR2	**c.4305A>G**	**p.(=)**		**Inf**	**1.0E+00**	**48,0,0**	**377,1,0**	**248,0,0**	**248,0,0**	**248,0,0**
ITPR2	c.4239C>T	p.(=)		1.07	5.7E-01	16,29,4	139,183,49	105,133,31	2087,2179,577	4220,2762,624
ITPR2	c.4125C>T	p.(=)		1.16	6.0E-01	49,1,0	355,32,3	278,21,1	4596,244,6	7288,272,6
ITPR2	c.4023C>T	p.(=)		Inf	1.0E+00	50,0,0	388,1,0	293,0,0	293,0,0	293,0,0
ITPR2	c.3824G>A	p.Arg1275Gln	0;0	0.28	4.1E-02	47,1,0	367,3,0	256,9,0	4771,51,0	7458,55,0
ITPR2	c.3635C>T	p.Ala1212Val	0;1	0.00	5.7E-02	50,0,0	373,0,0	261,3,0	261,3,0	261,3,0
ITPR2	c.3614G>A	p.Arg1205Gl	0;2	Inf	1.0E+0	50,0,0	370,1,0	266,0,0	266,0,0	266,0,0

		n			0					
ITPR2	c.3539G>A	p.Arg1180Gl n	0;0	2.0 7	5.0E-01	45,2,0	341,5,0	230,2,0	4631,68,0	7164,78,0
ITPR2	c.3485T>G	p.Val1162Gly	0;0	Inf	5.3E-01	44,1,0	354,1,0	249,0,0	4335,10,0	6173,12,0
ITPR2	c.3369A>G	p.(=)		0.8 0	6.4E-01	50,0,0	369,10,0	267,8,0	4629,150,1	7222,159,1
ITPR2	c.3358C>T	p.Arg1120X		0.0 0	3.8E-01	49,0,0	363,0,0	256,1,0	256,1,0	256,1,0
ITPR2	c.3117G>A	p.(=)		Inf	1.0E+0 0	47,0,0	368,1,0	245,0,0	4687,2,0	7202,2,0
ITPR2	c.2883C>T	p.(=)		1.0 1	1.0E+0 0	9,29,6	72,146,94	42,140,54	1234,2385,121 1	1857,3718,204 8
ITPR2	c.2854C>T	p.Pro952Ser	0;0	0.7 8	1.0E+0 0	48,0,0	304,1,0	274,1,0	4496,2,0	6598,2,0
ITPR2	c.2694C>T	p.(=)		0.9 0	4.4E-01	30,16,4	223,129,2 8	141,109,1 5	2689,1726,310	4946,1980,328
ITPR2	c.2358C>T	p.(=)		0.7 5	5.6E-02	36,14,0	284,83,12	196,92,7	3386,1347,135	5162,2144,306
ITPR2	c.2262A>C	p.(=)		0.7 5	1.4E-01	40,9,1	334,52,3	244,60,1	3989,833,50	6660,918,52
ITPR2	c.2055A>G	p.(=)		0.9 3	5.3E-01	15,27,8	183,154,5 1	113,137,3 3	1984,2198,593	3391,3137,860
ITPR2	c.2018T>C	p.Met673Thr	0;0	Inf	1.0E+0 0	50,0,0	390,1,0	287,0,0	4382,1,0	6240,1,0
ITPR2	c.2016A>C	p.(=)		0.7 4	3.8E-02	36,14,0	295,83,13	190,89,7	3294,1331,135	5101,1954,275
ITPR2	c.1905A>G	p.(=)		1.0 6	7.2E-01	34,13,1	254,113,1 3	177,75,7	3241,1332,144	4465,2379,406
ITPR2	c.1857C>T	p.(=)		Inf	1.0E+0	49,0,0	383,1,0	268,0,0	4702,22,0	7240,23,0

					0					
ITPR2	c.1674C>T	p.(=)		1.37	1.0E+00	48,0,0	354,2,0	276,1,0	4396,2,0	6289,2,0
ITPR2	c.1358C>T	p.Ala453Val	0;0	0.93	7.7E-01	41,6,1	358,24,0	242,19,1	4318,413,8	6837,446,8
ITPR2	c.1318C>T	p.Arg440X		0.00	3.9E-01	49,0,0	384,0,0	275,1,0	275,1,0	275,1,0
ITPR2	c.798G>A	p.(=)		1.87	5.7E-02	47,2,0	346,30,3	255,13,0	4573,176,1	6327,896,117
ITPR2	c.765A>G	p.(=)		1.04	7.5E-01	8,32,10	120,175,94	76,152,55	1341,2380,1055	3448,2822,1098
ITPR2	**c.570G>C**	**p.(=)**		**0.22**	**3.1E-01**	**48,0,0**	**382,1,0**	**281,3,0**	**281,3,0**	**281,3,0**
PRPH	c.26G>A	p.Arg9Gln	1;1	0.77	1.0E+00	18,1,0	43,2,0	31,2,0	4585,111,1	7459,145,2
PRPH	c.63C>T	p.(=)		0.66	5.3E-01	24,0,0	74,7,0	45,5,0	4516,169,0	7272,309,2
PRPH	**c.322T>C, c.79T>C**	**p.Phe108Leu, p.Phe27Leu**	**1;2**	**Inf**	**1.0E+00**	**2,0,0**	**4,1,0**	**5,0,0**	**5,0,0**	**5,0,0**
PRPH	**c.829G>A, c.14G>A**	**p.Ala277Thr, p.Ala6Thr**	**0;0**	**0.00**	**1.0E+00**	**2,0,0**	**2,0,0**	**2,0,0**	**4511,169,0**	**7276,319,1**
PRPH	**c.1083C>G, c.268C>G**	**p.(=)**		**Inf**	**1.0E+00**	**0,0,0**	**1,2,0**	**1,0,0**	**4535,129,0**	**7383,188,0**
PRPH	**c.1104A>G, c.289A>G**	**p.(=)**		**0.00**	**1.0E+00**	**0,0,0**	**9,0,0**	**7,0,0**	**4305,1,0**	**6507,1,0**
PRPH	**c.1107A>G, c.292A>G**	**p.(=)**		**0.86**	**8.3E-01**	**0,5,0**	**6,21,0**	**4,12,2**	**3018,1458,219**	**4431,2637,543**
PRPH	**c.1225G>A, c.410G>A**	**p.Val409Met,**	**0;0**	**0.16**	**1.5E-02**	**41,0,0**	**271,2,0**	**176,7,0**	**176,7,0**	**176,7,0**

PRPH	**c.1231G>A, c.416G>A**	**p.Val138Met p.Val411Ile, p.Val140Ile**	**0;0**	**0.00**	**8.3E-03**	**43,0,0**	**271,0,0**	**191,5,0**	**191,5,0**	**191,5,0**
PRPH	c.1303C>T, c.488C>T	p.Arg435Trp, p.Arg164Trp	1;2	0.76	1.0E+00	41,0,0	244,1,0	215,1,0	4889,5,0	7802,8,0
ANG	c.122A>T	p.Lys41Ile	1;0	0.69	7.5E-01	50,0,0	361,5,0	283,5,0	4938,28,0	7850,32,0
ANG	c.208A>G	p.Ile70Val	0;0	0.35	5.7E-01	50,0,0	387,1,0	301,2,0	4596,7,0	6797,9,0
ANG	c.330T>G	p.(=)		0.78	1.1E-01	32,15,3	304,76,7	220,76,10	3725,1169,90	5982,1771,147
ANG	c.363A>T	p.(=)		0.00	4.1E-01	50,0,0	388,0,0	305,1,0	4978,6,0	7669,223,8
NIPA1	c.441A>G, c.216A>G	p.(=)		1.02	9.0E-01	1,17,31	32,141,214	17,120,159	357,1937,2680	1027,3268,3595
NIPA1	**c.233T>G, c.8T>G**	**p.Val78Gly, p.Val3Gly**	**1;0**	**1.09**	**7.5E-01**	**32,1,0**	**174,51,0**	**200,46,0**	**200,46,0**	**200,46,0**
SPG11	c.7324G>C, c.6985G>C, c.1695G>C	p.Ala2442Pro, p.Ala2329Pro, p.Ala566Pro	1;2	Inf	1.0E+00	50,0,0	390,1,0	282,0,0	282,0,0	282,0,0
SPG11	c.7197G>A, c.1568G>A, c.6858G>A	p.(=)		1.13	1.0E+00	50,0,0	382,7,0	278,4,0	4838,120,0	7740,129,0
SPG11	c.7069C>T, c.6730C>T, c.1440C>T	p.Leu2357Phe, p.Leu2244Phe, p.Leu481Phe	1;2	1.91	1.0E+00	49,1,0	383,2,0	275,1,0	4929,23,0	7834,29,0

SPG11	c.7023C>T, c.714C>T, c.6684C>T, c.1394C>T	p.(=)		1.21	5.6E-01	46,4,0	353,30,0	257,18,0	4652,299,0	7533,328,1
SPG11	**c.6952dupC, c.643dupC, c.6613dupC, c.1323dupC**	**p.Arg2318ProfsX22, p.Arg215ProfsX22, p.Arg2205ProfsX22, p.Arg442ProfsX22**		**1.12**	**6.7E-01**	**30,5,0**	**156,51,0**	**166,44,0**	**166,44,0**	**166,44,0**
SPG11	c.6872G>A, c.6533G>A, c.1243G>A, c.563G>A	p.Cys2291Tyr, p.Cys2178Tyr, p.Cys415Tyr, p.Cys188Tyr	1;2	0.00	4.1E-01	50,0,0	383,0,0	302,1,0	302,1,0	302,1,0
SPG11	c.6632G>A, c.6293G>A, c.1003G>A, c.323G>A	p.Arg2211His, p.Arg2098His, p.Arg335His, p.Arg108His	0;2	Inf	1.5E-01	48,1,0	359,3,0	294,0,0	4577,15,0	6775,15,0
SPG11	c.6497T>C, c.868T>C, c.6158T>C	p.Ile2166Thr, p.Ile290Thr, p.Ile2053Thr	1;2	Inf	1.0E+00	50,0,0	386,1,0	303,0,0	4598,1,0	6794,1,0
SPG11	c.6330G>A, c.5991G>A, c.129G>A	p.(=)		0.40	2.8E-01	50,0,0	385,3,0	289,5,0	4917,53,0	7690,182,9
SPG11	c.6258G>T, c.57G>T, c.5919G>T	p.(=)		1.13	1.0E+00	49,1,0	385,4,0	294,3,0	4916,57,0	7816,68,0
SPG11	c.6224A>G, c.23A>G, c.5885A>G	p.Asn2075Ser, p.Asn8Ser, p.Asn1962Ser	0;0	0.23	3.1E-01	50,0,0	387,1,0	296,3,0	4925,50,0	7830,56,0

SPG11	c.5911A>G	p.Asn1971Asp	0;0	0.00	4.0E-01	50,0,0	383,0,0	293,1,0	293,1,0	293,1,0
SPG11	c.5796T>C, c.644T>C	p.(=)		0.00	4.1E-01	50,0,0	380,0,0	299,1,0	4596,2,0	6794,2,0
SPG11	**c.5757_5758delAG, c.605_606delAG**	**p.Glu1921SerfsX2, p.Glu204SerfsX2**		**Inf**	**1.0E+00**	**50,0,0**	**389,1,0**	**302,0,0**	**302,0,0**	**302,0,0**
SPG11	c.5361C>T, c.209C>T	p.(=)		Inf	5.3E-01	46,1,0	340,1,0	243,0,0	4916,3,0	7826,4,0
SPG11	c.5121G>T	p.Glu1707Asp	1;2	0.00	4.1E-01	50,0,0	389,0,0	302,1,0	4968,11,0	7878,12,0
SPG11	c.4343G>A	p.Cys1448Tyr	1;1	Inf	1.0E+00	50,0,0	387,1,0	300,0,0	300,0,0	300,0,0
SPG11	c.4261G>A	p.Asp1421Asn	0;0	0.00	1.6E-01	50,0,0	389,0,0	288,2,0	666,2,0	1378,3,0
SPG11	c.4260C>T	p.(=)		0.00	4.0E-01	50,0,0	389,0,0	295,1,0	4590,4,0	6788,4,0
SPG11	c.4026A>G	p.(=)		0.98	1.0E+00	50,0,0	388,3,0	287,2,0	4576,11,0	6774,11,0
SPG11	c.3988C>G	p.Gln1330Glu	0;1	1.92	1.0E+00	49,0,0	364,3,0	265,1,0	4562,2,0	6760,2,0
SPG11	c.3980T>C	p.Ile1327Thr	1;2	0.63	6.4E-01	50,0,0	369,2,0	263,2,0	263,2,0	263,2,0
SPG11	c.3680A>G, c.407A>G	p.Lys1227Arg, p.Lys136Arg	1;2	Inf	1.0E+00	49,1,0	390,0,0	302,0,0	302,0,0	302,0,0
SPG11	c.3446T>C, c.175T>C, c.173T>C	p.Leu1149Pro, p.Leu59Pro, p.Leu58Pro	1;2	0.00	3.9E-01	50,0,0	391,0,0	285,1,0	285,1,0	285,1,0
SPG11	c.3420G>A, c.149G>A,	p.(=)		0.9	7.1E-01	44,6,0	355,35,0	259,25,2	4455,490,17	7301,554,18

	c.147G>A			2						
SPG11	c.3037A>G	p.Lys1013Glu	0;0	2.52	3.3E-01	49,1,0	375,7,0	269,2,0	4811,135,1	7712,145,1
SPG11	c.2887A>C	p.(=)		0.64	6.5E-01	49,1,0	387,1,0	279,2,0	4943,14,0	7852,16,0
SPG11	c.2577A>C	p.Gln859His	0;1	Inf	1.0E+00	50,0,0	379,1,0	267,0,0	267,0,0	267,0,0
SPG11	**c.2377G>A**	**p.Val793Met**	**1;2**	**0.37**	**3.5E-01**	**44,0,0**	**345,2,0**	**212,3,0**	**212,3,0**	**212,3,0**
SPG11	c.2318T>G	p.Val773Gly	1;2	0.00	3.6E-01	29,0,0	272,0,0	166,1,0	4812,4,0	7683,4,0
SPG11	c.2083G>A	p.Ala695Thr	1;2	0.73	4.9E-01	48,0,0	343,11,0	234,7,1	4786,131,1	7681,147,1
SPG11	c.1930A>T	p.Thr644Ser	0;2	Inf	1.0E+00	48,1,0	382,0,0	265,0,0	265,0,0	265,0,0
SPG11	c.1698T>G, c.396T>G	p.Asp566Glu, p.Asp132Glu	0;2	1.36	5.5E-01	46,2,0	359,17,0	234,8,0	4740,177,0	7629,197,1
SPG11	c.1605C>T, c.303C>T	p.(=)		Inf	5.3E-01	48,0,0	356,2,0	236,0,0	4909,3,0	7820,3,0
SPG11	c.1529G>A, c.227G>A	p.Ser510Asn, p.Ser76Asn	0;2	Inf	1.0E+00	50,0,0	388,1,0	282,0,0	282,0,0	282,0,0
SPG11	c.1388T>C, c.86T>C	p.Phe463Ser, p.Phe29Ser	0;0	0.84	1.1E-01	22,21,7	119,211,57	81,158,55	1494,2435,1041	2163,3866,1852
SPG11	c.1348A>G, c.46A>G	p.Ile450Val, p.Ile16Val	0;0	0.40	2.8E-01	50,0,0	387,3,0	289,5,0	4925,45,0	7715,161,5
SPG11	c.1347C>T, c.45C>T	p.(=)		0.40	2.8E-01	50,0,0	387,3,0	288,5,0	4925,44,0	7720,155,5
SPG11	c.1108G>A	p.Glu370Lys	0;2	0.78	5.4E-01	50,0,0	376,13,0	280,11,0	4768,198,1	7661,216,1

SPG11	c.979C>G	p.Leu327Val	0;1	0.00	3.9E-01	50,0,0	386,0,0	274,1,0	651,2,0	1364,2,0
SPG11	c.808G>A	p.Val270Ile	0;2	0.16	8.2E-02	50,0,0	389,1,0	282,4,0	4865,97,0	7729,144,0
SPG11	c.394A>G	p.Ser132Gly	0;0	Inf	1.0E+00	50,0,0	390,1,0	279,0,0	279,0,0	279,0,0
FUS	c.74A>G	p.Tyr25Cys	1;	0.00	4.1E-01	49,0,0	368,0,0	286,1,0	286,1,0	286,1,0
FUS	c.132C>T	p.(=)		0.00	4.2E-01	50,0,0	376,0,0	303,1,0	4601,3,0	6798,3,0
FUS	c.147C>A	p.(=)		0.94	5.7E-01	29,13,6	173,158,44	133,137,33	2475,2086,420	4866,2551,474
FUS	c.153C>T	p.(=)		0.00	7.3E-02	49,0,0	376,0,0	303,3,0	4947,37,0	7383,480,31
FUS	**c.165_167delTTC**	**p.Ser56del**		**0.00**	**4.2E-01**	**48,0,0**	**372,0,0**	**304,1,0**	**304,1,0**	**304,1,0**
FUS	**c.231G>C, c.228G>C**	**p.(=)**		**0.76**	**6.0E-02**	**22,10,0**	**95,122,0**	**77,140,0**	**77,140,0**	**77,140,0**
FUS	**c.232A>C, c.229A>C**	**p.Thr78Pro, p.Thr77Pro**	**0;**	**0.93**	**8.1E-01**	**31,2,0**	**141,36,0**	**151,36,0**	**151,36,0**	**151,36,0**
FUS	**c.234T>C, c.231T>C**	**p.(=)**		**0.99**	**1.0E+00**	**33,0,0**	**173,10,0**	**184,9,0**	**184,9,0**	**184,9,0**
FUS	c.291C>T, c.288C>T	p.(=)		0.92	4.9E-01	8,19,15	79,139,72	60,141,71	1000,2421,1529	2398,3518,1944
FUS	c.339C>T, c.336C>T	p.(=)		0.00	1.9E-01	49,0,0	318,0,0	277,2,0	277,2,0	277,2,0
FUS	**c.422A>C, c.419A>C**	**p.Gln141Pro, p.Gln140Pro**	**0;**	**Inf**	**1.0E+00**	**47,0,0**	**332,1,0**	**285,0,0**	**285,0,0**	**285,0,0**

FUS	**c.423A>T, c.420A>T**	**p.Gln141His, p.Gln140His**	**0;**	**Inf**	**1.0E+00**	**47,0,0**	**332,1,0**	**286,0,0**	**286,0,0**	**286,0,0**
FUS	c.1080C>T, c.1077C>T, c.1083C>T	p.(=)		0.66	6.5E-01	38,0,0	213,2,0	166,2,0	4841,5,0	7750,5,0
FUS	**c.1204_1206delAGT, c.1207_1209delAGT, c.1201_1203delAGT**	**p.Ser402del, p.Ser403del, p.Ser401del**		**Inf**	**1.0E+00**	**50,0,0**	**364,1,0**	**296,0,0**	**296,0,0**	**296,0,0**
FUS	**c.1545T>G, c.1542T>G, c.1548T>G**	**p.(=)**		**0.77**	**7.7E-01**	**38,0,0**	**259,6,0**	**227,6,0**	**227,6,0**	**227,6,0**
FUS	c.1566G>A, c.1569G>A, c.1563G>A	p.(=)		Inf	1.0E+00	47,0,0	337,1,0	280,0,0	4940,18,0	7845,23,0
FUS	c.1574C>T, c.1577C>T, c.1571C>T	p.Pro525Leu, p.Pro526Leu, p.Pro524Leu	1;2	Inf	5.2E-01	43,0,0	317,2,0	250,0,0	250,0,0	250,0,0
GRN	c.99C>T	p.(=)		2.06	4.4E-01	6,0,0	50,4,0	59,2,0	4698,41,0	7612,43,0
GRN	c.384T>C, c.426T>C	p.(=)		0.85	7.2E-01	49,1,0	337,16,0	290,13,1	4708,269,5	7455,419,24
GRN	**c.693C>G, c.134C>G, c.495C>G**	**p.Cys231Trp, p.Cys45Trp, p.Cys165Trp**	**1;2**	**0.47**	**3.1E-01**	**40,0,0**	**229,3,0**	**207,5,0**	**207,5,0**	**207,5,0**
GRN	c.705C>A, c.146C>A, c.507C>A	p.(=)		0.00	4.3E-01	37,0,0	232,0,0	206,1,0	4506,1,0	6707,3,0
GRN	c.970G>A, c.499G>A, c.411G>A	p.Ala324Thr, p.Ala167Thr, p.Ala138Thr	0;0	0.34	5.7E-01	37,0,0	212,1,0	169,2,0	4457,14,0	6654,20,0
GRN	c.1297C>T, c.738C>T, c.826C>T	p.Arg433Trp, p.Arg247Trp, p.Arg276Trp	1;1	0.42	3.8E-01	26,0,0	111,2,0	85,3,0	4734,32,0	7644,38,0

GRN	**c.1695C>T, c.1224C>T**	**p.(=)**		**0.45**	**4.0E-01**	**5,0,0**	**32,2,0**	**24,3,0**	**24,3,0**	**24,3,0**
GRN	**c.1762G>T, c.1291G>T**	**p.Ala588Ser, p.Ala431Ser**	**1;1**	**0.63**	**1.0E+00**	**1,0,0**	**2,1,0**	**5,3,0**	**5,3,0**	**5,3,0**
MAPT	c.284C>T	p.Thr95Met	0;0	Inf	1.0E+00	10,0,0	49,1,0	50,0,0	4348,1,0	6548,1,0
MAPT	c.7A>T	p.Lys3X		0.95	7.6E-01	26,10,1	154,88,12	126,75,9	352,205,31	991,272,38
MAPT	**c.605C>T**	**p.Pro202Leu**	**0;1**	**2.20**	**1.8E-01**	**0,2,0**	**0,8,1**	**7,12,1**	**2822,1566,254**	**5405,1839,268**
MAPT	c.671T>G	p.Val224Gly	1;1	0.00	4.4E-01	8,0,0	52,0,0	47,1,0	4684,39,0	7593,43,0
MAPT	c.689A>G	p.Gln230Arg	1;2	1.15	8.2E-01	9,3,0	71,9,0	63,6,1	4218,495,9	7028,586,11
MAPT	c.698C>T	p.Pro233Leu	0;2	Inf	1.0E+00	14,0,0	78,1,0	80,0,0	80,0,0	80,0,0
MAPT	c.853G>A	p.Asp285Asn	0;0	0.83	2.5E-01	20,9,2	120,70,13	91,68,12	2910,1684,255	5532,1963,270
MAPT	c.855C>T	p.(=)		0.79	2.6E-01	27,7,0	147,48,2	125,45,5	3756,1021,76	6277,1393,99
MAPT	c.866T>C	p.Val289Ala	0;0	0.95	7.6E-01	24,9,2	124,74,14	119,80,11	2938,1696,254	5560,1976,268
MAPT	c.1108C>T	p.Arg370Trp	1;2	0.77	2.9E-01	2,2,1	28,37,5	24,44,9	2730,1552,217	5265,1777,229
MAPT	c.1280C>T	p.Ser427Phe	1;2	0.52	4.5E-01	50,0,0	388,3,0	303,4,0	4962,23,0	7874,27,0
MAPT	**c.1306C>A**	**p.Pro436Thr**	**0;0**	**0.00**	**4.1E-01**	**50,0,0**	**390,0,0**	**306,1,0**	**306,1,0**	**306,1,0**
MAPT	c.1321T>C	p.Tyr441His	0;0	1.0	8.4E-01	37,12,1	257,126,8	210,87,9	3204,1578,202	5108,2414,378

				4						
MAPT	c.1339T>C	p.Ser447Pro	0;0	0.93	6.0E-01	36,12,2	250,123,18	190,105,12	3019,1712,254	5642,1992,267
MAPT	**c.1405G>A, c.454G>A, c.367G>A, c.247G>A, c.280G>A, c.172G>A**	**p.Ala469Thr, p.Ala152Thr, p.Ala123Thr, p.Ala83Thr, p.Ala94Thr, p.Ala58Thr**	**0;1**	**Inf**	**5.2E-01**	**3,0,0**	**36,2,0**	**27,0,0**	**4679,25,1**	**7591,29,1**
MAPT	c.1479G>A, c.321G>A, c.246G>A, c.528G>A, c.441G>A, c.354G>A	p.(=)		0.98	9.2E-01	4,6,1	37,62,4	37,50,8	2856,1667,250	5518,1910,261
MAPT	c.1512T>C	p.(=)		0.95	7.1E-01	27,12,1	154,91,12	146,93,10	2556,1434,215	4607,1643,228
MAPT	c.1686A>G, c.1632A>G, c.681A>G, c.594A>G, c.507A>G, c.474A>G, c.399A>G	p.(=)		1.00	1.0E+00	9,7,0	57,31,1	52,24,3	2892,1625,240	5508,1912,253
MAPT	c.1720G>A, c.433G>A, c.628G>A, c.1666G>A, c.541G>A, c.715G>A, c.508G>A	p.Ala574Thr, p.Ala145Thr, p.Ala210Thr, p.Ala556Thr, p.Ala181Thr, p.Ala239Thr, p.Ala170Thr	0;0	Inf	5.3E-01	18,0,0	92,2,0	66,0,0	4354,12,0	6551,18,0
MAPT	c.1770T>C, c.765T>C, c.591T>C, c.678T>C, c.1716T>C, c.558T>C, c.483T>C	p.(=)		1.05	8.4E-01	16,5,2	75,60,5	66,52,2	2887,1667,244	5509,1948,257

