Genetic heterogeneity in amyotrophic lateral sclerosis and related neurological disorders

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

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

2022

Mark Doherty

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, except where otherwise stated.

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: Mark

Date: 21/04/2022

List of abbreviations

ACMG	American College of Medical Genetics
AF	Allele frequency
ALS	Amyotrophic lateral sclerosis
AOO	Age of onset
AZD	Alzheimer’s disease
B	Benign
bp	Base pair
BQSR	Base quality score recalibration
bvFTD	Behavioural variant frontotemporal dementia
CBS	Corticobasal syndrome
CCHS	Congenital central hypoventilation syndrome
CI	Confidence interval
CNS	Central nervous syndrome
dbNSFP	Database for nonsynonymous SNPs' functional predictions
DNA	Deoxyribonucleic acid
DOC	Depth of coverage
DPR	Dipeptide repeat
EE	Epileptic encephalopathy
EIEE1	Epileptic encephalopathy, early infantile, 1
fALS	Familial amyotrophic lateral sclerosis
fFTD	Familial frontotemporal dementia
FTD	Frontotemporal dementia
FTV	Frameshift or truncating variant
FXTAS	Fragile X–associated tremor and ataxia syndrome
gnomAD	Genome Aggregation Database
GOF	Gain-of-function
GOI	Gene of interest
HD	Huntington’s disease
HGMD	Human Gene Mutation Database
HPC	High performance computer

HSP	Hereditary spastic paraplegia
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
IBMPFD	Inclusion body myopathy with early-onset Paget disease and frontotemporal dementia
ID	Intellectual disability
INDEL	Insertion / deletion
IRR	In-repeat read
kb	kilobase
LB	Likely benign
LMN	Lower motor neurone
LOF	Loss-of-function
LP	Likely pathogenic
MAF	Minor allele frequency
MDS	Multidimensional scaling
MND	Motor neurone disease
MR	Mendelian randomisation
mRNA	Messenger ribonucleic acid
NGS	Next-generation sequencing
OR	Odds ratio
P	Pathogenic
PCR	Polymerase chain reaction
PDC	Parkinsonism-dementia complex
PE	Paired-end
pext	Proportion expression across transcripts
pLI	probability of being loss-of-function intolerant
PLS	Primary lateral sclerosis
PMA	Progressive muscular atrophy
PNFA	Progressive nonfluent aphasia
PSP	Progressive supranuclear palsy
QC	Quality control
RAN	Repeat-associated non-AUG
RE	Repeat expansion
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid

ROI	Region of interest
rpPCR	Repeat primed polymerase chain reaction
sALS	Sporadic amyotrophic lateral sclerosis
SCA	Spinocerebellar ataxia
SD	Semantic dementia
sFTD	Sporadic frontotemporal dementia
SKAT	Sequence kernel association test
SMA	Spinal muscular atrophy
SNP	Single nucleotide polymorphism
SNV	Single-nucleotide variant
STR	Short tandem repeat
UK	United Kingdom
UMN	Upper motor neurone
USA	United States of America
UTR	Untranslated region
VOI	Variant of interest
VQSR	Variant quality score recalibration
VUS	Variant(s) of uncertain significance
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Acknowledgments

I would firstly like to thank the ALS, epilepsy, FTD and PLS patients, their carers and relatives, and the healthy volunteers who kindly gave their time and energy to make this research possible.

I have been very fortunate to have benefitted from two great supervisors. I am very grateful to Russell McLaughlin and Orla Hardiman not only for their expertise, but also their excellent support, mentorship and guidance throughout my PhD.

This research wouldn't have been possible without many collaborators, both inside and outside of Trinity College Dublin. Thanks to Gianpiero Cavalleri, Katherine Benson and Hamidah Ghani in RCSI, for facilitating the study of epilepsy and for their help and expertise in all things epilepsy related. The study of ALS in Cuba would not have been possible without Tatiana Zaldívar Vaillant in the National Institute of Neurology, Havana, Cuba. I am very grateful to Sean McGrath in the TCD HPC for his help and patience with hosting the journALS data browser. Thanks to Ammar Al-Chalabi and Ahmad Al Khleifat from King's College, London, for their input on discordant families. Thanks to Peter Bede, Eoin Finegan, Rangariroyashe Chipika and Taha Omer for all their expertise in PLS and FTD.

I have been incredibly lucky with the friends and colleagues I have worked with in the Complex Trait Genomics lab. Ross Byrne is a very kind and smart person who is always happy to share his (seemingly endless) knowledge. Jennifer Hengeveld is a great friend who knows a worrying amount about PCR. I have a lot to learn from her patience, hard work and resilience. I am also very grateful to Jenny for kindly sharing PCR genotyping results. Ciarán Kelly's first introduction to the lab was to read and reread hundreds of papers for the journALS study in this thesis, I think he has forgiven me for this and I'm happy that we're now friends. With Laura O'Briain in the lab the future of repeat expansion genomics is in safe hands. Thanks also to the great final year students Fiana, Laura, Ivan and Fedor, who I was fortunate enough to help with their final year projects, and who I learned a lot from.

I am grateful to the whole team in the Academic Unit of Neurology, whose work is the foundation for this research. Particular thanks to Mark Heverin and Alice Vajda, to Emmet

Costello for collecting DNA samples and to Marie Ryan for her work on discordant families. I have also been very lucky to work closely with my friends in the Bradley lab. Great advice, conversation, laughs and pints were never far off thanks to Andrew, Bruno, Conor, Emily, Eppie, Iseult, Kevin, Lara, Marta, Matthew, Mikkel, Pier, Valeria and Victoria. I am grateful to Frank Wellmer and Pat Ryan for kindly giving me my first experience of bioinformatics. Without the excellent support staff in the genetics department, Brenda, Dave, Rachel, Sue and Oanh, things would run a lot less smoothly and all of this would have taken a lot more time. I would like to give an honorary mention to tea, NPR tiny desk concerts and the squirrels who like to play outside my window, for helping me to stay positive. Thanks to Science Foundation Ireland and FutureNeuro for funding this research.

I would not have been able to achieve this, or anything, without the endless support and encouragement from my friends and family, particularly my parents Fred and Mary, and my siblings Emma, Paul and Stephen. Thanks to James, Robbie, Danny, Shane and Julie for always making me smile.

Finally, I have too much to thank Louisse Mirabueno for; not least for her contribution to this work by reading hundreds of papers. More importantly, she has always been willing to listen to my rants and rambles and has always helped to put things in perspective. I’m excited to spend the rest of our lives together.

Table of contents

Table of figures

Table of tables

Summary

The overarching aim of this thesis is to clarify and further our understanding of the genetic causes of amyotrophic lateral sclerosis (ALS) and related diseases. It is hoped that achieving this can help bring clarity to patients, relatives and carers by improving genetic counselling and aiding in the design of clinical trials by improving patient stratification based on genetic background.

In the first research chapter of this thesis, a meta-analysis of all genetic variants previously reported in ALS and frontotemporal dementia (FTD) patients is performed. 3,114 variants in 356 genes were identified from a manual screen of the extant literature. Ultimately, 112 variants in 21 genes are found to cross the evidence threshold to be classified as pathogenic or likely pathogenic. This study also confirms the effect of reduced variant penetrance in ALS and FTD and finds that many variants exhibit significant geographic heterogeneity. A web application (alsftd.tcd.ie) is made available to provide all supporting evidence in an accessible format for clinicians, patients and researchers.

The second study in this thesis focuses on the identification of short tandem repeats (STRs) and repeat expansions (REs) in next-generation sequencing data. A benchmarking study of 7 tools is performed to assess their ability to correctly identify large REs, to accurately measure STRs and finally to compare results between whole-exome sequencing data and whole-genome sequencing data from the same patients. It is identified that many tools have good utility for identifying REs and accurately measuring STRs; however, no single tool provides perfect discrimination and the accuracy of results can be highly gene dependent. Consequently, it is advised that significant results observed from these tools should be subject to validation either with polymerase chain reaction or by taking a consensus approach with other tools. The lessons learned from the benchmarking study are applied to the study of 132 epilepsy patients, wherein no evidence is found supporting the pleiotropic role of REs known to cause other neurological diseases in the pathology of this disease.

Following the dual observations from the meta-analysis that the majority of ALS research has been performed in a small number of regions and that several genetic variants exhibit significant geographic heterogeneity, it is deemed beneficial to study the genetics of ALS in previously understudied populations. The third study in this thesis concerns the genetic

screening of 126 Cuban ALS patients and 111 controls for pathogenic genetic variants. A low rate of the *C9orf72* RE is observed. Interestingly the cohort does not carry *SOD1*, *TARDBP* or *VAPB* variants that are identified to be prevalent in North and South America.

The final research chapter examines the genetic basis of ALS and the related conditions FTD and primary lateral sclerosis (PLS) in Ireland. One PLS patient is found to harbour a previously unreported variant in the gene *SPAST*. Variants in the same amino acid have previously been reported to cause adult onset hereditary spastic paraplegia, a condition with significant clinical overlap with PLS. The genetics of ALS and FTD in Ireland are found to be distinct from the rest of the world by their absences. While rates of the *C9orf72* RE are found to be similar other European countries, Irish patients lack genetic variants that are commonly observed elsewhere. Finally a study of related individuals, who are similarly affected with ALS or FTD, but who have discordant *C9orf72* genotyping, is performed to further elucidate the basis of this discordance.

Publications from this thesis

Bede, P., Chipika, R. H., Christidi, F., Hengeveld, J. C., Karavasilis, E., Argyropoulos, G. D., Lope, J., Li Hi Shing, S., Velonakis, G., Dupuis, L., **Doherty, M. A.,** Vajda, A., McLaughlin, R. L., & Hardiman, O. (2021). Genotype-associated cerebellar profiles in ALS: focal cerebellar pathology and cerebro-cerebellar connectivity alterations. Journal of Neurology, Neurosurgery, and Psychiatry, 92(11), 1197–1205.

Bede, P., Chipika, R. H., Finegan, E., Li Hi Shing, S., Chang, K. M., **Doherty, M. A.,** Hengeveld, J. C., Vajda, A., Hutchinson, S., Donaghy, C., McLaughlin, R. L., & Hardiman, O. (2020). Progressive brainstem pathology in motor neuron diseases: Imaging data from amyotrophic lateral sclerosis and primary lateral sclerosis. Data in Brief, 29, 105229.

Bede, P., Chipika, R. H., Finegan, E., Li Hi Shing, S., **Doherty, M. A.,** Hengeveld, J. C., Vajda, A., Hutchinson, S., Donaghy, C., McLaughlin, R. L., & Hardiman, O. (2019). Brainstem pathology in amyotrophic lateral sclerosis and primary lateral sclerosis: A longitudinal neuroimaging study. NeuroImage. Clinical, 24, 102054.

Bede, P., Omer, T., Finegan, E., Chipika, R. H., Iyer, P. M., **Doherty, M. A.,** Vajda, A., Pender, N., McLaughlin, R. L., Hutchinson, S., & Hardiman, O. (2018). Connectivity-based characterisation of subcortical grey matter pathology in frontotemporal dementia and ALS: a multimodal neuroimaging study. Brain Imaging and Behavior. https://doi.org/10.1007/s11682-018-9837-9

Chipika, Rangariroyashe H., Christidi, F., Finegan, E., Li Hi Shing, S., McKenna, M. C., Chang, K. M., Karavasilis, E., **Doherty, M. A.,** Hengeveld, J. C., Vajda, A., Pender, N., Hutchinson, S., Donaghy, C., McLaughlin, R. L., Hardiman, O., & Bede, P. (2020). Amygdala pathology in amyotrophic lateral sclerosis and primary lateral sclerosis. Journal of the Neurological Sciences, 417, 117039.

Chipika, Rangariroyashe H., Siah, W. F., Shing, S. L. H., Finegan, E., McKenna, M. C., Christidi, F., Chang, K. M., Karavasilis, E., Vajda, A., Hengeveld, J. C., **Doherty, M. A.,** Donaghy, C., Hutchinson, S., McLaughlin, R. L., Hardiman, O., & Bede, P. (2020). MRI

data confirm the selective involvement of thalamic and amygdalar nuclei in amyotrophic lateral sclerosis and primary lateral sclerosis. Data in Brief, 32, 106246.

Finegan, E., Chipika, R. H., Li Hi Shing, S., **Doherty, M. A.,** Hengeveld, J. C., Vajda, A., Donaghy, C., McLaughlin, R. L., Pender, N., Hardiman, O., & Bede, P. (2019). The clinical and radiological profile of primary lateral sclerosis: a population-based study. Journal of Neurology, 266(11), 2718–2733.

Finegan, E., Hi Shing, S. L., Chipika, R. H., McKenna, M. C., **Doherty, M. A.,** Hengeveld, J. C., Vajda, A., Donaghy, C., McLaughlin, R. L., Hutchinson, S., Hardiman, O., & Bede, P. (2020). Thalamic, hippocampal and basal ganglia pathology in primary lateral sclerosis and amyotrophic lateral sclerosis: Evidence from quantitative imaging data. Data in Brief, 29, 105115.

Finegan, E., Li Hi Shing, S., Chipika, R. H., **Doherty, M. A.,** Hengeveld, J. C., Vajda, A., Donaghy, C., Pender, N., McLaughlin, R. L., Hardiman, O., & Bede, P. (2019). Widespread subcortical grey matter degeneration in primary lateral sclerosis: a multimodal imaging study with genetic profiling. NeuroImage. Clinical, 24, 102089.

Finegan, E., Shing, S. L. H., Chipika, R. H., Chang, K. M., McKenna, M. C., **Doherty, M. A.**, Hengeveld, J. C., Vajda, A., Pender, N., Donaghy, C., Hutchinson, S., McLaughlin, R. L., Hardiman, O., & Bede, P. (2021). Extra-motor cerebral changes and manifestations in primary lateral sclerosis. Brain Imaging and Behavior, 15(5), 2283–2296.

Finegan, E., Siah, W. F., Shing, S. L. H., Chipika, R. H., Chang, K. M., McKenna, M. C., **Doherty, M. A**., Hengeveld, J. C., Vajda, A., Donaghy, C., Hutchinson, S., McLaughlin, R. L., Hardiman, O., & Bede, P. (2020). Imaging and clinical data indicate considerable disease burden in “probable” PLS: Patients with UMN symptoms for 2-4 years. Data in Brief, 32, 106247.

McKenna, M. C., Chipika, R. H., Li Hi Shing, S., Christidi, F., Lope, J., **Doherty, M. A.,** Hengeveld, J. C., Vajda, A., McLaughlin, R. L., Hardiman, O., Hutchinson, S., & Bede, P. (2021). Infratentorial pathology in frontotemporal dementia: cerebellar grey and white matter alterations in FTD phenotypes. Journal of Neurology, 268(12), 4687–4697.

McKenna, M. C., Tahedl, M., Lope, J., Chipika, R. H., Li Hi Shing, S., **Doherty, M. A**., Hengeveld, J. C., Vajda, A., McLaughlin, R. L., Hardiman, O., Hutchinson, S., & Bede, P. (2021). Mapping cortical disease-burden at individual-level in frontotemporal dementia: implications for clinical care and pharmacological trials. Brain Imaging and Behavior. https://doi.org/10.1007/s11682-021-00523-7

Ryan, M., Heverin, M., **Doherty, M. A**., Davis, N., Corr, E. M., Vajda, A., Pender, N., McLaughlin, R., & Hardiman, O. (2018). Determining the incidence of familiality in ALS. Neurology: Genetics, 4(3). https://doi.org/10.1212/NXG.0000000000000239

Ryan, Marie, Zaldívar Vaillant, T., McLaughlin, R. L., **Doherty, M. A**., Rooney, J., Heverin, M., Gutierrez, J., Lara-Fernández, G. E., Pita Rodríguez, M., Hackembruch, J., Perna, A., Vazquez, M. C., Musio, M., Ketzoian, C. N., Logroscino, G., & Hardiman, O. (2019). Comparison of the clinical and genetic features of amyotrophic lateral sclerosis across Cuban, Uruguayan and Irish clinic-based populations. Journal of Neurology, Neurosurgery, and Psychiatry. https://doi.org/10.1136/jnnp-2018-319838

Chapter 1

Introduction

Amyotrophic lateral sclerosis (ALS) is a devastating and fatal neurological disease and is the primary focus of this thesis. ALS onset typically occurs between age 50 and 65 (O'Toole *et al.* 2008; Giancarlo Logroscino *et al.* 2010), at which point a formerly healthy individual will begin to experience muscle wasting, stiffness and weakness. This is followed by paralysis of the voluntary and respiratory muscles. Average survival is typically between 20 and 36 months with just 5 to 10% of patients surviving more than ten years from first symptom onset (Adriano Chiò *et al.* 2009).

ALS treatment

ALS was first described in the mid-19th century by Jean-Martin Charcot (Charcot and Joffroy 1869), however despite over 150 years of research there is still no cure. Currently Riluzole is the only drug approved for the treatment of ALS in Europe (Petrov *et al.* 2017). Riluzole extends life by 2-3 months (Miller, Mitchell, and Moore 2012); however, this extension occurs primarily in the later stages of disease when disability is already high (Fang *et al.* 2018). Riluzole was first brought to the market in 1995; however, the therapeutic mechanism is still unknown. Over 60 other molecules have now been investigated, with all failing to reach the market (Petrov *et al.* 2017). In progressing drugs to human clinical trials there is increasing recognition of the potential importance of stratifying patients based on genetic background. This is true for therapies which may target a specific gene (Lagier-Tourenne *et al.* 2013), but also for treatments which may target a specific pathway (Broce *et al.* 2018). However, in order for this to be effective we must first have a good understanding of which genes, variants, and polygenic burdens are truly causing or increasing risk for ALS and how this differs across populations and phenotypes.

Biological processes in ALS

Causative pathogenic mechanisms in ALS still remain unclear; however, disruption of several processes that are essential to neuronal functional have been observed (Mejzini *et al.* 2019). Affected processes include altered ribonucleic acid (RNA) metabolism, nucleocytoplasmic transport defects, impaired proteostasis, impaired deoxyribonucleic acid (DNA) repair, mitochondrial disfunction and oxidative stress, axonal transport defects, vesicular transport defects, neuroinflammation, excitotoxicity, and oligodendryte dysfunction (Mejzini *et al.* 2019).

ALS epidemiology

An individual has a 1 in 400 likelihood of developing ALS in their lifetime (A. Chiò *et al.* 2009; Alonso *et al.* 2009; Johnston *et al.* 2006). With an incidence rate of 3.1 cases per 100,000 people per year (Ryan, Heverin, *et al.* 2019), over 150 people are expected to be diagnosed with ALS in Ireland this year. However, given its relatively late age of onset and poor prognosis, the number of people living with ALS at any given time is low, with prevalence estimates of between 4.1 and 8.4 per 100,000 people (Longinetti and Fang 2019). For comparison, multiple sclerosis has an incidence rate of 2.1 cases per 100,000 people but a prevalence of 35.9 people per 100,000 (Walton *et al.* 2020).

With the notable exception of some clusters of high incidence (discussed below), lower ALS incidence is observed in non-Caucasian populations (Africa 0.41 (95% (CI: 0.34-0.5)), Asia (0.55 (95% CI: 0.46-0.66)), South America (1.1 (95 % CI: (1-1.2))) than in Europe (2 (95% CI: 1.9-2.1)) (GBD 2016 Motor Neuron Disease Collaborators 2018). This difference is not explained by the rate of surveillance, socioeconomic status or lifespan in these regions (GBD 2016 Motor Neuron Disease Collaborators 2018). There is evidence that populations that have undergone recent admixture may have a reduced risk of ALS, with one study finding that Cuban individuals with self-reported admixed ancestry may have reduced ALS mortality relative to people who self-report as white or black (Zaldivar *et al.* 2009).