MAPT	c.1815G>A, c.636G>A, c.810G>A, c.1761G>A, c.723G>A, c.528G>A, c.603G>A	p.(=)		1.62	3.1E-01	8,3,0	69,16,0	55,8,0	4319,414,8	7017,631,9
MAPT	c.2178G>A, c.999G>A, c.2124G>A, c.1086G>A, c.906G>A, c.1080G>A, c.873G>A, c.1173G>A, c.798G>A	p.(=)		0.78	1.0E+00	45,0,0	325,1,0	287,1,0	4585,3,0	6788,3,0
UNC13A	**c.5079C>A, c.5061C>A, c.5136C>A, c.4998C>A, c.5343C>A**	**p.(=)**		**0.00**	**1.0E+00**	**0,0,0**	**0,1,0**	**0,0,0**	**3702,654,30**	**5596,1285,89**
UNC13A	c.4497G>A, c.4422G>A, c.4416G>A, c.4761G>A	p.(=)		1.23	9.7E-02	24,19,4	165,148,28	161,99,23	2827,1789,286	5371,2036,296
UNC13A	c.4380G>C, c.4644G>C, c.4305G>C, c.4299G>C	p.(=)		0.00	4.3E-01	45,0,0	299,0,0	255,1,0	255,1,0	255,1,0
UNC13A	**c.4259C>A, c.4523C>A**	**p.Pro1420Gln, p.Pro1508Gln**	**0;0**	**0.52**	**4.2E-05**	**30,7,0**	**114,86,1**	**63,117,0**	**63,117,0**	**63,117,0**
UNC13A	**c.4251G>A, c.4515G>A**	**p.(=)**		**0.68**	**2.5E-01**	**33,1,0**	**171,20,0**	**123,19,0**	**123,19,0**	**123,19,0**
UNC13A	c.4029C>T, c.4023C>T, c.4293C>T	p.(=)		0.65	1.3E-01	40,0,0	237,30,1	153,26,1	4251,541,18	7061,602,18
UNC13A	c.3687C>T, c.3681C>T, c.3951C>T	p.(=)		0.79	3.7E-01	50,0,0	346,33,1	273,31,0	4328,552,17	7029,650,17
UNC13A	c.3576T>C, c.3840T>C, c.3570T>C	p.(=)		0.96	7.4E-01	24,23,3	182,162,44	140,130,35	2152,2111,638	2520,3078,2119
UNC13A	c.3552A>G, c.3816A>G,	p.(=)		0.7	1.0E+0	49,1,0	377,1,0	301,2,0	4864,44,0	7694,49,0

	c.3546A>G			1	0				
UNC13A	**c.3461A>G, c.3725A>G, c.3455A>G**	**p.Asp1154Gly, p.Asp1242Gly, p.Asp1152Gly**	**1;1**	**0.40**	**2.1E-01**	**37,0,0**	**215,3,0**	**231,7,0**	**231,7,0**	**231,7,0**
UNC13A	c.3397G>T, c.3391G>T, c.3661G>T	p.Ala1133Ser, p.Ala1131Ser, p.Ala1221Ser	0;1	1.36	1.5E-01	40,9,1	306,67,0	260,42,0	4122,784,28	6898,851,28
UNC13A	c.3108G>A, c.3372G>A, c.3102G>A	p.(=)		0.71	1.5E-01	46,4,0	345,29,3	255,37,0	4245,548,16	6795,624,17
UNC13A	c.3101T>C, c.3365T>C, c.3095T>C	p.Leu1034Pro, p.Leu1122Pro, p.Leu1032Pro	0;0	0.00	1.0E+00	0,0,50	0,0,373	0,0,287	0,0,665	0,0,1378
UNC13A	c.3098T>A, c.3362T>A, c.3092T>A	p.Val1033Asp, p.Val1121Asp, p.Val1031Asp	0;0	Inf	1.0E+00	49,0,0	371,1,0	286,0,0	286,0,0	286,0,0
UNC13A	c.3080C>T, c.3344C>T, c.3074C>T	p.Pro1027Leu, p.Pro1115Le	0;2	0.67	7.2E-01	45,2,0	336,2,0	253,4,0	4383,58,0	6369,65,0

		u, p.Pro1025Leu								
UNC13A	**c.2626C>A, c.2890C>A, c.2620C>A**	**p.Gln876Lys, p.Gln964Lys, p.Gln874Lys**	**1;**	**0.77**	**4.8E-01**	**36,2,0**	**214,15,0**	**180,16,0**	**180,16,0**	**180,16,0**
UNC13A	c.2289T>C, c.2553T>C, c.2283T>C	p.(=)		Inf	1.0E+00	48,0,0	333,1,0	294,0,0	4559,1,0	6709,1,0
UNC13A	**c.2068G>A, c.2332G>A, c.2062G>A**	**p.Ala690Thr, p.Ala778Thr, p.Ala688Thr**	**1;2**	**0.00**	**4.5E-01**	**45,0,0**	**291,0,0**	**269,1,0**	**269,1,0**	**269,1,0**
UNC13A	c.1767C>G, c.2031C>G, c.1761C>G	p.(=)		1.23	5.7E-01	25,5,0	108,38,3	64,12,3	3943,781,33	6017,1536,119
UNC13A	c.1368C>T, c.1632C>T	p.(=)		0.71	1.0E+00	20,2,0	81,0,0	71,2,0	4334,20,0	6519,20,0
UNC13A	**c.1188C>T, c.1452C>T**	**p.(=)**		**0.00**	**1.0E+00**	**1,0,0**	**5,0,0**	**3,0,0**	**4155,6,0**	**6144,6,0**
UNC13A	c.1075G>A, c.1339G>A	p.Ala359Thr, p.Ala447Thr	0;1	1.09	6.6E-01	7,14,3	38,85,17	32,70,10	1562,2142,979	3649,2792,1092
UNC13A	**c.1049_1051delAGG, c.1313_1315delAGG**	**p.Glu350del, p.Glu438del**		**0.57**	**3.9E-01**	**14,0,0**	**89,6,0**	**67,7,0**	**67,7,0**	**67,7,0**
UNC13A	c.1031T>C, c.1295T>C	p.Leu344Pro, p.Leu432Pro	0;2	2.29	4.8E-01	15,2,0	109,4,0	96,2,0	4219,15,0	6313,16,0
UNC13A	**c.907T>C, c.1171T>C**	**p.Ser303Pro, p.Ser391Pro**	**0;2**	**1.00**	**1.0E+00**	**11,1,0**	**70,8,0**	**72,8,0**	**72,8,0**	**72,8,0**
UNC13A	**c.905A>C, c.1169A>C**	**p.His302Pro, p.His390Pro**	**0;2**	**1.68**	**5.2E-01**	**14,1,0**	**75,5,0**	**76,3,0**	**76,3,0**	**76,3,0**
UNC13A	c.771C>G, c.1035C>G	p.(=)		0.65	6.8E-01	6,0,0	49,3,0	35,3,0	4432,85,0	7098,97,0

UNC13A	c.694C>T, c.958C>T	p.Pro232Ser, p.Pro320Ser	0;1	0.00	4.2E-01	50,0,0	367,0,0	305,1,0	305,1,0	305,1,0
UNC13A	c.474G>C, c.738G>C	p.(=)		0.70	2.6E-02	1,10,39	12,100,270	4,61,233	89,1095,3618	214,1818,5405
UNC13A	c.182C>T, c.446C>T	p.Thr61Met, p.Thr149Met	0;2	0.71	1.0E+00	50,0,0	366,1,0	296,1,0	4810,19,0	7475,24,0
UNC13A	c.153C>T, c.417C>T	p.(=)		0.74	7.0E-01	48,1,0	360,2,0	302,3,0	4424,12,0	6330,12,0
ALS2	c.4957C>T	p.Arg1653Cys	1;1	0.00	3.7E-01	50,0,0	356,0,0	239,1,0	4337,2,0	6197,3,0
ALS2	c.4641G>A	p.(=)		0.32	5.6E-01	49,0,0	387,1,0	277,2,0	277,2,0	277,2,0
ALS2	c.4382G>A	p.Arg1461Gln	0;0	0.65	1.0E+00	50,0,0	384,1,0	282,1,0	282,1,0	282,1,0
ALS2	c.4135C>A	p.Pro1379Thr	1;1	0.00	1.6E-01	50,0,0	386,0,0	296,2,0	296,2,0	296,2,0
ALS2	c.4119A>G	p.Ile1373Met	1;0	Inf	5.3E-01	46,0,0	342,2,0	243,0,0	4664,43,0	7189,47,0
ALS2	c.4107G>A	p.(=)		Inf	1.0E+00	46,0,0	352,1,0	245,0,0	245,0,0	245,0,0
ALS2	c.4015C>T	p.(=)		0.63	6.7E-02	0,3,44	1,57,288	0,25,218	25,682,3997	83,1354,5829
ALS2	c.3885G>A	p.(=)		0.68	2.9E-01	45,3,0	349,14,0	266,15,1	4404,366,7	6998,399,7
ALS2	c.3863C>T	p.Pro1288Leu	0;0	Inf	1.0E+00	50,0,0	375,1,0	302,0,0	4403,3,0	6299,3,0
ALS2	c.3741T>G	p.(=)		Inf	1.0E+00	50,0,0	385,1,0	282,0,0	4728,20,0	7267,20,0
ALS2	c.3307C>A, c.1243C>A	p.His1103As	1;0	0.0	3.9E-01	49,0,0	384,0,0	277,1,0	277,1,0	277,1,0

		n, p.His415Asn		0						
ALS2	**c.3250T>G, c.1186T>G**	**p.Tyr1084Asp, p.Tyr396Asp**	**1;1**	**0.76**	**4.1E-01**	**40,3,0**	**280,20,0**	**187,18,0**	**187,18,0**	**187,18,0**
***ALS2**	**c.3182+2T>G, c.1118+2T>G**			**0.31**	**9.8E-02**	**46,0,0**	**361,3,0**	**254,6,0**	**254,6,0**	**254,6,0**
ALS2	c.3129C>G, c.1065C>G	p.(=)		Inf	1.0E+00	50,0,0	387,1,0	294,0,0	294,0,0	294,0,0
ALS2	c.3094C>T, c.1030C>T	p.Arg1032Cys, p.Arg344Cys	1;2	Inf	1.0E+00	49,0,0	374,1,0	274,0,0	274,0,0	274,0,0
ALS2	c.3046C>G, c.982C>G	p.Pro1016Ala, p.Pro328Ala	0;0	Inf	1.0E+00	50,0,0	386,1,0	291,0,0	4445,1,0	6365,17,0
ALS2	**c.2992C>A, c.928C>A**	**p.(=)**		**0.34**	**5.7E-01**	**48,0,0**	**352,1,0**	**268,2,0**	**268,2,0**	**268,2,0**
ALS2	c.2796C>T, c.732C>T	p.(=)		1.19	3.5E-01	34,15,1	308,75,4	241,54,2	4029,745,31	6262,1120,53
ALS2	c.2606A>C, c.542A>C	p.Gln869Pro, p.Gln181Pro	1;2	Inf	1.0E+00	50,0,0	389,1,0	290,0,0	290,0,0	290,0,0
ALS2	c.2566A>G, c.502A>G	p.Thr856Ala, p.Thr168Ala	0;1	Inf	1.0E+00	47,0,0	372,1,0	253,0,0	253,0,0	253,0,0
ALS2	c.2466G>A, c.402G>A	p.(=)		0.94	6.0E-01	8,25,15	57,175,145	37,117,102	750,2244,1719	2360,2984,1894
ALS2	c.2408A>G, c.344A>G	p.Lys803Arg, p.Lys115Arg	0;1	Inf	1.0E+00	50,0,0	390,1,0	289,0,0	289,0,0	289,0,0
ALS2	c.2241C>T, c.177C>T	p.(=)		0.67	1.0E+00	50,0,0	388,1,0	295,1,0	4445,25,0	6403,28,0

ALS2	**c.2216C>T, c.152C>T**	**p.Ala739Val, p.Ala51Val**	**0;0**	**0.63**	**2.4E-02**	**46,2,0**	**304,52,0**	**210,54,0**	**210,54,0**	**210,54,0**
ALS2	c.2155C>T, c.91C>T	p.Pro719Ser, p.Pro31Ser	1;2	0.00	3.9E-01	50,0,0	389,0,0	282,1,0	282,1,0	282,1,0
ALS2	c.2098A>G, c.34A>G	p.Thr700Ala, p.Thr12Ala	0;0	Inf	5.2E-01	50,0,0	384,2,0	279,0,0	279,0,0	279,0,0
ALS2	c.2028A>G	p.(=)		0.00	3.8E-01	50,0,0	383,0,0	267,1,0	267,1,0	267,1,0
ALS2	c.1641G>A	p.(=)		Inf	1.0E+00	50,0,0	380,1,0	277,0,0	4388,4,0	6255,9,0
ALS2	c.1627G>A	p.Asp543Asn	1;2	0.00	3.3E-02	47,0,0	290,0,0	247,4,0	4389,11,0	6357,11,0
ALS2	**c.1283C>A**	**p.Thr428Asn**	**0;0**	**Inf**	**1.0E+00**	**50,0,0**	**388,1,0**	**302,0,0**	**302,0,0**	**302,0,0**
ALS2	c.1265T>C	p.Met422Thr	1;0	Inf	1.0E+00	50,0,0	388,1,0	302,0,0	4454,1,0	6399,1,0
ALS2	**c.1226C>G**	**p.Ala409Gly**	**0;0**	**0.33**	**2.4E-03**	**48,0,0**	**358,12,0**	**252,23,0**	**252,23,0**	**252,23,0**
ALS2	c.1115C>G	p.Pro372Arg	0;1	0.34	5.7E-01	50,0,0	389,1,0	294,2,0	4742,22,0	7306,23,0
ALS2	c.1102G>A	p.Val368Met	0;0	0.93	7.8E-01	0,4,46	5,70,315	2,48,240	45,871,3894	145,1662,5633
ALS2	c.280A>G	p.Ile94Val	0;0	0.98	1.0E+00	47,3,0	371,15,1	285,14,0	4596,232,4	7069,445,7
DCTN1	**c.3788T>G, c.3371T>G, c.3386T>G, c.3722T>G, c.3767T>G, c.3713T>G, c.3662T>G**	**p.Val1263Gly, p.Val1124Gly, p.Val1129Gl**	**1;2**	**0.76**	**2.1E-01**	**32,3,0**	**135,52,0**	**114,52,0**	**114,52,0**	**114,52,0**

		y, p.Val1241Gly, p.Val1256Gly, p.Val1238Gly, p.Val1221Gly								
DCTN1	c.3746C>T, c.3344C>T, c.3620C>T, c.3725C>T, c.3329C>T, c.3680C>T, c.3671C>T	p.Thr1249Ile, p.Thr1115Ile, p.Thr1207Ile, p.Thr1242Ile, p.Thr1110Ile, p.Thr1227Ile, p.Thr1224Ile	1;0	0.64	7.1E-01	41,0,0	272,3,0	266,4,0	4905,43,0	7816,48,0
DCTN1	c.3501C>T, c.3435C>T, c.3426C>T, c.3084C>T, c.3099C>T, c.3480C>T, c.3375C>T	p.(=)		0.44	3.9E-01	45,0,0	299,2,0	224,3,0	224,3,0	224,3,0
DCTN1	c.3146G>A, c.2744G>A, c.3125G>A, c.3035G>A, c.3095G>A, c.3086G>A	p.Arg1049Gln, p.Arg915Gln, p.Arg1042Gln, p.Arg1012Gln, p.Arg1032Gln,	0;2	1.51	7.0E-01	47,0,0	313,4,0	273,2,0	4930,23,0	7845,24,0

		p.Arg1029Gln								
DCTN1	c.2952C>G, c.2892C>G, c.2901C>G, c.2841C>G, c.2550C>G, c.2931C>G	p.(=)		0.00	4.6E-01	46,0,0	287,0,0	280,1,0	4579,2,0	6782,2,0
*DCTN1	c.2887-2A>G, c.2836-2A>G, c.2827-2A>G, c.2866-2A>G, c.2485-2A>G, c.2776-2A>G			Inf	5.0E-01	39,0,0	237,2,0	246,0,0	246,0,0	246,0,0
DCTN1	c.2559C>T, c.2448C>T, c.2157C>T, c.2499C>T, c.2508C>T, c.2538C>T	p.(=)		0.86	1.0E+00	30,0,0	222,1,0	216,1,0	216,1,0	216,1,0
DCTN1	c.2551C>G, c.2149C>G, c.2440C>G, c.2530C>G, c.2500C>G, c.2491C>G	p.Leu851Val, p.Leu717Val, p.Leu814Val, p.Leu844Val, p.Leu834Val, p.Leu831Val	0;0	0.00	4.8E-01	27,0,0	212,0,0	222,1,0	600,1,0	1312,2,0
DCTN1	c.2448A>G, c.2046A>G, c.2397A>G, c.2337A>G, c.2427A>G, c.2388A>G	p.(=)		1.44	1.0E+00	48,0,0	369,4,0	301,2,0	4907,74,0	7613,281,3
DCTN1	c.2353C>T, c.1951C>T, c.2332C>T, c.2242C>T, c.2302C>T, c.2293C>T	p.Arg785Trp, p.Arg651Trp, p.Arg778Trp, p.Arg748Trp, p.Arg768Trp, p.Arg765Trp	1;2	Inf	1.0E+00	50,0,0	377,1,0	301,0,0	4594,7,0	6795,9,0
DCTN1	c.2339T>C, c.1937T>C, c.2288T>C, c.2279T>C, c.2318T>C, c.2228T>C	p.Ile780Thr, p.Ile646Thr, p.Ile763Thr, p.Ile760Thr,	1;0	Inf	1.0E+00	50,0,0	384,1,0	304,0,0	4603,1,0	6806,1,0

		p.Ile773Thr, p.Ile743Thr								
DCTN1	**c.2209G>C, c.2149G>C, c.2098G>C, c.1807G>C, c.2158G>C, c.2188G>C**	**p.Glu737Gln, p.Glu717Gln, p.Glu700Gln, p.Glu603Gln, p.Glu720Gln, p.Glu730Gln**	**1;1**	**0.71**	**1.0E+00**	**49,0,0**	**360,1,0**	**289,1,0**	**289,1,0**	**289,1,0**
DCTN1	c.2019C>T, c.1998C>T, c.1968C>T, c.1959C>T, c.1617C>T, c.1908C>T	p.(=)		0.00	4.1E-01	49,0,0	380,0,0	300,1,0	300,1,0	300,1,0
DCTN1	c.1484G>A, c.1433G>A, c.1424G>A, c.1463G>A, c.1373G>A, c.1082G>A	p.Arg495Gln, p.Arg478Gln, p.Arg475Gln, p.Arg488Gln, p.Arg458Gln, p.Arg361Gln	0;0	2.73	5.3E-02	34,7,0	202,9,1	183,5,0	4687,178,1	7578,203,1
DCTN1	**c.1480G>A, c.1429G>A, c.1420G>A, c.1369G>A, c.1459G>A, c.1078G>A**	**p.Ala494Thr, p.Ala477Thr, p.Ala474Thr, p.Ala457Thr, p.Ala487Thr, p.Ala360Thr**	**0;2**	**0.74**	**1.0E+00**	**37,0,0**	**203,1,0**	**178,1,0**	**4478,1,0**	**6676,6,0**
DCTN1	**c.1231C>A, c.829C>A,**	**p.Arg411Ser,**	**0;0**	**Inf**	**5.4E-01**	**41,0,0**	**246,2,0**	**167,0,0**	**167,0,0**	**167,0,0**

	c.1210C>A, c.1171C>A, c.1120C>A, c.1180C>A	**p.Arg277Ser, p.Arg404Ser, p.Arg391Ser, p.Arg374Ser, p.Arg394Ser**								
DCTN1	c.1060G>A, c.1000G>A, c.658G>A, c.949G>A, c.1039G>A, c.1009G>A	p.Ala354Thr, p.Ala334Thr, p.Ala220Thr, p.Ala317Thr, p.Ala347Thr, p.Ala337Thr	1;2	1.22	1.0E+00	45,1,0	272,1,0	194,1,0	194,1,0	194,1,0
DCTN1	c.1059C>T, c.1038C>T, c.657C>T, c.999C>T, c.1008C>T, c.948C>T	p.(=)		0.66	1.0E+00	46,0,0	274,1,0	211,1,0	211,1,0	211,1,0
DCTN1	**c.801G>A, c.399G>A, c.741G>A, c.780G>A, c.750G>A, c.690G>A**	**p.(=)**		**Inf**	**1.0E+00**	**16,0,0**	**128,1,0**	**143,0,0**	**143,0,0**	**143,0,0**
DCTN1	**c.713G>T, c.692G>T, c.662G>T, c.311G>T, c.602G>T, c.653G>T**	**p.Arg238Leu, p.Arg231Leu, p.Arg221Leu, p.Arg104Leu, p.Arg201Leu, p.Arg218Leu**	**1;2**	**0.00**	**4.4E-01**	**26,0,0**	**164,0,0**	**146,1,0**	**146,1,0**	**146,1,0**
DCTN1	c.597G>A, c.546G>A,	p.(=)		Inf	1.0E+0	18,0,0	121,1,0	97,0,0	4393,0,0	6583,7,0

	c.537G>A, c.195G>A, c.486G>A, c.576G>A				0					
DCTN1	c.586A>G, c.184A>G, c.526A>G, c.475A>G, c.565A>G, c.535A>G	p.Ile196Val, p.Ile62Val, p.Ile176Val, p.Ile159Val, p.Ile189Val, p.Ile179Val	0;0	0.76	1.0E+00	18,0,0	120,3,0	104,3,0	4711,69,2	7617,76,2
DCTN1	c.82C>T, c.103C>T, c.94C>T, c.31C>T	p.Arg28Trp, p.Arg35Trp, p.Arg32Trp, p.Arg11Trp	1;2	0.00	1.6E-01	45,0,0	308,0,0	235,2,0	235,2,0	235,2,0
DCTN1	c.42C>T, c.63C>T, c.54C>T	p.(=)		0.71	1.0E+00	37,0,0	275,1,0	221,1,0	221,1,0	221,1,0
VAPB	**c.474_476delTTC**	**p.Ser159del**		**Inf**	**5.2E-01**	**50,0,0**	**386,2,0**	**278,0,0**	**278,0,0**	**278,0,0**
VAPB	c.510G>A	p.Met170Ile	0;0	2.56	6.5E-01	50,0,0	385,4,0	279,1,0	4939,19,0	7854,20,0
NEFH	c.1054C>A	p.Arg352Ser	1;	0.14	5.0E-03	31,0,0	184,2,0	142,10,0	4778,52,0	7687,59,0
NEFH	c.1105G>A	p.Ala369Thr	0;	0.00	5.2E-02	43,0,0	254,0,0	175,3,0	175,3,0	175,3,0
NEFH	c.1200C>T	p.(=)		0.76	8.7E-02	28,16,0	172,91,7	125,62,18	3464,1310,109	6039,1631,129
NEFH	c.1387G>A	p.Glu463Lys	0;	0.90	6.1E-01	45,5,0	324,52,3	244,45,1	4066,852,50	6892,940,52
NEFH	c.1580C>T	p.Pro527Leu	0;	0.35	5.7E-01	47,0,0	354,1,0	279,2,0	279,2,0	279,2,0
NEFH	c.1740C>T	p.(=)		0.52	2.8E-01	50,0,0	370,6,0	291,8,0	4857,119,1	7763,129,1

NEFH	c.1844C>T	p.Pro615Leu	1;2	0.97	8.5E-01	30,17,2	217,147,12	170,109,13	3028,1715,227	5081,2490,315
NEFH	**c.1933G>A**	**p.Glu645Lys**	**0;2**	**2.22**	**2.6E-02**	**41,2,0**	**314,30,0**	**253,10,0**	**253,10,0**	**253,10,0**
NEFH	**c.1935_1936insG**	**p.Ala646GlyfsX5**		**1.17**	**4.9E-01**	**34,9,0**	**237,63,0**	**198,44,0**	**198,44,0**	**198,44,0**
NEFH	**c.1935A>G**	**p.(=)**		**1.59**	**9.8E-02**	**39,5,0**	**292,41,0**	**246,21,0**	**246,21,0**	**246,21,0**
NEFH	**c.1938A>C**	**p.(=)**		**1.02**	**8.9E-01**	**27,18,0**	**163,147,0**	**137,115,0**	**137,115,0**	**137,115,0**
NEFH	**c.1965A>T**	**p.(=)**		**1.43**	**6.2E-03**	**16,29,0**	**118,206,0**	**129,125,0**	**129,125,0**	**129,125,0**
NEFH	**c.1970_1975delAGGAAG**	**p.Glu658_Glu659del**		**1.64**	**9.7E-04**	**24,23,0**	**189,148,0**	**187,79,0**	**187,79,0**	**187,79,0**
NEFH	**c.1973A>C**	**p.Glu658Ala**	**0;0**	**0.58**	**5.7E-02**	**40,1,0**	**298,26,0**	**227,32,0**	**227,32,0**	**227,32,0**
NEFH	**c.1974A>C**	**p.Glu658Asp**	**0;2**	**0.78**	**5.7E-01**	**45,1,0**	**323,15,0**	**248,14,0**	**248,14,0**	**248,14,0**
NEFH	**c.1975G>A**	**p.Glu659Lys**	**0;2**	**0.83**	**6.8E-01**	**46,1,0**	**325,12,0**	**258,11,0**	**258,11,0**	**258,11,0**
NEFH	**c.1989T>A**	**p.(=)**		**Inf**	**1.9E-06**	**43,2,0**	**332,22,0**	**300,0,0**	**300,0,0**	**300,0,0**
NEFH	**c.2082C>T**	**p.(=)**		**Inf**	**5.1E-01**	**47,0,0**	**330,0,1**	**304,0,0**	**304,0,0**	**304,0,0**
NEFH	c.2232T>C	p.(=)		0.94	6.6E-01	2,18,29	14,145,216	12,107,179	230,1711,3035	321,2459,5112
NEFH	**c.2368_2370delAAG**	**p.Lys790del**		**Inf**	**1.0E+00**	**48,0,0**	**324,1,0**	**278,0,0**	**278,0,0**	**278,0,0**
NEFH	c.2414A>C	p.Glu805Ala	1;1	1.02	9.4E-01	31,14,0	200,92,4	177,62,10	3518,1307,101	5785,1900,157
NEFH	c.2646C>T	p.(=)		0.0	2.3E-02	37,0,0	281,0,0	199,2,1	199,2,1	199,2,1