Between 5 and 20% of patients present with a family history of ALS (fALS), the remaining 80-95% of cases are defined as sporadic ALS (sALS) (Ryan *et al.* 2018; Byrne *et al.* 2011).

Importantly, a classification of ‘sporadic ALS’ is not an indication of a patient’s genetic background, merely the family history they have presented with. Ryan *et al.* (2018) found that longitudinal surveillance of ALS registers over a 23 year period increased the percentage of cases identified as having a familial background from 5% to 20%. If a true family history of ALS does exist, the problem of correctly identifying this is compounded by the late onset of ALS and reduced penetrance of some ALS variants.

The heritability of a trait is the proportion of phenotypic variance within a population which is attributable to genetic variance. Estimates for the heritability of ALS are between 52% (Ryan, Heverin, *et al.* 2019) and 76% (A. Al-Chalabi *et al.* 2010). A heritability of 61% (95% CI 38–78%) has been estimated for solely sporadic cases (A. Al-Chalabi *et al.* 2010) and 36.9% (95% CI, 19.8%-53.9%) for patients with no known genetic risk (Ryan, Heverin, *et al.* 2019). Evidently genetic factors play a large role in ALS pathogenesis, not just for familial cases but also for patients with no reported family history.

Environmental risk factors

The fact that ALS is not entirely heritable implicates the contribution of non-genetic factors. Several lifestyle factors have been studied as potentially increasing ALS risk (Ingre *et al.* 2015). Many famous athletes, including American baseball player Lou Gehrig, have developed ALS. Several observational studies have investigated a potential correlation between high levels of physical activity or low Body Mass Index and ALS risk. While some studies have found this to be a negative correlation (V. Gallo *et al.* 2016; Pupillo *et al.* 2014), the majority of observational studies have found a positive relationship (A. E. Visser *et al.* 2018; Eaglehouse *et al.* 2016; Harwood *et al.* 2016; Huisman *et al.* 2013; Lehman *et al.* 2012; Ettore Beghi *et al.* 2010; Chio *et al.* 2009; Okamoto *et al.* 2009; E. L. Abel 2007; Taioli 2007; Belli and Vanacore 2005).

Mendelian randomisation (MR) is a method of inferring the true causality of potential risk factors (Davey Smith and Ebrahim 2003). Genetic alleles that are correlated with potential risk factors are randomly assorted in a population. In MR studies this random assortment among cases and controls is utilised to identify if the risk of developing disease is affected by genetic liability to be exposed to risk factor of interest. Not only do MR studies not suffer from many of the confounding issues that observational studies traditionally have, but they can also be performed on previously generated data such as summary statistics from genome-wide association studies (GWAS). To date over 20 MR studies have been conducted in ALS (Julian *et al.* 2021). These studies have shown that LDL cholesterol level (odds ratio (OR): 1.12 (95% CI: 1.03–1.20)), coronary heart disease (OR: 1.06 (95% CI: 1.0–1.13)) and self-reported high cholesterol (OR: 2.39 (95% CI: 1.48–3.84)) are likely to be causative ALS risk factors (Bandres-Ciga *et al.* 2019). MR studies have found that smoking is not a causative risk factor for ALS (Opie-Martin *et al.* 2020; van Rheenen *et al.* 2021), despite observational studies to the contrary (H. Wang *et al.* 2011). MR studies are in concurrence both that a sedentary lifestyle is not protective against ALS and that low intensity exercise does not increase risk, but there are conflicting results in studies of high intensity exercise (Julian *et al.* 2021).

There have been geographic clusters of high ALS incidences in Guam, New Guinea and the Kii Peninsula in Japan. Patients began presenting in the 1950s with ALS/ Parkinsonism Dementia Complex (ALS/PDC) at rates up to 100 times higher than the rate of ALS elsewhere (G. Logroscino and Piccininni 2019). Incidence rates in New Guinea are still elevated; however, since the 1960s rates in Guam and Kii have fallen and are now approaching comparable incidences to the rest of the world. The decreased incidences over such a short period suggests that this is not a genetic effect. Studies have suggested that this is due to a reduction in dietary intake of β-N-methylamino-l-alanine, a chemical present in the roots of cycad trees (P. A. Cox and Sacks 2002; Banack and Cox 2003; Murch, Cox, and Banack 2004), but this remains contentious (Chernoff *et al.* 2017).

In addition to affecting the overall risk of developing ALS, lifestyle factors may be modifiers of disease. Byrne *et al.* (2013) regressed the reported mean age of ALS for a region against the population life expectancy within the same region. A positive correlation was observed (r=0.91, p=0.01), indicating either that environmental conditions that are conducive to a longer life delay ALS onset, or that environmental conditions that shorten lifespan also accelerate ALS onset. Analysis in this study was based on the mean age of onset for each region. As there is large variability in ALS onset within a population, it remains to be seen whether this result is replicable when including all available ages of onset for a given region; this question is addressed in Chapter 2.

Sex as a risk factor

Sex is an ALS risk factor, with males being 1.3 times more likely to develop ALS than females (Giancarlo Logroscino *et al.* 2010). Additionally, in a study of sex-specific heritability, heritability was higher among mother-daughter pairings than father-son or mixed sex pairings, suggesting a sex specific inheritance of risk factors (Ryan, Heverin, *et al.* 2019). Sex is also a modifier of disease phenotype with males exhibiting earlier onset than females (McCombe and Henderson 2010). Males are more likely to present with spinal symptoms regardless of their age of onset while females are more likely to present with spinal onset when young and bulbar onset with increasing age (Giancarlo Logroscino *et al.* 2010).

Frontotemporal dementia

It is now understood that ALS is a disease which does not solely rob patients of their physical capabilities. Approximately 15% of ALS patients develop concomitant frontotemporal dementia (FTD), and a similar percentage of FTD patients develop ALS (Phukan *et al.* 2012; Lomen-Hoerth, Anderson, and Miller 2002). With an incidence rate of 1.61 (95% CI: 1.14-1.99) cases/100,000 per year, a prevalence of 10.84 (95% CI: 9.27-12.42) and a lifetime risk of 1/742, FTD is the second most common form of dementia in people under the age of 65 (after Alzheimer's disease (AZD)) (Onyike and Diehl-Schmid 2013). FTD is highly heritable with approximately 40% of patients reporting a significant family history (Goldman *et al.* 2005).

FTD results from atrophy of the frontal and temporal lobes and is distinguished from other early-onset dementias as behavioural changes or language dysfunction typically precede memory loss (Warren, Rohrer, and Rossor 2013). Patients experience a progressive decline in interpersonal and executive skills and often develop unusual behaviours such as apathy, disinhibition and new obsessions. The clinical presentation of FTD is highly heterogenous with several clinical subphenotypes and closely related conditions (table 1.1). The genetics of FTD globally are examined in Chapter 2 and the genetics of FTD in Ireland are examined in Chapter 3.

Table 1.1: FTD subphenotypes and related conditions

	Subphenotype	Abbreviation	Description
Frontotemporal dementia subphenotypes	Behavioural variant FTD	bvFTD	Early symptoms may include switching jobs or partners, reduced social awareness or altered preferences and tastes.
	Progressive non-fluent aphasia	PNFA	Patients lose the ability to make fluent conversation.
	Semantic dementia	SD	Patients experience a progressive decline in vocabularly, forgetting the meaning of words.
Other conditions associated with frontotemporal cognitive change	Corticobasal syndrome	CBS	Early symptoms include stiffness or tremors in a particular limb or the feeling that a limb doesn't belong to you. This can progress to other limbs. Patients may experience problems with memory loss, planning or coping with new situations.
	Progressive supranuclear palsy	PSP	Patients experience a decline in balance and mobility and an inability to maintain gaze on an object. This is often accompanied by changes in behaviour such as irritability and apathy.

Table based on information from Warren et al. (2013)

Motor neurone diseases

ALS sits within the ALS-FTD phenotypic continuum, but also within a phenotypically and genetically heterogenous spectrum of motor neurone diseases (MNDs). ALS is characterised by the loss of both lower motor neurones (LMNs), which are present in the brainstem and spinal cord and innervate the somatic musculature, and upper motor neurones (UMNs), which are present in the motor cortex and brainstem and provide input to the LMNs. Loss of UMNs prevents signalling to the LMNs, resulting in muscle stiffness and weakness, while LMN degeneration prevents muscles from receiving signals, leading to weakness and muscular atrophy (Kent-Braun *et al.* 1998).

ALS is the most common and devastating adult-onset MND and is typically distinguishable from other MNDs due to its aggressive nature and associated LMN and UMN degeneration. Other MNDs are generally classified by whether patients experience selective LMN degeneration (progressive muscular atrophy (PMA), spinal muscular atrophy (SMA)), or UMN degeneration (hereditary spastic paraplegia (HSP), primary lateral sclerosis (PLS)).

Lower motor neurone disorders

The two most common purely LMN disorders are SMA and PMA.

SMA is an autosomal recessive MND that is one of the leading causing of infant mortality and is estimated to affect approximately 10 births per 100,000 (Jedrzejowska *et al.* 2010; Arkblad *et al.* 2009; Prior *et al.* 2010). SMA is caused by recessive LOF SMN1 variants resulting in low levels of the SMN protein, causing LMNs to deteriorate and muscles to atrophy. Even within SMA the phenotypic spectrum is broad. At its most severe (SMA type 0), patients exhibit symptoms at birth, are never able to sit and typically survive less than 6 months; in contrast, patients with SMA type IV have onset in adulthood, may walk independently and have normal life expectancy (Nicolau *et al.* 2021). Recently developed antisense oligonucleotides (ASOs) have been shown to halve the risk of death or permanent mechanical ventilation in infant patients (Finkel *et al.* 2017, 2021).

While SMA patients usually have childhood onset and causative mutations segregate strongly in their pedigrees, PMA onset is generally in adulthood and there is rarely a family history of MND. PMA patients typically experience weakness and muscle wasting in the hands which spreads to the lower body as LMNs continue to degenerate. PMA is estimated to account for 7-8% of adult-onset MNDs (W.-K. Kim *et al.* 2009). By definition PMA patients do not show UMN signs at onset; however, many develop UMN symptoms at later stages of disease (J. Visser *et al.* 2007), further highlighting the complexity of the MND spectrum. The age of onset and prognosis is similar in both ALS and PMA patients (Riku *et al.* 2014; J. Visser *et al.* 2007). PMA is often described as a 'sporadic' disease; however, multiple members of families carrying *SOD1* variants have been reported to have solely LMN symptoms, ruling out a diagnosis of ALS (Wen *et al.* 2021; Cervenakova *et al.* 2000) and LMN-predominant adult onset patients have been observed to carry variants in *CHMP2B* (L. E. Cox *et al.* 2010) and *VAPB* variants (Nishimura *et al.* 2004).

Upper motor neurone disorders

HSP results from the loss of UMNs and has a mean global prevalence of 1.8 patients per 100,000 people (Ruano *et al.* 2014). The condition is typically characterized by gradual onset and slow progression, with patients experiencing stiffness and weakness of the lower extremities. Onset can occur at any time from childhood to adulthood and patients do not usually experience a reduced lifespan. A hallmark of HSP is that it has distinct autosomal dominant, recessive or X-linked inheritance in pedigrees, and consequently variants in over 70 genes have been associated with HSP inheritance (de Souza *et al.* 2017; Parodi *et al.* 2017; Klebe, Stevanin, and Depienne 2015; Lo Giudice *et al.* 2014).

PLS is an adult onset UMN disorder which causes patients to experience stiffness in their arms and legs and often progresses to difficulty in swallowing. While there is considerable overlap between the phenotypes of adult onset HSP and PLS, the upper body stiffness and bulbar symptoms often observed in PLS are rarely a feature of HSP (Frans Brugman *et al.* 2009). PLS is estimated to account for 7% of adult onset MNDs (W.-K. Kim *et al.* 2009).

A PLS diagnosis is made based on the elimination of other possibilities based on consensus criteria (M. R. Turner *et al.* 2020). Patients must have onset after 25 years of age, UMN symptoms for at least two years, and UMN symptoms in two of three regions (upper limb, lower limb and bulbar). For a diagnosis, patients must also lack sensory symptoms, LMN degeneration and an alternative diagnosis. A diagnosis of probable PLS is made if symptoms are present for 2-4 years and definite PLS if patients have symptoms for more than four years. Despite these careful diagnostic criteria, many patients with a PLS diagnosis subsequently develop UMN symptoms and their diagnosis is re-evaluated (Gordon *et al.* 2006). The genetics of PLS are studied in Chapter 5.

ALS genetics

Epidemiological evidence shows that developing ALS occurs as a six-step process. Al-Chalabi *et al.* (2014) interrogated the population based ALS registers of five countries. An observed linear relationship between log incidence and log age demonstrates increased risk of developing ALS with age; this is consistent with a multistep model. Further to this, Chio *et al.* (2018) found that patients carrying a known ALS genetic variant still conformed to the multistep model, however required fewer steps than patients lacking an established mutation, providing further proof that ALS development is a complex interplay between genetic and either environmental or developmental factors, or both.

The first known genetic causes of ALS were discovered in 1993 when 11 mutations in the gene *SOD1* were identified in thirteen families (Rosen *et al.* 1993). In the intervening three decades much research has been done to identify ALS associated genes and variants. There is no agreed panel of genes that are truly associated with ALS. Different reviews have cited 29 genes (Chia, Chiò, and Traynor 2018) or more than 40 (Peters, Ghasemi, and Brown 2015). In Chapter 2 it is identified that at least 356 genes and over 3,000 variants have been reported in patients with either ALS or FTD. The supporting evidence for each variant has not previously been assessed in a comprehensive and uniform manner. Many previously reported variants are too common in the population to be highly penetrant ALS variants (Kenna, McLaughlin, Hardiman, *et al.* 2013). It is likely that many of these genes and variants represent only spurious associations with disease aetiology, however much clarity is required in the field; this is a topic which is explored extensively in Chapter 2.

There is increasing evidence that ALS susceptibility may increase with mutational burden. Genome-wide association studies (GWAS) have identified a polygenic architecture in ALS (van Rheenen *et al.* 2016, 2021), wherein the contribution of many single nucleotide polymorphisms (SNPs) across the genome contribute to ALS genetic architecture. Additionally, van Blitterswijk *et al.* (2012) screened fALS and sALS cases for mutations in five reliably associated ALS genes. Multiple mutations were observed in fALS cases more often than expected by chance ($p = 1.57 \times 10^{-7}$), supporting an oligogenic basis for ALS. However, these findings do not mean that any observation of two rare variants in an ALS patient can be designated as an oligogenic case of ALS; as described previously, variants may be rare individually, but the occurrence of individual rare variants is common. Therefore assertions of oligogenic causes of ALS should be accompanied by statistically significant support, this is further outlined in Chapter 4.

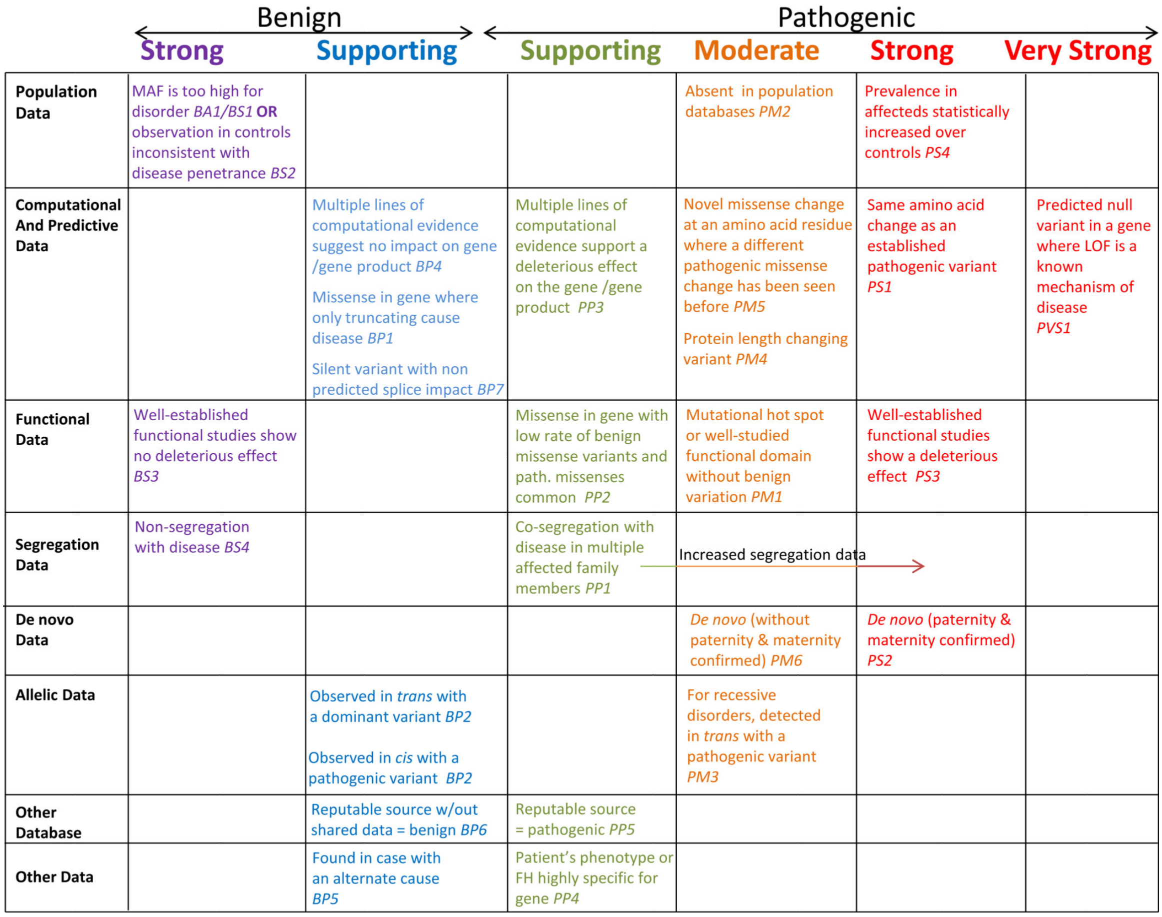
Variant pathogenicity

Genetic counselling can help both patients and potentially presymptomatic relatives make informed decisions regarding family planning and disease management and may impact enrolment in clinical trials. However, with the uncertainty over which genes and variants are truly likely to be pathogenic, there is no current consensus on what genetic testing and counselling should be offered to patients (Vajda *et al.* 2017). Clarifying this should be a major priority in ALS research.

Distinguishing pathogenic variants from non-pathogenic is not a straightforward task and its difficulty is exacerbated in ALS due to genetic heterogeneity, late age-of-onset (AOO), incomplete variant penetrance and a high proportion of sporadic cases. On average, each individual inherits 74 *de novo* single nucleotide variants (SNVs) that were not present in their parent's germline (Veltman and Brunner 2012). We also inherit half of our parent's *de novo* variants that were absent in our grandparents, and so on. The result is that we are all a collage of rare variation that may be unique to us, our immediate families or individuals we share an ancestor with somewhere on our family tree, but the majority of this rare variation is not pathogenic. While individual variants may be rare, the overall presence of rare variants is exceedingly common. Over 241 million small variants are identified in the Genome Aggregation Database (gnomAD), a collection of 141,436 exomes and genomes, with the vast majority of these being rare variants (Karczewski *et al.* 2020). Identifying a rare variant in a patient, even within a previously associated gene, is not sufficient evidence to infer causality.

The American College of Medical Genetics (ACMG) have proposed guidelines for determining whether a variant should be interpreted as pathogenic (P), likely pathogenic (LP), benign (B), likely benign (LB) or a variant of uncertain significance (VUS) (Richards *et al.* 2015). These guidelines take multiple strands of evidence into account to arrive at a variant's classification. The guidelines account for the frequency of the variant in the population, *in silico* pathogenicity prediction tools, functional studies, segregation data, whether the variant is *de novo* in a gene that is susceptible to *de novo* variation, whether a variant matches the proposed pattern of inheritance for a disease, previous reports for the variant, whether the phenotype of the patient is highly specific to what would be expected for the variant in question and whether a carrier has any other likely pathogenic variant (figure 1.1). The strength of a particular piece of evidence determines whether it is assigned as supporting, moderate, strong, very strong or stand-alone evidence (figure 1.1). The various strands of evidence accounted for in figure 1.1 are assessed cumulatively (figure 1.2) to ultimately determine at a variant's designation.

While the ACMG guidelines provide clear recommendations for variant classification, there are categories outlined in figure 1.1 that have a degree of ambiguity and require interpretation. For example, category BA1 provides clear guidance that an allele frequency (AF) above 5% in large population datasets such as gnomAD is stand-alone evidence in favour of benignity, this is unambiguous and does not require interpretation by a researcher assessing variants. In contrast, category PM1 states that a variant being located in a mutational hotspot without benign variation, counts as moderate evidence of pathogenicity; however, the classification of a mutational hotspot is very gene and disease dependent, leaving the assessment of this category open to ambiguous interpretation. Many papers have been published describing further guidelines for the interpretation of specific categories (Abou Tayoun *et al.* 2018; Harrison, Biesecker, and Rehm 2019; Cho *et al.* 2020; Wilcox *et al.* 2021; Brnich *et al.* 2019; Jarvik and Browning 2016). It is also recognised that the ACMG guidelines often require gene- or disease-specific modification as certain categories will be more or less relevant for a given condition or may provide stronger or weaker evidence in a given context (Morales Ana *et al.* 2020; Kelly *et al.* 2018; Oza *et al.* 2018; Romanet *et al.* 2019; Feliubadaló *et al.* 2021; Maxwell *et al.* 2016; Fortuno *et al.* 2021). There are currently no agreed guidelines for the interpretation of the ACMG guidelines for ALS.

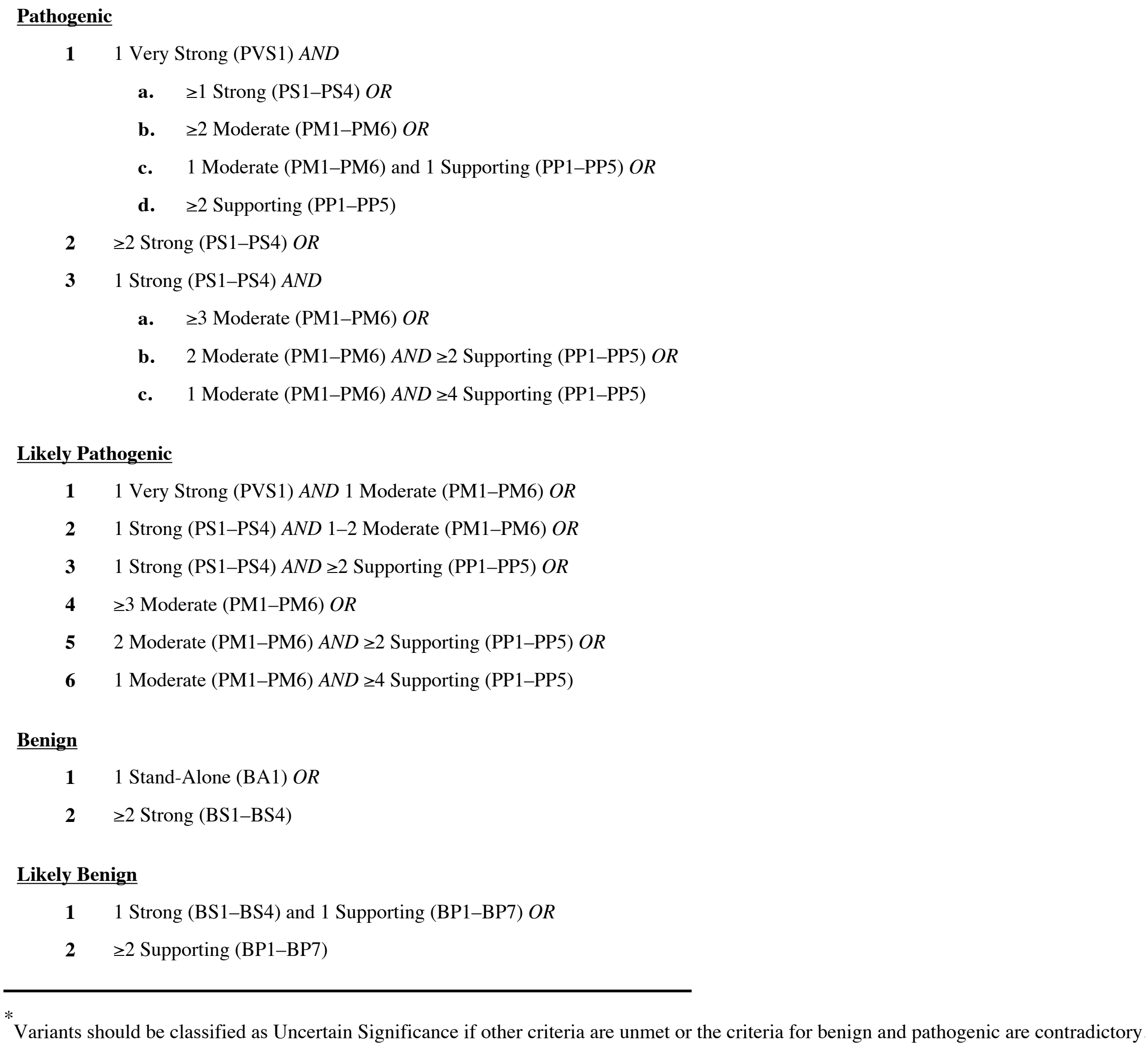


	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product *BP4* Missense in gene where only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1* →	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed in *trans* with a dominant variant *BP2* Observed in *cis* with a pathogenic variant *BP2*		For recessive disorders, detected in *trans* with a pathogenic variant *PM3*		
Other Database		Reputable source w/out shared data = benign *BP6*	Reputable source = pathogenic *PP5*			
Other Data		Found in case with an alternate cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Figure 1.1: ACMG evidence framework

This chart displays the organisation of evidence categories for determining a variant's pathogenicity. Evidence can either support a pathogenic or benign variant annotation with different strengths of evidence being designated as supporting, moderate, strong or stand-alone. Evidence categories are further described in Chapter 2. Abbreviations: BS, benign strong; BP, benign supporting; FH, family history; LOF, loss-of-function; MAF, minor allele frequency; path., pathogenic; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong

Figure reproduced from Richards et al. (2015)



Pathogenic

1 1 Very Strong (PVS1) *AND*

- **a.** ≥1 Strong (PS1–PS4) *OR*
- **b.** ≥2 Moderate (PM1–PM6) *OR*
- **c.** 1 Moderate (PM1–PM6) and 1 Supporting (PP1–PP5) *OR*
- **d.** ≥2 Supporting (PP1–PP5)

2 ≥2 Strong (PS1–PS4) *OR*

3 1 Strong (PS1–PS4) *AND*

- **a.** ≥3 Moderate (PM1–PM6) *OR*
- **b.** 2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR*
- **c.** 1 Moderate (PM1–PM6) *AND* ≥4 Supporting (PP1–PP5)

Likely Pathogenic

1 1 Very Strong (PVS1) *AND* 1 Moderate (PM1–PM6) *OR*

2 1 Strong (PS1–PS4) *AND* 1–2 Moderate (PM1–PM6) *OR*

3 1 Strong (PS1–PS4) *AND* ≥2 Supporting (PP1–PP5) *OR*

4 ≥3 Moderate (PM1–PM6) *OR*

5 2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR*

6 1 Moderate (PM1–PM6) *AND* ≥4 Supporting (PP1–PP5)

Benign

1 1 Stand-Alone (BA1) *OR*

2 ≥2 Strong (BS1–BS4)

Likely Benign

1 1 Strong (BS1–BS4) and 1 Supporting (BP1–BP7) *OR*

2 ≥2 Supporting (BP1–BP7)

* Variants should be classified as Uncertain Significance if other criteria are unmet or the criteria for benign and pathogenic are contradictory

Figure 1.2: ACMG rules for combining criteria to classify sequence variants

Figure reproduced from Richards et al. (2015)

Repeat expansions

Short tandem repeats (STRs), are short repeated DNA motifs, typically of 1-6 bps in length, that comprise 3% of the human genome (Lander *et al.* 2001). Due to errors during DNA replication, these repeats are highly polymorphic in length. Although loosely defined, a repeat expansion (RE) occurs when an STR is expanded beyond the normal length observed in the healthy population.

STR variability in more than 50 genes has now been linked to various neurological disorders (Depienne and Mandel 2021). There is little in common across all disease-associated REs. They vary in their composition and the number of repeats required for pathogenesis. They are observed in coding regions, introns and untranslated regions (UTRs). Even the likely/proposed pathogenic mechanism of action differs across STR loci (Chintalaphani *et al.* 2021; Malik *et al.* 2021; Khristich and Mirkin 2020; Paulson 2018; Depienne and Mandel 2021).

Repeat expansion pathogenic mechanisms

LOF can occur due to epigenetic gene silencing, such as in fragile X syndrome (FXTAS) (Oberlé *et al.* 1991; Verkerk *et al.* 1991) and Friedreich's ataxia (FRDA) (V. Campuzano *et al.* 1996). Expansions present in the genic regulatory regions experience hypermethylation of the expanded allele, preventing gene expression.

One gain-of-function (GOF) mechanism is protein misfolding. Many of the CAG repeats including those responsible for Huntington's disease (HD) and several of the spinocerebellar ataxias (SCAs) cause long polyglutamine (polyQ) tracts in the gene transcript. These polyQ tracts result in protein misfolding and lead to protein aggregates which disrupt cellular control mechanisms and lead to neuronal cell death. Similar pathogenesis can arise from certain polyalanine (polyA) repeats (congenital central hypoventilation syndrome (CCHS) and early infantile epileptic encephalopathy EIEE1).

Specific expanded motifs are capable of forming stable secondary structures when transcribed. These hairpin or G-quadruplex structures sequester RNA-binding proteins into RNA foci. Patients with myotonic dystrophy 1 (DM1), have a CTG expansion in the 3'UTR. This expansion does not affect expression; however, the formed RNA foci sequester essential splicing factors leading to aberrant splicing of essential muscle genes (Kanadia *et al.* 2003; Philips, Timchenko, and Cooper 1998).

The final proposed GOF pathogenic RE mechanism is repeat associated non-ATG (RAN) translation. The three-dimensional structures formed by certain GC motifs result in translation machinery being operational at the locus even in the absence of a start codon. The RE is transcribed bidirectionally and in all frames; producing peptide repeats which aggregate throughout the central nervous system (CNS). RAN translation has been observed in DM1 (Zu *et al.* 2011) , FXTAS (Todd *et al.* 2013), HD (Bañez-Coronel *et al.* 2015), SCA2 (Scoles *et al.* 2015), SCA8 (Zu *et al.* 2011) and ALS (Zu *et al.* 2013).

Additional STR Features

The fact that large REs are pathogenic is not the only phenotypic effect of STRs. Repeat variability in the normal range at the population level is linked to variability in complex human traits, with STRs being enriched in human promoters and enhancers and recurrently being found to affect the expression of neighbouring genes (Gymrek 2017).

STR disease loci have been shown to exhibit pleiotropic effects, wherein variants in the same gene can result in disparate traits. For example, the *C9orf72* RE, discussed below, causes both ALS and FTD and has been observed as a rare cause of Parkinson disease, Huntington disease-like syndrome and AZD (Woollacott and Mead 2014). While large expansions in *ATXN2* cause SCA2, intermediate length expansions of 27-32 repeats are an ALS risk factor (Sproviero *et al.* 2017; M.-D. Wang *et al.* 2014; Daoud *et al.* 2011). *HTT* expansions which cause HD have also recently been linked to ALS in a similar manner (Dewan *et al.* 2021), although this result is disputed (Thomas *et al.* 2021). More broadly, pleiotropy is a common feature of neurological disorders (Polushina *et al.* 2021).

C9orf72

In 2010, Laaksovirta *et al.* identified a 232 kilobase (kb) haplotype on chromosome 9 that is statistically enriched in fALS patients (OR: 21·0 (95% CI 11·2-39·1)). In 2011, repetition of 29 or more hexanucleotide (G_4C_2) motif units in the gene *C9orf72*, which lies in this haplotypic region, was found to be the most common cause of ALS and FTD in Europe (DeJesus-Hernandez *et al.* 2011; Renton *et al.* 2011). The expansion is believed to have arisen once on this specific haplotype; thus, while many controls and patients without the

RE share the same haplotype, the RE has never been observed in the absence of this haplotype (Smith *et al.* 2013; Laaksovirta *et al.* 2010; Mok *et al.* 2012).

The RE displays significant population-specific heterogeneity, explaining 34% and 5% of European fALS and sALS cases respectively, but only 2% and less than 1% respectively in Asia (Zou *et al.* 2017). This population-specific heterogeneity highlights the importance of studying populations from diverse ethnic backgrounds, as variants that are rare in one ALS or FTD population may be common elsewhere, Chapter 4 aims to address this issue. The extent of population-specific genetic heterogeneity has not been assessed for the vast majority of previously reported variants and is explored in Chapter 2.

C9orf72 pathogenic mechanism

The pathogenic mechanism associated with the *C9orf72* RE in ALS and FTD is still an open debate.

C9orf72 RE pathogenicity may arise from heterozygous loss of gene function. *C9orf72* regulates autophagy and endolysosomal trafficking and function (Woollacott and Mead 2014). Decreased messenger ribonucleic acid (mRNA) and protein levels have been observed in *C9orf72* RE positive central nervous system (CNS) tissues and induced pluripotent stem cell derived neuronal cell lines (Sivadasan *et al.* 2016; van Blitterswijk *et al.* 2015; Xiao *et al.* 2015; Waite *et al.* 2014; Belzil *et al.* 2013; Ciura *et al.* 2013; Donnelly *et al.* 2013; Fratta *et al.* 2013; Mori, Weng, *et al.* 2013; Xi *et al.* 2013; Gijselinck *et al.* 2012; DeJesus-Hernandez *et al.* 2011). Additionally, in gnomAD the ratio of observed loss-of-function (LOF) variants to the number that would be expected for a gene this size is 0.58, indicating a lack of tolerance for LOF variants. On the other hand; to date only a single reported sALS case harbouring a LOF splice-acceptor variant has been reported and the identified variant is classified as a VUS (F. Liu *et al.* 2016). Additionally, patients who are homozygous do not show increased disease severity (Fratta *et al.* 2013). Reduction of endogenous *C9orf72* function has produced neuronal defects in C. elegans (Therrien *et al.* 2013) and zebrafish (Ciura *et al.* 2013); however, this has not replicated in mice (Koppers *et al.* 2015; Lagier-Tourenne *et al.* 2013).

A potentially more robust explanation for *C9orf72* pathogenicity is toxic RNA GOF. Several studies have observed the presence of nuclear RNA foci throughout the CNS in *C9orf72* RE positive patients (Cooper-Knock *et al.* 2014; Cooper-Knock, Shaw, and Kirby 2014;

Donnelly *et al.* 2013; Gendron *et al.* 2013; Lagier-Tourenne *et al.* 2013; Lee *et al.* 2013; Mizielinska *et al.* 2013; Mori, Arzberger, *et al.* 2013; Zu *et al.* 2013; DeJesus-Hernandez *et al.* 2011). Similar to DM1 discussed above, evidence suggests these foci act as 'protein sinks', sequestering RBPs, preventing them functioning elsewhere in the body. Chew *et al.* (2015) induced the expression of $(G_4C_2)_{66}$ throughout the mouse CNS, causing neuronal loss and behavioural deficits.

There is also evidence supporting the role of pathogenic dipeptide repeats (DPRs) in ALS. In ALS the hexanucleotide repeat undergoes bidirectional RAN translation, producing 6 alternate DPRs which aggregate throughout the CNS (Ash *et al.* 2013; Mackenzie *et al.* 2013). Evidence from Drosophila has indicated that these DPRs rather than RNA foci may be responsible for pathogenesis. Mizielinska *et al.* (2014) produced two transgenic fly lines; one line carried the RE and the other carried the expansion but with stop codon interruptions, thus preventing translation. RNA foci formed in both lines; however, DPRs and subsequent early lethality were only observed in the absence of stop codons. Tran *et al.* (2015) induced an intronic $(G_4C_2)_{160}$ repeat in Drosophila. Unlike in the Mizielinska *et al.* model, this repeat was not accompanied by a polyA tail, so the resulting mRNA could not be transported to the cytoplasm for RAN translation to occur. Consequently, high levels of nuclear RNA foci were observed but with low levels of RAN translation and very little toxicity, indicating that DPRs rather than RNA foci are responsible for pathogenesis.

Summary

ALS is a disease that continues to destroy lives. Currently there is much still unknown about the genetic causes of ALS and related disorders. Ultimately the goal of any research in ALS or related diseases is to help patients. It is hoped that further elucidating and clarifying the genetic causes of these diseases will help patients in the short term by improving genetic counselling, and in the long term by aiding and improving the design, enrolment and identification of targets for clinical trials. It is the aim of this thesis to be a step in this direction.

Aims

The overarching aim of this thesis is to clarify and further our understanding of the genetic causes of ALS and related diseases. This research is presented over four chapters:

- Chapter 2: A comprehensive uniform analysis of three decades of ALS and FTD genetics research
    - It is the aim of this chapter to amalgamate and perform a uniform analysis of research from all previous genetic screening studies in ALS and FTD. Following this, it is the aim to develop a web interface to make this research accessible to patients, clinicians and researchers.
- Chapter 3: Identifying repeat expansions in neurological disorders
    - The aims of this chapter are firstly to utilise ALS data to evaluate tools designed for the characterisation of STRs and REs from next-generation sequencing (NGS) data, and secondly to utilise the results of this evaluation to interrogate an epilepsy cohort that has not previously been studied for REs.
- Chapter 4: The genetic profile of ALS in Cuba
    - This study aims to characterise the profile of ALS genetics in Cuba, a population that has not previously been screened for ALS genetic variants.
- Chapter 5: The broader spectrum of motor neurone disease genetics in Ireland
    - This study aims to examine the genetic landscape of ALS, PLS and FTD in Ireland.