			0							
NEFH	c.2740G>A	p.Val914Met	0;	0.75	1.0E+00	41,0,0	301,1,0	258,1,0	258,1,0	258,1,0
NEFH	c.2757C>T	p.(=)		1.53	7.5E-01	39,1,0	266,6,0	201,3,0	4827,55,0	7724,74,0
NEFH	c.2784A>G	p.(=)		0.94	6.8E-01	2,16,23	9,129,182	9,98,156	221,1706,3014	308,2459,5090
NEFH	c.2977A>G	p.Lys993Glu	1;	0.00	4.3E-01	38,0,0	315,0,0	265,1,0	265,1,0	265,1,0
CHMP2B	c.27C>T	p.(=)		1.19	5.4E-01	24,4,0	135,42,3	104,26,1	4016,752,41	6298,1343,84
CHMP2B	c.118A>G	p.Lys40Glu	1;1	Inf	1.0E+00	50,0,0	372,1,0	261,0,0	261,0,0	261,0,0
CHMP2B	c.123G>T	p.Gln41His	1;2	Inf	1.0E+00	50,0,0	372,1,0	263,0,0	263,0,0	263,0,0
CHMP2B	c.312T>C, c.189T>C	p.(=)		1.60	1.1E-01	0,3,43	1,23,336	0,26,214	16,506,4396	467,1777,5590
CHMP2B	c.372A>C, c.249A>C, c.282A>C	p.(=)		1.23	3.5E-01	33,8,1	248,53,4	176,36,0	4084,760,38	6382,1338,78
FIG4	c.122T>C, c.59T>C	p.Ile41Thr, p.Ile20Thr	1;2	0.29	5.6E-01	46,0,0	360,1,0	238,2,0	4899,18,0	7814,19,0
***FIG4**	**c.447-2A>T, c.384-2A>T, c.9-2A>T, c.216-2A>T**			**4.21**	**2.7E-01**	**43,1,0**	**320,6,0**	**220,1,0**	**220,1,0**	**220,1,0**
FIG4	c.1090A>T, c.259A>T	p.Met364Leu, p.Met87Leu	0;0	0.81	4.8E-01	47,3,0	363,24,1	271,24,0	4611,355,7	7272,592,25
FIG4	c.1305G>C, c.474G>C	p.(=)		Inf	1.0E+00	50,0,0	388,1,0	281,0,0	4579,2,0	6782,2,0
FIG4	c.1426C>T, c.595C>T	p.Arg476Cys,	1;0	Inf	1.0E+0	47,0,0	354,1,0	244,0,0	622,0,0	1334,1,0

		p.Arg199Cys			0					
FIG4	c.1525C>T, c.694C>T	p.(=)		Inf	1.0E+00	49,0,0	387,1,0	287,0,0	4586,1,0	6789,1,0
FIG4	c.1527G>A, c.696G>A	p.(=)		0.00	4.0E-01	50,0,0	387,0,0	287,1,0	287,1,0	287,1,0
FIG4	c.1793T>A, c.962T>A, c.210T>A	p.Phe598Tyr, p.Phe321Tyr, p.Phe71Tyr	0;0	0.00	3.7E-01	38,0,0	336,0,0	219,1,0	219,1,0	219,1,0
FIG4	c.1961T>C, c.1130T>C	p.Val654Ala, p.Val377Ala	0;0	0.78	9.6E-02	37,11,2	287,96,7	184,77,9	3552,1263,132	4030,2504,1329
FIG4	**c.2200G>A, c.121G>A, c.1369G>A**	**p.Glu734Lys, p.Glu41Lys, p.Glu457Lys**	**;1**	**0.84**	**1.0E+00**	**41,0,0**	**216,1,0**	**216,1,0**	**4514,3,0**	**6717,3,0**
FIG4	c.2559G>A, c.1641G>A	p.(=)		1.21	9.0E-02	13,22,15	107,190,79	92,139,49	1738,2355,865	3648,3195,1031
HFE	**c.18G>C**	**p.Arg6Ser**	**1;**	**Inf**	**5.3E-01**	**12,0,0**	**61,2,0**	**43,0,0**	**4336,6,0**	**6538,7,0**
HFE	**c.50C>T**	**p.Thr17Ile**	**1;**	**0.54**	**1.0E+00**	**9,1,0**	**45,0,0**	**29,1,0**	**4317,12,0**	**6519,12,0**
HFE	**c.68G>A**	**p.Arg23His**	**1;**	**Inf**	**1.0E+00**	**7,1,0**	**23,0,0**	**17,0,0**	**4690,4,0**	**7600,9,0**
HFE	c.138G>A	p.(=)		0.00	4.1E-01	50,0,0	388,0,0	305,1,0	305,1,0	305,1,0
HFE	c.157G>A, c.88G>A	p.Val53Met, p.Val30Met	1;2	0.00	4.1E-01	50,0,0	387,0,0	301,1,0	4979,1,0	7892,4,0
HFE	c.187C>G, c.118C>G	p.His63Asp, p.His40Asp	1;0	1.02	8.9E-01	33,16,1	255,124,9	201,97,6	3558,1323,101	6275,1515,108
HFE	c.193A>T, c.124A>T	p.Ser65Cys, p.Ser42Cys	1;2	0.96	1.0E+00	48,2,0	371,16,0	291,13,0	4829,150,3	7728,167,3
HFE	c.502G>C, c.238G>C,	p.Glu168Gln,	1;2	0.0	1.8E-01	44,0,0	264,0,0	224,2,0	4897,7,0	7813,7,0

	c.433G>C	p.Glu80Gln, p.Glu145Gln		0						
HFE	c.670C>T, c.352C>T, c.661C>T, c.394C>T, c.601C>T, c.130C>T, c.364C>T, c.406C>T, c.628C>T	p.Arg224Trp, p.Arg118Trp, p.Arg221Trp, p.Arg132Trp, p.Arg201Trp, p.Arg44Trp, p.Arg122Trp, p.Arg136Trp, p.Arg210Trp	1;2	Inf	5.2E-01	49,1,0	385,1,0	296,0,0	4593,3,0	6795,4,0
HFE	c.766G>A, c.226G>A, c.502G>A, c.724G>A, c.460G>A, c.448G>A, c.697G>A, c.490G>A, c.757G>A	p.Val256Ile, p.Val76Ile, p.Val168Ile, p.Val242Ile, p.Val154Ile, p.Val150Ile, p.Val233Ile, p.Val164Ile, p.Val253Ile	1;1	Inf	1.0E+00	50,0,0	387,1,0	295,0,0	295,0,0	295,0,0
HFE	c.829G>A, c.565G>A, c.523G>A, c.553G>A, c.820G>A, c.760G>A, c.787G>A, c.289G>A, c.511G>A	p.Glu277Lys, p.Glu189Lys, p.Glu175Lys, p.Glu185Lys, p.Glu274Lys, p.Glu254Lys, p.Glu263Lys, p.Glu97Lys, p.Glu171Lys	1;2	0.69	1.0E+00	50,0,0	382,1,0	299,1,0	4970,8,0	7883,11,0

HFE	c.845G>A, c.539G>A, c.569G>A, c.836G>A, c.581G>A, c.527G>A, c.803G>A, c.305G>A, c.776G>A	p.Cys282Tyr, p.Cys180Tyr, p.Cys190Tyr, p.Cys279Tyr, p.Cys194Tyr, p.Cys176Tyr, p.Cys268Tyr, p.Cys102Tyr, p.Cys259Tyr	1;2	0.78	1.5E-01	42,8,0	302,70,3	234,60,7	4345,608,26	7187,681,27
DPP6	**c.114C>T**	**p.(=)**		**0.00**	**1.0E+00**	**0,1,0**	**0,1,0**	**0,0,0**	**0,0,0**	**0,0,0**
DPP6	**c.785C>T**	**p.Ser262Leu**		**0.00**	**1.0E+00**	**0,0,0**	**0,0,0**	**0,1,0**	**1946,400,23**	**3326,597,34**
DPP6	**c.982T>G**	**p.Cys328Gly**		**0.00**	**1.0E+00**	**0,0,0**	**0,0,0**	**0,0,0**	**0,0,0**	**0,0,0**
DPP6	c.657C>T, c.465C>T, c.471C>T	p.(=)		1.06	6.0E-01	4,28,17	39,167,182	31,127,124	419,1939,2397	910,3103,3299
DPP6	c.666C>T, c.474C>T, c.480C>T	p.(=)		2.39	2.7E-01	48,2,0	382,9,0	282,3,0	4707,46,0	7261,52,0
DPP6	c.723A>G, c.537A>G, c.531A>G	p.(=)		1.19	3.9E-01	0,8,41	4,52,324	3,49,245	51,812,3908	297,1822,5229
DPP6	c.879T>C, c.687T>C, c.693T>C, c.558T>C	p.(=)		Inf	1.0E+00	50,0,0	388,1,0	296,0,0	296,0,0	296,0,0
DPP6	c.883G>A, c.697G>A, c.562G>A, c.691G>A	p.Glu295Lys, p.Glu233Lys, p.Glu188Lys, p.Glu231Lys	1;2	Inf	1.0E+00	50,0,0	389,1,0	296,0,0	296,0,0	296,0,0
DPP6	c.945C>T, c.759C>T, c.624C>T, c.753C>T	p.(=)		0.69	7.7E-02	45,3,0	325,44,2	248,49,1	4080,736,32	6300,1226,61

DPP6	c.948C>T, c.756C>T, c.627C>T, c.762C>T	p.(=)		Inf	1.0E+00	47,0,0	371,1,0	292,0,0	4460,0,0	6472,3,0
DPP6	c.1458C>T, c.1272C>T, c.1266C>T, c.1137C>T	p.(=)		2.14	6.5E-01	46,1,0	364,2,0	293,1,0	4469,8,0	6544,11,0
DPP6	c.1459G>A, c.1273G>A, c.1138G>A, c.1267G>A	p.Val487Met, p.Val425Met, p.Val380Met, p.Val423Met	1;2	0.00	4.2E-01	49,0,0	364,0,0	298,1,0	298,1,0	298,1,0
DPP6	**c.1616T>C, c.1295T>C, c.1430T>C, c.1424T>C**	**p.Phe539Ser, p.Phe432Ser, p.Phe477Ser, p.Phe475Ser**	**1;2**	**0.00**	**4.3E-01**	**50,0,0**	**351,0,0**	**296,1,0**	**296,1,0**	**296,1,0**
DPP6	c.1711A>C, c.1519A>C, c.1390A>C, c.1525A>C	p.Lys571Gln, p.Lys507Gln, p.Lys464Gln, p.Lys509Gln	0;0	0.00	4.0E-01	50,0,0	389,0,0	294,1,0	4770,8,0	7337,25,0
DPP6	c.1896A>G, c.1575A>G, c.1710A>G, c.1704A>G	p.(=)		1.63	2.9E-02	2,9,0	18,56,10	29,49,2	1087,2329,1226	1343,3452,2626
DPP6	c.1911G>A, c.1725G>A, c.1590G>A, c.1719G>A	p.(=)		1.37	1.2E-01	5,10,1	33,72,14	35,58,4	1971,2118,580	3968,2838,652
DPP6	c.2127T>C, c.1941T>C, c.1935T>C, c.1806T>C	p.(=)		0.99	9.5E-01	28,16,2	197,118,14	170,90,17	3275,1447,165	4371,2723,627
DPP6	c.2205C>T, c.1884C>T,	p.(=)		0.3	1.3E-01	49,0,0	344,4,0	261,7,0	4756,43,0	7352,114,7

	c.2013C>T, c.2019C>T			8						
DPP6	**c.2295A>G, c.2103A>G, c.1974A>G, c.2109A>G**	**p.(=)**		**Inf**	**5.1E-01**	**49,0,0**	**356,2,0**	**296,0,0**	**296,0,0**	**296,0,0**
DPP6	c.2331C>T, c.2139C>T, c.2010C>T, c.2145C>T	p.(=)		0.00	1.0E+00	48,0,0	372,0,0	286,0,0	4841,6,0	7577,13,0
DPP6	c.2373C>T, c.2181C>T, c.2187C>T, c.2052C>T	p.(=)		1.01	1.0E+00	36,13,0	313,58,3	236,43,4	3799,975,60	6182,1258,109
***DPP6**	**c.2378-1G>A, c.2192-1G>A, c.2057-1G>A, c.2186-1G>A**			**1.87**	**7.2E-01**	**48,0,0**	**358,6,0**	**254,2,0**	**254,2,0**	**254,2,0**
DPP6	c.2427G>A, c.2106G>A, c.2241G>A, c.2235G>A	p.(=)		1.08	6.9E-01	34,14,2	299,86,6	226,62,5	3578,1116,101	4732,2251,422
DPP6	c.2561T>C, c.2375T>C, c.2369T>C, c.2240T>C	p.Leu854Pro, p.Leu792Pro, p.Leu790Pro, p.Leu747Pro	1;0	0.94	5.9E-01	23,21,6	211,144,29	154,122,24	2637,1875,406	3479,3231,1056
PON1	c.953C>T	p.Thr318Ile	1;0	0.00	4.0E-01	50,0,0	390,0,0	289,1,0	4588,2,0	6791,2,0
PON1	c.603G>A	p.(=)		2.59	6.5E-01	50,0,0	386,4,0	283,1,0	4946,16,0	7860,18,0
PON1	c.602C>T	p.Ala201Val	0;0	0.43	3.8E-01	50,0,0	386,2,0	278,3,0	4931,28,0	7844,31,0
PON1	c.575A>G	p.Gln192Arg	0;0	0.97	8.1E-01	24,23,3	211,152,28	150,115,21	2544,2013,407	2905,3306,1669
PON1	c.468A>G	p.(=)		Inf	1.0E+00	50,0,0	388,1,0	280,0,0	280,0,0	280,0,0
PON1	**c.391dupG**	**p.Val131GlyfsX15**		**0.00**	**3.9E-01**	**50,0,0**	**391,0,0**	**282,1,0**	**282,1,0**	**282,1,0**
PON1	c.163T>A	p.Leu55Met	1;2	0.94	5.8E-01	18,26,6	159,176,54	109,128,43	1982,2264,712	4041,3035,798

PON3	c.971G>A	p.Gly324Asp	1;2	0.00	4.0E-01	49,0,0	382,0,0	292,1,0	4935,35,0	7850,36,0
PON3	c.954C>T	p.(=)		Inf	1.0E+00	49,0,0	376,1,0	285,0,0	4584,1,0	6787,1,0
PON3	**c.819T>C**	**p.(=)**		**0.83**	**5.6E-01**	**47,1,0**	**330,28,0**	**247,23,0**	**247,23,0**	**247,23,0**
PON3	c.609T>C	p.(=)		1.22	8.2E-01	49,1,0	379,12,0	281,7,0	4908,57,1	7814,67,1
PON3	c.408G>T	p.Met136Ile	0;0	0.30	5.6E-01	49,0,0	380,1,0	259,2,0	4554,7,0	6756,8,0
PON3	c.297G>A	p.(=)		1.06	5.9E-01	17,20,12	93,199,91	85,121,72	1446,2454,1056	1930,3805,2137
PON3	c.262A>G	p.Met88Val	1;0	0.65	1.0E+00	50,0,0	383,1,0	280,1,0	4943,16,0	7859,16,0
PON1, PON3	c.94C>T	p.Arg32X		Inf	2.8E-01	48,0,0	371,3,0	280,0,0	4937,21,0	7849,25,0
PON1, PON3	c.63C>T	p.(=)		1.14	7.6E-01	1,3,0	10,34,1	12,28,1	2607,1810,302	3945,3085,605
PON2	c.932C>G, c.896C>G, c.995C>G	p.Ser311Cys, p.Ser299Cys, p.Ser332Cys	1;0	1.04	8.0E-01	32,17,1	210,157,23	166,95,22	2934,1762,265	4447,2950,480
PON2	c.913C>T, c.976C>T, c.877C>T	p.Arg305Cys, p.Arg326Cys, p.Arg293Cys	1;2	Inf	2.9E-01	48,1,0	384,2,0	276,0,0	4573,3,0	6776,3,0
PON2	c.827C>T, c.791C>T, c.890C>T	p.Ser276Leu, p.Ser264Leu, p.Ser297Leu	1;0	0.63	1.0E+00	50,0,0	381,1,0	273,1,0	4572,2,0	6775,2,0
PON2	c.661T>G, c.625T>G, c.724T>G	p.Ser221Ala, p.Ser209Ala,	0;0	Inf	1.0E+00	49,0,0	380,1,0	257,0,0	257,0,0	257,0,0

		p.Ser242Ala								
PON2	c.443C>G, c.506C>G, c.407C>G	p.Ala148Gly, p.Ala169Gly, p.Ala136Gly	1;0	1.04	8.0E-01	31,17,0	205,158,22	155,90,20	2924,1756,263	4445,2939,475
PON2	c.359T>G, c.422T>G	p.Ile120Arg, p.Ile141Arg	1;1	0.00	4.0E-01	50,0,0	389,0,0	287,1,0	287,1,0	287,1,0
PON2	**c.286delA, c.349delA**	**p.Arg96GlyfsX5, p.Arg117GlyfsX5**		**2.00**	**1.0E+00**	**49,1,0**	**386,2,0**	**290,1,0**	**290,1,0**	**290,1,0**
ELP3	c.139G>A, c.97G>A	p.Ala47Thr, p.Ala33Thr	1;0	Inf	2.7E-01	50,0,0	380,3,0	302,0,0	4600,2,0	6803,2,0
ELP3	c.206G>T, c.164G>T	p.Arg69Leu, p.Arg55Leu	1;1	Inf	1.0E+00	50,0,0	380,1,0	294,0,0	294,0,0	294,0,0
ELP3	c.326G>A, c.284G>A, c.110G>A, c.50G>A, c.3G>A, c.33G>A	p.Cys109Tyr, p.Cys95Tyr, p.Cys37Tyr, p.Cys17Tyr, p.Met1?, p.Met11Ile	;2	Inf	5.3E-01	49,1,0	389,1,0	278,0,0	4577,1,0	6780,1,0
ELP3	c.1068G>A, c.852G>A, c.792G>A, c.711G>A, c.1026G>A, c.681G>A	p.(=)		0.68	1.0E+00	50,0,0	386,1,0	298,1,0	298,1,0	298,1,0
ELP3	c.1318G>A, c.97G>A, c.115G>A, c.1042G>A, c.1102G>A, c.1276G>A, c.961G>A, c.931G>A	p.Glu440Lys, p.Glu33Lys, p.Glu39Lys, p.Glu348Lys, p.Glu368Lys, p.Glu426Lys, p.Glu321Lys, p.Glu311Lys	1;2	0.32	5.6E-01	49,0,0	363,1,0	259,2,0	4936,3,0	7852,3,0

ELP3	c.1459C>T, c.1183C>T, c.1417C>T, c.1072C>T, c.1243C>T, c.256C>T, c.1102C>T	p.Arg487Trp, p.Arg395Trp, p.Arg473Trp, p.Arg358Trp, p.Arg415Trp, p.Arg86Trp, p.Arg368Trp	1;2	0.53	1.0E+00	46,0,0	330,1,0	199,1,0	4870,8,0	7786,8,0
ELP3	**c.366G>A**	**p.(=)**		**0.00**	**1.0E+00**	**0,0,0**	**0,0,0**	**0,0,0**	**287,82,9**	**833,223,35**
SETX	c.7944C>T, c.7845C>T, c.2757C>T, c.8031C>T	p.(=)		Inf	1.0E+00	37,0,0	220,1,0	180,0,0	4479,1,0	6682,1,0
SETX	c.7834A>G, c.7735A>G, c.2647A>G, c.7921A>G	p.Ser2612Gly, p.Ser2579Gly, p.Ser883Gly, p.Ser2641Gly	0;0	1.14	7.4E-01	34,2,0	202,19,1	176,15,0	4611,252,6	6595,1076,114
SETX	**c.7787C>G, c.7874C>G, c.2600C>G, c.7688C>G**	**p.Ala2596Gly, p.Ala2625Gly, p.Ala867Gly, p.Ala2563Gly**	**1;0**	**1.48**	**2.5E-01**	**9,4,0**	**125,19,0**	**142,16,0**	**142,16,0**	**142,16,0**
SETX	c.7759A>G, c.7846A>G, c.7660A>G, c.2572A>G	p.Ile2587Val, p.Ile2616Val, p.Ile2554Val, p.Ile858Val	0;0	1.24	9.7E-02	22,20,4	160,154,28	145,115,15	2472,2065,416	2770,3162,1937

SETX	c.7709C>T, c.7610C>T, c.2522C>T, c.7796C>T	p.Pro2570Leu, p.Pro2537Leu, p.Pro841Leu, p.Pro2599Leu	1;0	0.00	4.2E-01	50,0,0	362,0,0	297,1,0	297,1,0	297,1,0
SETX	c.7682C>T, c.7583C>T, c.2495C>T, c.7769C>T	p.Ser2561Leu, p.Ser2528Leu, p.Ser832Leu, p.Ser2590Leu	0;0	Inf	5.3E-01	50,0,0	356,2,0	253,0,0	4552,1,0	6755,1,0
SETX	c.7645G>A, c.7546G>A, c.2458G>A, c.7732G>A	p.Val2549Ile, p.Val2516Ile, p.Val820Ile, p.Val2578Ile	1;0	Inf	5.2E-01	50,0,0	387,2,0	305,0,0	4605,0,0	6807,1,0
SETX	c.7640T>C, c.7727T>C, c.2453T>C, c.7541T>C	p.Ile2547Thr, p.Ile2576Thr, p.Ile818Thr, p.Ile2514Thr	0;0	Inf	7.0E-03	50,0,0	378,10,0	305,0,0	4913,69,1	7824,74,1
SETX	c.7590T>A, c.2403T>A, c.7677T>A, c.7491T>A	p.(=)		Inf	1.0E+00	50,0,0	388,1,0	303,0,0	303,0,0	303,0,0
SETX	c.7524A>G, c.7611A>G, c.2337A>G, c.7425A>G	p.(=)		0.00	4.1E-01	50,0,0	390,0,0	303,1,0	303,1,0	303,1,0
SETX	c.7406T>C, c.7307T>C, c.7493T>C, c.2219T>C	p.Leu2469Pro, p.Leu2436Pr	0;0	0.00	4.1E-01	50,0,0	390,0,0	302,1,0	4601,2,0	6804,2,0

		o, p.Leu2498Pro, p.Leu740Pro								
SETX	c.2095C>T, c.7369C>T	p.Leu699Phe, p.Leu2457Phe	0;0	1.12	1.0E+00	6,0,0	58,3,0	72,3,0	2409,34,1	3995,37,1
SETX	c.7139G>A, c.1865G>A	p.Arg2380Gln, p.Arg622Gln	0;2	0.33	5.7E-01	49,1,0	387,0,0	283,2,0	4582,3,0	6785,3,0
SETX	c.7119T>C, c.1845T>C	p.(=)		Inf	1.0E+00	50,0,0	389,1,0	293,0,0	293,0,0	293,0,0
SETX	**c.6848C>A, c.1574C>A**	**p.Thr2283Lys, p.Thr525Lys**	**1;2**	**0.23**	**2.5E-02**	**48,0,0**	**357,3,0**	**241,8,0**	**241,8,0**	**241,8,0**
SETX	c.6507G>A, c.1233G>A	p.(=)		0.29	4.8E-03	50,0,0	383,7,0	279,16,0	4760,208,5	7653,231,5
SETX	**c.6049A>G, c.775A>G**	**p.Met2017Val, p.Met259Val**	**0;2**	**Inf**	**1.0E+00**	**50,0,0**	**390,1,0**	**285,0,0**	**285,0,0**	**285,0,0**
SETX	c.6024T>C, c.750T>C	p.(=)		0.00	3.9E-01	50,0,0	389,0,0	285,1,0	285,1,0	285,1,0
SETX	**c.6005G>A, c.731G>A**	**p.Arg2002His, p.Arg244His**	**1;2**	**0.32**	**5.6E-01**	**50,0,0**	**375,1,0**	**268,2,0**	**268,2,0**	**268,2,0**
SETX	c.5998C>G, c.724C>G	p.Gln2000Glu, p.Gln242Glu	1;1	Inf	2.8E-01	49,1,0	386,2,0	285,0,0	4959,4,0	7874,5,0