Chapter 2

journALS: a comprehensive, uniform analysis of three decades of ALS and FTD genetics research

Introduction

Both ALS and FTD have a significant genetic component and a large proportion of patients presenting with a family history of disease (Rohrer *et al.* 2009; Ryan, Heverin, *et al.* 2019). In 1993, the identification of segregating variants in *SOD1* marked the discovery of the first ALS-associated gene (Rosen *et al.* 1993). In the intervening three decades thousands of variants in hundreds of genes have been implicated in ALS or FTD with varying degrees of supporting evidence. In recent years, next-generation DNA sequencing has led to a deluge of reported rare variants in previously linked ALS and FTD genes; however, without additional supporting evidence the identification of patients carrying a rare variant in a putative or established ALS or FTD gene is not sufficient evidence to determine the variant's significance (Richards *et al.* 2015).

Relevant factors in assessing the clinical significance of genetic variants include: confidence that the variant is causative for the disease (pathogenicity), the probability that a carrier of the variant will develop the disease over the course of their lifetime (penetrance) and the proportion of cases carrying the variant (prevalence). The difficulty of interpreting the clinical significance of potentially pathogenic variants is exacerbated in ALS and FTD due to genetic heterogeneity, late AOO, incomplete variant penetrance and a high proportion of sporadic cases, wherein patients present with no apparent family history of disease. The

ACMG has provided guidelines for assessing variant pathogenicity (Richards *et al.* 2015); however these have been shown to require disease-specific interpretation and modification (Kelly *et al.* 2018; Oza *et al.* 2018; Romanet *et al.* 2019; Feliubadaló *et al.* 2021) and no consensus has yet been reached for the application of these guidelines to ALS or FTD.

In this study, an extensive review of the extant literature is combined with the most recent genetics and genomics guidelines and datasets to develop the journALS data browser. This data browser is simultaneously a catalogue of 30 years' genetic research in ALS and FTD, a uniform analysis to assess the pathogenicity, penetrance and prevalence of all previously reported ALS- and FTD-associated genetic variants and a framework for the future interpretation of novel variants or variants with additional available evidence. As routine genetic testing is becoming more widely available (Vajda *et al.* 2017), and ALS clinical trials are beginning to enrol based on genetic status ("ALS Signal Dashboard" n.d.), it is now essential that we are able to separate truly pathogenic variants from variants with insufficient supporting evidence.

Methods

Article identification

Four methods were employed to identify all pertinent genetic studies of ALS or FTD from the first published study in 1993 (Rosen *et al.* 1993) to July 2020. The Human Gene Mutation Database (HGMD) v2017.4 (Stenson *et al.* 2017) was utilised by identifying all papers listing any variant falling in any gene linked to any ALS, MND or FTD phenotype (supplementary table S2.1), and excluding papers denoted as "functional characterisation". Secondly, the reference lists of recent reviews and meta-analyses (supplementary table S2.2) were mined to identify key papers which were absent from HGMD. Thirdly, ClinVar (GRCh37_clinvar_20200615) (Landrum *et al.* 2018) was parsed for any variants linked to ALS, FTD or unspecified MND. All previously unscreened papers listing these variants in the ClinVar variant citations file were included. Finally, to redress the fact that the reviews, meta-analyses and HGMD v2017.4 have missed very recent articles, a PubMed screen was carried out on 24/06/2020 using the search terms ("genetic analysis" OR "genetic screen" OR "next-generation" OR "sequencing") AND ("amyotrophic lateral sclerosis" OR "motor neuron disease" OR "frontotemporal dementia"), results were filtered to the previous four years and reviews, clinical trials and studies that did not include patient screening were omitted. These combined searches provided a shortlist of 'potentially relevant papers'.

Article screening

Articles were screened by a team of three users. Using Python v 2.7.9 (Van Rossum and Drake 1995) and Tkinter (Lundh 1999), a custom graphic user interface was created to ensure that papers were assessed and data was output in a uniform manner. Potentially relevant papers were first screened for inclusion based on whether they were a genetic study of patients with ALS, FTD or unspecified MND. Studies which screened more than one unrelated individual for at least the exons of one entire gene were marked as 'potential population studies'.

For each patient carrying an identified variant, the following information was recorded where available: nationality, ethnicity, site of onset, age of onset, disease duration, the presence of cognitive impairment, primary phenotype (ALS, FTD, ALS-FTD), detailed phenotype, variant zygosity, *de novo* status, concurrent variants and family history. Segregation information was also recorded where available.

Screening error

150 population studies were independently screened twice by separate users. These independent screens were compared and any conflicts were resolved to form a consensus. Three measures of interobserver concordance were assessed. Firstly, how successfully did users identify variants in the literature; secondly, how accurately did users identify the frequency of correctly identified variants; and finally, how accurately did users identify phenotype and genotype data for correctly identified individuals.

Population frequencies

Studies marked as 'potential population studies' were collectively screened to find the most representative population studies for a country. To avoid inflation or deflation of calculated AFs, articles were excluded as population studies if the cohort was selected for being negative or positive for previously screened variants. For each clinical ascertainment centre only the most representative study for each gene was included, to avoid patients being included as part of multiple studies. If ascertainment centre was unavailable or uncertain then the most representative study for the country for each gene was included. Studies screening multiple genes simultaneously were prioritised for inclusion over single gene studies from the same ascertainment centre. Justifications and numbers for all 'potential population studies' are included in Supplementary File S1 (available at https://github.com/dohertymark/journALS/Supplementary_Material). Global and region-specific AFs were calculated from the variants observed in population studies.

The AF of the *C9orf72* RE in the general population was calculated by combining the control cohorts of identified *C9orf72* population studies (supplementary table S2.3).

Data processing and annotation

Variants reported in the literature were manually converted to GRCh37 coordinates. Variants were normalised and annotated using Variant Tools v0.5772 (Tan, Abecasis, and Kang 2015), SnpEff v4.3s (Cingolani *et al.* 2012) and GEMINI v0.30.2 (Paila *et al.* 2013). Following annotation, variants in all genes identified in the literature were extracted from available ALS specific datasets including the fALS browser of the ALS Variant Server (ALSVS) ("ALS Variant Server, Worcester, MA" n.d.), the ALS Data Browser (ALSdb) ("ALSdb, New York City, New York" n.d.; Cirulli *et al.* 2015) and the Project MinE Data Browser (van der Spek, van Rheenen, Pulit, Kenna, McLaughlin, *et al.* 2019). The Project MinE AFs were converted from minor AFs to alternate AF. All variants in these genes were merged with the variants from the literature and annotated with dbNSFP 4.0a (X. Liu, Jian, and Boerwinkle 2011, 2013; X. Liu *et al.* 2016), spidex 1.0 (Xiong *et al.* 2015), dbscSNV1.1 (Jian, Boerwinkle, and Liu 2014), the University of California Santa Cruz (UCSC) RepeatMasker tract (W. J. Kent *et al.* 2002) and gnomAD v2.1.1.1 (Karczewski *et al.* 2020) exome and genome AFs, probability of loss of function intolerance (pLI) scores, gene constraint scores, coverage and proportion expressed across transcripts (pext) scores. Insertions and deletions (INDELs) were annotated using PROVEAN v1.1 (Choi *et al.* 2012; Choi 2012), SIFT (Sim *et al.* 2012) and VEST4 (Douville *et al.* 2016).

Several intermediate-length repeat expansions and copy number variants have been associated with ALS or FTD. Where reported in the literature these were annotated in the database; however, with the exception of the *C9orf72* RE, these were not included in the analysis for several reasons: it is typically a range of repeat lengths that is implicated rather than a single variant, these variants typically increase risk rather than are strictly causative, and finally, these variants are typically not annotated in genomics databases and are therefore unable to be integrated in the uniform analysis.

JournALS data browser

The journALS data browser is available at alsftd.tcd.ie, detailing analyses described in the following sections. The interface was designed and built using R v3.6.1 (R Core Team 2019) and Shiny v1.4 (Chang *et al.* 2019). Unless otherwise stated, analysis and plotting is conducted with base R. Data is managed using R packages data.table v1.12.8 (Dowle and Srinivasan 2019), dplyr v0.8.5 (Hadley Wickham *et al.* 2020), DT v0.12 (Xie, Cheng, and Tan 2020), plyr v1.8.6 (H. Wickham 2011), R.utils v2.10.1 (Bengtsson 2020), stringr v1.4 (Hadley Wickham 2019) and tidyr v1.0.2 (Hadley Wickham and Henry 2020). Some aspects of plotting are achieved using gridExtra v2.3 (Auguie 2017), ggvis v0.4.5 (Chang and Wickham 2019) and berryFunctions v1.18.2 (Boessenkool 2019). Code for data pre-processing, the data browser and all figures and statistics from this paper is open source and available at https://github.com/dohertymark/journALS. The features available on the journALS data browser are outlined in table 2.1.

gnomAD allele frequency

The primary gnomAD dataset contains 4,243 ALS samples from the ALSgen consortium and is thus not representative of the general population as an ALS or FTD control cohort. Hereafter, references to gnomAD AFs refer to the non-neuro subset of gnomAD, a collection which includes 104,068 exomes and 10,636 genomes. GnomAD AFs were calculated by summing the number of alleles observed in the gnomAD exome and genome subsets and dividing by the sum of the number of alleles sequenced in the respective subsets. For SNVs and single base INDELs which were absent in either subset, if median coverage at the site was greater than 29 in non-neuro exomes or genomes the variant was assumed to be non-variant in all non-neuro exomes or non-neuro genomes respectively, otherwise absent variants in either subset were not assigned an allele frequency.

Penetrance

Variant penetrance is the probability that a variant carrier will develop disease during their lifetime. Where sufficient data is available, variant penetrance was estimated by two alternative means. Penetrance estimates were calculated for a dominant (heterozygous) form of disease. The first method is referred to as the ‘population penetrance’, following the Bayesian method proposed by (Minikel *et al.* 2016).

Table 2.1: Features available on the journALS data browser

Section	Feature	Description
Variant Browser		
Analysis for a variant of interest (VOI)	Pathogenicity	A display describing the fulfilled ACMG criteria and the overall variant classification for the VOI
	General Information	VOI annotations include variant impact, allele frequencies in publicly available datasets, dbNSFP annotations and journALS annotations such as pathogenicity and penetrance
	Phenotype Information	Phenotype information manually curated from the literautre for all identified carriers of the VOI
	Geographic heterogeneity	A display and test to determine if the VOI exhibits geographic heterogeneity. Comparisons can be made between or within continents, for both ALS and FTD, and for familial or sporadic forms of disease
	Age of Onset	A display and test to determine if the ages of onset for carriers of the VOI differ significantly from the age of onset of the rest of the cohort. Comparisons can be made across phenotype, sex and family history
	Pedigrees	Plots of each pedigree found in the literature carrying the VOI and quantification of the level of segregation across all pedigrees
Region Browser		
Analysis of a region of interest (ROI) which may be globally, a continent or a country of interest	Individuals	Phenotype information for all individuals from the ROI
	Analysis	A visual display of the proportion of cases explained by each gene for the ROI. Analysis is available for P variant, P or LP variants or all variants, for both ALS and FTD, and for familial and sporadic forms of disease
	Population Studies	A list and description of populaiton studies included in the analyis of the ROI
Gene Browser		
Analysis for a particular gene of interest (GOI)	Comparison Plot	Comparse features such as allele frequency, proportion of familial cases and penetrance of variants in the GOI
	Gene Plot	A visual display of all variants observed in the GOI. Variants are coloured by their pathogenicity
	Variant Table	Phenotype information and annotations for all observed variants in the GOI
Summary		
	Summary	Summary information of overall results and details for each gene found to contain pathogenic or likely pathogenic variants
Downloads		
	Downloads	All of our data and code are open source and available for easy download and analysis
Annotate		
	Annotate	Users can annotate their VOIs from our dataset and download the results directly

Here, population penetrance (or the probability of disease given the allele) was calculated via equation 2.1:

$$P(D|A) = \frac{P(A|D) \times P(D)}{P(A)} \qquad \textit{Equation 2.1}$$

where P(D|A) is the likelihood of developing the disease for allele carriers; P(A|D) is the probability of having the allele given the disease, defined by the overall case AF calculated from the literature; P(D) is the probability of having the disease, defined by the lifetime risk; P(A) is the probability of having the allele, defined by the AF in the general population.

The gnomAD cohort was assumed to contain individuals presymptomatic for ALS and FTD, thereby representing the AF in the general population. The overall lifetime risk of disease (P(D)) is 1/400 for ALS (McGuire *et al.* 1996; Traynor *et al.* 1999; E. Beghi *et al.* 2007; Vázquez *et al.* 2008; Ryan, Heverin, *et al.* 2019), and 1/742 for FTD (Coyle-Gilchrist *et al.* 2016). Wilson 95% confidence intervals were calculated from the upper and lower bounds of the binomial proportions of P(A|D) and P(A). Where variants were observed in population studies and had an available gnomAD AF, population penetrance was separately calculated for the lifetime risk of developing ALS, FTD and ALS or FTD. Penetrance and confidence intervals were calculated using the R package binom v1.1-1 (Dorai-Raj 2014).

Variant penetrance was also calculated for ALS, FTD and ALS or FTD using the 'familial penetrance' method (Spargo *et al.* 2021), wherein a variant with increased lifetime penetrance will result in an increased proportion of variant carriers presenting with a positive family history of disease. Sibship size was estimated from the global 2018 total world fertility rate ("Databank.Worldbank.Org" 2021). The proportion of variant carriers presenting with a positive family history and AFs in familial and sporadic cases were calculated from the literature. The familial rate of ALS was estimated at 11.1% (Ryan *et al.* 2018) and the familial rate of FTD was estimated at 20.1% (Goldman *et al.* 2005).

Geographic heterogeneity

For each variant, between- and within-continent geographic heterogeneity was calculated from the region-specific variant AFs. Countries Iran, Israel and Turkey were assigned to the 'Middle East'. Russia was assigned to Europe. Sardinia in Italy and the Kii Peninsula in Japan were treated as stand-alone regions due to their historically unique ALS epidemiologies (A. Chiò *et al.* 2013; G. Logroscino and Piccininni 2019). Overall proportion across regions were calculated using both random and fixed effects models. Heterogeneity between regions was assessed via I^2 (the percentage of variation across regions attributable to heterogeneity) (Higgins and Thompson 2002; Higgins *et al.* 2003) and likelihood-ratio test p-value.

Significant heterogeneity may result from real geographic heterogeneity but may also indicate a differential reporting of, for example, common or intronic variants across studies. Analysis and visualisation was conducted using the R package meta v4.16-2 (Balduzzi, Rücker, and Schwarzer 2019). The journALS data browser has heterogeneity statistics available for all variants across all categories (country, phenotype, family history). In this study the geographic heterogeneity for pathogenic (P) or likely pathogenic (LP) variants in any category which has at least one variant carrier and at least two groups was tested. Heterogeneity was considered significant if I^2 was greater than 0.5 and the p-value was below the Bonferroni corrected p-value.

Segregation

Segregation was calculated via the counting meioses method of Jarvik and Browning (2016). The full-likelihood Bayesian (FLB) (Thompson, Easton, and Goldgar 2003) and cosegregation likelihood ratio (CLR) (Mohammadi *et al.* 2009) methods are more nuanced, accounting for reduced penetrance and age of onset; however they require gene penetrance classes and ages of onset for everyone in the pedigree. Gene penetrance classes would require assumptions which violate our agnostic approach and age of onset was not available for all family members for pedigrees collected from the literature. Counting meioses has been found to perform similarly to FLB and CSLR for identifying P variants (Rañola *et al.* 2018) and was compatible with the available data. Given the incomplete and age related penetrance of known ALS related variants, as per Jarvik and Browning (2016) a conservative approach was used which only takes into account affected individuals. A homozygous model of segregation was assumed if all affected genotyped individuals were homozygous, otherwise a heterozygous model of segregation was assumed. Meioses were calculated using the CoSeg R package v 0.49 (Rañola *et al.* 2018) and pedigrees were plotted using kinship2 v 1.8.5 (Sinnwell, Therneau, and Schaid 2014).

Age of onset

Where reported, the AOO of all variant carriers in the literature were collected. A variant associated with significantly early or late disease onset is indicative of a likely common underlying molecular mechanism in carriers of that variant, providing strong evidence of the pathogenicity of the variant in question. Kruskal-Wallis tests were conducted to identify whether the age of carriers of the variant of interest significantly differed from the reported ages of all other variant carriers reported in the literature. Variant carriers can be categorised based on phenotype (all, ALS, FTD), sex (all, male, female) and family history (all, sporadic, familial); thus there were 27 tests possible per variant.

Testing 27 categories for each variant would generate many tests containing zero individuals and many tests for which a statistically significant result is impossible. Therefore, only categories with the potential to yield a significant result were tested and corrected for. In any category with below six variant carriers, even if these are the earliest onset individuals in the data, it is impossible to return a statistically significant result after correcting for the number of tests that would be performed. Therefore, only categories with six or more individuals were tested and corrected for. This resulted in a p-value threshold of 9.75x10^{-5} (supplementary figure S2.1).

To reduce potential confounding factors only the index case from each family was considered in age comparisons. Density plots of AOO display the median AOO and confidence intervals were calculated with bootstrapping.

ACMG categorisation

Variant pathogenicity was assessed in accordance with the ACMG guidelines for variant interpretation (Richards *et al.* 2015). Many studies of specific conditions have previously outlined the necessity to modify ACMG guidelines and add interpretation where guidelines are non-specific (Kelly *et al.* 2018; Oza *et al.* 2018; Romanet *et al.* 2019; Brandt *et al.* 2020; O. Campuzano *et al.* 2020; Morales Ana *et al.* 2020).

To take an agnostic approach to variant categorisation, ACMG categories were treated in three classes. First, categories deemed to not be applicable to the current study were excluded (supplementary table S2.4). The second class represented categories which could be assessed independently (supplementary table S2.5), and third are dependent categories which relied on prior assessment of independent categories (supplementary table S2.6). Detailed methods for interpreting categories are outlined in supplementary tables S2.4-2.8 and supplementary figure S2.2.

Life expectancy

The life expectancy of each country for which a reported ALS or FTD patient had an available AOO was downloaded from the World Health Organisation (www.who.int). For both conditions, the AOO for all patients were regressed against the 2019 life expectancy at birth for the patient's country of origin. This analysis was also performed using the patient's reported sex and genes as covariates.

Results

Article screening

Initial assembly of literature from PubMed, ClinVar and HGMD identified 2,914 potentially relevant articles for further screening (supplementary figure S2.3). 1,028 of these were found to be relevant genetic studies of ALS or FTD. Potential population studies were manually filtered to find the most representative study for each country. 244 articles were designated as population studies (Supplementary File S1). Supplementary file S2 details the treatment of each article (supplementary files available at https://github.com/dohertymark/journALS/Supplementary_Material).

Within the literature 3,114 variants were reported in 356 genes and 479 pedigrees were recorded. After extracting variants present in these 356 genes from ALSdb, ALSVS and Project MinE, this study represents a complete analysis of 1,469,421 variants (supplementary figure S2.3). Full data and analysis of each variant is available on the journALS data browser (alsftd.tcd.ie).

Interobserver concordance

Following a double screen of 150 population studies and the generation of a consensus, three measures of concordance were calculated. Firstly, as a measure of how accurately observers identified variants in the literature, it was found that while there was an average of 7.19 variants per study in the consensus, independent screens had a false inclusion rate of 0.06 variants per screen and a false exclusion rate of 0.35 variants per screen. Secondly, the average AF of variants in the consensus was 1.58% while the average AF discordance was found to be 0.01% (standard deviation (SD) 0.12%). Finally, in identifying phenotype and genotype data of correctly identified individuals, independent screens were 97.14% accurate when excluding omitted data points and 91.5% accurate when treating omitted data as inaccurate.