SETX	c.5842A>G, c.568A>G	p.Met1948Val, p.Met190Val	1;2	Inf	1.0E+00	50,0,0	387,1,0	279,0,0	279,0,0	279,0,0
SETX	c.5811T>C, c.537T>C	p.(=)		1.00	1.0E+00	33,15,1	274,106,8	191,69,8	3532,1301,113	4070,2603,1189
SETX	c.5587A>G, c.313A>G	p.Thr1863Ala, p.Thr105Ala	1;2	Inf	1.0E+00	49,1,0	391,0,0	278,0,0	278,0,0	278,0,0
SETX	c.5563A>G, c.289A>G	p.Thr1855Ala, p.Thr97Ala	0;0	1.02	9.4E-01	34,15,1	275,107,9	197,72,8	3541,1301,113	4078,2603,1190
SETX	c.5283A>G, c.9A>G	p.(=)		Inf	1.0E+00	49,0,0	375,1,0	263,0,0	4558,4,0	6761,4,0
SETX	c.5271A>G	p.(=)		0.61	1.0E+00	50,0,0	387,1,0	268,1,0	268,1,0	268,1,0
SETX	c.4755T>G	p.(=)		1.76	5.4E-01	50,0,0	384,8,0	287,3,0	4888,80,0	7796,88,0
SETX	c.4707T>C	p.(=)		0.00	4.0E-01	50,0,0	391,0,0	288,1,0	288,1,0	288,1,0
SETX	c.4660T>G	p.Cys1554Gly	1;2	4.55	1.6E-01	49,1,0	385,6,0	284,1,0	4922,41,0	7831,48,0
SETX	c.4156A>G	p.Ile1386Val	0;0	1.04	8.2E-01	0,15,35	6,107,277	8,68,191	112,1292,3541	1190,2591,4080
SETX	**c.3930dupA**	**p.Arg1311ThrfsX2**		**Inf**	**1.0E+00**	**49,1,0**	**391,0,0**	**293,0,0**	**293,0,0**	**293,0,0**
SETX	c.3754G>A	p.Gly1252Arg	0;0	0.97	8.8E-01	0,15,35	6,107,276	8,70,212	118,1289,3561	1198,2588,4098
SETX	c.3576T>G	p.Asp1192Glu	0;0	0.97	8.7E-01	0,15,35	5,93,286	3,66,203	93,1168,3689	533,2448,4885
SETX	c.3455T>G	p.Phe1152Cy	1;1	1.3	4.2E-01	47,3,0	362,25,0	263,13,0	4629,320,5	7166,679,25

		s		7						
SETX	c.3336T>C	p.(=)		1.23	1.0E+00	50,0,0	383,4,0	267,2,0	4940,7,0	7853,10,0
SETX	c.3229G>A	p.Asp1077Asn	1;1	Inf	5.3E-01	50,0,0	383,2,0	267,0,0	4936,9,0	7850,10,0
SETX	c.3221A>G	p.Glu1074Gly	1;2	0.00	3.8E-01	50,0,0	384,0,0	264,1,0	264,1,0	264,1,0
SETX	c.3147C>T	p.(=)		0.95	8.3E-01	42,8,0	336,53,1	228,36,2	4199,718,27	6767,1048,45
SETX	c.2975A>G	p.Lys992Arg	1;0	3.45	5.5E-02	48,1,1	375,13,0	277,3,0	4799,159,0	7698,176,0
SETX	c.2964A>G	p.(=)		0.00	3.9E-01	50,0,0	389,0,0	278,1,0	4577,2,0	6780,2,0
SETX	c.2842C>A	p.Pro948Thr	0;0	Inf	1.0E+00	49,1,0	387,0,0	282,0,0	282,0,0	282,0,0
SETX	c.2755G>C	p.Val919Leu	0;0	Inf	1.0E+00	50,0,0	383,1,0	266,0,0	266,0,0	266,0,0
SETX	c.2717C>T	p.Ser906Leu	0;0	Inf	1.0E+00	49,0,0	380,1,0	258,0,0	4557,1,0	6760,1,0
SETX	c.2479A>G	p.Lys827Glu	0;0	Inf	1.0E+00	49,0,0	378,1,0	246,0,0	4914,8,0	7828,9,0
SETX	c.2411T>C	p.Leu804Ser	1;0	Inf	1.0E+00	49,0,0	378,1,0	253,0,0	4928,2,0	7843,2,0
SETX	c.1979C>G	p.Ala660Gly	0;0	0.99	1.0E+00	43,7,0	339,44,1	234,33,0	4384,539,21	6103,1543,214
SETX	c.1869A>C	p.Glu623Asp	1;2	0.00	3.8E-01	50,0,0	385,0,0	267,1,0	4564,4,0	6767,4,0
SETX	c.1750C>G	p.Leu584Val	1;2	0.32	5.6E-01	50,0,0	390,1,0	278,2,0	278,2,0	278,2,0

SETX	c.1655A>C	p.Gln552Pro	0;1	0.00	4.0E-01	50,0,0	390,0,0	290,1,0	290,1,0	290,1,0
SETX	c.1077T>C	p.(=)		0.98	9.4E-01	0,15,35	6,104,276	6,64,187	110,1280,3544	588,2589,4673
SETX	c.814C>G	p.His272Asp	1;2	Inf	1.0E+00	50,0,0	386,1,0	280,0,0	280,0,0	280,0,0
SETX	c.710A>G	p.Tyr237Cys	1;2	0.62	1.0E+00	50,0,0	386,1,0	272,1,0	4570,3,0	6773,3,0
SETX	c.654G>C	p.Lys218Asn	1;2	0.00	3.9E-01	50,0,0	391,0,0	279,1,0	4951,7,0	7867,7,0
SETX	c.540A>G	p.(=)		Inf	1.0E+00	50,0,0	385,1,0	279,0,0	279,0,0	279,0,0
SETX	**c.506G>A**	**p.Arg169His**	**1;2**	**0.89**	**1.0E+00**	**49,0,0**	**372,3,0**	**249,2,0**	**249,2,0**	**249,2,0**
SETX	c.472T>G	p.Leu158Val	1;2	0.39	2.8E-01	50,0,0	383,3,0	280,5,0	4910,53,0	7821,58,0
SETX	c.234G>A	p.(=)		0.00	1.4E-01	50,0,0	384,0,0	262,2,0	4559,5,0	6761,6,0
SETX	c.192A>G	p.(=)		0.90	1.0E+00	48,1,0	377,5,0	256,2,1	4910,24,1	7824,26,1
SETX	c.59G>A	p.Arg20His	0;0	0.38	9.9E-02	49,0,0	380,5,0	288,9,0	4866,108,1	7771,119,1
MOB3B	c.615G>A	p.(=)		0.00	4.1E-01	50,0,0	390,0,0	303,1,0	303,1,0	303,1,0
MOB3B	c.414C>T	p.(=)		1.11	4.6E-01	28,19,3	243,128,12	195,84,13	3304,1488,178	5880,1815,191
MOB3B	c.98G>A	p.Arg33Gln	0;0	0.00	4.1E-01	50,0,0	354,0,0	277,1,0	277,1,0	277,1,0
IFNK	**c.30_31insTGTT**	**p.Trp13PhefsX4**		**1.2**	**5.2E-01**	**44,6,0**	**366,21,0**	**269,14,0**	**269,14,0**	**269,14,0**

IFNK	**c.43G>T**	**p.Glu15X**		**6 0.52**	**1.1E-02**	**46,3,0**	**330,31,0**	**205,37,0**	**205,37,0**	**205,37,0**
IFNK	c.263A>G	p.Tyr88Cys	1;2	0.65	1.0E+00	50,0,0	390,1,0	287,1,0	4961,5,0	7832,50,0
IFNK	c.279G>A	p.(=)		0.00	4.0E-01	50,0,0	391,0,0	287,1,0	287,1,0	287,1,0
IFNK	c.282C>T	p.(=)		0.00	4.0E-01	50,0,0	391,0,0	287,1,0	287,1,0	287,1,0
IFNK	c.397A>G	p.Lys133Glu	0;0	0.79	1.0E+00	0,0,49	0,4,386	0,2,274	0,83,4870	111,885,6873
IFNK	c.486G>C	p.(=)		0.00	3.9E-01	49,0,0	385,0,0	280,1,0	280,1,0	280,1,0
C9orf72	c.1404C>T	p.(=)		0.42	2.0E-01	49,1,0	383,3,0	270,6,0	4898,56,0	7809,61,0
C9orf72	c.1275G>A	p.(=)		0.00	3.8E-01	48,0,0	362,0,0	252,1,0	4552,1,0	6752,2,0
C9orf72	c.870C>T	p.(=)		0.88	4.0E-01	35,12,3	275,95,16	193,84,11	3232,1534,200	5090,2459,333
C9orf72	c.620A>G	p.Asn207Ser	0;0	0.93	7.2E-01	38,9,1	301,69,6	194,47,5	3788,1048,87	6523,1227,89
SIGMAR1	c.622C>T, c.355C>T, c.529C>T	p.Arg208Trp, p.Arg119Trp, p.Arg177Trp	1;2	0.37	5.8E-01	13,0,0	161,1,0	127,2,0	4788,19,0	7599,123,1
SIGMAR1	c.595C>T, c.328C>T, c.502C>T	p.Leu199Phe, p.Leu110Phe, p.Leu168Phe	1;0	0.00	4.3E-01	14,0,0	169,0,0	139,1,0	4437,3,0	6640,3,0

VCP	**c.2406T>C**	**p.(=)**		**0.00**	**1.0E+00**	**45,0,0**	**282,0,0**	**262,0,0**	**4559,3,0**	**6762,3,0**
VCP	c.2249A>G	p.Asn750Ser	0;0	Inf	1.0E+00	50,0,0	388,1,0	306,0,0	306,0,0	306,0,0
VCP	**c.2092G>C**	**p.Ala698Pro**	**1;2**	**0.70**	**5.9E-01**	**49,0,0**	**341,7,0**	**272,7,0**	**272,7,0**	**272,7,0**
VCP	c.1704A>G	p.(=)		0.85	7.7E-01	48,1,0	373,5,0	298,5,0	4930,50,1	7835,61,1
VCP	**c.1641G>A**	**p.(=)**		**Inf**	**1.0E+00**	**50,0,0**	**386,1,0**	**299,0,0**	**299,0,0**	**299,0,0**

SIFT predition values of "0" and "1" denote "tolerated" and "deleterious" predictions respectively. PolyPhen (PP) prediction values of "0", "1" and "2" denote "benign", "possibly damaging" and "probably damaging" predictions respectively. Mutations failing variant quality control are shown in bold. OR and p refer to the odds ratio and p-value returned from allelic tests of association with case-control status (tests included internally sequenced samples only). Prefixing of a gene with an "" implies variant occurs within s consensus splice site.*

PARK7:ENST00000338639,ENST00000493373,ENST00000377493,ENST00000493678,ENST00000377488,ENST00000377491,ENST00000469225;TARDBP:ENST00000240185,ENST00000473118,ENST00000315091,ENST00000476201,ENST00000439080;FGGY:ENST00000371218,ENST00000303721,ENST00000371212,ENST00000582567,ENST00000413489,ENST00000371210;DCTN1:ENST00000361874,ENST00000409438,ENST00000407639,ENST00000409868,ENST00000394003,ENST00000409567,ENST00000409240,ENST00000458655,ENST00000417090,ENST00000413111,ENST00000437375,ENST00000421392,ENST00000440727,ENST00000454119,ENST00000449655;ALS2:ENST00000264276,ENST00000457679,ENST00000467448,ENST00000409632,ENST00000410052;CHMP2B:ENST00000263780,ENST00000494980,ENST00000471660;HFE:ENST00000357618,ENST00000353147,ENST00000309234,ENST00000488199,ENST00000349999,ENST00000352392,ENST00000461397,ENST00000470149,ENST00000336625,ENST00000317896,ENST00000397022;FIG4:ENST00000230124,ENST00000454215,ENST00000441478,ENST00000368941,ENST00000415980,ENST00000419951;PON1:ENST00000222381,ENST00000542556;PON3:ENST00000265627,ENST00000427422;PON1,PON3:ENST00000265627,ENST00000427422,ENST00000542556;PON2:ENST00000222572,ENST00000433091,ENST00000536183;DPP6:ENST00000377770,ENST00000406326,,ENST00000404039,ENST00000332007,ENST00000427557;ELP3:ENST00000256398,ENST00000521015,ENST00000521570,ENST00000524024,ENST00000521099,ENST00000520288,ENST00000520270,ENST00000524103,ENST00000380353,EN

ST00000542181,ENST00000537665,ENST00000517975,ENST00000523357,;MOB3B:ENST00000262244;IFNK:ENST00000276943;C9orf72:ENST00000380003,ENST00000379995,ENST00000379997;SIGMAR1:ENST00000277010,ENST00000378892,ENST00000477726;VCP:ENST00000358901;SETX:ENST00000224140,ENST00000393220,ENST00000436441,ENST00000372169,;OPTN:ENST00000263036,ENST00000378764,ENST00000378757,ENST00000378747,ENST00000378752,ENST00000378748,ENST00000430081;ITPR2:ENST00000381340;PRPH:ENST00000257860,ENST00000451891,ENST00000532332;ATXN2:ENST00000377617,ENST00000389154,ENST00000550844,ENST00000542287,ENST00000389153,ENST00000535949,ENST00000482777,ENST00000550889,ENST00000550104,ENST00000471866,ENST00000548492;ANG:ENST00000336811,ENST00000397990;NIPA1:ENST00000337435,ENST00000437912,ENST00000561183;SPG11:ENST00000261866,ENST00000535302,ENST00000559511,ENST00000558138,ENST00000427534,ENST00000558319,ENST00000558988,ENST00000559754,ENST00000559193,ENST00000557866;FUS:ENST00000254108,ENST00000380244,ENST00000568685;GRN:ENST00000053867,ENST00000585512,ENST00000589265,ENST00000593167,ENST00000589536,ENST00000587518,ENST00000587109,ENST00000591740,ENST00000588143,ENST00000588237,ENST00000587387,ENST00000592783,ENST00000586443;MAPT:ENST00000344290,ENST00000431008,ENST00000262410,ENST00000571987,ENST00000535772,ENST00000574436,ENST00000351559,ENST00000415613,,ENST00000576518,ENST00000340799,ENST00000420682,ENST00000446361,ENST00000347967,ENST00000334239;UNC13A:ENST00000519716,ENST00000552293,ENST00000551649,ENST00000252773,ENST00000550896,ENST00000428389;VAPB:ENST00000475243;NEFH:ENST00000310624

Table S5.2: Carriers of known and candidate ALS variants

Gene	Mutation	C9orf72 Expansion	fALS	Sex	Site	Cognitive Status	Age (Years)	Survival (Months)
ALS2	c.2098A>G(p.Thr700Ala)	-	-	M	spinal	NA	67	39
ALS2	c.3094C>T(p.Arg1032Cys)	-	-	M	bulbar	co-morbid FTD	57	>61
ALS2	c.2408A>G(p.Lys803Arg)	-	-	M	spinal	NA	58	>44
ALS2	c.2606A>C(p.Gln869Pro)	-	-	M	spinal	normal	57	>55
ALS2, SETX	c.2098A>G(p.Thr700Ala), c.7682C>T(p.Ser2561Leu)	-	-	M	spinal	NA	84	58
C9orf72		+	+	M	spinal	NA	35	70
C9orf72		+	+	M	bulbar	NA	NA	NA
C9orf72		+	-	M	both	NA	34	67
C9orf72		+	-	F	spinal	NA	51	13
C9orf72		+	-	F	bulbar	NA	60	20
C9orf72		+	-	M	bulbar	NA	47	30
C9orf72		+	-	M	spinal	NA	66	8
C9orf72		+	-	M	spinal	NA	44	23
C9orf72		+	-	F	bulbar	NA	43	36
C9orf72		+	-	F	bulbar	NA	52	17
C9orf72		+	-	M	spinal	NA	68	NA
C9orf72		+	+	F	bulbar	NA	82	12
C9orf72		+	-	F	spinal	NA	42	54
C9orf72		+	-	M	spinal	NA	NA	NA
C9orf72		+	-	F	spinal	normal	55	30

C9orf72	+	+	F	both	normal	44	50
C9orf72	+	-	M	spinal	cognitively impaired	55	24
C9orf72	+	-	M	spinal	co-morbid FTD	62	52
C9orf72	+	+	M	spinal	co-morbid FTD	56	16
C9orf72	+	-	F	spinal	normal	44	37
C9orf72	+	+	F	spinal	co-morbid FTD	65	16
C9orf72	+	+	M	spinal	co-morbid FTD	65	25
C9orf72	+	-	M	spinal	behaviourally impaired	48	>66
C9orf72	+	-	F	bulbar	normal	63	14
C9orf72	+	-	F	spinal	cognitively impaired	49	37
C9orf72	+	-	F	spinal	cognitively impaired	62	18
C9orf72	+	+	M	bulbar	normal	59	19
C9orf72	+	+	F	bulbar	normal	64	41
C9orf72	+	+	F	bulbar	cognitively impaired	65	33
C9orf72	+	+	M	bulbar	normal	66	32
C9orf72	+	-	M	spinal	co-morbid FTD	36	46
C9orf72	+	-	M	bulbar	NA	59	25
C9orf72	+	-	F	spinal	NA	72	>87
C9orf72	+	-	M	bulbar	No	44	29
C9orf72	+	+	M	spinal	cognitively and behaviourally impaired	59	17
C9orf72	+	+	F	spinal	NA	52	>41

C9orf72, CHMP2B	c.123G>T(p.Gln41His)	+	-	M	spinal	NA	55	48
C9orf72, SETX	c.2842C>A(p.Pro948Thr)	+	+	M	bulbar	co-morbid FTD	58	17
C9orf72, SPG11	c.3680A>G(p.Lys1227Arg)	+	+	M	spinal	NA	65	43
CHMP2B	c.118A>G(p.Lys40Glu)	-	-	M	spinal	NA	61	NA
DCTN1	c.2887-2A>G()	-	-	M	spinal	NA	63	26
DCTN1, SPG11	c.2887-2A>G(), c.1529G>A(p.Ser510Asn)	-	-	M	spinal	NA	63	42
DPP6	c.883G>A(p.Glu295Lys)	-	-	F	bulbar	normal	62	21
ELP3	c.326G>A(p.Cys109Tyr)	NA	-	F	spinal	NA	68	12
ELP3	c.206G>T(p.Arg69Leu)	-	-	F	bulbar	cognitively impaired	68	21
ELP3	c.326G>A(p.Cys109Tyr)	-	+	M	bulbar	normal	67	NA
FGGY	c.1716G>A(p.Met572Ile)	-	-	M	spinal	NA	75	11
FUS	c.1574C>T(p.Pro525Leu)	NA	-	M	spinal	NA	13	17
FUS	c.1574C>T(p.Pro525Leu)	-	-	F	bulbar	normal	21	11
HFE	c.766G>A(p.Val256Ile)	-	-	M	spinal	cognitively impaired	76	7
ITPR2	c.3614G>A(p.Arg1205Gln)	-	-	F	bulbar	NA	79	19
MAPT	c.698C>T(p.Pro233Leu)	-	-	M	spinal	normal	57	>55
MAPT	c.284C>T(p.Thr95Met)	-	-	F	spinal	normal	81	>48
OPTN	c.1192C>G(p.Gln398Glu)	-	-	F	spinal	NA	NA	NA
PON2	c.661T>G(p.Ser221Ala)	-	-	F	spinal	co-morbid FTD	58	12
SETX	c.7645G>A(p.Val2549Ile)	-	-	F	spinal	NA	64	24
SETX	c.5842A>G(p.Met1948Val)	-	-	M	spinal	NA	47	NA
SETX	c.2755G>C(p.Val919Leu)	-	-	M	bulbar	co-morbid FTD	68	15

SETX	c.7682C>T(p.Ser2561Leu)	-	-	M	spinal	cognitively impaired	57	33
SETX	c.5587A>G(p.Thr1863Ala)	-	+	F	other	NA	61	>103
SETX	c.7645G>A(p.Val2549Ile)	-	-	F	bulbar	cognitively and behaviourally impaired	76	26
SETX	c.2975A>G(p.Lys992Arg)	-	+	M	bulbar	NA	78	14
SPG11	c.4343G>A(p.Cys1448Tyr)	-	-	M	spinal	NA	63	33
SPG11	c.1930A>T(p.Thr644Ser)	-	+	F	bulbar	NA	68	27
SPG11	c.394A>G(p.Ser132Gly)	-	-	M	spinal	normal	51	>42
SPG11	c.2577A>C(p.Gln859His)	-	-	F	bulbar	NA	59	NA
SPG11	c.7324G>C(p.Ala2442Pro)	-	-	M	spinal	NA	60	21
TARDBP, *ALS2*	c.859G>A(p.Gly287Ser), c.2566A>G(p.Thr856Ala)	-	-	M	bulbar	normal	67	>51
TARDBP, *SETX*	c.859G>A(p.Gly287Ser), c.814C>G(p.His272Asp)	-	-	F	bulbar	NA	66	49
UNC13A	c.3098T>A(p.Val1033Asp)	-	-	M	spinal	NA	80	25
VCP	c.2249A>G(p.Asn750Ser)	-	-	F	bulbar	NA	63	>35

ALS2 - ENST00000264276; CHMP2B - ENST00000263780; DCTN1 - ENST00000361874; DPP6 - ENST00000377770; ELP3 - ENST00000256398; FGGY - ENST00000371218; FUS - ENST00000254108; HFE - ENST00000357618; ITPR2 - ENST00000381340; MAPT - ENST00000344290; OPTN - ENST00000263036; PON2 - ENST00000222572; SETX - ENST00000224140; SPG11 - ENST00000261866; TARDBP - ENST00000240185; UNC13A - ENST00000519716; VCP - ENST00000358901

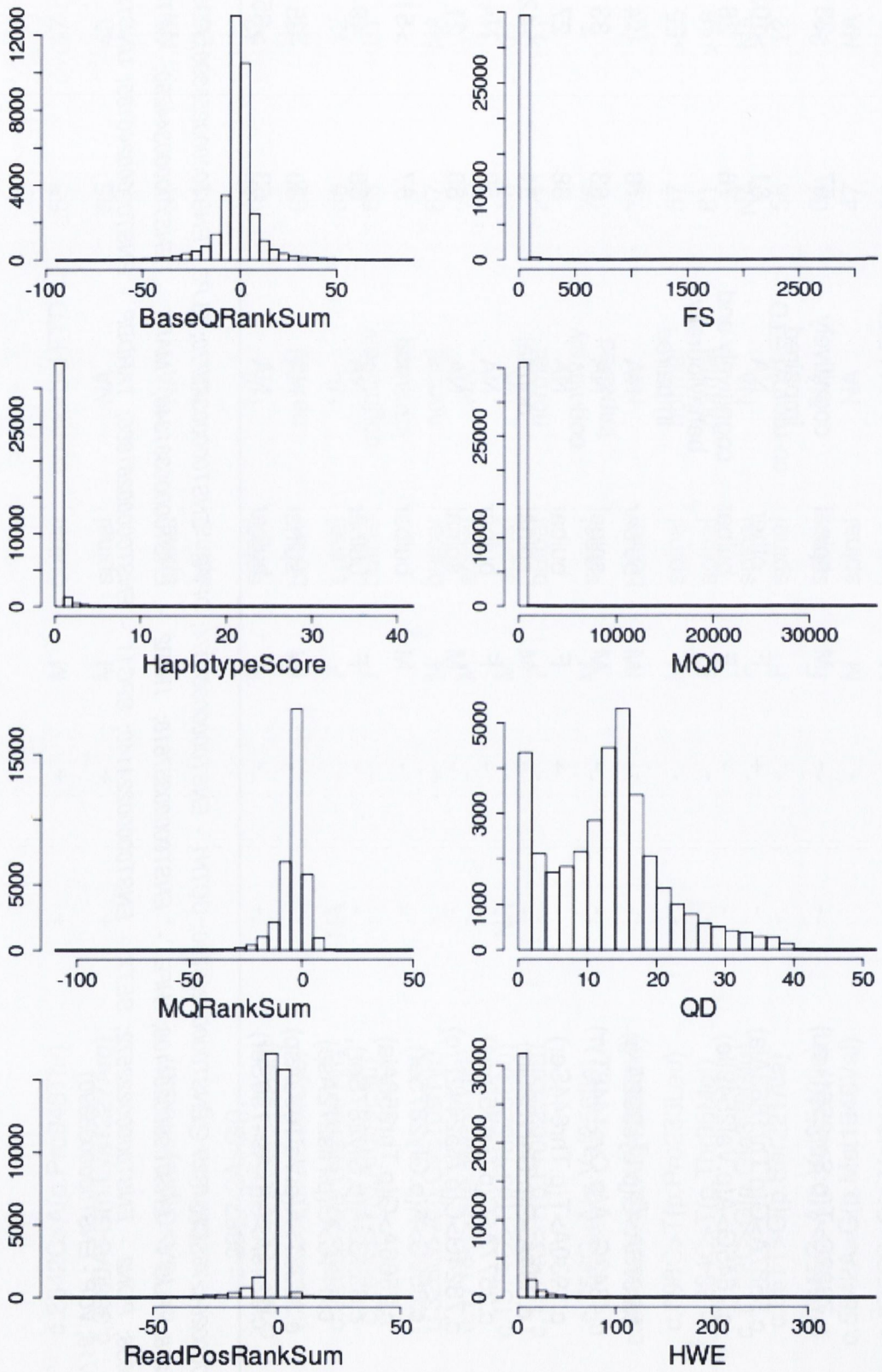
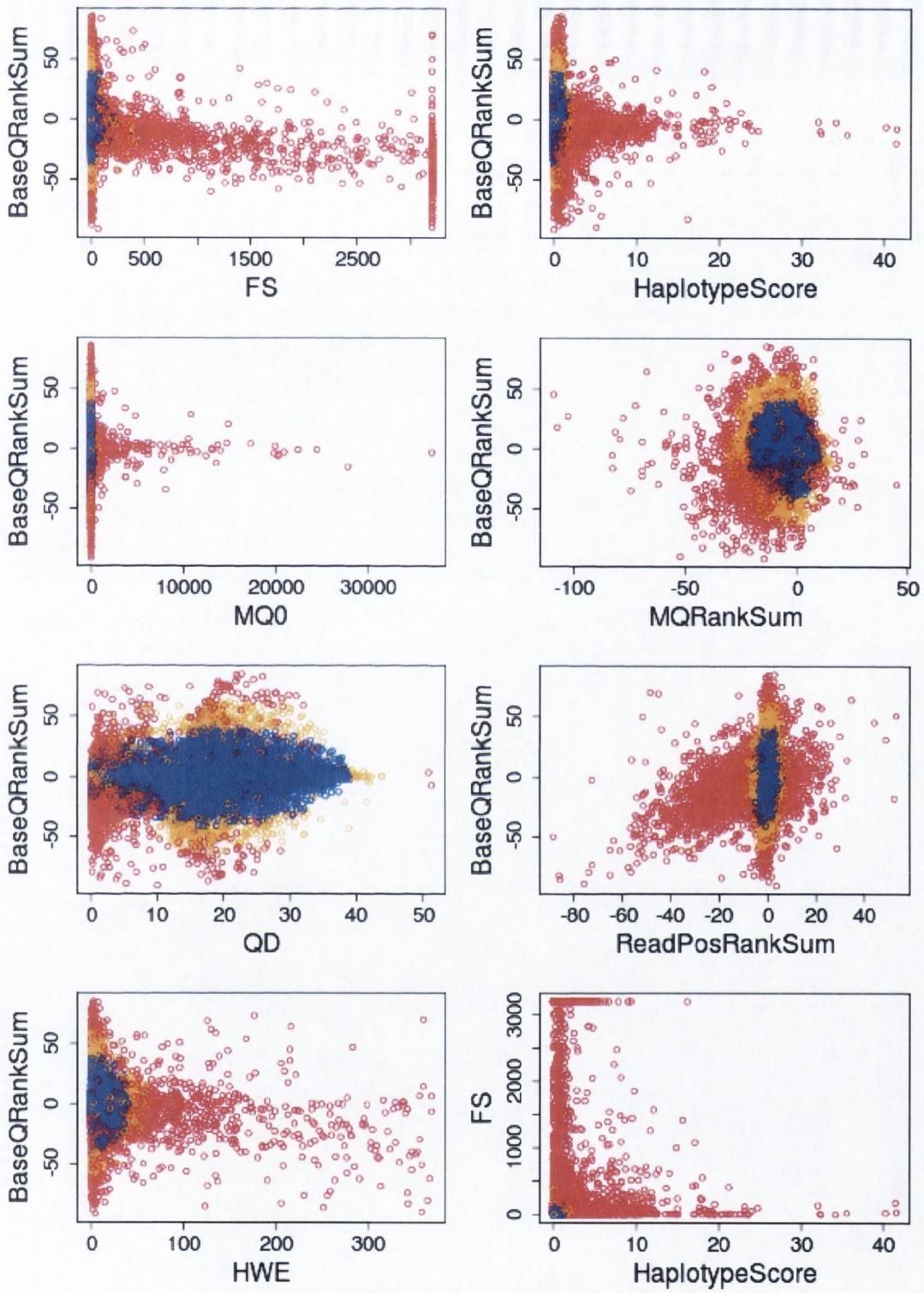
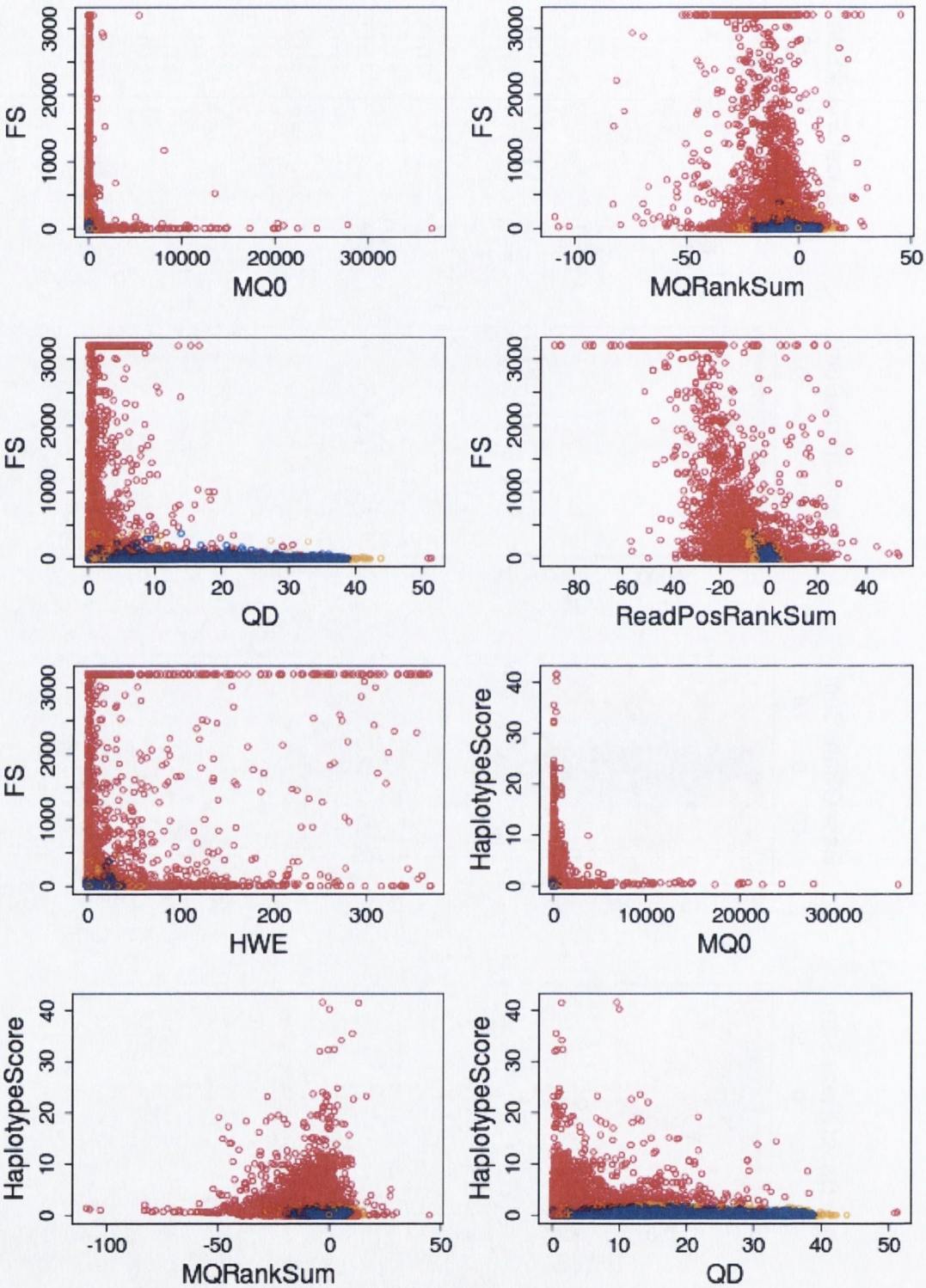
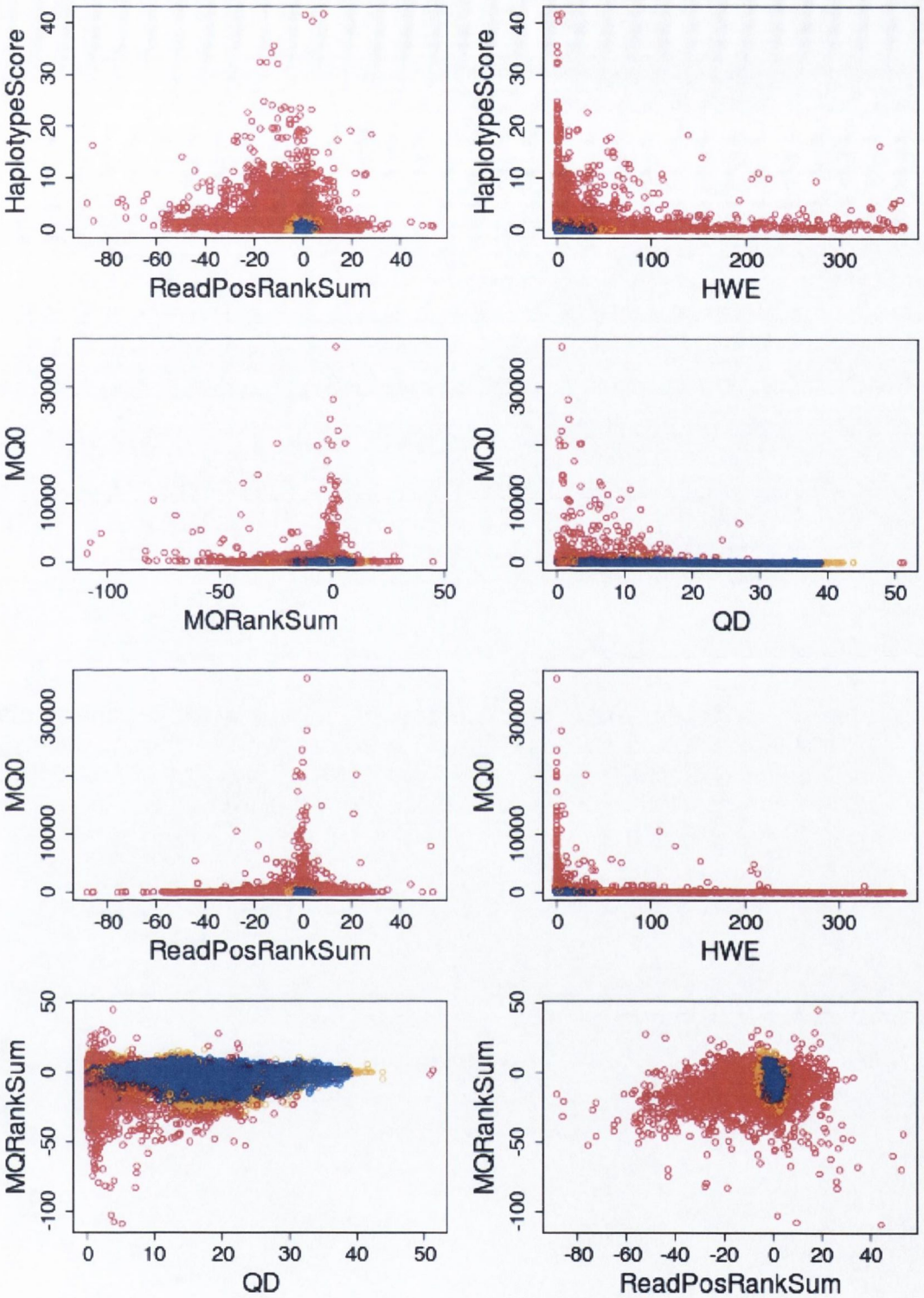


Figure S6.1: Histograms outlining distribution of variant quality control statistics.


BaseQRankSum
FS
HaplotypeScore
MQ0
MQRankSum
QD
ReadPosRankSum
HWE
FS
HaplotypeScore


FS
MQ0
FS
MQRankSum
FS
QD
FS
ReadPosRankSum
FS
HWE
HaplotypeScore
MQ0
HaplotypeScore
MQRankSum
HaplotypeScore
QD


HaplotypeScore
ReadPosRankSum
HaplotypeScore
HWE
MQ0
MQRankSum
MQ0
QD
MQ0
ReadPosRankSum
MQ0
HWE
MQRankSum
QD
MQRankSum
ReadPosRankSum

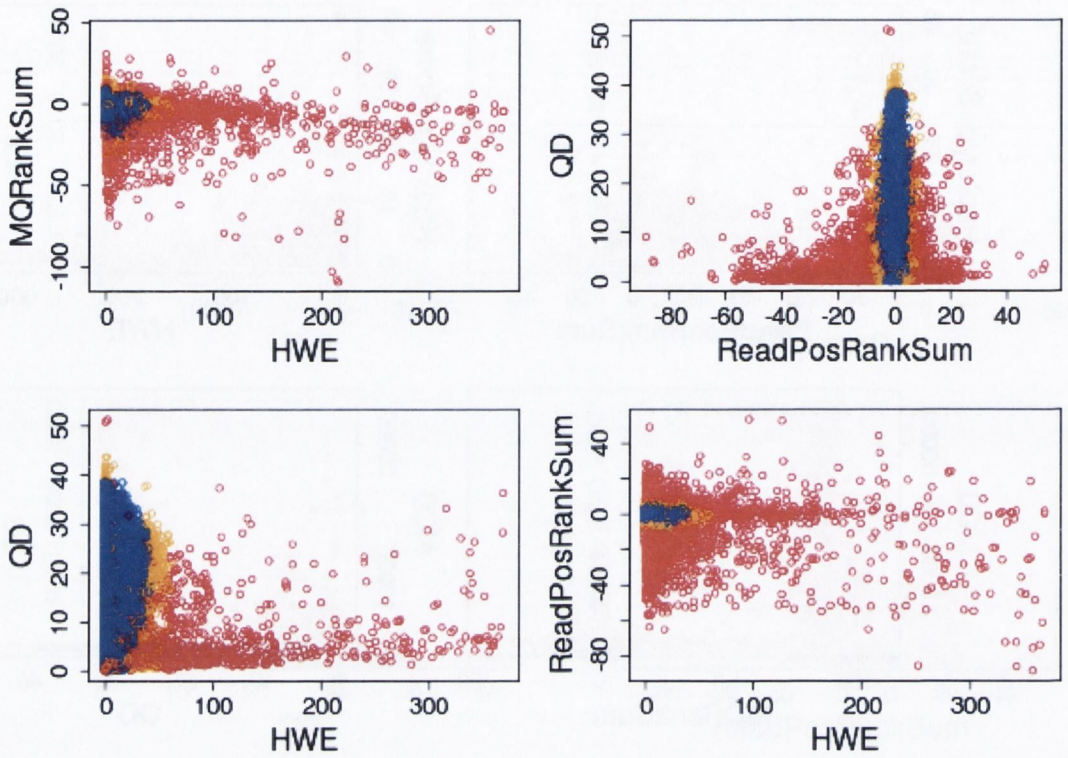


Figure S6.2: Pair-wise distributions of variant quality control statistics. Variants assigned a status of "PASS" before analysis of missing genotype rates are denoted by blue circles, while those assigned a status of "SUSPICIOUS" or "FAIL" are denoted by orange and red circles respectively.

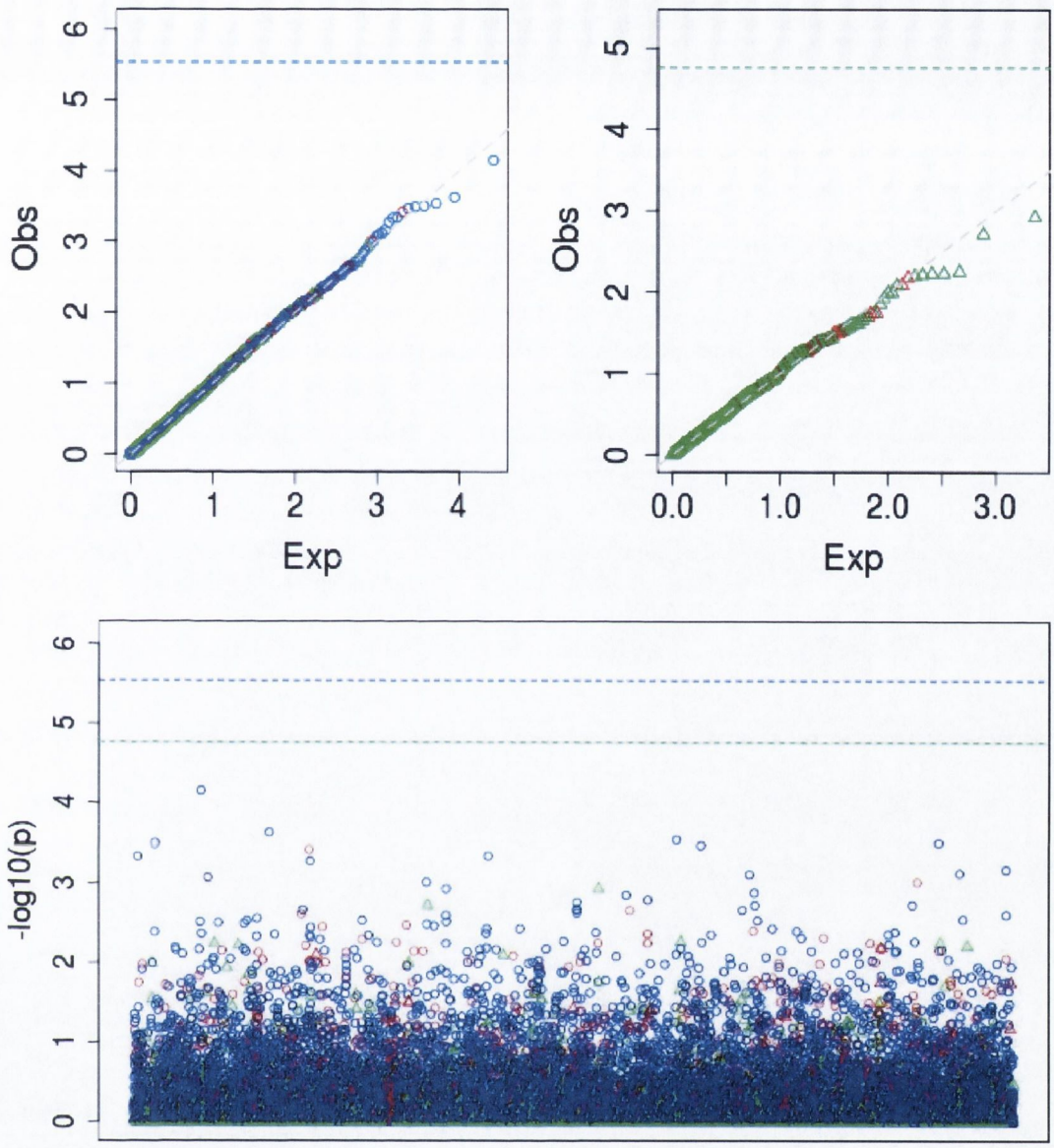


Figure S6.3: Single variant analyses: age of disease onset. Plots denote p-values for linear regression of age of onset with respect to variant minor allele counts. Gender, status for known ALS mutations and sequencing centre were included as sample covariates. Blue circles and green triangles denote non-functional and functional variants respectively. Variants have been re-coloured red where the status assigned during variant QC was "SUSPICIOUS" and grey where they did not meet required minP thresholds. Horizontal blue and green lines denote the thresholds for statistically significant associations among non-functional and functional variants respectively.

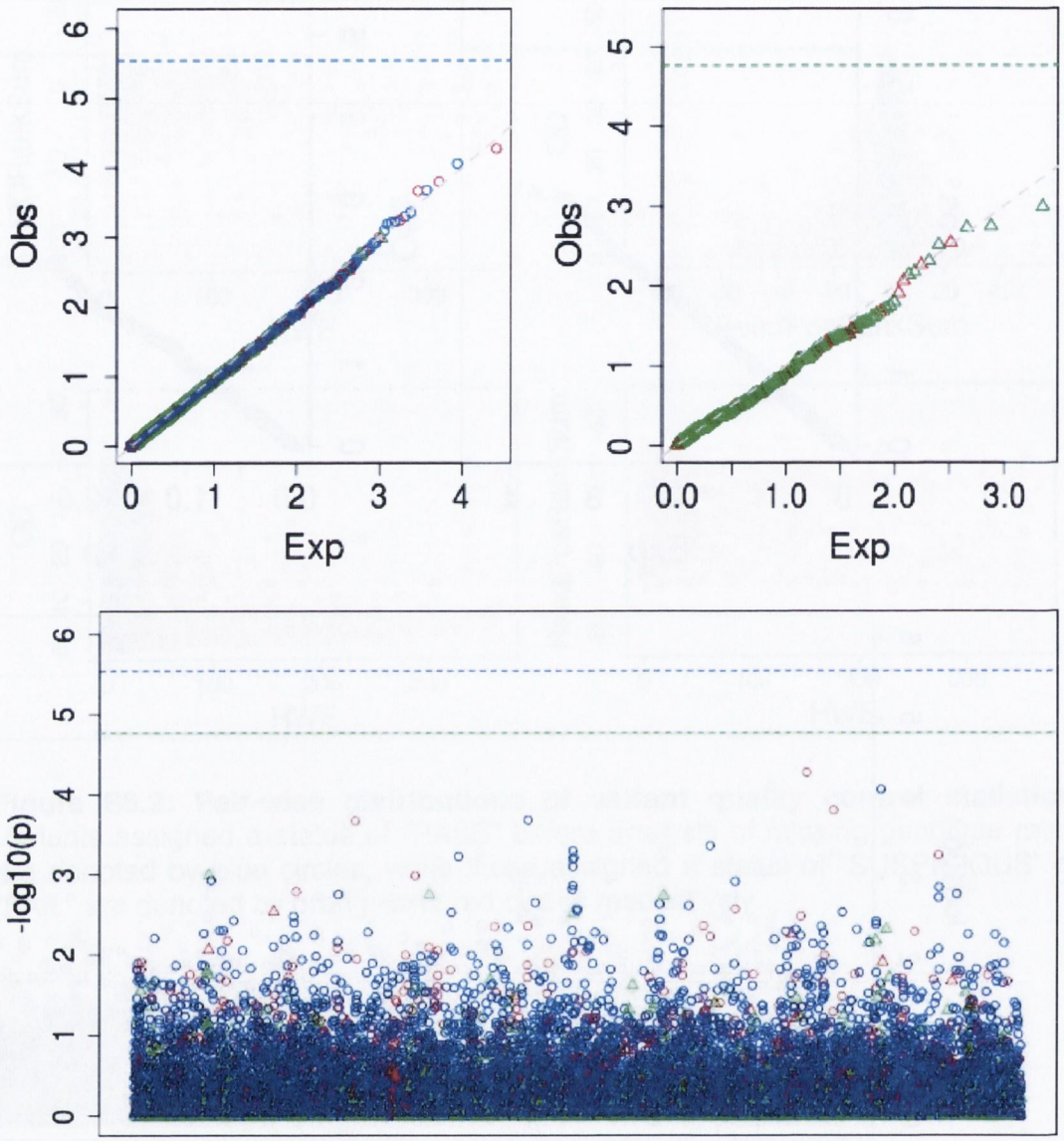


Figure S6.4: Single variant analyses: site of disease onset. Plots denote p-values for logistic regression of site of onset (1 – bulbar, 0 - spinal) with respect to variant minor allele counts. Gender, status for known ALS mutations and sequencing centre were included as sample covariates. Blue circles and green triangles denote non-functional and functional variants respectively. Variants have been re-coloured red where the status assigned during variant QC was "SUSPICIOUS" and grey where they did not meet required minP thresholds. Horizontal blue and green lines denote the thresholds for statistically significant associations among non-functional and functional variants respectively.

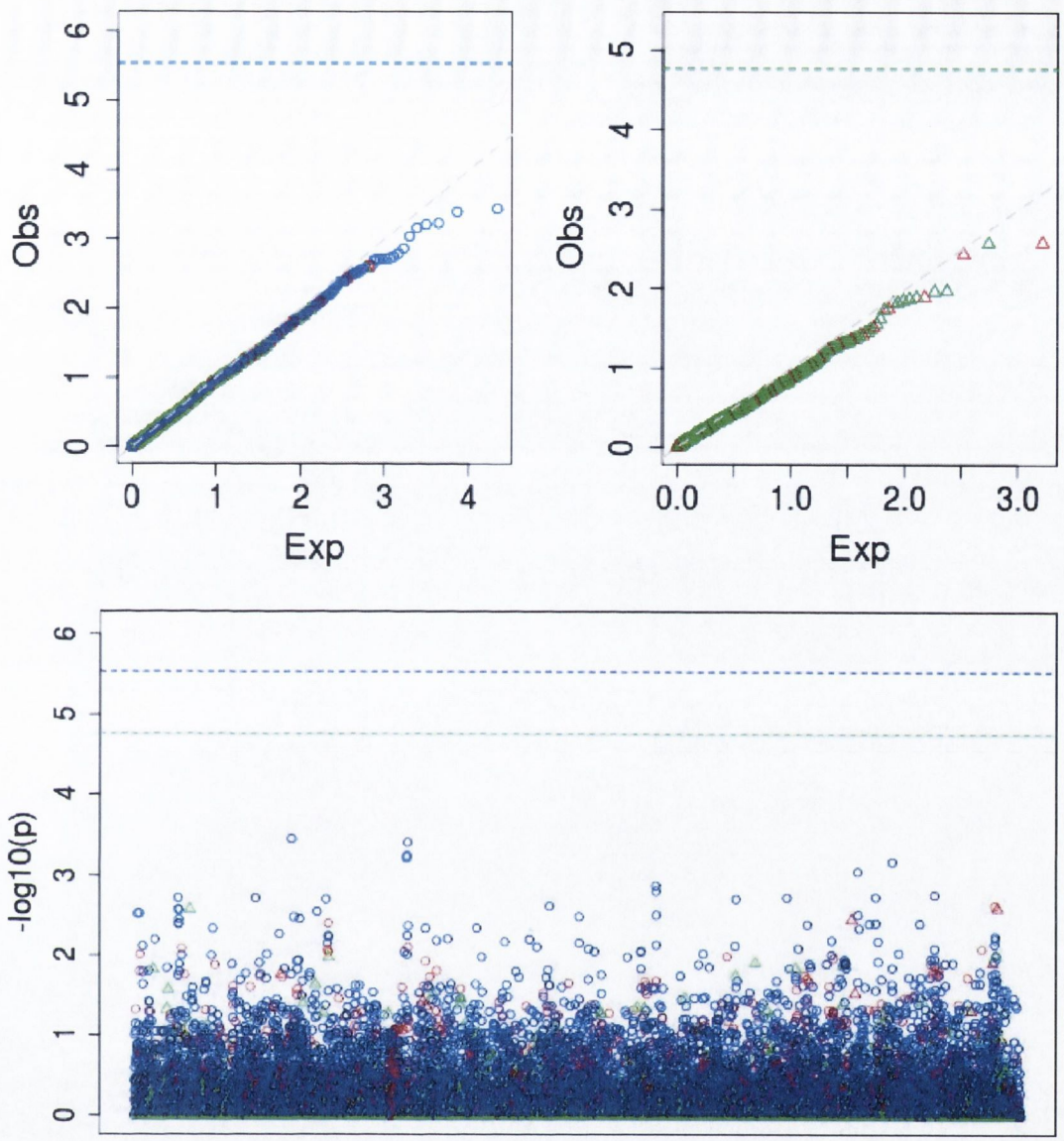


Figure S6.5: Single variant analyses: status for cognitive impairment. Plots denote p-values for logistic regression of status for cognitive impairment (1 – impairment, 0 – no impairment) with respect to variant minor allele counts. Gender, status for known ALS mutations and sequencing centre were included as sample covariates. Blue circles and green triangles denote non-functional and functional variants respectively. Variants have been re-coloured red where the status assigned during variant QC was "SUSPICIOUS" and grey where they did not meet required minP thresholds. Horizontal blue and green lines denote the thresholds for statistically significant associations among non-functional and functional variants respectively.