Pathogenicity

112 variants in 21 genes were identified to have sufficient evidence to be classified as pathogenic or likely pathogenic. Of the original 28 ACMG categories, 3 pathogenic and 5 benign rules were deemed to not be applicable (supplementary table S2.4), 8 pathogenic and 7 benign rules were found to be capable of independent assessment on a first-round screen (supplementary table S2.5) and 5 pathogenic rules relied on the first-round screen for their assessment (supplementary table S2.6). 5 rules justified modified strength categories. All considered rules were applied at least once (supplementary figure S2.4).

Using the modified ACMG criteria, non-VUS classification was successfully applied to 10.6% of the 3,114 variants reported in the literature (supplementary figure S2.3). 1.1% are classified as Pathogenic, 2.5% are Likely Pathogenic, 3.1% are Benign and 2.1% are Likely Benign.

The ability to accurately classify variants improves as supporting and opposing evidence increases. Of the 2,844 VUS identified in the literature, 75% were reported in a single proband, with little other supporting or opposing evidence. When considering variants identified in more than one proband, 20.8% of variants received a non-VUS classification, rising to 62.8% when variants were identified in 10 or more probands. Of the 1,466,307 additional variants identified in Project MinE, ALSdb or ALSVS, 8.5% were found to be Benign and 0.55% are found to be Likely Benign.

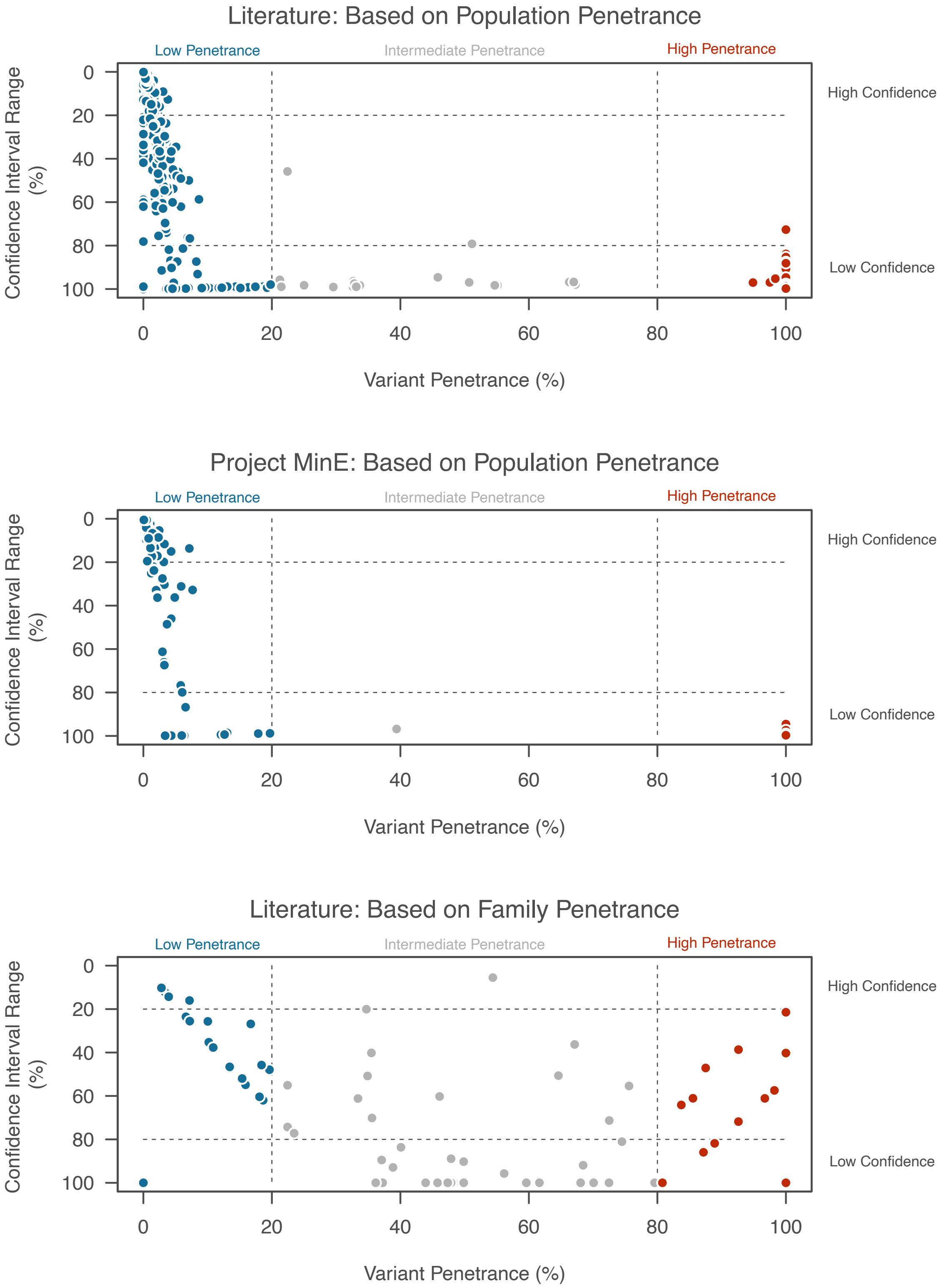
Penetrance

Variant penetrance is the probability that a variant carrier will develop disease during their lifetime. Where sufficient data were available, variant penetrance was estimated by two alternative means. The population penetrance method was found to produce higher confidence estimates of low penetrance variants in ALS and FTD and the familial penetrance produced higher confidence estimates of intermediate and high penetrance variants.

Population penetrance estimates were calculated to assess the likelihood of a variant carrier developing either ALS (1,253 variants) or FTD (791 variants), or the cumulative risk of developing ALS or FTD (649 variants). 57% of calculated variants were found to have low estimated lifetime penetrance for developing ALS, with 66% of these having high confidence, providing strong evidence against these being heterozygous, stand-alone pathogenic variants (figure 2.1.A). Of the remaining variants, 96% were found to be highly penetrant and 4% have intermediate penetrance; however, these were associated with large confidence intervals. Similar patterns were observed when penetrance was calculated based on the AFs in the Project MinE cohort (figure 2.1.B); or for the likelihood of developing FTD (supplementary figure S2.5.A) or the cumulative risk of developing ALS or FTD (supplementary figure S2.5.B). There is strong correlation between the penetrance of variants as calculated from the literature and from the Project MinE dataset, highlighting the reliability of our calculated AFs (supplementary figure S2.6). While penetrance estimates calculated from Project MinE, ALSdb or ALSVS have stronger correlations to each other than to penetrance estimates calculated from the literature (supplementary figure S2.7), this likely reflects that these datasets are uniformly of European ancestry while our literature data has a larger global component.

To examine the potential of large datasets to identify intermediate penetrance variants via the population penetrance method, the penetrance and confidence intervals that are expected to be obtained from a study as large as the target 15,000 cases of the Project MinE cohort were calculated. It was found that even a study of this size will struggle to confidently identify the penetrance of these variants (supplementary figure S2.8). While it can be difficult to significantly increase patient numbers in studies of rare diseases, it was found that increasing the size of the control cohort which is available improves the accuracy with which penetrance can be predicted (supplementary figure S2.9). Increasing the size of large publicly available genomics cohorts would not only benefit the study of ALS and FTD, but all genetic disorders.

There were a further 1,719 variants identified in the literature for which population penetrance was not calculated for any phenotype; either because they are too rare to appear in the designated population studies or do not have an associated gnomAD AF (e.g. for large INDELs). Consequently, the fact that 57% of variants were found to be low penetrance (with varying confidence) is biased towards variants with a higher AF and does not necessarily represent the entire dataset.



Literature: Based on Population Penetrance

Project MinE: Based on Population Penetrance

Literature: Based on Family Penetrance

Figure 2.1: ALS penetrance estimates

A) The ALS population penetrance estimates are shown here for 1,253 variants that had an AF calculated from the literature and an available gnomAD AF. 57% of these variants have low penetrance (below 20%) with 66% of these having high confidence. Due to the high lifetime risk of ALS and the low AF of each variant, this method struggles to confidently identify intermediate and high penetrance variants. B) The ALS population penetrance estimates calculated from the Project MinE case AF are shown for 372 variants which are present in the Project MinE data and the literature. C) The ALS familial penetrance estimates are shown for 534 variants which have a calculated AF in fALS and sALS cases.

The familial penetrance method of Spargo *et al.* (2021) was used to calculate variant penetrance for ALS (534 variants) (figure 2.1.C), FTD (104 variants) and ALS or FTD (10 variants) (supplementary figure S2.6). Rather than relying on the variant AF in the general population this method is instead based on the proportion of variant carriers that present with a positive family history. A similar 60% of variants were predicted to be low penetrance but generally lack confidence. However, while this method was less successful than the population penetrance method at confidently predicting low penetrance variants, it was more confident in predicting intermediate and high penetrance variants.

In examining the penetrance estimates of the *C9orf72* RE, these two alternative methods, with differing underlying data and assumptions, concurrently showed that carriers of a *C9orf72* RE have an approximate 50% chance of developing ALS during their lifetime (population penetrance method: 0.511 (95% CI: 0.208-1); family penetrance method: 0.5439 (95% CI: 0.5164-0.5714). This is in line with the observation that carriers of the *C9orf72* RE may instead experience cognitive impairment or FTD. Indeed, when considering the lifetime risk of RE carriers developing either ALS or FTD, both estimates were much closer to one (population penetrance method: 0.796 (95% CI: 0.319-1); family penetrance method: 1 (95% CI: 1-1)). Carriers of the *C9orf72* RE may not develop ALS but are likely to develop a disease along the ALS-FTD spectrum in the course of their lifetime.

Prevalence

Globally, it was found that reported variants in ALS and FTD P or LP genes can currently explain up to 68.7% of fALS, 51.2% of fFTD, 21.4% of sALS and 9.6% of sFTD; however, these figures are considerably lower when considering strictly P or LP variants (figure 2.2). Considering that most cases of both ALS and FTD are sporadic, a clear picture emerges that despite the high heritability of ALS and FTD the majority of cases still lack a clear genetic diagnosis.

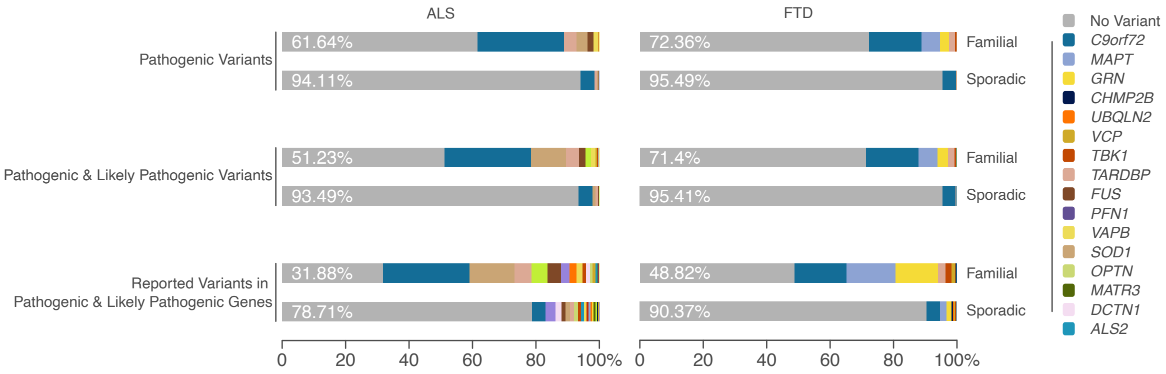


Figure 2.2: Global proportion of cases explained for ALS and FTD

The overall proportion of global ALS and FTD cases with an explained genetic cause varies if considering pathogenic variants, pathogenic and likely pathogenic variants, or all reported variants in genes with observed pathogenic or likely pathogenic variants.

Geographic distribution

While P and LP variants may be individually rare on a global scale, it is not uncommon for variants to form local hotspots, where a large proportion of cases are explained by a single variant. Between- or within-continent geographic heterogeneity was observed for 11% of the P or LP variants which were capable of being tested (figure 2.3.A). These variants either exhibit a gradient in their geographic distribution or are responsible for a large proportion of cases in a given area and were found to be virtually absent throughout the rest of the world. With a few notable exceptions, the majority of ALS and FTD genetic studies have come from countries with a majority European ancestry (figure 2.3 B). The same is true of current large ALS genomics efforts. Increasing the diversity in ALS and FTD studies would provide an opportunity to include these countries in future clinical trials and, given the observed geographic heterogeneity, to learn more about the biology underlying these conditions.

Oligogenic inheritance

There is strong evidence supporting the role of oligogenic inheritance in ALS; wherein ALS patients regularly harbour multiple variants in ALS associated genes (van Blitterswijk *et al.* 2012; Nguyen *et al.* 2018; Kuuluvainen *et al.* 2019; McCann *et al.* 2020). Recently, Nguyen *et al.* (2018) reported that a patient's development to either ALS or FTD is influenced by their combination of variants. Based on the extant literature they observed that patients with a *C9orf72* RE and a further variant in either *FUS*, *OPTN*, ANG or *SOD1* always presented with ALS while patients with a *C9orf72* RE and a *GRN* variant always presented with FTD. Several patients who contradict these observations were identified in the journALS database (supplementary table S2.9).

Discordant pedigrees

9 pedigrees which have a segregating P or LP variant, but in which there is an affected individual who does not have the variant in question were identified (supplementary table S2.10). Three of these pedigrees were discordant for the *C9orf72*:c.-45+163GGGGCC[>24] repeat expansion, one of these was also discordant for *TARDBP*:c.1144G>A(p.[A382T]). 5 pedigrees were discordant for segregating *SOD1* variants and the final had an incompletely segregating *TARDBP*:c.1055A>G(p.[N352S]) variant.

Life expectancy

An ALS patient's AOO was found to be significantly correlated with their country's life expectancy (slope=1.16, p-value=9.45x10^{-20}); however the same pattern was not observed for patients with FTD (slope=0.27, p-value=0.142) (figure 2.4). This indicates that a one year increase in a country's life expectancy delays ALS onset by an average of one year and 19 days. The same patterns are observed for both ALS and FTD when sex and gene are included as covariates (supplementary figure S2.10).

Genes carrying pathogenic and likely pathogenic variants

112 P or LP variants were identified in 21 genes (supplementary figure S2.11). The key features of each gene and its supporting evidence are outlined below.

ALS associated genes: dominant and recessive

SOD1

Of 244 reported *SOD1* variants, 49 were identified as P or LP causes of ALS. Variants are present in every exon and all are missense, with no B or LB missense variants observed (supplementary figure S2.11/S2.12 A). These variants are a major global cause of fALS (11.0%: 95% CI 9.7-12.5%) and a minor cause of sALS (0.9%: 95% CI 0.8-1.2%) (supplementary figure S2.13 B); and while several variants are rare globally, they can explain large proportions of cases in a particular region and thus have significant within- or between-continent geographic heterogeneity (figure 2.3 A).

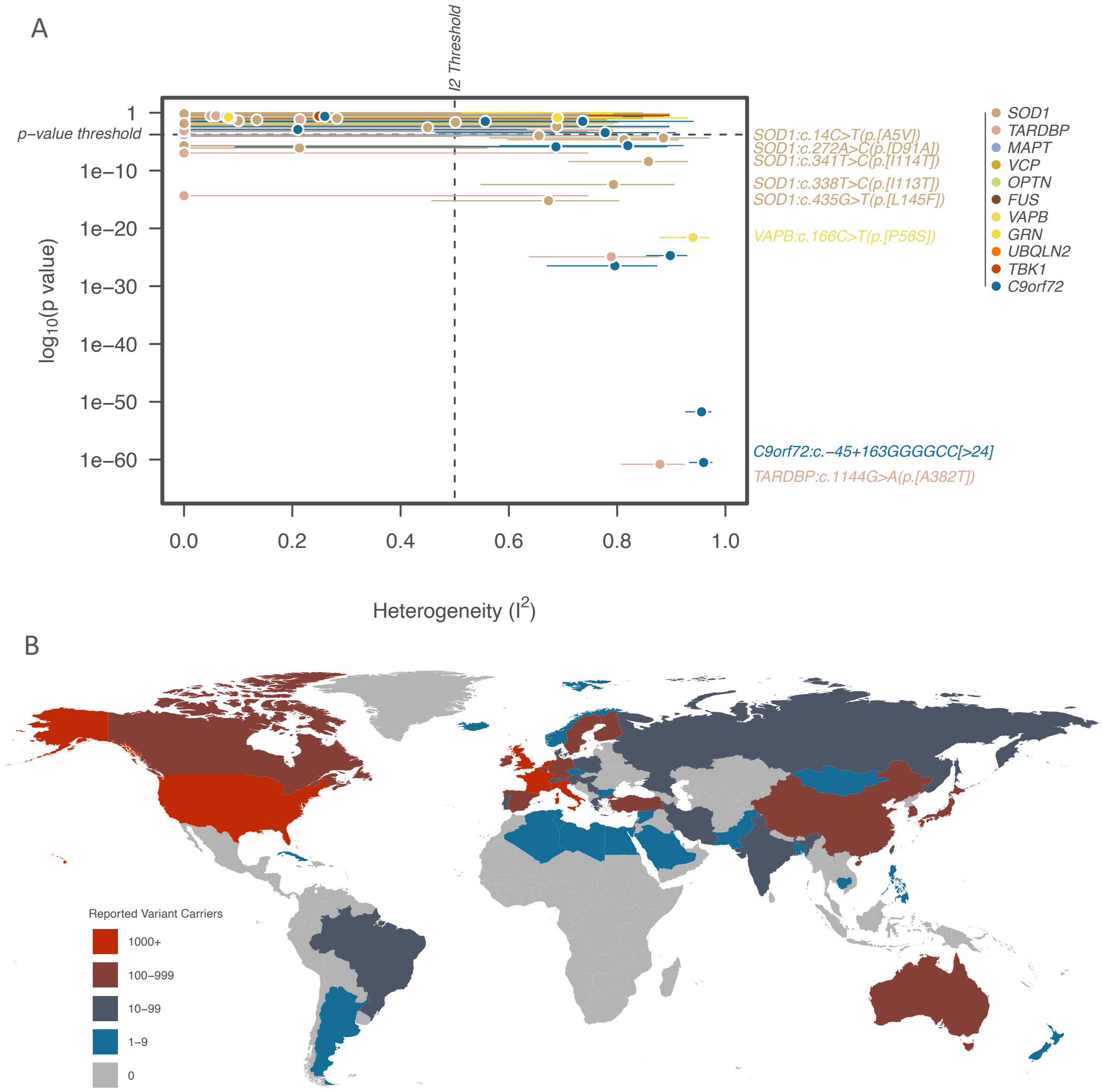


Figure 2.3: Geographic heterogeneity and distribution

A) Geographic heterogeneity is observed for 11% (8/72) of the pathogenic or likely pathogenic variants which were tested. Variants are tested in each category for which they have more than one variant carrier in population studies. Categories were defined on family histories (familial/sporadic), phenotype (ALS/FTD) and between and within continents. A variant may achieve significance in multiple categories and if so is only labelled once. Annotations appear for significant variants beside their lowest p-value in any category. B) The distribution of reported variant carriers is not evenly distributed globally. With the exceptions of China, South Korea and Japan, the majority of reported variant carriers are from countries primarily of European ancestry. (Note: carrier counts may include the same individuals across multiple studies).