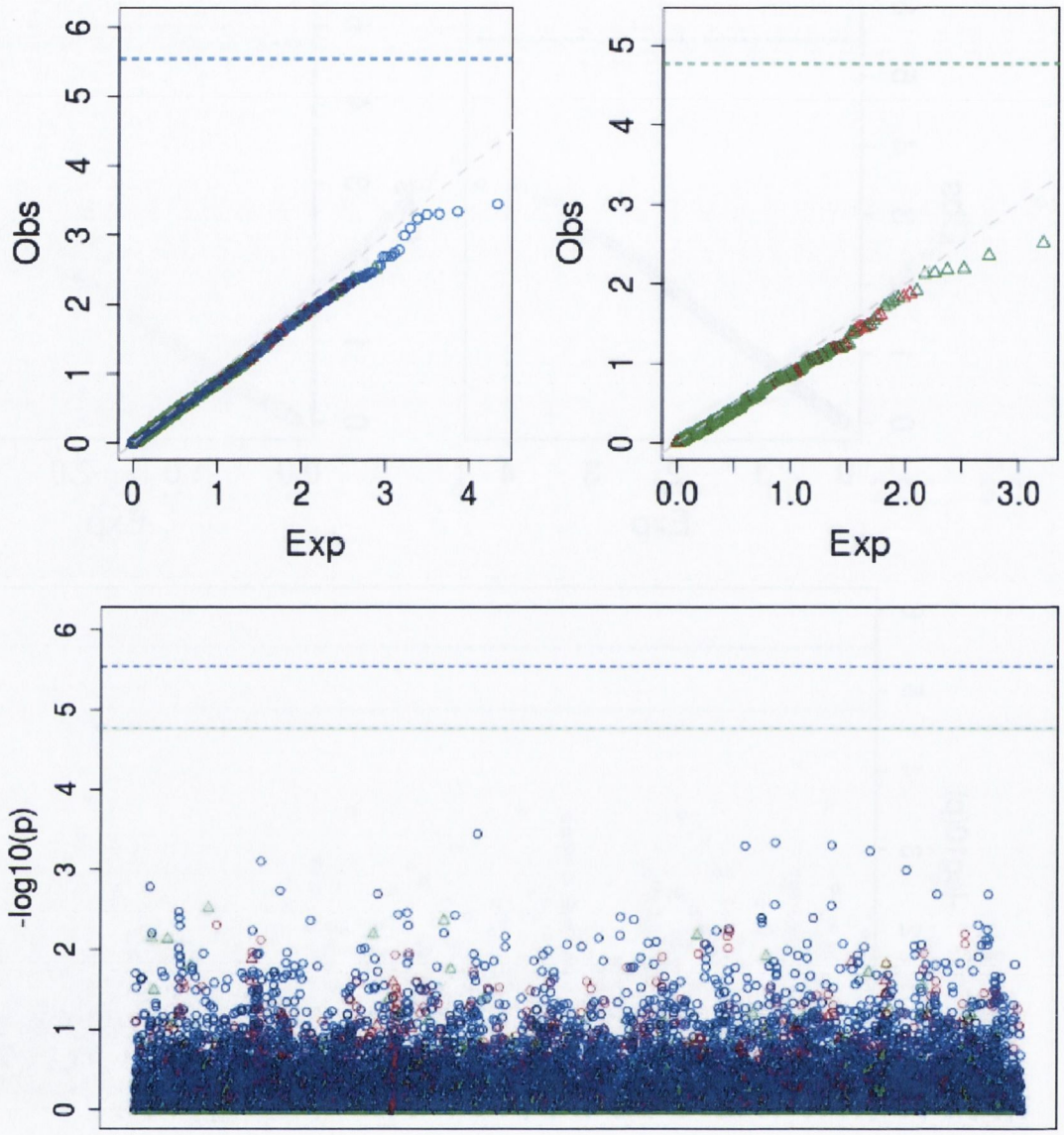


Figure S6.6: Single variant analyses: status for executive dysfunction. Plots denote p-values for logistic regression of status for executive cognitive impairment (1 – impairment, 0 – no impairment) with respect to variant minor allele counts. Gender, status for known ALS mutations and sequencing centre were included as sample covariates. Blue circles and green triangles denote non-functional and functional variants respectively. Variants have been re-coloured red where the status assigned during variant QC was "SUSPICIOUS" and grey where they did not meet required minP thresholds. Horizontal blue and green lines denote the thresholds for statistically significant associations among non-functional and functional variants respectively.

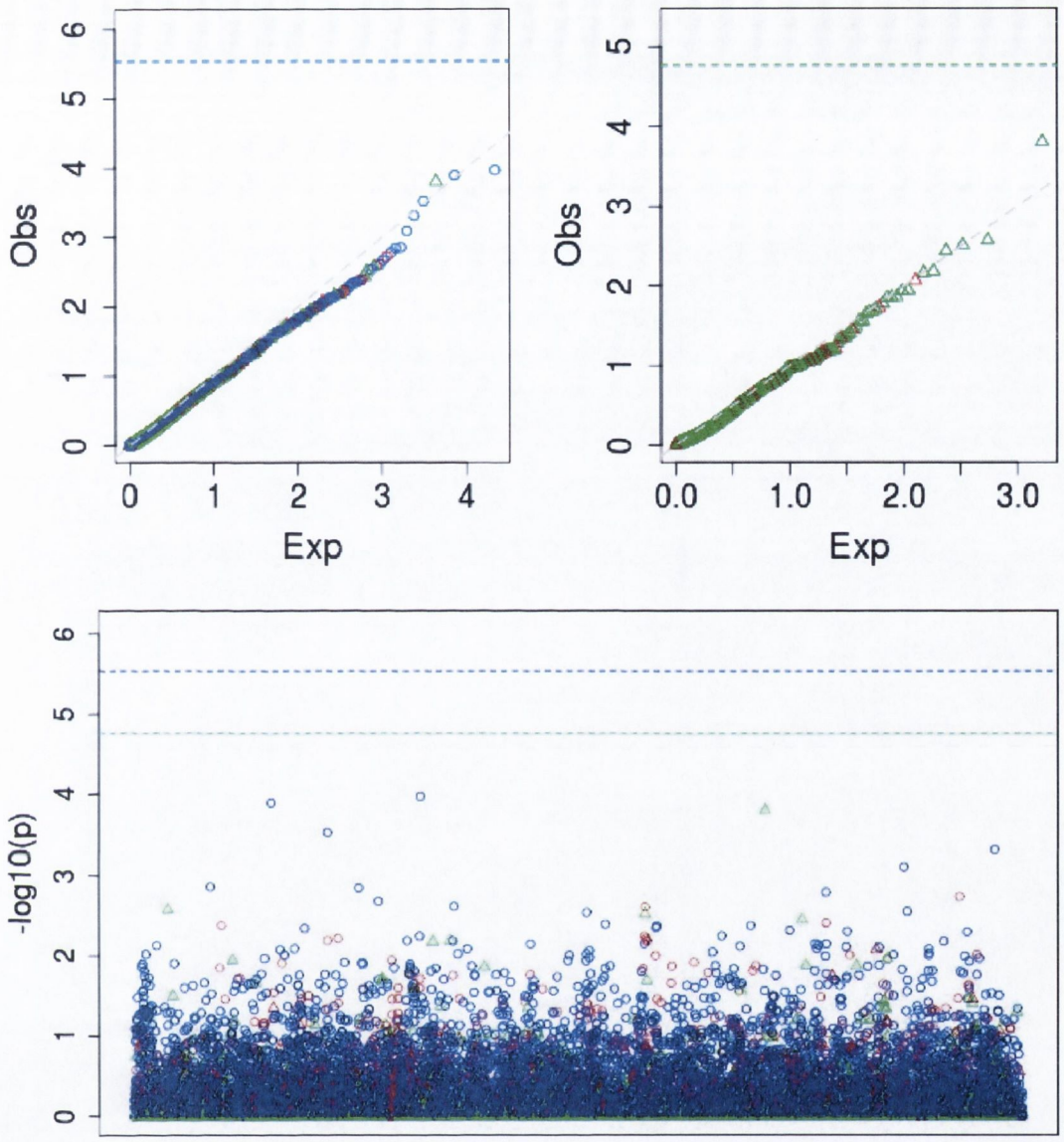


Figure S6.7: Single variant analyses: status for non-executive cognitive impairment. Plots denote p-values for logistic regression of status for non-executive cognitive impairment (1 – impairment, 0 – no impairment) with respect to variant minor allele counts. Gender, status for known ALS mutations and sequencing centre were included as sample covariates. Blue circles and green triangles denote non-functional and functional variants respectively. Variants have been re-coloured red where the status assigned during variant QC was "SUSPICIOUS" and grey where they did not meet required minP thresholds. Horizontal blue and green lines denote the thresholds for statistically significant associations among non-functional and functional variants respectively.

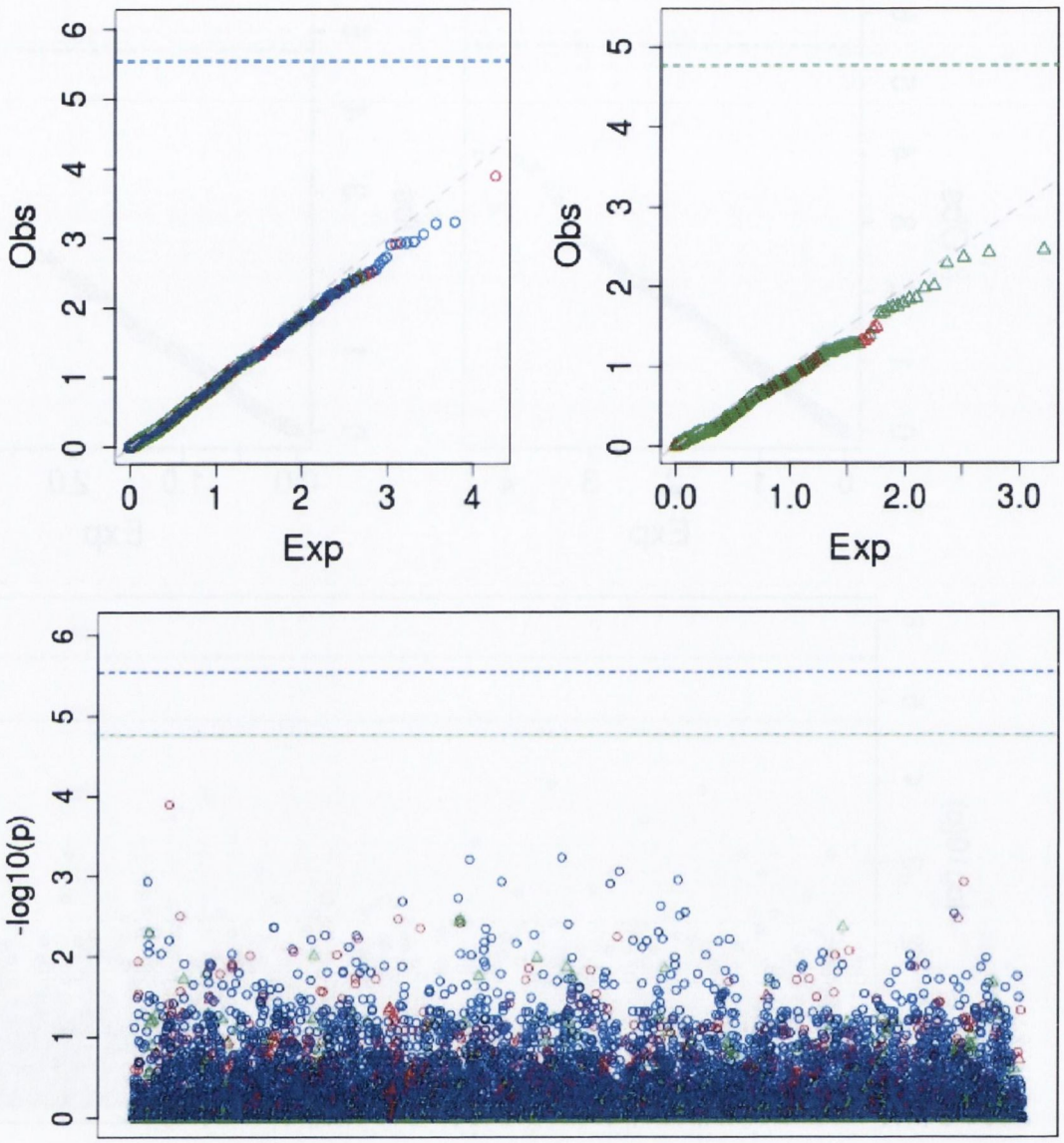


Figure S6.8: Single variant analyses: FTD. Plots denote p-values for logistic regression of status for frontotemporal dementia (1 – present, 0 – absent) with respect to variant minor allele counts. Gender, status for known ALS mutations and sequencing centre were included as sample covariates. Blue circles and green triangles denote non-functional and functional variants respectively. Variants have been re-coloured red where the status assigned during variant QC was "SUSPICIOUS" and grey where they did not meet required minP thresholds. Horizontal blue and green lines denote the thresholds for statistically significant associations among non-functional and functional variants respectively.

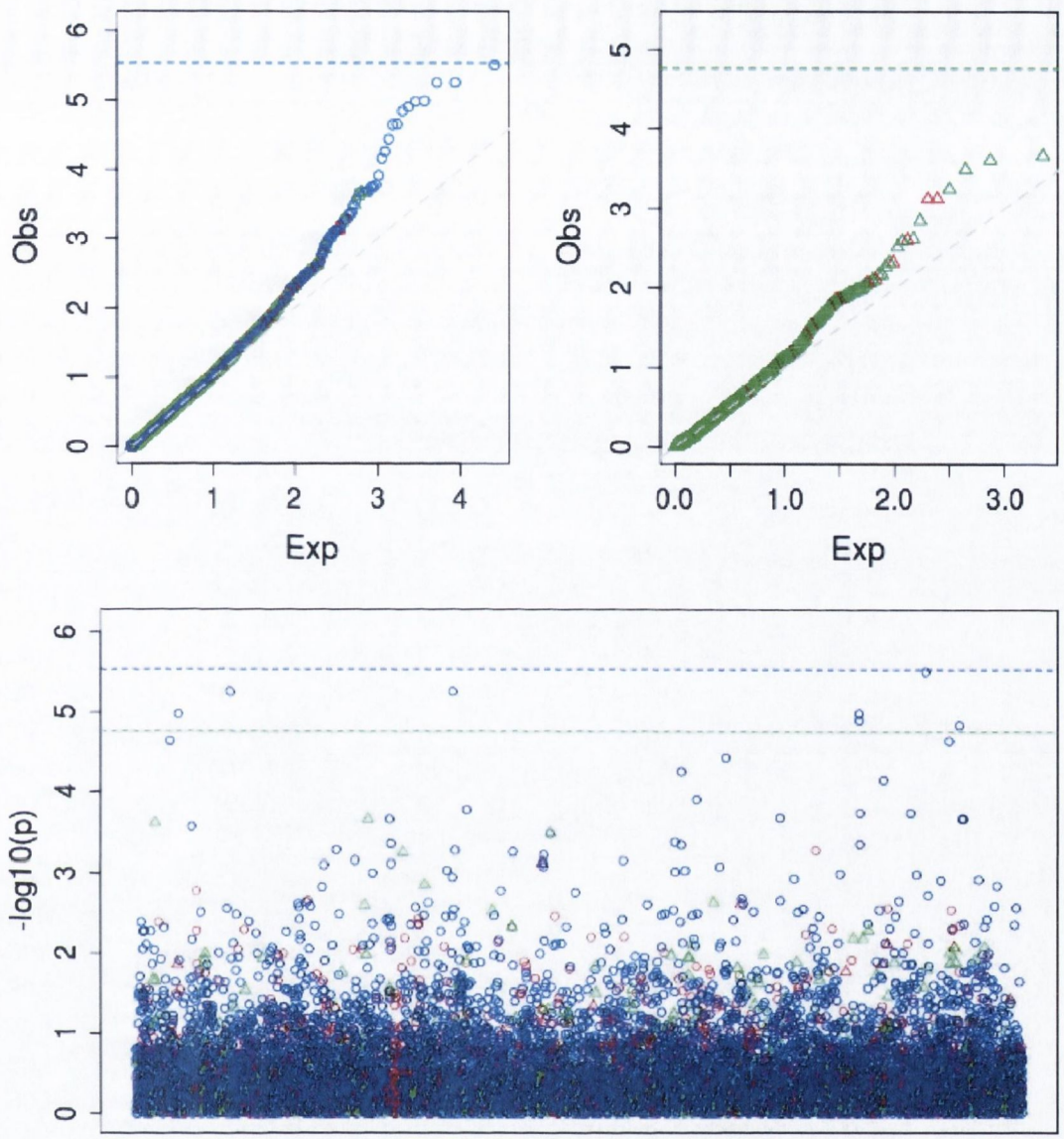


Figure S6.9 Single variant analyses: patient Survival. QQ and Manhattan plots outlining the null probabilities of observed associations between patient survival times and variant minor allele counts. Tests were based on the fitting of Cox-proportional hazard models where gender, status for known ALS mutations and sequencing centre were incorporated as sample covariates. Blue circles and green triangles denote non-functional and functional variants respectively. Variants have been re-coloured red where the status assigned during variant QC was "SUSPICIOUS" and grey where they did not meet required minP thresholds. Horizontal blue and green lines denote the thresholds for statistically significant associations among non-functional and functional variants respectively

Table S6.1 Aggregate Analyses

Locus	Risk (EMP1)	Risk (EMP2)	Phenotype (EMP1)	Phenotype (EMP2)
ABL1	4.9E-01	1.0E+00	1.4E-01	1.0E+00
ACTA1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ACTA2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ACTB	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ACTC1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ACTN2	6.4E-01	1.0E+00	1.0E+00	1.0E+00
ACTR1A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ACTR1B	2.6E-01	1.0E+00	7.0E-01	1.0E+00
ADAP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ADRB2	4.0E-01	1.0E+00	7.3E-01	1.0E+00
AHR	9.7E-01	1.0E+00	7.6E-01	1.0E+00
AKAP13	9.7E-01	1.0E+00	2.3E-01	1.0E+00
AKAP9	2.7E-01	1.0E+00	3.6E-01	1.0E+00
AKT1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
AKT2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
AKTIP	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ALB	8.5E-01	1.0E+00	2.0E-01	1.0E+00
ALMS1	9.1E-01	1.0E+00	1.2E-01	1.0E+00
ALPK2	5.7E-01	1.0E+00	8.1E-02	1.0E+00
ALS2	6.1E-01	1.0E+00	7.4E-01	1.0E+00
ALS2CL	5.4E-01	1.0E+00	9.2E-01	1.0E+00
ALYREF	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ANG	2.3E-02	1.0E+00	1.0E+00	1.0E+00
ANXA1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ANXA2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
AP2B1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
AP2M1	2.9E-01	1.0E+00	1.0E+00	1.0E+00
APAF1	7.3E-01	1.0E+00	5.3E-01	1.0E+00
APBA1	6.0E-01	1.0E+00	1.0E+00	1.0E+00
APBB1	5.8E-01	1.0E+00	5.4E-01	1.0E+00
APC	9.4E-01	1.0E+00	9.9E-01	1.0E+00
APEX1	8.3E-02	1.0E+00	4.9E-01	1.0E+00
APOA1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
APOB	8.2E-02	1.0E+00	5.8E-02	1.0E+00
APOE	1.0E+00	1.0E+00	1.0E+00	1.0E+00
APP	2.3E-01	1.0E+00	2.2E-01	1.0E+00
APTX	1.1E-01	1.0E+00	6.6E-01	1.0E+00
AR	2.1E-01	1.0E+00	9.3E-01	1.0E+00
ARF6	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ARHGDIA	4.3E-02	1.0E+00	9.0E-02	1.0E+00
ARHGEF10	6.5E-02	1.0E+00	8.9E-01	1.0E+00
ARHGEF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ARHGEF6	6.2E-02	1.0E+00	7.5E-01	1.0E+00

ARHGEF7	9.4E-01	1.0E+00	8.4E-01	1.0E+00
ARIH2	3.3E-02	1.0E+00	1.0E+00	1.0E+00
ARRB2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ATF2	1.5E-01	1.0E+00	3.9E-02	1.0E+00
ATF6	6.9E-01	1.0E+00	7.6E-01	1.0E+00
ATM	2.9E-01	1.0E+00	6.2E-01	1.0E+00
ATN1	2.0E-01	1.0E+00	3.2E-01	1.0E+00
ATP2A2	5.9E-01	1.0E+00	6.0E-01	1.0E+00
ATP6AP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ATXN1	3.7E-01	1.0E+00	5.7E-01	1.0E+00
ATXN10	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ATXN2	3.3E-01	1.0E+00	9.3E-01	1.0E+00
ATXN3	5.0E-01	1.0E+00	6.6E-01	1.0E+00
B4GALT1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
BAD	1.0E+00	1.0E+00	1.0E+00	1.0E+00
BAG1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
BAX	8.2E-01	1.0E+00	1.0E+00	1.0E+00
BBS1	3.4E-01	1.0E+00	6.9E-01	1.0E+00
BBS4	6.8E-01	1.0E+00	7.8E-02	1.0E+00
BCAR3	4.7E-01	1.0E+00	5.8E-01	1.0E+00
BCL2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
BCL2L1	1.5E-01	1.0E+00	1.4E-01	1.0E+00
BCL2L11	1.0E+00	1.0E+00	1.0E+00	1.0E+00
BCL6	5.0E-02	1.0E+00	2.1E-01	1.0E+00
BICD2	6.0E-01	1.0E+00	2.4E-01	1.0E+00
BID	1.0E+00	1.0E+00	1.0E+00	1.0E+00
BIN1	2.4E-01	1.0E+00	1.0E+00	1.0E+00
BLM	8.4E-01	1.0E+00	9.1E-01	1.0E+00
BMP7	1.0E+00	1.0E+00	1.0E+00	1.0E+00
BMPR2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
BNIP3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
BOLL	1.0E+00	1.0E+00	1.0E+00	1.0E+00
BRCA1	6.7E-01	1.0E+00	1.9E-01	1.0E+00
BSCL2	2.9E-01	1.0E+00	2.9E-01	1.0E+00
BUB1	8.1E-01	1.0E+00	2.0E-01	1.0E+00
C10orf2	9.6E-01	1.0E+00	5.0E-01	1.0E+00
C9orf72	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CACNA1A	4.1E-01	1.0E+00	8.5E-01	1.0E+00
CALCOCO1	1.0E-02	9.7E-01	7.0E-01	1.0E+00
CALCOCO2	1.1E-01	1.0E+00	8.1E-01	1.0E+00
CALM1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CALM2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CALM3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CALR	4.2E-01	1.0E+00	3.6E-02	1.0E+00
CASP1	3.4E-01	1.0E+00	5.3E-01	1.0E+00
CASP3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CASP4	4.4E-01	1.0E+00	9.8E-01	1.0E+00
CASP7	5.8E-01	1.0E+00	7.3E-01	1.0E+00

CASP8	1.9E-01	1.0E+00	5.7E-02	1.0E+00
CASP9	6.4E-01	1.0E+00	6.5E-01	1.0E+00
CAV1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CBS	9.8E-01	1.0E+00	7.1E-01	1.0E+00
CCAR1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CCDC53	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CCND1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CCNE1	7.1E-01	1.0E+00	1.9E-02	1.0E+00
CCP110	5.2E-01	1.0E+00	7.9E-01	1.0E+00
CCS	2.2E-01	1.0E+00	7.3E-01	1.0E+00
CCT5	7.9E-01	1.0E+00	3.0E-01	1.0E+00
CD2BP2	1.3E-01	1.0E+00	1.5E-01	1.0E+00
CD4	1.3E-01	1.0E+00	1.0E+00	1.0E+00
CD40	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CDC37	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CDC42	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CDC5L	3.8E-01	1.0E+00	1.0E+00	1.0E+00
CDH1	1.0E+00	1.0E+00	2.0E-02	1.0E+00
CDH2	6.2E-01	1.0E+00	6.8E-01	1.0E+00
CDK1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CDK2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CDK4	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CDK5	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CDK5R1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CDK5RAP2	5.0E-01	1.0E+00	2.8E-01	1.0E+00
CDKN1A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CDKN1B	8.7E-01	1.0E+00	4.3E-01	1.0E+00
CDKN2A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CELF1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CELF4	4.0E-02	1.0E+00	1.0E+00	1.0E+00
CENPJ	8.4E-01	1.0E+00	5.1E-01	1.0E+00
CEP250	6.3E-01	1.0E+00	6.2E-01	1.0E+00
CEP290	9.0E-01	1.0E+00	1.0E+00	1.0E+00
CEP57	4.2E-01	1.0E+00	7.2E-01	1.0E+00
CEP63	4.1E-01	1.0E+00	7.3E-02	1.0E+00
CEP76	5.3E-01	1.0E+00	3.7E-01	1.0E+00
CEP78	9.5E-01	1.0E+00	9.5E-01	1.0E+00
CETN2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CFTR	1.2E-01	1.0E+00	8.8E-01	1.0E+00
CHM	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CHMP2B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CHMP5	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CIAO1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CIRBP	7.4E-01	1.0E+00	7.4E-01	1.0E+00
CKAP5	4.4E-01	1.0E+00	9.1E-01	1.0E+00
CLASP1	6.3E-01	1.0E+00	5.4E-01	1.0E+00
CLIP1	5.9E-01	1.0E+00	9.4E-01	1.0E+00
CLIP2	3.4E-01	1.0E+00	7.8E-01	1.0E+00

CLTC	7.1E-01	1.0E+00	4.4E-02	1.0E+00
COIL	1.0E+00	1.0E+00	1.0E+00	1.0E+00
COL4A3BP	2.3E-01	1.0E+00	7.0E-01	1.0E+00
CP	8.7E-01	1.0E+00	4.4E-01	1.0E+00
CPNE4	6.5E-01	1.0E+00	7.9E-01	1.0E+00
CREB1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CREBBP	4.9E-01	1.0E+00	4.3E-01	1.0E+00
CSNK1D	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CSNK1E	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CSNK2A1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CSTF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CSTF2T	7.0E-01	1.0E+00	4.2E-01	1.0E+00
CTCF	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CTDP1	3.4E-01	1.0E+00	1.4E-01	1.0E+00
CTGF	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CTNNB1	7.6E-01	1.0E+00	4.1E-01	1.0E+00
CTNND1	3.8E-02	1.0E+00	4.6E-01	1.0E+00
CTTNBP2	1.1E-01	1.0E+00	9.3E-01	1.0E+00
CXCR4	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CYCS	1.0E+00	1.0E+00	1.0E+00	1.0E+00
CYLD	1.0E+00	1.0E+00	1.0E+00	1.0E+00
DAXX	4.0E-01	1.0E+00	6.8E-01	1.0E+00
DCC	7.4E-01	1.0E+00	3.0E-01	1.0E+00
DCTN1	5.6E-01	1.0E+00	7.7E-01	1.0E+00
DCTN2	5.4E-01	1.0E+00	8.0E-01	1.0E+00
DCTN3	4.8E-01	1.0E+00	4.8E-01	1.0E+00
DDIT3	3.1E-01	1.0E+00	4.9E-01	1.0E+00
DDOST	7.9E-01	1.0E+00	2.2E-01	1.0E+00
DDX1	5.3E-01	1.0E+00	8.2E-01	1.0E+00
DDX17	3.3E-01	1.0E+00	1.0E+00	1.0E+00
DDX3X	1.0E+00	1.0E+00	1.0E+00	1.0E+00
DDX5	4.6E-01	1.0E+00	1.3E-02	1.0E+00
DDX50	3.1E-01	1.0E+00	8.7E-01	1.0E+00
DDX58	4.1E-01	1.0E+00	1.7E-01	1.0E+00
DHX15	1.0E+00	1.0E+00	1.0E+00	1.0E+00
DHX38	6.5E-01	1.0E+00	1.4E-01	1.0E+00
DHX9	1.0E+00	1.0E+00	1.0E+00	1.0E+00
DLG4	2.8E-01	1.0E+00	3.3E-01	1.0E+00
DMD	2.0E-01	1.0E+00	2.6E-01	1.0E+00
DNAJA1	9.6E-01	1.0E+00	1.0E+00	1.0E+00
DNAJB1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
DNM2	3.4E-01	1.0E+00	3.8E-01	1.0E+00
DNMT3A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
DPP6	1.0E+00	1.0E+00	1.0E+00	1.0E+00
DRD2	9.2E-01	1.0E+00	7.2E-01	1.0E+00
DROSHA	5.9E-01	1.0E+00	9.6E-01	1.0E+00
DST	2.3E-01	1.0E+00	5.2E-01	1.0E+00
DTNBP1	7.8E-01	1.0E+00	3.7E-01	1.0E+00

DUSP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
DVL1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
DYNC1H1	1.3E-01	1.0E+00	2.3E-01	1.0E+00
DYNC1I2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
DYNLL1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
EEA1	1.1E-01	1.0E+00	4.9E-01	1.0E+00
EEF1A1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
EEF1G	1.0E+00	1.0E+00	1.0E+00	1.0E+00
EEF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
EFNB2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
EGFR	3.7E-01	1.0E+00	9.4E-01	1.0E+00
EIF2C2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
EIF6	9.6E-01	1.0E+00	2.2E-01	1.0E+00
ELANE	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ELAVL1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ELAVL4	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ELP3	3.1E-01	1.0E+00	2.6E-01	1.0E+00
EP300	8.1E-01	1.0E+00	3.2E-01	1.0E+00
EPB41	7.6E-01	1.0E+00	6.5E-01	1.0E+00
EPHA4	1.0E+00	1.0E+00	1.0E+00	1.0E+00
EPS15	3.0E-01	1.0E+00	5.6E-03	1.0E+00
ERBB2	1.2E-01	1.0E+00	4.4E-01	1.0E+00
ERBB3	5.1E-01	1.0E+00	6.9E-01	1.0E+00
ERCC2	3.3E-01	1.0E+00	2.9E-01	1.0E+00
ESR1	2.2E-01	1.0E+00	9.4E-01	1.0E+00
ESR2	9.2E-01	1.0E+00	3.7E-01	1.0E+00
ETS1	4.8E-01	1.0E+00	2.8E-01	1.0E+00
ETS2	6.5E-01	1.0E+00	8.8E-01	1.0E+00
ETV6	4.4E-01	1.0E+00	1.0E-01	1.0E+00
EWSR1	2.6E-01	1.0E+00	8.2E-01	1.0E+00
EXOSC10	8.8E-01	1.0E+00	5.4E-01	1.0E+00
EZR	1.0E+00	1.0E+00	1.0E+00	1.0E+00
FAM110A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
FAS	1.0E+00	1.0E+00	1.0E+00	1.0E+00
FASLG	8.7E-01	1.0E+00	1.9E-01	1.0E+00
FBLN1	4.5E-02	1.0E+00	2.5E-01	1.0E+00
FECH	9.4E-01	1.0E+00	2.2E-03	9.8E-01
FGD4	1.3E-01	1.0E+00	1.0E+00	1.0E+00
FGFR1	4.9E-01	1.0E+00	8.8E-01	1.0E+00
FGFR1OP	6.3E-01	1.0E+00	8.4E-01	1.0E+00
FGFR3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
FGGY	4.9E-01	1.0E+00	2.5E-01	1.0E+00
FIG4	8.4E-01	1.0E+00	4.2E-01	1.0E+00
FLII	4.3E-01	1.0E+00	7.2E-01	1.0E+00
FLNA	3.3E-02	1.0E+00	1.0E+00	1.0E+00
FN1	3.3E-01	1.0E+00	2.2E-01	1.0E+00
FOS	9.7E-01	1.0E+00	2.2E-01	1.0E+00
FST	8.7E-01	1.0E+00	2.3E-01	1.0E+00

FUS	7.3E-01	1.0E+00	2.1E-02	1.0E+00
FYN	4.6E-01	1.0E+00	7.2E-01	1.0E+00
G3BP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
G3BP2	1.6E-01	1.0E+00	7.1E-01	1.0E+00
GAN	9.0E-02	1.0E+00	2.2E-01	1.0E+00
GAPDH	1.0E+00	1.0E+00	1.0E+00	1.0E+00
GARS	4.2E-02	1.0E+00	8.5E-01	1.0E+00
GATA1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
GH1	7.8E-01	1.0E+00	2.8E-01	1.0E+00
GHR	1.2E-01	1.0E+00	2.5E-01	1.0E+00
GJB1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
GJC2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
GNAI2	4.8E-01	1.0E+00	4.8E-01	1.0E+00
GRB2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
GRIK2	1.0E+00	1.0E+00	9.0E-01	1.0E+00
GRIN1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
GRIN2A	6.4E-01	1.0E+00	3.4E-01	1.0E+00
GRIN2B	4.3E-01	1.0E+00	6.6E-01	1.0E+00
GRM1	2.3E-01	1.0E+00	3.0E-02	1.0E+00
GRM5	1.0E+00	1.0E+00	1.0E+00	1.0E+00
GRN	3.2E-01	1.0E+00	4.4E-01	1.0E+00
GRP	4.7E-01	1.0E+00	5.8E-01	1.0E+00
GSK3B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
GSN	5.6E-01	1.0E+00	8.2E-01	1.0E+00
GTF3A	8.0E-01	1.0E+00	7.1E-02	1.0E+00
H2AFX	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HAP1	7.9E-01	1.0E+00	9.4E-01	1.0E+00
HAUS2	4.6E-01	1.0E+00	1.0E+00	1.0E+00
HDAC1	5.3E-01	1.0E+00	1.0E+00	1.0E+00
HDAC2	9.3E-01	1.0E+00	1.0E+00	1.0E+00
HDAC3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HDAC6	4.4E-01	1.0E+00	3.8E-01	1.0E+00
HECW1	1.9E-01	1.0E+00	1.6E-01	1.0E+00
HERC3	4.3E-01	1.0E+00	9.6E-01	1.0E+00
HFE	9.3E-01	1.0E+00	2.2E-01	1.0E+00
HGS	5.5E-01	1.0E+00	9.5E-01	1.0E+00
HIF1A	9.5E-01	1.0E+00	5.3E-01	1.0E+00
HLA-B	2.4E-01	1.0E+00	7.9E-01	1.0E+00
HLA-DRB1	2.7E-01	1.0E+00	9.3E-01	1.0E+00
HMGB1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HMOX1	7.2E-01	1.0E+00	1.8E-01	1.0E+00
HNRNPA0	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPA1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPA2B1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPA3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPAB	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPC	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPD	1.0E+00	1.0E+00	1.0E+00	1.0E+00

HNRNPF	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPH1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPH2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPH3	6.6E-01	1.0E+00	1.9E-01	1.0E+00
HNRNPK	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPL	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPM	7.8E-01	1.0E+00	3.6E-01	1.0E+00
HNRNPR	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRNPU	8.4E-01	1.0E+00	3.2E-01	1.0E+00
HNRNPUL1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HNRPDL	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HOMER1	3.3E-01	1.0E+00	1.0E+00	1.0E+00
HOXB2	1.7E-01	1.0E+00	1.0E+00	1.0E+00
HOXD10	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HPRT1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HRAS	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HSF1	2.1E-01	1.0E+00	1.7E-01	1.0E+00
HSP90AA1	4.6E-01	1.0E+00	9.0E-01	1.0E+00
HSP90AB1	6.1E-01	1.0E+00	5.7E-01	1.0E+00
HSPA5	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HSPA8	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HSPB1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HSPB8	7.2E-01	1.0E+00	6.5E-01	1.0E+00
HSPD1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HTATSF1	5.5E-01	1.0E+00	9.3E-01	1.0E+00
HTRA2	4.3E-02	1.0E+00	1.0E+00	1.0E+00
HTT	7.9E-01	1.0E+00	4.4E-01	1.0E+00
IFNK	9.4E-01	1.0E+00	6.0E-01	1.0E+00
IFT74	9.4E-01	1.0E+00	9.0E-02	1.0E+00
IGF1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
IGF1R	9.2E-01	1.0E+00	4.7E-01	1.0E+00
IGHMBP2	1.9E-01	1.0E+00	4.5E-01	1.0E+00
IKBKAP	9.7E-02	1.0E+00	9.3E-01	1.0E+00
IKBKE	2.9E-01	1.0E+00	1.4E-01	1.0E+00
IKBKG	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ILF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ILF3	6.6E-01	1.0E+00	7.7E-01	1.0E+00
ILK	3.2E-01	1.0E+00	9.7E-01	1.0E+00
IMMT	6.9E-01	1.0E+00	8.3E-02	1.0E+00
INSR	3.6E-01	1.0E+00	5.0E-01	1.0E+00
IRS1	1.1E-01	1.0E+00	9.0E-01	1.0E+00
ITGB1	7.3E-01	1.0E+00	1.7E-01	1.0E+00
ITM2B	7.2E-01	1.0E+00	7.5E-01	1.0E+00
ITPR1	3.4E-01	1.0E+00	5.0E-01	1.0E+00
ITPR2	1.2E-01	1.0E+00	7.0E-01	1.0E+00
JAK2	4.1E-01	1.0E+00	1.8E-01	1.0E+00
JPH2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
JPH3	1.0E+00	1.0E+00	1.0E+00	1.0E+00

JUN	1.0E+00	1.0E+00	1.0E+00	1.0E+00
KALRN	7.3E-01	1.0E+00	5.2E-01	1.0E+00
KARS	8.2E-01	1.0E+00	3.6E-01	1.0E+00
KDR	1.3E-01	1.0E+00	3.2E-01	1.0E+00
KHDRBS1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
KIF11	1.5E-01	1.0E+00	4.7E-01	1.0E+00
KIF1B	1.7E-01	1.0E+00	1.0E-01	1.0E+00
KIF3B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
KIF5A	6.2E-01	1.0E+00	2.4E-01	1.0E+00
KIF5B	1.6E-01	1.0E+00	1.0E+00	1.0E+00
KIFAP3	9.9E-01	1.0E+00	1.0E+00	1.0E+00
KIT	1.4E-01	1.0E+00	4.1E-02	1.0E+00
KPNA2	1.2E-01	1.0E+00	7.1E-02	1.0E+00
KRAS	1.0E+00	1.0E+00	1.0E+00	1.0E+00
L1CAM	7.4E-01	1.0E+00	4.2E-01	1.0E+00
LCK	1.0E+00	1.0E+00	1.0E+00	1.0E+00
LDHA	7.3E-01	1.0E+00	3.1E-01	1.0E+00
LMAN1	2.5E-01	1.0E+00	1.0E+00	1.0E+00
LMNA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
LMNB1	7.9E-01	1.0E+00	8.1E-02	1.0E+00
LRPPRC	3.3E-02	1.0E+00	5.2E-01	1.0E+00
LRRK2	4.7E-01	1.0E+00	9.3E-01	1.0E+00
LYAR	4.1E-01	1.0E+00	5.8E-01	1.0E+00
MAF	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MALT1	3.6E-01	1.0E+00	4.7E-01	1.0E+00
MAP1B	4.2E-01	1.0E+00	3.6E-01	1.0E+00
MAP2K1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAP2K3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAP2K6	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAP3K10	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAP3K11	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAP3K5	7.5E-01	1.0E+00	5.9E-01	1.0E+00
MAP3K7	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAPK1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAPK10	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAPK14	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAPK8	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAPK9	3.5E-01	1.0E+00	1.0E+00	1.0E+00
MAPRE1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAPT	5.5E-01	1.0E+00	9.9E-01	1.0E+00
MBD2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MCC	9.4E-01	1.0E+00	6.5E-01	1.0E+00
MCM3	9.6E-01	1.0E+00	9.9E-01	1.0E+00
MDM2	7.0E-02	1.0E+00	8.7E-01	1.0E+00
MECP2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MED25	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MEF2C	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MLH1	3.4E-01	1.0E+00	8.2E-01	1.0E+00

MMP2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MOB3B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MOV10	9.7E-01	1.0E+00	3.7E-01	1.0E+00
MOV10L1	4.5E-01	1.0E+00	4.1E-02	1.0E+00
MPHOSPH9	2.9E-01	1.0E+00	5.8E-01	1.0E+00
MPZ	4.7E-01	1.0E+00	2.3E-01	1.0E+00
MRE11A	5.0E-01	1.0E+00	6.0E-01	1.0E+00
MSH2	2.0E-01	1.0E+00	1.0E+00	1.0E+00
MSI1	5.8E-01	1.0E+00	1.9E-02	1.0E+00
MSI2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MSN	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MTNR1A	4.0E-01	1.0E+00	6.4E-01	1.0E+00
MTOR	1.6E-01	1.0E+00	7.7E-01	1.0E+00
MYC	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MYCBP2	1.2E-01	1.0E+00	9.4E-01	1.0E+00
MYH11	6.9E-01	1.0E+00	4.2E-01	1.0E+00
MYH9	2.1E-01	1.0E+00	9.5E-01	1.0E+00
MYO1E	7.1E-01	1.0E+00	2.0E-01	1.0E+00
MYO6	2.1E-01	1.0E+00	7.7E-01	1.0E+00
MYOC	9.8E-02	1.0E+00	1.3E-01	1.0E+00
NCBP2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NCK1	1.4E-01	1.0E+00	8.8E-01	1.0E+00
NCL	2.8E-01	1.0E+00	3.4E-01	1.0E+00
NCOA2	3.8E-01	1.0E+00	1.2E-01	1.0E+00
NCOR1	6.7E-01	1.0E+00	3.9E-01	1.0E+00
NDC80	5.1E-01	1.0E+00	6.6E-02	1.0E+00
NDE1	6.8E-01	1.0E+00	8.6E-01	1.0E+00
NDEL1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NDN	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NDRG1	2.6E-01	1.0E+00	6.5E-01	1.0E+00
NEDD1	4.4E-01	1.0E+00	1.0E+00	1.0E+00
NEDD4L	7.3E-01	1.0E+00	1.1E-01	1.0E+00
NEDD9	7.9E-01	1.0E+00	9.4E-01	1.0E+00
NEFH	6.2E-02	1.0E+00	7.1E-01	1.0E+00
NEFL	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NEFM	1.9E-01	1.0E+00	6.9E-01	1.0E+00
NEK2	4.0E-01	1.0E+00	1.3E-01	1.0E+00
NF1	6.7E-01	1.0E+00	2.5E-01	1.0E+00
NF2	9.9E-01	1.0E+00	1.0E+00	1.0E+00
NFKB1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NFKB2	7.5E-01	1.0E+00	4.6E-01	1.0E+00
NFKBIA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NFX1	7.0E-01	1.0E+00	5.6E-01	1.0E+00
NGF	5.4E-01	1.0E+00	2.0E-01	1.0E+00
NIPA1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NOL6	2.6E-01	1.0E+00	6.1E-01	1.0E+00
NOS1	6.4E-01	1.0E+00	4.4E-01	1.0E+00
NOS2	2.8E-01	1.0E+00	7.3E-01	1.0E+00

NOS3	2.2E-01	1.0E+00	5.8E-02	1.0E+00
NOX1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NPM1	4.4E-01	1.0E+00	1.0E+00	1.0E+00
NR2F2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NR3C1	3.2E-01	1.0E+00	3.9E-01	1.0E+00
NRAS	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NRIP1	1.9E-01	1.0E+00	4.6E-01	1.0E+00
NTRK1	3.3E-03	3.7E-01	9.1E-01	1.0E+00
NTRK2	2.8E-01	1.0E+00	7.6E-01	1.0E+00
NUP98	8.2E-01	1.0E+00	3.8E-01	1.0E+00
NXF1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
OCRL	8.5E-01	1.0E+00	1.0E+00	1.0E+00
OFD1	9.9E-01	1.0E+00	1.0E+00	1.0E+00
OPTN	1.0E+00	1.0E+00	1.0E+00	1.0E+00
OSBPL9	8.5E-02	1.0E+00	1.0E+00	1.0E+00
OSTF1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PABPC1	1.3E-02	1.0E+00	1.0E+00	1.0E+00
PABPC4	7.1E-01	1.0E+00	1.0E+00	1.0E+00
PABPN1	1.0E+00	1.0E+00	3.3E-01	1.0E+00
PAFAH1B1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PARD3	4.1E-01	1.0E+00	6.5E-01	1.0E+00
PARK2	6.1E-01	1.0E+00	8.7E-02	1.0E+00
PARK7	6.4E-01	1.0E+00	6.0E-01	1.0E+00
PARP1	1.6E-01	1.0E+00	6.3E-02	1.0E+00
PCBP1	1.5E-01	1.0E+00	1.0E+00	1.0E+00
PCBP2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PCM1	1.9E-01	1.0E+00	9.7E-01	1.0E+00
PCNA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PCNT	7.9E-01	1.0E+00	1.7E-01	1.0E+00
PDGFRA	8.0E-01	1.0E+00	1.1E-01	1.0E+00
PDS5B	4.1E-01	1.0E+00	1.5E-01	1.0E+00
PHB2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PICK1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PIK3R1	5.3E-01	1.0E+00	9.2E-02	1.0E+00
PITPNM1	6.8E-01	1.0E+00	5.7E-01	1.0E+00
PITPNM2	3.9E-01	1.0E+00	4.6E-01	1.0E+00
PITPNM3	1.1E-01	1.0E+00	6.6E-01	1.0E+00
PKD2	7.3E-02	1.0E+00	3.8E-01	1.0E+00
PLAA	6.8E-01	1.0E+00	5.5E-01	1.0E+00
PLEKHF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PLEKHG4	4.3E-01	1.0E+00	5.4E-01	1.0E+00
PLK1	7.7E-01	1.0E+00	1.8E-01	1.0E+00
PLK4	2.2E-02	1.0E+00	3.7E-01	1.0E+00
POLI	9.9E-01	1.0E+00	7.2E-01	1.0E+00
POLR2A	2.2E-01	1.0E+00	4.5E-01	1.0E+00
POLR2B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PON1	1.7E-01	1.0E+00	3.5E-01	1.0E+00
PON2	1.0E+00	1.0E+00	6.7E-03	1.0E+00

PON3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PPARG	8.2E-02	1.0E+00	7.2E-02	1.0E+00
PPARGC1A	9.3E-01	1.0E+00	6.3E-01	1.0E+00
PPM1L	7.1E-01	1.0E+00	4.2E-01	1.0E+00
PPP1CA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PPP1CC	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PPP2CA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PPP2R1A	1.0E+00	1.0E+00	3.2E-01	1.0E+00
PPP2R2B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PPP3CA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PPP3R1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PQBP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PRDX1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PRKACA	1.1E-01	1.0E+00	1.0E+00	1.0E+00
PRKAR2B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PRKCA	9.3E-01	1.0E+00	2.7E-02	1.0E+00
PRKCB	1.6E-01	1.0E+00	6.2E-01	1.0E+00
PRKCD	1.5E-01	1.0E+00	9.2E-01	1.0E+00
PRKCE	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PRKCG	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PRKDC	2.2E-01	1.0E+00	1.9E-01	1.0E+00
PRLR	7.1E-01	1.0E+00	1.1E-03	2.6E-01
PRPH	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PRSS3	2.5E-01	1.0E+00	5.6E-01	1.0E+00
PRUNE2	4.5E-01	1.0E+00	1.0E-01	1.0E+00
PSEN1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PSEN2	6.8E-01	1.0E+00	7.5E-01	1.0E+00
PSMF1	4.7E-01	1.0E+00	7.8E-01	1.0E+00
PTBP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PTBP2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PTEN	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PTK2B	2.9E-01	1.0E+00	4.1E-01	1.0E+00
PTPN1	9.1E-01	1.0E+00	1.0E+00	1.0E+00
PTPN11	1.0E+00	1.0E+00	1.0E+00	1.0E+00
PUF60	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RAB11B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RAB11FIP3	6.0E-02	1.0E+00	1.0E+00	1.0E+00
RAB14	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RAB5A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RAB5B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RAB5C	1.4E-01	1.0E+00	5.9E-01	1.0E+00
RAB7A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RAB8A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RABGEF1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RAC1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RAC2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RACGAP1	1.6E-01	1.0E+00	1.0E+00	1.0E+00
RAD50	4.6E-01	1.0E+00	5.4E-01	1.0E+00