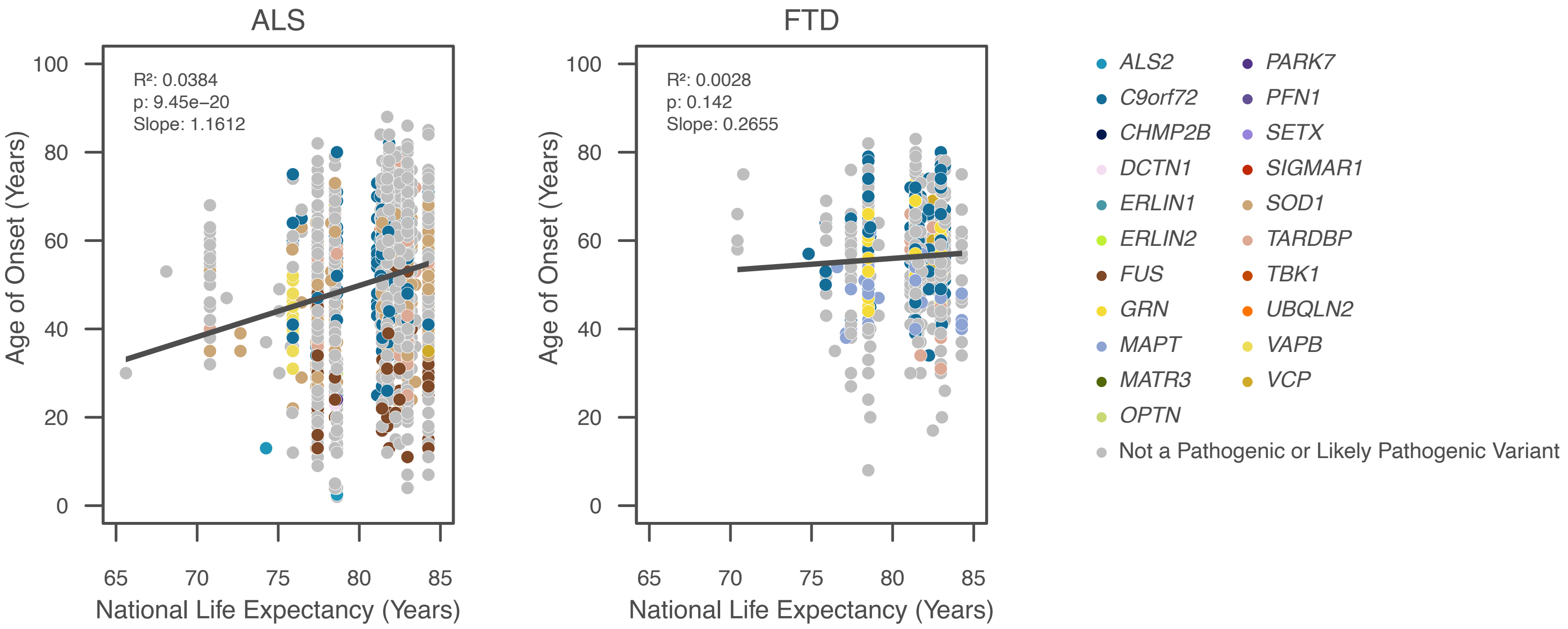


Figure 2.4: Regression of life expectancy and age of onset

When patient's reported AOO is regressed against the life expectancy in their reported country a significant correlation is observed for ALS (slope=1.16, p-value=9.45x10^{-20}); however, no significant relationship is observed for FTD (slope=0.27, p-value=0.14).

ALS patients with P or LP variants in *SOD1* have an earlier median AOO (48.5: 95% CI 46.5-50) than non-*SOD1* variants (55: 95% CI 55-56) (supplementary figure S2.14 A), although this does not appear to be the case for all *SOD1* P or LP variants. Carriers of *SOD1* VUS also present with moderately early AOO, indicating the presence of further P variants currently with insufficient supporting evidence. With the exception of the homozygous *SOD1*:c.272A>C(p.[D91A]), all *SOD1* variants were found to be dominant. While three ALS-FTD *SOD1* variant carriers are reported, two of these individuals carry a LB or intronic variant, indicating an alternative genetic cause.

OPTN

A heterozygous dominant missense variant and a homozygous frameshift variant in *OPTN* were identified as LP causes of ALS (supplementary figure S2.11/S2.12 B); explaining below 1% of global fALS cases (supplementary figure S2.13 B). All 44 reported missense VUS are reported in heterozygosity. While there is evidence that *OPTN* frameshift and truncating variants (FTVs) only cause ALS when homozygous, this is inconclusive as only 21% of carriers of LOF VUS are observed in homozygosity.

ALS associated genes: dominant

FUS

11 P or LP heterozygous variants were identified in, or bordering, the nuclear-localisation sequence in the final two exons of *FUS* (supplementary figure S2.11/S2.12 C). These variants are frequently *de novo* and are associated with early AOO (27 95% CI: 31-35), although in rare instances healthy carriers have been observed into their 70s (Yan *et al.* 2010). *FUS* P and LP variants are observed in fALS (2.3% 95% CI: 1.7-3.0) and sALS (0.2% 95% CI: 0.2-0.4%) (supplementary figure S2.13 B). Carriers of P or LP variants present with ALS with the exception of one ALS-FTD patient and one FTD patient (supplementary figure S.2.12 C). *FUS* variants are associated with significantly early onset (supplementary figure S2.14 C).

VAPB

A single *VAPB* P variant (supplementary figure S2.11/S2.12 D) was found to be a highly geographically heterogeneous (figure 2.3) cause of ALS; explaining 33% (95% CI 21-48%) of fALS cases in Brazil and rarely observed in the rest of the world. This variant is associated with significantly early AOO (median AOO 42 95% CI: 41-46) (supplementary figure S2.14 D). While 5 additional missense VUS have been reported, the presence of B and LB missense variants indicates that *VAPB* is tolerant of missense variants and pathogenicity should not be assumed.

SETX

A *SETX* heterozygous missense variant (supplementary figure S2.11/S2.12 E) was identified as a rare LP cause of ALS. While AOO is only given for one of the six identified carriers of this variant, all are described as having onset before age 30 and very slow progression. While *SETX* VUS variants are reported in approximately 3% of fALS and sALS cases (supplementary figure s11 B); the observation, in reference datasets, of B and LB missense variants throughout the gene, demonstrate that pathogenicity of these VUS variants should not be assumed. The LP variant itself is a rare cause of ALS, explaining below 0.2% of fALS cases globally.

MATR3

A *MATR3* heterozygous missense variant (supplementary figure S2.11/S2.12 F) was found to be a LP cause of slowly progressive ALS in a North American pedigree (Johnson *et al.* 2014). The absence of the LP variant in population studies (supplementary figure S2.13 B), namely Project MinE, ALSdb and ALSVS, demonstrates that this is a rare cause of ALS. Missense VUS have been reported in additional ALS cases (supplementary figure S2.14 F); however, the identification of both rare non-segregating missense variants (Saez-Atienzar *et al.* 2020) and LB missense variants is evidence that pathogenicity should not be assumed for rare missense *MATR3* variants.

ERLIN2

A heterozygous missense variant (supplementary figure S2.11/S2.12 G) was found to be the LP cause of ALS in a French family presenting with spastic paraplegia progressing to ALS (Muratet *et al.* 2019). The same study reported one homozygous and two heterozygous additional *ERLIN2* VUS variants; two of which were identified in individuals with ALS preceded by spastic paraplegia. This LP variant is absent in ALSdb and ALSVS and present in a single individual in Project MinE indicating that it is not a common cause of ALS in European populations.

DCTN1

A LP *DCTN1* variant (supplementary figure S2.11/S2.12 H) was found to segregate in a large North-American ALS pedigree (Puls *et al.* 2003). A further 61 primarily missense VUS variants are reported in *DCTN1*. While two of these variants (*DCTN1*:c.175G>C(p.[G59R]) and *DCTN1*:c.3302G>A(p.[R1101K])) have weak and supporting segregation evidence respectively, the presence of B missense variants in *DCTN1* demonstrates that pathogenicity should not be assumed even for rare missense *DCTN1* variants. Population studies have confirmed that this LP variant is an infrequent cause of ALS (supplementary figure S2.13 B).

PFN1

Two heterozygous missense *PFN1* variants (supplementary figure S2.11/S2.12 I), were identified as the LP cause of ALS in four adult onset ALS pedigrees and a further fALS case (C.-H. Wu *et al.* 2012). *PFN1* variants are a rare cause of ALS as both variants are absent in our designated population studies, Project MinE and ALSVS, while *PFN1*:c.448T>G(p.[C150G]) is present in a single patient in ALSdb. Additionally, 13 *PFN1* VUS including 8 missense variants are reported; however, these should be interpreted with caution as benign *PFN1* missense variants are also identified.

ALS associated genes: recessive

ALS2

Three frameshift, one stop-gain and one splice donor variant were identified as *ALS2* LP variants; all are homozygous and associated with extremely early onset ALS (supplementary figure S2.11/S2.12 J). 11 further stop-gain, splice site or frameshift VUS are reported in ALS patients; typically in homozygosity and with early AOO. No B or LB stop-gain or frameshift variants are present; however, a B splice donor variant is identified which is frequently homozygous in gnomAD. While this variant (ENST00000496244.1:*ALS2*:n.352+1A>G) is expressed in a non-protein coding transcript, future splice donor variants should nonetheless be interpreted with caution. 34 typically heterozygous missense *ALS2* VUS present with an age profile resembling typical adult onset ALS. The presence of B and LB missense variants in *ALS2* encourages very cautious interpretation of missense variants.

PARK7

A homozygous stop gain variant in *PARK7* (supplementary figure S2.11/S2.12 K) was found to be a LP cause of Parkinsonism and ALS in a Turkish pedigree with a history of consanguinity (Özoğuz *et al.* 2015; Hanagasi *et al.* 2016). Two additional homozygous VUS in this gene show strong segregation in an Italian family presenting with early-onset Parkinsons, dementia and ALS (Annesi *et al.* 2005). The absence of this LP variant Project MinE, ALSdb and ALSVS data indicate that this is not a common cause of ALS in populations of European ancestry.

SIGMAR1

A homozygous *SIGMAR1* variant (supplementary figure S2.11/S2.12 L) has been identified as the LP cause of slowly progressive juvenile ALS in a Saudi Arabian family with a history of consanguinity (Al-Saif, Al-Mohanna, and Bohlega 2011). This variant has not been observed elsewhere in the literature or in the Project MinE, ALSdb and ALSVS datasets. While other exonic and 3'UTR *SIGMAR1* variants have been observed in ALS and FTD cases (supplementary figure S2.14 L), these have mostly been reported in heterozygosity and have lacked sufficient evidence to be deemed P or LP.

ERLIN1

A homozygous *ERLIN1* variant (supplementary figure S2.11/S2.12 M) strongly segregates in a consanguineous Turkish pedigree exhibiting early-onset, slowly progressive ALS (Tunca *et al.* 2018).This variant is absent from the Project MinE, ALSdb and ALSVS data, and is homozygous in a single ALSVS patient.

ALS associated genes: X-linked

UBQLN2

UBQLN2:c.1490C>A(p.[P497H]) is as an X-linked LP variant in a large pedigree presenting without male-to-male transmission (H.-X. Deng *et al.* 2011). Male carriers in this family had an early median AOO of 33 (95% CI 25-47) and all had developed ALS by age 49, while female carriers had a median AOO of 49.5 (95% CI 42-60) and only 83% had developed ALS by age 71. This variant has additionally been identified in a patient from the UK and a 33 year old male with ALS-FTD from Italy; however, it explains less than 0.1% of familial cases globally (supplementary figure S2.13 B). The 41 additional *UBQLN2* VUS in the literature are primarily reported in individuals with ALS. There are 6 VUS displaying some level of segregation and all fall between amino acids 487 and 509. The AOO of male VUS carriers (30 95% CI: 26-54); although not statistically significant, appears earlier than female VUS carriers (53 95% 43-58). Collectively this indicates the presence of additional potentially pathogenic variants in the region flanking *UBQLN2*:c.1490C>A(p.[P497H]); however, individually these variants currently lack sufficient supporting evidence.

ALS and FTD associated genes: dominant

C9orf72

The hexanucleotide *C9orf72* RE was found to be a major cause of both ALS and FTD (supplementary figure S2.11/S2.12 O). The RE exhibits significant geographic heterogeneity (figure 2.3), explaining above 30% of fALS and 5% of sALS cases in countries with primarily European ancestry while being virtually absent in Asia. ALS patients carrying a *C9orf72* RE have a delayed AOO (56 95% CI: 54.6-57) relative to other ALS patients in the database (52 95% 51-53), while *C9orf72* carriers presenting with FTD have the same AOO (58 95% CI: 57-59), as other FTD patients in the database (57 95% CI: 56-58). While missense variants have been observed in *C9orf72*, there is insufficient evidence supporting their pathogenicity.

TBK1

In *TBK1* a P disruptive in-frame deletion and a LP splice donor variant were identified in cases along the ALS-FTD spectrum (supplementary figure S2.11/S2.12 P). There are a further 42 LOF VUS in the literature, four of which demonstrate some degree of segregation. Rare variant burden analysis has previously identified *TBK1* LOF variants as being significantly enriched in cases along the ALS-FTD spectrum (Cirulli *et al.* 2015; Freischmidt *et al.* 2015); however, the observation of three Project MinE LOF variants and one in-frame deletion with higher control than case AF, demonstrates that pathogenicity of individual *TBK1* LOF variants should not be assumed. *TBK1* P and LP variants explain below 1% of fALS and fFTD and below 0.1% of sALS cases globally (supplementary figure S2.13 B).

TARDBP

10 P or LP missense variants were found in the C-terminal glycine rich final exon of *TARDBP* (supplementary figure S2.11). These variants present with phenotypes spanning

the ALS-FTD spectrum; although this may be variant dependent (supplementary figure S2.12 Q).

P and LP *TARDBP* variants are a global causes of fALS (4.0%: 95% CI 3.2%-5.0%) and fFTD (2.0% 95% CI 1.0-3.8%) and are also observed in sALS (0.9% 95% CI 0.7-1.1%) and sFTD (0.2% 95% CI 0.03-0.9%). Geographic heterogeneity is observed for *TARDBP*:c.1144G>A(p.[A382T]) which is present in a large proportion of Sardinian fALS (32% 95% CI 23.0-42.1%) and sALS (20.4% 95% CI 15.8-25.6%) cases but is virtually absent throughout the rest of the world.

VCP

Patients carrying one of the seven heterozygous LP *VCP* missense variants present with ALS, FTD, inclusion body myopathy, Paget disease of bone, or various combinations of these phenotypes (supplementary figure S2.11/S2.12 R). Phenotype can vary for carriers of the same variant and even within pedigrees. Globally these variants explain below 1% of fALS and fFTD (supplementary figure S2.13 B). While these variants are clustered in three exons, a LB missense variant is also identified in this region, prompting cautious interpretation of *VCP* missense variants, particularly in the absence of a family history on the inclusion body myopathy with Paget disease of bone and frontotemporal dementia (IBMPFD) spectrum.

FTD associated genes: dominant

GRN

10 P or LP *GRN* variants were identified almost exclusively in FTD patients (supplementary figure S2.11/S2.12 T), explaining 3.3% of fFTD and 0.1% of sFTD (supplementary figure S2.13 B). As much as 13.5% of fFTD and 1.5% of sFTD are potentially explained when considering all reported *GRN* VUS (supplementary figure S2.13 C). A single patient presenting with ALS is found to carry a *GRN* P variant (Cannon *et al.* 2013) and a pedigree presenting with FTD-MND is identified; however, they also carry a concurrent *C9orf72* pathogenic repeat expansion (Lashley *et al.* 2014). With the exception of one missense variant, all P and LP variants identified in *GRN* are LOF variants and are identified all throughout the gene (supplementary figure S2.11/S2.12).

MAPT

Six heterozygous P or LP *MAPT* variants, including five missense and one intronic variant, were found to explain 6% of fFTD and below 0.1% of sFTD (supplementary figure S2.11/S2.12 T). As much as 15% of fFTD and 2% of sFTD are potentially explained when considering the additional 64 reported VUS, of which 15 have segregation evidence ranging from supporting to strong. All P and LP variants are reported in the microtubule-binding domain. FTD patients carrying *MAPT* variants experience moderately early AOO (47 95% CI: 46-50) (supplementary figure S2.14 T).

CHMP2B

A C-terminal truncating LP *CHMP2B* splice acceptor variant was identified in a well-characterised Danish FTD pedigree (J. Brown *et al.* 1995; Skibinski *et al.* 2005; Holm *et al.* 2007; Urwin *et al.* 2010; Stokholm *et al.* 2013). While other missense, intronic and UTR VUS variants are reported throughout *CHMP2B* in ALS and FTD patients (supplementary figure S2.11-S2.12.P), not much inference can be made from these, as missense, intronic and UTR B and LB variants are also observed throughout the gene.

Study limitations

This study represents a catalog and analysis of 30 years of genetics research in ALS and FTD and a framework for the future interpretation of novel variants and variants with additional available evidence; nonetheless, there are limitations to this research.

Common variants are inconsistently reported across studies, are more susceptible to population stratification, and, where reliable associations are observed, tend to represent risk factors rather than definitively causal variants. Consequently, our analysis is biased towards the interpretation of rare rather than common variants, and GWAS studies remain the most appropriate reference for the interpretation of common variants.

With the exception of the *C9orf72* RE, complex variants such as REs, chromosomal rearrangements and copy number variants have been annotated in our dataset but have been omitted from analysis. These variants are typically not annotated in genomics population databases and do not have *in silico* predictions and therefore cannot be integrated in our analysis pipeline. Where expansions such as the *ATXN2* RE have reliably been associated with ALS, these typically represent risk factors rather than causal variants and this increased risk is typically over a range of allele lengths rather than a single variant.

The analysis presented here is biased against more recently reported variants which have not had the same length of time to accumulate supporting and conflicting evidence.

The database relies on accurate reporting in the literature. If a mistake is present in a study this will be reflected in our analysis; however, it is hoped that this is negated by the accumulation of evidence as the same error is unlikely to pervade multiple studies.

In pursuing accuracy and clarity it has been necessary to omit phenotype data when it is reported for a cohort and cannot be deduced for an individual variant carrier.

This study is constrained by what is reported in the literature. There are, for example, instances where a variant is reported to segregate fully in a pedigree but this is not shown. These instances are noted in the database but cannot be included in this analysis without further information.

This study represents an analysis of variants rather than genes; consequently, the 21 genes harbouring P or LP variants should not be considered an exhaustive list of ALS and FTD genes. *NEK1* and *KIF5A* have both recently been identified as ALS associated genes (Kenna *et al.* 2016; Nicolas *et al.* 2018); however, these have been identified through exome burden studies, meaning that while these genes are reliably associated, there is not a particular definitive identifiable pathogenic variant.

Future integration

The aim of this study is to create a useful, manually curated and uniformly analysed synopsis of the last 30 years of genetics research in ALS and FTD. Supplementary table S2.11 provides a set of suggested minimal reporting guidelines that would greatly aid in incorporating future genetic studies into the journALS data browser, enabling it to become a regularly updated database and analysis of the most up-to-date research in the field.