RAF1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RALBP1	2.9E-01	1.0E+00	1.1E-02	1.0E+00
RALY	8.5E-01	1.0E+00	3.4E-01	1.0E+00
RAN	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RAP1A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RASA1	3.0E-01	1.0E+00	7.5E-01	1.0E+00
RB1	1.6E-01	1.0E+00	1.0E+00	1.0E+00
RBCK1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RBFOX1	5.7E-01	1.0E+00	2.9E-01	1.0E+00
RBM10	1.2E-01	1.0E+00	1.0E+00	1.0E+00
RBM14	6.5E-01	1.0E+00	2.1E-01	1.0E+00
RBM3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RBM39	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RBM4B	4.3E-01	1.0E+00	1.0E+00	1.0E+00
RBM5	7.2E-01	1.0E+00	1.0E+00	1.0E+00
RBM8A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RBMS1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RBMS2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RBMX	3.7E-01	1.0E+00	1.0E+00	1.0E+00
RBMY1A1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RCBTB2	8.6E-01	1.0E+00	5.7E-01	1.0E+00
RCC1	3.8E-01	1.0E+00	4.2E-02	1.0E+00
RELA	4.2E-01	1.0E+00	3.2E-01	1.0E+00
RHOA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RIN2	8.9E-01	1.0E+00	2.2E-01	1.0E+00
RIN3	4.5E-02	1.0E+00	4.7E-01	1.0E+00
RIPK1	3.2E-01	1.0E+00	7.6E-02	1.0E+00
RNASE1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RNASE2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RNASE4	7.1E-01	1.0E+00	1.4E-01	1.0E+00
RNASE6	3.8E-01	1.0E+00	5.4E-01	1.0E+00
RNF11	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RNF19A	4.2E-01	1.0E+00	9.7E-01	1.0E+00
RNH1	5.9E-01	1.0E+00	3.3E-01	1.0E+00
RNPS1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ROCK1	7.2E-01	1.0E+00	9.4E-01	1.0E+00
RORB	4.2E-01	1.0E+00	1.0E+00	1.0E+00
RPA1	7.0E-02	1.0E+00	1.0E+00	1.0E+00
RPGR	8.1E-01	1.0E+00	8.7E-01	1.0E+00
RPL10A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RPL13	6.6E-01	1.0E+00	1.0E+00	1.0E+00
RPL18A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RPL4	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RPL5	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RPL6	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RPL7	8.8E-02	1.0E+00	8.6E-01	1.0E+00
RPN1	3.4E-01	1.0E+00	1.0E+00	1.0E+00
RPS16	1.0E+00	1.0E+00	1.0E+00	1.0E+00

RPS3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RPS5	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RPS6	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RSPH1	5.1E-01	1.0E+00	4.4E-01	1.0E+00
RTN4	8.4E-01	1.0E+00	8.0E-01	1.0E+00
RUNX1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RUVBL1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RUVBL2	4.2E-01	1.0E+00	7.3E-01	1.0E+00
RXRA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SAFB	1.7E-01	1.0E+00	6.4E-01	1.0E+00
SAFB2	8.4E-01	1.0E+00	8.1E-02	1.0E+00
SAR1A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SART3	2.8E-01	1.0E+00	8.0E-01	1.0E+00
SCHIP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SCNN1A	3.0E-01	1.0E+00	8.1E-01	1.0E+00
SCP2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SERGEF	6.7E-01	1.0E+00	9.0E-01	1.0E+00
SETX	4.2E-01	1.0E+00	2.6E-01	1.0E+00
SF1	3.0E-01	1.0E+00	1.0E+00	1.0E+00
SF3A2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SF3B1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SF3B4	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SFPQ	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SIRT1	8.2E-01	1.0E+00	8.8E-01	1.0E+00
SLC16A2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SLC25A6	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SLC2A4	4.6E-01	1.0E+00	2.9E-01	1.0E+00
SMAD1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SMAD3	5.5E-01	1.0E+00	1.6E-01	1.0E+00
SMAD4	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SMARCA4	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SMC1A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SMC3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SMN1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SMN2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SMU1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SNAP23	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SNAP25	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SNAP29	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SNCA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SNRPA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SNRPB2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SOD1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SOD2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SOS1	2.1E-01	1.0E+00	1.5E-01	1.0E+00
SOX10	3.7E-01	1.0E+00	9.3E-01	1.0E+00
SP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SPAST	9.2E-01	1.0E+00	1.6E-01	1.0E+00

SPEN	6.8E-01	1.0E+00	7.8E-01	1.0E+00
SPG11	7.5E-01	1.0E+00	8.6E-01	1.0E+00
SPG20	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SPG7	4.2E-01	1.0E+00	4.1E-01	1.0E+00
SPI1	2.2E-01	1.0E+00	2.5E-01	1.0E+00
SPTBN2	4.0E-01	1.0E+00	7.5E-01	1.0E+00
SRC	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SREBF1	1.8E-01	1.0E+00	4.9E-01	1.0E+00
SREK1	8.4E-01	1.0E+00	1.0E+00	1.0E+00
SRPK1	5.2E-01	1.0E+00	7.8E-01	1.0E+00
SRRM1	5.4E-01	1.0E+00	3.7E-01	1.0E+00
SRSF1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SRSF10	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SRSF11	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SRSF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SRSF3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SRSF4	3.5E-01	1.0E+00	9.1E-02	1.0E+00
SRSF5	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SRSF6	1.6E-01	1.0E+00	6.2E-01	1.0E+00
SRSF7	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SRSF9	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SRXN1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SSB	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SSR4	1.0E+00	1.0E+00	1.0E+00	1.0E+00
STAT1	3.4E-01	1.0E+00	6.2E-01	1.0E+00
STAT3	8.3E-03	1.0E+00	6.4E-01	1.0E+00
STK11	1.0E+00	1.0E+00	1.0E+00	1.0E+00
STK24	6.2E-02	1.0E+00	2.8E-01	1.0E+00
STMN1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
STX1A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
STX1B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
STX2	5.4E-01	1.0E+00	3.3E-01	1.0E+00
STX4	1.0E+00	1.0E+00	1.0E+00	1.0E+00
STX5	3.1E-01	1.0E+00	3.4E-01	1.0E+00
SYK	1.1E-01	1.0E+00	6.6E-02	1.0E+00
SYNCRIP	1.9E-01	1.0E+00	1.6E-01	1.0E+00
TAB2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TAF15	3.5E-01	1.0E+00	6.2E-01	1.0E+00
TAF1L	1.7E-03	3.1E-01	9.6E-01	1.0E+00
TARDBP	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TAX1BP1	1.0E+00	1.0E+00	1.7E-01	1.0E+00
TBC1D17	5.8E-01	1.0E+00	1.0E+00	1.0E+00
TBCE	2.7E-01	1.0E+00	9.1E-01	1.0E+00
TBK1	2.2E-01	1.0E+00	8.3E-01	1.0E+00
TBP	3.2E-01	1.0E+00	1.0E+00	1.0E+00
TCERG1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TCF4	6.7E-03	1.0E+00	7.0E-01	1.0E+00
TDGF1	7.1E-01	1.0E+00	1.1E-01	1.0E+00

TEK	4.4E-01	1.0E+00	2.7E-01	1.0E+00
TF	5.7E-01	1.0E+00	8.9E-02	1.0E+00
TFRC	3.0E-01	1.0E+00	4.2E-01	1.0E+00
TGFB1	3.5E-01	1.0E+00	6.2E-01	1.0E+00
TGFBR1	6.8E-01	1.0E+00	7.1E-01	1.0E+00
THRA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TIAL1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TLE1	3.5E-01	1.0E+00	7.4E-01	1.0E+00
TNF	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TNFAIP3	8.7E-02	1.0E+00	3.3E-01	1.0E+00
TNFRSF1A	5.4E-01	1.0E+00	1.0E+00	1.0E+00
TNFRSF1B	8.5E-02	1.0E+00	6.0E-02	1.0E+00
TNFSF8	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TOP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TOP2A	1.8E-01	1.0E+00	2.5E-01	1.0E+00
TOPORS	2.7E-01	1.0E+00	3.3E-01	1.0E+00
TP53	7.9E-01	1.0E+00	1.0E+00	1.0E+00
TP63	2.4E-01	1.0E+00	1.0E+00	1.0E+00
TP73	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TPM3	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TRA2A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TRA2B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TRAF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TRAF3	3.1E-01	1.0E+00	5.0E-01	1.0E+00
TRAF6	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TRIB3	9.4E-01	1.0E+00	8.3E-01	1.0E+00
TRIM28	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TRIP13	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TRPV4	4.9E-01	1.0E+00	2.3E-01	1.0E+00
TSC1	4.9E-01	1.0E+00	3.5E-01	1.0E+00
TSG101	1.8E-01	1.0E+00	1.0E+00	1.0E+00
TUBA1A	8.3E-01	1.0E+00	8.9E-01	1.0E+00
TUBA4A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TUBB	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TUBB4A	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TUBB4B	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TUBG1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TXN	1.0E+00	1.0E+00	1.0E+00	1.0E+00
U2AF1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
U2AF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
UBA1	8.0E-01	1.0E+00	1.0E+00	1.0E+00
UBB	1.0E+00	1.0E+00	1.0E+00	1.0E+00
UBC	1.0E+00	1.0E+00	1.0E+00	1.0E+00
UBE2I	3.2E-01	1.0E+00	5.7E-01	1.0E+00
UBE2R2	1.7E-01	1.0E+00	1.0E+00	1.0E+00
UCHL1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
UNC13A	3.7E-01	1.0E+00	3.8E-01	1.0E+00
UPF1	1.0E+00	1.0E+00	1.0E+00	1.0E+00

USF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
USO1	7.6E-01	1.0E+00	7.6E-01	1.0E+00
USP20	5.6E-01	1.0E+00	4.2E-01	1.0E+00
VAMP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
VAMP2	1.0E+00	1.0E+00	1.0E+00	1.0E+00
VAMP8	1.0E+00	1.0E+00	1.0E+00	1.0E+00
VAPA	4.3E-01	1.0E+00	7.3E-01	1.0E+00
VAPB	5.4E-01	1.0E+00	8.3E-01	1.0E+00
VAV1	9.6E-01	1.0E+00	7.4E-01	1.0E+00
VCAN	9.7E-01	1.0E+00	5.7E-01	1.0E+00
VCL	3.5E-01	1.0E+00	5.0E-01	1.0E+00
VCP	1.0E+00	1.0E+00	1.0E+00	1.0E+00
VDAC1	1.0E+00	1.0E+00	1.0E+00	1.0E+00
VDR	6.9E-01	1.0E+00	7.2E-01	1.0E+00
VEGFA	1.0E+00	1.0E+00	1.0E+00	1.0E+00
VHL	1.0E+00	1.0E+00	1.0E+00	1.0E+00
VIM	1.0E+00	1.0E+00	1.0E+00	1.0E+00
VPS13A	6.7E-01	1.0E+00	7.2E-01	1.0E+00
WBSCR16	1.0E+00	1.0E+00	1.0E+00	1.0E+00
WDR20	6.8E-01	1.0E+00	7.1E-01	1.0E+00
WDR36	6.8E-01	1.0E+00	6.5E-01	1.0E+00
WNK1	1.9E-01	1.0E+00	2.2E-01	1.0E+00
WWP1	7.7E-01	1.0E+00	5.0E-01	1.0E+00
XBP1	4.0E-01	1.0E+00	4.2E-01	1.0E+00
Xg-ALS2	6.9E-01	1.0E+00	9.3E-01	1.0E+00
Xg-ANG	1.9E-01	1.0E+00	3.7E-01	1.0E+00
Xg-DCTN1	8.5E-01	1.0E+00	9.2E-01	1.0E+00
Xg-FUS	8.7E-01	1.0E+00	7.0E-01	1.0E+00
Xg-Hom	6.5E-02	1.0E+00	8.0E-01	1.0E+00
Xg-OPTN	9.3E-01	1.0E+00	4.2E-01	1.0E+00
Xg-SETX	1.2E-01	1.0E+00	4.1E-01	1.0E+00
Xg-SOD1	3.2E-01	1.0E+00	5.0E-01	1.0E+00
Xg-tALSp1	3.0E-01	1.0E+00	1.0E+00	1.0E+00
Xg-tALSp2	6.4E-01	1.0E+00	9.3E-01	1.0E+00
Xg-TARDBP	2.5E-01	1.0E+00	5.6E-02	1.0E+00
Xg-VAPB	4.8E-01	1.0E+00	8.0E-01	1.0E+00
XPO1	4.3E-01	1.0E+00	2.3E-01	1.0E+00
XRCC5	9.4E-01	1.0E+00	1.0E+00	1.0E+00
XRCC6	7.3E-01	1.0E+00	3.2E-01	1.0E+00
XRN2	1.4E-01	1.0E+00	6.6E-01	1.0E+00
YARS	6.5E-01	1.0E+00	5.4E-01	1.0E+00
YBX1	2.9E-01	1.0E+00	1.0E+00	1.0E+00
YWHAB	1.0E+00	1.0E+00	1.0E+00	1.0E+00
YWHAE	1.0E+00	1.0E+00	1.0E+00	1.0E+00
YWHAG	1.0E+00	1.0E+00	1.0E+00	1.0E+00
YWHAQ	1.0E+00	1.0E+00	1.0E+00	1.0E+00
YWHAZ	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ZBTB16	1.0E+00	1.0E+00	1.0E+00	1.0E+00

ZFAND5	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ZMYM2	7.6E-01	1.0E+00	5.6E-01	1.0E+00

The EMP1 columns denote the minimum p-value observed per grouping following correction for the number of statistical tests performed per grouping. In the case of tests of association with patient phenotype, this includes correction for the number of phenotypic categories analyzed. The EMP2 columns denote the fully corrected p-values, where correction for the number of groupings analyzed has also been performed. Intervals denoted with the prefix "Xg" represent gene sets. Xg-Hom includes candidate genes reported to exhibit protein sequence similarity with a known ALS gene in section 3. Xg-tALSp1 and Xg-tALSp2 denote genes encoding proteins observed to occur within 1 or 2 interactions of an ALS protein in the molecular interaction networks described in section 3. All other gene sets represent genes encoding proteins observed to interact directly with a known ALS protein in section 3.

Appendix 2 - Example Scripts

Contents:

Script S1: PageRank-Priors.r

Script S2: SequenceAlignment.py

Script S1: PageRank-Priors.r

```
# Parse arguments from commandline
args=commandArgs(T)

seeds.path=args[1]
network.path=args[2]
B=as.numeric(args[3])
nt=as.numeric(args[4])
iter=as.numeric(args[5])
perm=as.numeric(args[6])
out.path=args[7]

# Load libraries for multiprocessing
library(doMC)
registerDoMC(nt)


# Load data
S=scan(seeds.path,comment.char="#",what="character")
N=as.matrix(read.table(network.path,as.is=T,h=F,sep="\t"))

# Remove redundant and self interactions from network
N=unique(N)
N=N[N[,1]!=N[,2],]

# Function to return interactors for a single node
getInteractors=function(query,N)
        {return(unique(c(N[N[,2]==query,1],N[N[,1]==query,2])))}

# Function to generate hyperlink matrix
getH=function(N)
        {
        # Features in output matrix listed columnwise
        nodes=unique(as.vector(N))
        K=table(N);K=as.vector(K[match(nodes,names(K))])
        H=c()
        for (i in nodes)
                {H=cbind(H,nodes %in% getInteractors(i,N))}
        H=H*K^-1
        colnames(H)=nodes
        rownames(H)=nodes
        return(H)
        }

getImportance=function(H,B,S,iter,out.path)
        {
        # Calculate relative importance - observed data
        pv=(colnames(H)%in%S)*length(S)^-1
        a=1-B
        Bpv=B*pv
        Ii=pv
        dev=c()
        for (i in 1:iter)
                {
                I0=Ii
                Ii=a*(colSums(H*Ii))+Bpv
                dev[i]=mean((Ii-I0)^2)
                }
        tiff(out.path)
        plot(dev,xlab="Iteration",ylab="Mean Square Deviation",main="")
        dev.off()
```

```
        Ii[pv>0]=NA
        return(Ii)
        }

getImportance.pb=function(H,B,S,iter,pb,perm)
        {
        # Calculate relative importance - permutation
        setTxtProgressBar(pb, perm) # update progress bar
        pv=(colnames(H)%in%S)*length(S)^-1
        a=1-B
        Bpv=B*pv
        Ii=pv
        for (i in 1:iter)
                {Ii=a*(colSums(H*Ii))+Bpv}
        Ii[pv>0]=NA
        return(Ii)
        }

# Function to perform PageRank-Priors analysis
pageRankPriors=function(out.path,S,N,B=0.3,perm=0,iter=100)
        # Input paramaters:
        # S     Vector of seed nodes
        # N     Matrix of binary interactions
        # B     Back probability
        # perm  Number of permutations to perform to assess ranking
significance.  Each permutation will rank nodes by their relative
importance to a vector of randomly selected set nodes. The length of
this vector will equal that of the input vector S.
        {
        # Establish Hyperlink matrix
        H=getH(N)
        write.table(H,file=paste(out.path,".H.txt",sep=""),quote=F,sep="\
t")

        # Analysis
        I=getImportance(H,B,S,iter=iter,out.path=paste(out.path,".tiff",s
ep=""))
        write.table(I,file=paste(out.path,".results.txt",sep=""),quote=F,
sep="\t")

        # Permutations
        if (perm>0)
                {
                pb = txtProgressBar(min = 0, max = perm, style = 3)
                null=foreach(i=1:perm,.combine="rbind") %dopar%

        {getImportance.pb(H,B,sample(colnames(H),length(S)),iter=iter,pb,
i)}
                }
        write.table(null,file=paste(out.path,".null.txt",sep=""),quote=F,
sep="\t")

        return(rbind(I,null))
        }


X=pageRankPriors(out.path,S,N,B=0.3,perm=perm,iter=iter)
```

Script S2: SequenceAlignment.py

```
#!/usr/bin/env python

# Prints bash scripts to process a batch of sequencing data (alignment
and alignment QC)
# The co-odrinated execution of these bash scripts on a local server was
then performed using a master python script
# It was decided to print intermediate bash scripts as this provided a
simple mechanism to stop, update and restart analyses and to pause and
skip over lanes running into problems with memmory allocation.

# Import modules
import os, sys, threading, re, time

# Enviromental variables
os.environ['threads']='3'
os.environ['RAM']='Xmx20g'
os.environ['ReadGroup']='/media/kevin/Passport/Reference/CandidateGeneRe
sequencing_corrected_further.3.rg'
os.environ['bwa']='/kennak/bwa-0.7.4/bwa'
os.environ['GATK']='/media/kevin/Passport/TargetedResequencing_Reanalysi
s/Scripts/GenomeAnalysisTK-2.4-9-g532efad/GenomeAnalysisTK.jar'
os.environ['picard']='/media/kevin/Passport/TargetedResequencing_Reanaly
sis/Scripts/picard-tools-1.60/picard-tools-1.60'
os.environ['addRG']='/media/kevin/Passport/TargetedResequencing_Reanalys
is/Scripts/addRG.py'
os.environ['ref']='/media/kevin/Passport/Reference/human_g1k_v37_decoy.f
asta'
os.environ['dbSNP']='/media/kevin/Passport/Reference/00-All.LA.vcf'
os.environ['indels']='/media/kevin/Passport/Reference/1000G_phase1.indel
s.b37.vcf'
os.environ['indels2']='/media/kevin/Passport/Reference/Mills_and_1000G_g
old_standard.indels.b37.vcf'
os.environ['ESP6500SI']='/media/kevin/Passport/Reference/ESP6500SI.LA.vc
f'
os.environ['1000Genomes']='/media/kevin/Passport/Reference/ALL.wgs.phase
1_release_v3.20101123.snps_indels_sv.sites.LA.vcf'

# Load input files
fastq=[i.strip() for i in open(sys.argv[1],'r').readlines()] #       File
containing list of fastq files to analyze
lanes=[i.strip() for i in open(sys.argv[2],'r').readlines()] #       File
outlining lane ID assigned to each fastq file
script_outdir=sys.argv[3]                                    #  Directory
to print bash scripts
file_outdir=sys.argv[4]                                              #
Directory to write output files.


# Processing Functions

def align(fastq,lane,bam):

        # Align reads with bwa, add read group annotation using a custom
python script, convert to binary format and sort using samtools
        # M option engaged for picard compatibility; http://bio-
bwa.sourceforge.net/bwa.shtml
        cmd="(%s mem -CMt%s %s %s | python %s %s %s | samtools view -uS -
|          samtools          sort          -          %s)          &>>
%s"%(os.environ['bwa'],os.environ['threads'],os.environ['ref'],fastq,os.
```

```
environ['addRG'],lane,os.environ['ReadGroup'],os.path.splitext(bam)[0],o
s.environ['output_log'])
        os.system('echo "%s" >> "%s"'%(cmd,os.environ["cmd_log"]))

def index_bam(bam):
        # Index bam file
        cmd='samtools index %s &>> %s'%(bam,os.environ['output_log'])
        os.system('echo "%s" >> "%s"'%(cmd,os.environ["cmd_log"]))

def MarkDuplicates(bam):
        # Remove PCR duplicates
        out=bam.replace(".bam",".dedup.bam")
        cmd='java   -jar   -%s   %s/MarkDuplicates.jar   I="%s"   O="%s"
METRICS_FILE=%s         REMOVE_DUPLICATES=true         ASSUME_SORTED=true
VALIDATION_STRINGENCY=SILENT              VERBOSITY=ERROR              &>>
%s'%(os.environ['RAM'],os.environ["picard"],bam,out,bam.replace(".bam","
.pcr"),os.environ['output_log'])
        os.system('echo "%s" >> "%s"'%(cmd,os.environ["cmd_log"]))
        return out

def merge(bams,merged):
        # Merge bam files
        bams=map(lambda bam: 'I="'+bam+'"',bams)
        bams=" ".join(bams)
        cmd='java    -jar    -%s    %s/MergeSamFiles.jar    %s    O="%s"
USE_THREADING=true   VALIDATION_STRINGENCY=SILENT   VERBOSITY=ERROR   &>>
%s'%(os.environ['RAM'],os.environ["picard"],bams,merged,os.environ['outp
ut_log'])
        os.system('echo "%s" >> "%s"'%(cmd,os.environ["cmd_log"]))

def RealignerTargetCreator(bam):
        # Identify loci for local indel realignment
        out=bam.replace(".bam",".indel.intervals")
        cmd='java   -jar   -%s   %s   -nt   %s   -I   "%s"   -R   "%s"   -T
RealignerTargetCreator -l "INFO" --known %s --known %s -o "%s" &>>
%s'%(os.environ['RAM'],os.environ["GATK"],os.environ['threads'],bam,os.e
nviron['ref'],os.environ['indels'],os.environ['indels2'],out,os.environ[
'output_log'])
        os.system('echo \'%s\' >> "%s"'%(cmd,os.environ["cmd_log"]))
        return out

def IndelRealigner(bam,indel_intervals,tgt=None):
        # Perform local indel realignment
        out=bam.replace(".bam",".ra.bam")
        cmd='java -jar -%s %s -I "%s" -R "%s" -T IndelRealigner -l "INFO"
-targetIntervals   "%s"   -known   "%s"   -known   "%s"   -o   "%s"   &>>
%s'%(os.environ['RAM'],os.environ["GATK"],bam,os.environ['ref'],indel_in
tervals,os.environ['indels'],os.environ['indels2'],out,os.environ['outpu
t_log'])
        if tgt!=None:
                tgt=map(lambda i:"-L \""+i+'"',tgt)
                tgt=" ".join(tgt)
                cmd+=" "+tgt
        os.system('echo \'%s\' >> "%s"'%(cmd,os.environ["cmd_log"]))
        return out

def BaseRecalibration(bam):
        # Base quality score recalibration
        out=bam.replace(".bam",".recal")
        intermediate=bam.replace(".bam",".int.recal")
        cmd='java -jar -%s %s -nct %s -I "%s" -R "%s" -T BaseRecalibrator
-intermediate  %s  --default_platform  illumina  -l  "INFO"  -o  "%s"  --
knownSites "%s" --knownSites "%s" --knownSites "%s" --knownSites "%s" --
```

```
knownSites "%s" --plot_pdf_file %s &>>
%s'%(os.environ['RAM'],os.environ["GATK"],os.environ['threads'],bam,os.e
nviron['ref'],intermediate,out,os.environ["dbSNP"],os.environ['indels'],
os.environ['indels2'],os.environ['1000Genomes'],os.environ['ESP6500SI'],
out+".pdf",os.environ['output_log'])
        os.system('echo \'%s\' >> "%s"'%(cmd,os.environ["cmd_log"]))
        return out

def PrintReads(bam,out,recal=None):
        # Incorporate recalibrated base quality scores
        cmd='java -jar -%s %s -nct %s -I "%s" -R "%s" -T PrintReads -l
"INFO" -o "%s" &>>
%s'%(os.environ['RAM'],os.environ["GATK"],os.environ['threads'],bam,os.e
nviron['ref'],out,os.environ['output_log'])
        if recal!=None:
                cmd+=" -BQSR \""+recal+"\""
        os.system('echo \'%s\' >> "%s"'%(cmd,os.environ["cmd_log"]))
        return out

def main(fastq,lane,outdir):
        # Print bash script for a given lane of sequencing data
        os.system("echo > %s"%(os.environ["cmd_log"]))
        bam=outdir+"/"+re.split("\.",os.path.basename(fastq))[0]+".bam" #
Use original file name rather than lane name to cater for the
independent analysis of files relating to the same lane (PE data)
        os.environ['output_log']=outdir+"/"+re.split("\.",os.path.basenam
e(fastq))[0]+".log"
        align(fastq,lane,bam)
        index_bam(bam)
        bam=MarkDuplicates(bam)
        index_bam(bam)
        indel_intervals=RealignerTargetCreator(bam)
        bam=IndelRealigner(bam,indel_intervals,tgt=None)
        index_bam(bam)
        recal=BaseRecalibration(bam)
        bam=PrintReads(bam,bam.replace(".bam",".rc.bam"),recal)
        index_bam(bam)


# Print scripts for the analysis of a batch of sequencing data, one
script per lane
for i in range(len(lanes)):
        os.environ['cmd_log']=script_outdir+"/"+re.split("\.",os.path.bas
ename(fastq[i]))[0]+".sh"
        main(fastq[i],lanes[i],file_outdir)
```