Discussion

The journALS data browser aims to serve three functions: to be a useful database of reported ALS and FTD genetic variants, to be an analysis of the last 30 years of research in the field and to continue as a resource which can be rapidly updated as new genetic studies, annotations, or analyses emerge. 1,028 relevant genetic studies of ALS or FTD are amalgamated, annotated and analysed, identifying 112 P or LP variants in 21 genes.

A number of interesting features emerge from this analysis, serving to highlight the complexity of the genetics of these two conditions.

Despite the fact that the extant literature has been amalgamated, annotated with modern references datasets and analysed in accordance with the ACMG guidelines, the majority of observed variants remain classified as VUS. It is demonstrated that the likelihood of variant classification increases with increasing case reports. To aid this, a set of suggested minimal reporting guidelines is provided, so that future studies can be integrated with this work, to aid in further classifying novel and previously reported variants. Even variants that have previously been found to segregate in large pedigrees can be found to be likely benign as new evidence emerges (Johnson *et al.* 2014; Saez-Atienzar *et al.* 2020); therefore, it is vital to constantly reassess the evidence and analysis of ALS and FTD genes and variants.

It is important to note that the 21 genes identified do not represent an exhaustive list of ALS and FTD genes. It is likely that some VUS variants in the database represent truly pathogenic variants which do not yet have sufficient evidence. Additionally, our pathogenicity analysis is performed on a per-variant rather than per-gene basis. Genes such as *NEK1* and *KIF5A* have recently been reliably identified as ALS associated genes through exome burden studies. While the genes themselves are reliably associated, no particular variant is identified to have sufficient evidence to be reported as P or LP.

It is found that the high lifetime risk for ALS and FTD, the low AF of each variant and the typically small reported case numbers for each variant create difficulties in confidently calculating variant penetrance in ALS and FTD. Indeed, even a study as large as the 15,000 cases targeted by the Project MinE cohort will struggle in confidently estimating the penetrance of rare variants even with a control cohort as large as gnomAD. While increasing the number of patients in a study may not always be feasible, this method of estimating penetrance is found to benefit significantly from an increase in the size of available population control cohorts. Increasing the size of these publicly available genomics resources would benefit not only the study of ALS and FTD but all genetic diseases.

The familial penetrance method outperforms the population penetrance method in identifying intermediate penetrance variants but with significant caveats. This method relies on accurate classification of the proportion of variant carriers presenting with a positive family history. The manually curated dataset enables this to be calculated from the literature; however, this information will be harder to accurately ascertain for future novel variants. This is further complicated by the observation that as more detailed registers are kept for longer periods, the proportion of cases classified as familial increases (Ryan *et al.* 2018); making this penetrance estimate prone to underestimation. This method does indicate that intermediate penetrance variants play a greater role in the genetic architecture of ALS than FTD and may therefore explain the higher proportion of fFTD than fALS cases.

Several discordant pedigrees are observed, wherein one or more affected relatives do not carry the P or LP variant that otherwise segregates in the family. In some cases this is explained by an alternately segregating variant in the family. Where there is no other segregating variant in the pedigree, possible explanations include the presence of unidentified variants, or the presence of an environmental factor which increases the likelihood of developing ALS in the family, either the discordantly segregating variant or the environmental factor may be sufficient by themselves and together they greatly increase the likelihood of developing ALS. The results may also be attributable to somatic mosaicism wherein individuals carry the variant in some tissues, such as in the nervous system, but not others, such as in the blood. There is the possibility of human error in processing these samples; however, in several cases the chances of this have been reduced by resampling individuals and testing samples independently at multiple labs. It is also possible that these are not truly causative variants; however, the weight of evidence in their favour makes this unlikely.

It is common for ALS and FTD variants to exhibit significant geographic heterogeneity. This is important because, although a variant may be rare globally, it can be responsible for a large number of cases in a particular region, influencing the planning and execution of clinical trials in that region.

It is observed that the genetics underlying ALS and FTD remain understudied in large portions of the world. Increasing genetic screening and studies in these areas will not only improve global parity but will also, given the commonly observed global heterogeneity, provide an opportunity to identify new ALS and FTD genes and pathways.

Byrne *et al.* (2013) observed a relationship between the mean ALS AOO in a country and the life expectancy in that country. Given the large variability in ALS onset within any country, the current research provides an important demonstration that this result is also obtained when using all available patient AOOs rather than mean AOO. A significant relationship is observed, with an increase of one year in a country's life expectancy delaying ALS onset by an average of one year and 19 days. This indicates either that healthy environmental factors which extend life expectancy delay ALS onset, or that unhealthy environmental factors which shorten life expectancy accelerate ALS onset. While this result does demonstrate a significant overall trend, there is still significant variability in the data that is not explained by life expectancy (R^2=0.0384); consequently this is not a useful metric for predicting an individual's AOO. Interestingly, the same trend is not observed for FTD, indicating a different etiology that is unaffected by life expectancy.

This analysis provides a reorientated view of several ALS and FTD genes. Unfortunately, it remains the case that the majority of cases of ALS and FTD lack a genetic diagnosis. As we move into a future where genetic counselling becomes an increasingly common and clinical trials and enrolling based on genetic status, this research will hopefully be a supportive tool in these endeavours, a tool with the ability to be adapted into the future.

Chapter 3

Identifying repeat expansions in neurological disorders

Introduction

Chapter 2 explored the role of pathogenic variants in ALS and FTD. Of the 112 pathogenic or likely pathogenic variants identified in that research, one RE, *C9orf72* GGGGCC, was identified as a pathogenic cause of ALS and FTD. Historically, due to the nature of their structure, it has not been possible to accurately classify REs from sequencing data; consequently, one limitation of the previous chapter is that, with the exception of the *C9orf72* RE, it was necessary to omit analysis of REs and STRs. Regardless, other additional STRs, including in genes *ATXN2* (Sproviero *et al.* 2017; Adriano Chiò *et al.* 2015; M.-D. Wang *et al.* 2014; Daoud *et al.* 2011), *NIPA1* (Tazelaar *et al.* 2019; Blauw *et al.* 2012), *ATXN1* (Lattante *et al.* 2018; Tazelaar *et al.* 2020) and *HTT* (Dewan *et al.* 2021), have been identified as ALS risk factors.

As outlined in Chapter 1, REs are a major cause of not only ALS and FTD, but many neurological diseases, with STR variability in more than 50 genes now linked to various neurological disorders (table 3.1).

Table 3.1: Disease-associated repeat expansions

			Repeat			
Gene	**Disease**	**Motif**	**Threshold**	**Genomic Location**	**hg19 Coordinates**	**Reference**
AFF2	FRAXE	CCG	≥200	5’ UTR	chrX:147582151-147582211	Knight *et al.* 1993
AR	SBMA	CAG	≥38	Coding	chrX:66765160-66765226	La Spada *et al.* 1991
ARX	EIEE1	GCG	≥23	Coding	chrX:25031777-25031806	Strømme *et al.* 2002
ATN1	DRPLA	CAG	≥48	Coding	chr12:7045879-7045936	Koide *et al.* 1994
ATXN1	SCA1	CAG	≥39	Coding	chr6:16327864-16327954	Orr *et al.* 1993
ATXN10	SCA10	ATTCT	≥280	Intron	chr22:46191234-46191304	Matsuura *et al.* 2000
ATXN2	SCA2/ALS	CAG	≥32	Coding	chr12:112036753-112036822	Pulst *et al.* 1996 Daoud *et al.* 2011
ATXN3	SCA3	CAG	≥55	Coding	chr14:92537353-92537386	Kawaguchi *et al.* 1994
ATXN7	SCA7	CAG	≥36	Coding	chr3:63898360-63898390	Lindblad *et al* . 1996
ATXN8OS	SCA8	CTG.CAG	≥74	3’UTR	chr13:70713485-70713515	Koob *et al.* 1999
C9orf72	ALS	GGGGCC	≥30	5’UTR/ Intronic	chr9:27573526-27573544	Renton *et al.* 2011 DeJesus-Hernandez *et al.* 2011
CACNA1A	SCA6	CAG	≥20	Coding	chr19:13318672-13318711	Zhuchenko *et al.* 1997
CBL	JS	CCG	≥100	5’UTR	chr11:119077000-119077032	Jones *et al* . 1995
CNBP	DM2	CCTG/CAGG	≥50	Intron	chr3:128891419-128891499	Liquori *et al* . 2001
CSTB	EPM1	C4GC4GCG	≥30	Promoter	chr21:45196324-45196360	Lalioti *et al* . 1997
DIP2B	ID	GGC	≥150	5-UTR	chr12:50898785-50898805	Winnepenninckx *et al* . 2007
DMPK	DM1	CAG	≥50	3’UTR	chr19:46273462-46273522	Mahadevan *et al* . 1992
FMR1	FXTAS	CGG	≥200	5’UTR	chrX:146993568-146993628	Oberlé *et al* . 1991 Verkerk *et al* . 1991
FOXL2	BPES	GCN	≥19	Coding	chr3:138664864-138664977	De Baere *et al.* 2001
FXN	FRDA	GAA	≥66	Intron	chr9:71652203-71652220	Campuzano *et al* . 1996
GIPC1	OPDM2	CCG	≥97	5’UTR	chr19:14606854-14606886	Deng *et al.* 2020
GLS	GDPAG	GCA	≥680	5’UTR	chr2:191745598-191745646	van Kuilenburg *et al.* 2019
HOXA13	HFG	GCN	≥24	5’UTR	chr7:27239299-27239410	Deng *et al.* 2020 Goodman *et al.* 2000 Utsch *et al.* 2002
HOXD13	SD1	GCN	≥22	5’UTR	chr2:176957787-176957831	Akarsu *et al.* 1996
HTT	HD	CAG	≥35	Coding	chr4:3076603-3076660	MacDonald *et al* . 1993
JPH3	HDL2	CTG/CAG	≥41	3’UTR	chr16:87637893-87637935	Margolis *et al.* 2001
LRP12	OPDM1	CGG	≥90	5’UTR	chr8:105601281-105601290	Ishiura *et al.* 2019
MARCHF6	FAME3	TTTTA(TTTCA)$_N$TTTTA	≥660	Intron	chr5:10356460-10356519	Florian *et al.* 2019
NIPA1	ALS	CGC	≥8*	Coding	chr15:23086367-23086390	Blauw *et al.* 2012
NOP56	SCA36	GGCCTG	≥650	Intron	chr20:2633386-2633403	Kobayashi *et al.* 2011
NOTCH2NLA	NIID	CGG	≥61	5’UTR	chr1:145209324-145209344	Ishiura *et al* . 2019 Sone *et al* . 2019 Tian *et al* . 2019
NUTM2B	OPML1	CGG/CCG	≥40	Noncoding gene	chr10:81586140-81586160	Ishiura *et al* . 2019
PABPN1	OPMD	GCN	≥12	Coding	chr14:23790681-23790701	Richard *et al.* 2017 Brais *et al.* 1998
PHOX2B	CCHS	GCN	≥24	Coding	chr4:41747989-41748049	Amiel *et al.* 2003
PPP2R2B	SCA12	CAG	≥51	5’UTR	chr5:146258290-146258320	Holmes *et al.* 1999
RAPGEF2	FAME7	TTTTA(TTTCA)$_N$TTTTA	≥22	Intron	chr4:160263709-160263768	Ishiura *et al.* 2018
RFC1	CANVAS	AAAAG	≥400	Intron	chr4:39350044-39350099	Cortese *et al* . 2019
RUNX2	CCD	GCN	≥27	Coding	chr6:45390433-45390486	Mundlos *et al.* 1997
SAMD12	FAME1	TTTTA(TTTCA)$_N$TTTTA	≥440	Intron	chr8:119379055-119379157	Zeng *et al* . 2019 Ishiura *et al.* 2018
SOX3	XLMR	GCN	≥26	Coding	chrX:139586483-139586527	Laumonnier *et al.* 2002
STARD7	FAME2	TTTTA(TTTCA)$_N$TTTTA	≥40	Coding	chr2:96862809-96862858	Corbett *et al.* 2019
TBP	SCA17	CAN	≥47	Coding	chr6:170870994-170871105	Koide *et al* . 1999
TCF4	FECD	CTG	≥70	Intron	chr18:53253386-53253461	Mootha *et al.* 2015 Mootha *et al* . 2014
TNRC6A	FAME6	TTTTA(TTTCA)$_N$TTTTA	NA	Intron	chr16:24624761-24624850	Ishiura *et al* . 2018
YEATS2	FAME4	TTTTA(TTTCA)$_N$TTTTA	≥800	Intron	chr3:183429976-183430091	Yeetong *et al* . 2019
ZIC2	HPE5	GCN	≥25	Coding	chr13:100637703-100637746	Brown 2001

* The *NIPA1* allele is an ALS risk factor rather than being a strict pathogenic cutoff

Pathogenic thresholds listed are the pathogenic range in the literature; however, some are subject to debate

Detecting repeat expansions

PCR in combination with capillary electrophoresis remains the gold-standard means of accurately classifying the length of STRs and short REs. However, for many REs, including *C9orf72*, the pathogenic repeat range can be kilobases (kbs) long. This range exceeds what is possible to amplify with standard PCR, and so requires the application of either Southern blotting or repeat-primed PCR (rpPCR).

To perform Southern blotting, the DNA is enzymatically fragmented and fractionated by electrophoresis. The fragments are blotted onto a porous membrane, maintaining their relative positions. The membrane is submerged in a solution containing a labelled probe which will bind to its complementary DNA sequence and the position can be visualised (figure 3.1). While Southern blotting is capable of measuring kilobase size REs, the method is time consuming and low throughput; additionally, the *C9orf72* RE is known to be unstable in blood, exhibiting somatic mosaicism, and can result in unclear Southern blots (Beck *et al.* 2013; van Blitterswijk *et al.* 2013; DeJesus-Hernandez *et al.* 2011).

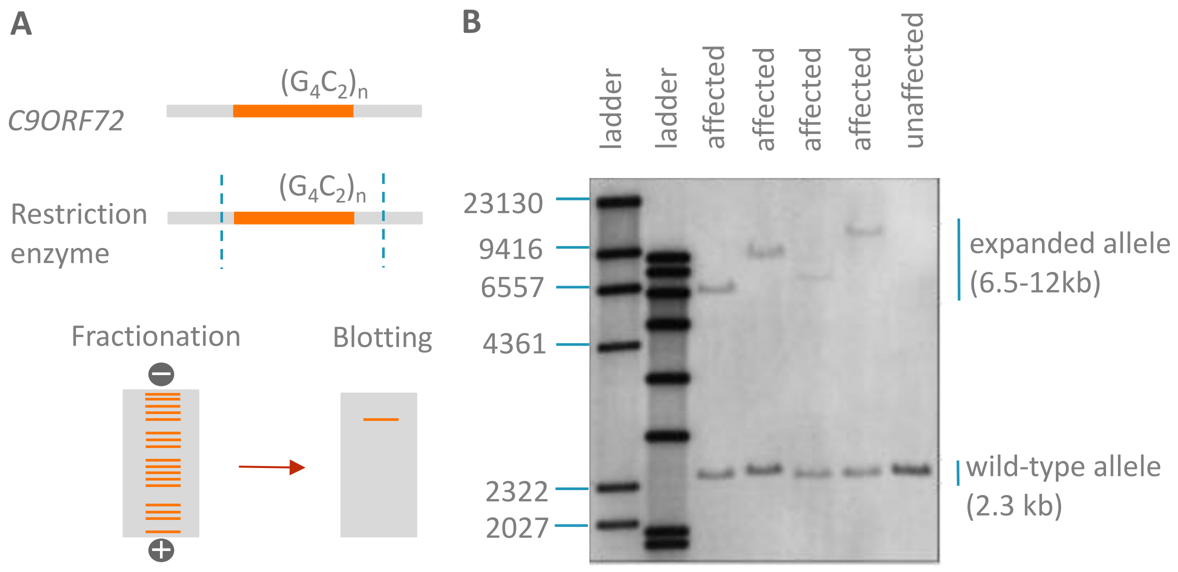


Figure 3.1: Southern blotting

A) The DNA is fragmented using restriction enzymes and separated using gel electrophoresis. The gel is soaked in alkali to denature the DNA. Fragments are transferred to a positively charged gel where they hybridize, maintaining their position. The membrane is incubated with nonspecific probes to saturate nonspecific binding sites. The membrane is incubated with labelled probe DNA with sequence complementary to the sequence of interest. The labelled DNA can be visualised. B) A Southern blot of 4 *C9orf72* positive individuals and one *C9orf72* negative individual (adapted from (DeJesus-Hernandez et al. 2011)).

An alternative means of detecting the *C9orf72* RE is through repeat-primed PCR (rpPCR) (figure 3.2). In rpPCR an anchor and fluorescently labelled forward primer operate with a reverse primer which binds to multiple sites within the RE (figure 3.2.A). A characteristic sawtooth pattern indicates the presence of the RE (figure 3.2.B). While rpPCR has higher throughput than Southern blotting, it loses resolution past 30 repeats. Consequently, positive cases can be identified but accurate allele sizes cannot.

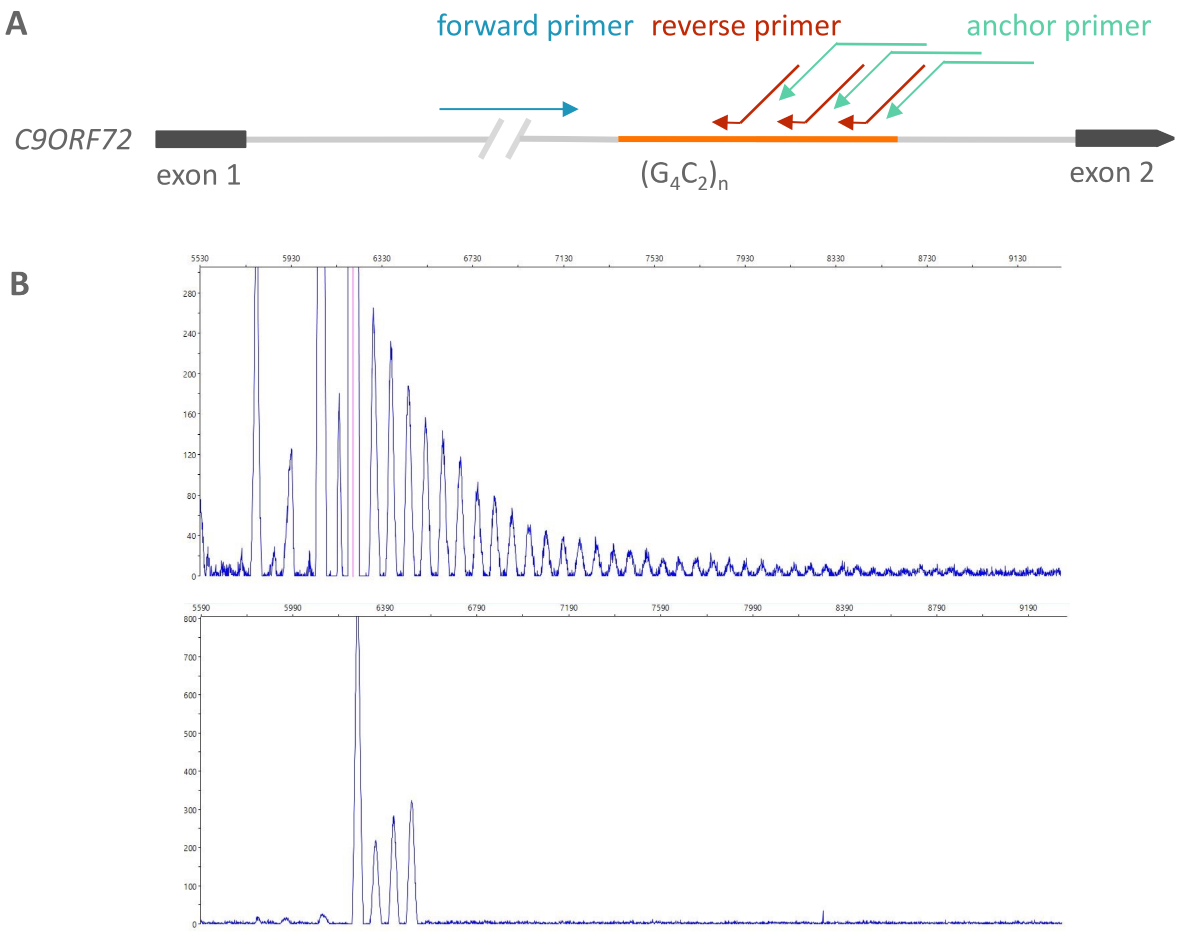


Figure 3.2: Repeat-primed PCR

A) An anchor and fluorescently labelled (FAM) forward primer are present in addition to a reverse primer which binds to multiple sites within the repeat. Amplified fragments are measured by capillary electrophoresis. B) A characteristic sawtooth pattern in the upper image indicates the presence of the RE, and the bottom panel indicates a sample that is negative for the repeat.

Third-generation DNA sequencing

It is now possible to directly measure long REs using third-generation sequencing. One method of third-generation sequencing is to utilise small protein channels called nanopores. Electrophoresis is used to pass DNA through a biological pore embedded on a membrane over an electrical grid. As DNA passes through the nanopore, different DNA base 6-mers cause characteristic changes in electrical current density which can be measured and converted to DNA bases (Stoddart *et al.* 2009). This differs from Sanger sequencing and NGS in that DNA strands can be read without the need for fragmentation; and thus, a single continuous molecule can be sequenced without the need for PCR. Read lengths exceeding 2 million base pairs (bps) have been reported (Payne *et al.* 2018); while the maximum read length with Sanger and NGS is on the order of 500-1500 bp. At an individual base level, these platforms have a higher error rate than Sanger or NGS; however, they allow direct measurement of chromosomal aberrations, CNVs and REs that have not traditionally been directly measurable. Third-generation sequencing has to-date been used to sequence REs in FXTAS (Grosso *et al.* 2021) and ALS (Ebbert *et al.* 2018). While this technology has great promise, it is more expensive and less high-throughput than NGS sequencing and the higher per-base error makes it less amenable to studying other types of variants in the genome.

Next-generation DNA sequencing

The development of next-generation sequencing (NGS) in the mid-2000s (Margulies *et al.* 2005) massively reduced the cost and time required to sequence a human genome. Preparing DNA for NGS requires shearing the DNA into fragments of 150 to 500bp in length. This fragmentation poses challenges to the calculation of repeat-lengths for two reasons. Firstly, the fragment lengths are below the pathogenic range of many REs. Secondly, fragmentation was historically accompanied by PCR amplification, which itself poses two issues. PCR amplification does not occur uniformly; low-GC regions and non-repetitive regions are preferentially amplified, thus uniform genomic coverage cannot be used to predict the depth of sequencing at a repeat. Secondly, PCR amplification introduces stutter noise at STRs as the DNA polymerase slips (Hauge and Litt 1993), creating artificial variability. Illumina have developed a PCR-free method of whole-genome sequencing (WGS), which both creates uniform genomic coverage during sequencing, and removes stutter-error (Kozarewa *et al.* 2009). However; a lot of sequencing data has already been generated using PCR. Whole-exome sequencing (WES) also relies on PCR amplification and is still commonplace as it yields higher coverage of the exome at lower cost.

Several tools have been developed in recent years, designed to utilise the features of NGS to either form an estimate of repeat length or to perform a statistical test to determine the likelihood that a repeat is present based on the presence of certain features at a locus (figure 3.3).

These tools vary in their utility and approach to measuring REs and STRs (figure 3.4). Both RepeatSeq (Highnam *et al.* 2013) and HipSTR (Willems *et al.* 2017) only take into account enclosed repeats. Functioning on the assumption that a repeat is below the length of a read; these tools can theoretically accurately genotype short repeats but fail to estimate long expansions.

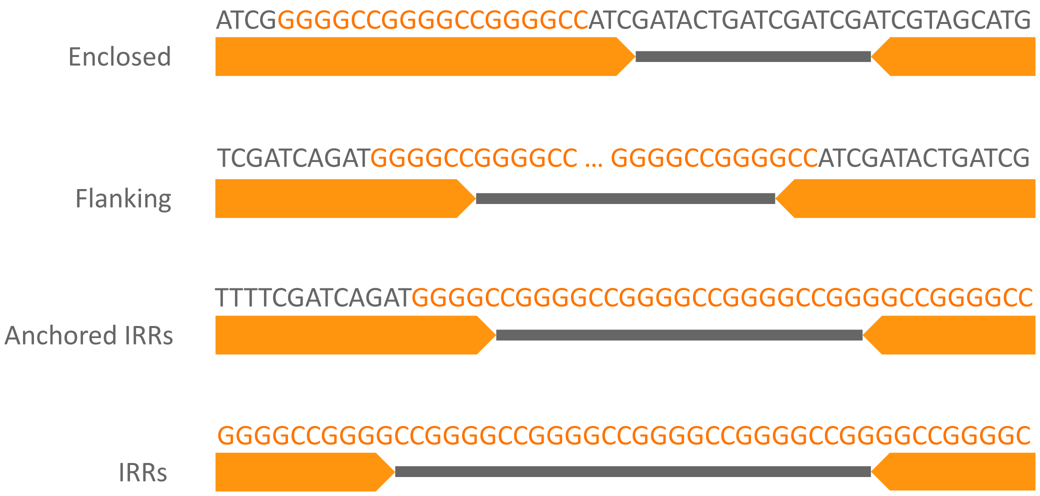
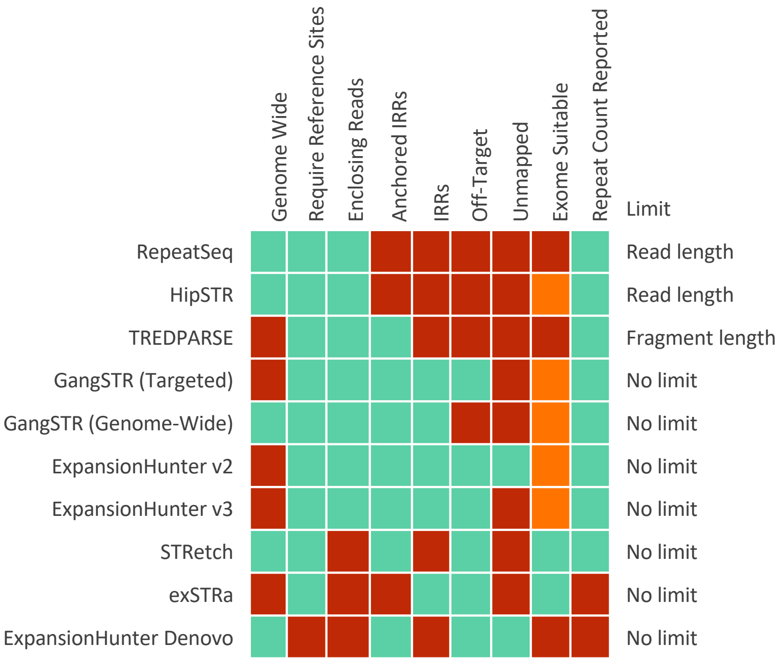

ATCGGGGGCCGGGGCCGGGGCCATCGATACTGATCGATCGATCGTAGCATG
Enclosed
TCGATCAGATGGGGCCGGGGCC … GGGGCCGGGGCCATCGATACTGATCG
Flanking
TTTTCGATCAGATGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCC
Anchored IRRs
GGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGC
IRRs

Figure 3.3: Genomic features available to RE genotyping tools

An STR can be entirely enclosed in an NGS read. Alternatively a pair of NGS reads may flank a repeat. An RE may also fall entirely within a read, this could be anchored to the correct genomic location by a partner read or both pairs of reads may be in the repeat.



	Genome Wide	Require Reference Sites	Enclosing Reads	Anchored IRRs	IRRs	Off-Target	Unmapped	Exome Suitable	Repeat Count Reported	Limit
RepeatSeq	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟩	Read length
HipSTR	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟧	🟩	Read length
TREDPARSE	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟩	Fragment length
GangSTR (Targeted)	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟧	🟩	No limit
GangSTR (Genome-Wide)	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟧	🟩	No limit
ExpansionHunter v2	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟧	🟩	No limit
ExpansionHunter v3	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟧	🟩	No limit
STRetch	🟩	🟩	🟥	🟩	🟥	🟩	🟥	🟩	🟩	No limit
exSTRa	🟥	🟩	🟥	🟥	🟩	🟩	🟥	🟩	🟥	No limit
ExpansionHunter Denovo	🟩	🟥	🟥	🟩	🟥	🟩	🟩	🟥	🟥	No limit

Figure 3.4: Features of repeat genotyping tools

Green squares indicate that the category is utilised in a corresponding tool. Red squares indicate that a category is not utilised in a given category. Orange indicates that for several tools the applicability in exome sequencing is still uncertain.

TREDPARSE (Tang *et al.* 2017) utilises the features of paired-end (PE) sequencing to extend the possible repeat-length estimate beyond the length of an individual read, to the length of the DNA fragment that undergoes PE sequencing. GangSTR (Mousavi *et al.* 2019) and ExpansionHunter (Dolzhenko *et al.* 2019, 2017) further extend this by utilising reads which are entirely composed of the repeat motif (in-repeat reads (IRRs) (figure 3.3)). Both tools additionally account for the genomic context surrounding a repeat; in particular coverage and fragment length, and allow the specification of 'off-target loci'; these are loci in the genome where IRRs may have mistakenly aligned. GangSTR can be run in 'Target' or 'Genome-Wide' mode. The 'Genome-Wide' mode does not screen off-target IRRs but comes with considerable speed increases and scalability. A change from ExpansionHunter version2 to ExpansionHunter3 saw the tool no longer screen unaligned reads, providing moderate increase in speeds. ExpansionHunter3 also takes a graph-based approach which allows the reconstruction of complex alleles where the repeat is not a straightforward expansion of a single motif. The disadvantage of TREDPARSE, GangSTR and ExpansionHunter is that they are not scalable at the level of the genome (with the exception of running GangSTR in targeted mode). Additionally; while target loci for TREDPARSE can be easily generated provided they are simple repeats and reside in the reference genome, both GangSTR and ExpansionHunter require specialist analysis to generate the necessary target files with specified off-target loci.

While all the tools previously mentioned are capable of providing a repeat length estimate for an individual sample, alternative methods have been developed that perform an outlier detection test between case and control cohort. exSTRa determines the repeat content of each read aligning to a target STR locus (Tankard *et al.* 2018). The STR content of all reads for an individual are then compared to the rest of the cohort and a statistical test for outliers is performed. STRetch (Dashnow *et al.* 2018) derives decoy chromosomes that include artificially long versions of target repeat loci. Reads are realigned to the decoy chromosome and a likelihood ratio test is performed comparing the original read alignment to the alignment at the decoy chromosome, with subsequent results compared between cases and controls. ExpansionHunter Denovo (Dolzhenko *et al.* 2020) is the only tool mentioned so far that does not require a predetermined list of target repeat loci (either disease loci or genome-wide). ExpansionHunter Denovo identifies anchored IRRs whose mates (including unaligned and misaligned mates) contain repetitive motifs. Clusters of reads sharing a similar profile denote a locus harbouring a large repeat.

Unsurprisingly the publication of each software has presented the tool in question as outperforming other software across their chosen metric. This can partially be explained by newer tools outperforming older, but may also be a result of different measurement objectives. At the time of writing I am aware of two objective benchmarking studies.

Halman and Oshlack (2020) screened the X chromosome of 433 male samples to identify which genotyping tool made the fewest erroneous heterozygous calls. This study used only software which utilised enclosed reads, thus focusing on STRs which fell below read length. The authors identified that RepeatSeq and HipSTR had the lowest heterozygous error rate. This study is limited in that it does not address REs which exceed the read length, it did not perform confirmatory PCR genotyping and due to the design of the study a limited number of tools were possible to study.

Rajan-Babu *et al.* (2020) studied WGS PCR free data from 118 patients with an expansion in either *AR*, *ATN1*, *ATXN1*, *ATXN3*, *DMPK*, *FMR1*, *FXN*, or *HTT* and simulated genomes of patients with *C9orf72*, *FMR1* or *FMR2* expansions. They identified that no individual tool provides perfect identification and that an ensemble approach combining the results of tools is optimal. The limitations of this study are that PCR genotypes were not available at the unexpanded loci in each sample (so no benchmarking of unexpanded STR genotyping was obtained), only a small number of loci were studied, and the study included simulated data which is limited in its capacity to accurately recreate either biological or DNA sequencing complexity.

There are a number of open questions with regards to genotyping software. How does the accuracy of tools compare in WES and WGS data? How do tools compare when genotyping PCR validated alleles in the normal range? Are sensitivity and specificity consistent across a larger number of loci?

Epilepsy

As previously described, REs are implicated in the pathology of many neurological conditions. Epilepsy is a group of heterogeneous neurological conditions characterised by a predisposition to seizures, with more than 50 million people affected worldwide (Covanis *et al.* 2015). 20-30% of cases have a definitive extraneous cause such as head trauma, but the remaining cases have some degree of a genetic basis (Hildebrand *et al.* 2013).

REs have previously been observed as a cause of epilepsy. Familial Adult Myoclonic Epilepsy (FAME) is an autosomal dominant condition with adult onset. Patients typically experience hand tremors, myoclonic jerks and rare seizures (Lagorio *et al.* 2019). Ishiura *et al.* (2018) identified that an intronic TTTCA/TTTTA expansion in a number of genes

(*SAMD12*, *TNRC6A*, *RAPGEF2*) was sufficient to create a GOF effect creating RNA foci that sequester RNA binding proteins resulting in the observed phenotype. A coding GCG RE in the gene ARX has also been found to cause early infantile epileptic encephalopathy (EIEE1) (Strømme *et al.* 2002).

One subgroup of patients who are believed to have primarily monogenic causes are patients with severe childhood epilepsies, often with concomitant intellectual disability (Perucca, Bahlo, and Berkovic 2020). Benson *et al.* (2020) performed WES and array-comparative genomic hybridisation on 96 such trios and 5 further siblings to identify small variants and large chromosomal aberrations in Irish patients with *de novo* epilepsy and intellectual disability. A genetic diagnosis was made in 31% of these patients. The remaining 69% of patients who lack a genetic diagnosis are likely to have monogenic causes which either reside outside the exome, or which are complex variants such inversions, CNVs or REs.

Research Aims

1. Utilise gold-standard PCR genotypes to determine the sensitivity and specificity of several RE genotyping tools across a larger range of genes than has previously been studied.
2. Compare the accuracy of RE genotyping tools in WES and WGS data.
3. Identify if any previously reported disease-associated RE loci exhibit a pleiotropic effect, causing *de novo* cases of epilepsy in the Irish population.

Note: it is outside the scope of this study to identify novel loci that may cause epilepsy in the Irish population or to identify pleiotropic loci that may lead to ALS as these are the subjects of ongoing research.

Methods

Study participants

This study includes data from patients living with ALS, PLS, or epilepsy as well as relatives and control individuals.

ALS and PLS patients attended the national specialist MND clinic at Beaumont Hospital Dublin. All ALS patients were diagnosed as definite, probable or possible ALS by specialist neurologists in accordance with the El Escorial criteria (Brooks *et al.* 2000). A PLS diagnosis was made if patients had progressive UMN signs for four years, no LMN signs to eliminate the possibility of ALS, and the patients were over 40 to rule out HSP. The PLS cohort is described in greater detail in Chapter 5. Control individuals were age and location matched to ALS patients and were neurologically normal at the time of blood sampling.

Patients with epilepsy and their relatives were recruited via tertiary referral clinical centres throughout Ireland, specifically Beaumont Hospital, Cork University Hospital, Galway University Hospital, Our Lady's Children's Hospital Crumlin, St. James' Hospital, and the Daughters of Charity (St. Vincent De Paul). Patients were recruited and deeply phenotyped by an advanced nurse practitioner. Many patients also experienced Intellectual Disability (ID). This cohort was collected and DNA sequenced as part of the RCSI FutureNeuro / Lighthouse Project and has been previously described in detail by Benson *et al.* (2020), wherein 101 trios were screened exome-wide for SNVs and chromosomal abnormalities, providing 31% with a molecular diagnosis; however, the potential impact of REs has not been studied in this cohort.

DNA sequencing

The cohorts sequenced in this study are outlined in table 3.2. Briefly; 150bp PE PCR-free WGS sequencing was performed for 272 Irish ALS cases and 136 Irish controls to a depth of 40X. This data was sequenced as part of Project MinE and has been described previously (Project MinE ALS Sequencing Consortium 2018).

Table 3.2: Datasets In This Study

Dataset ID	Phenotype	Patients (n)	Controls (n)	Trios	Individual Parents	Sequencing Type	Sequencing Platform	Sequencing	Exome Enrichement Kit	Target Coverage	Source
ALS_WGS	ALS	272	136	0	0	WGS PCR Free	Illumina HiSeq 2000	150bp PE	N/A	40X	ProjectMinE
ALS_WES	ALS/PLS	66	0	0	0	WES	Illumina NovaSeq	150bp PE	Agilent SureSelect	90X	This Thesis
EE_WGS_PCR_FREE	Epilepsy	30	0	0	0	WGS PCR Free	Illumina HiSeq	150bp PE	N/A	30X	RCSI FutureNeuro/ Lighthouse Project
EE_WGS_PCR	Epilepsy	11	0	10	1	WGS with PCR	Illumina HiSeq	150bp PE	N/A	30X	RCSI FutureNeuro/ Lighthouse Project
EE_WES	Epilepsy	114	0	106	0	WES	Illumina NextSeq	75bp PE	SeqCap EZ Exome	45X	RCSI FutureNeuro/ Lighthouse Project

29 proband overlap between RCSI_WES & EE_WGS_PCR_FREE

Three methods of DNA sequencing were carried out for epilepsy patients. 114 patients underwent 75bp WES sequencing, including four pairs of siblings. Both parents were also sequenced for 106 of these patients. 30 patients underwent 30X PCR-free WGS, 29 of whom had previously undergone exome sequencing. The final epilepsy cohort consisted of 11 patients, of which ten had both parents also sequenced and 1 had a single parent sequenced. This cohort underwent WGS with PCR. Data for all epilepsy cohorts were provided in FASTQ format.

The final cohort in this study is described in further detail in Chapters 4 and 5. It includes a large Irish pedigree, a Cuban ALS pedigree and Irish PLS samples who underwent 150bp PE exome sequencing. In this chapter this dataset is primarily used to draw comparison to the epilepsy WES data.

Data processing

With the exception of the ALS WGS samples and controls which were available as pre-processed bam files (Project MinE ALS Sequencing Consortium 2018), all data were aligned from PE FASTQ files. Reads were aligned to the GRCh37 version of the human reference genome (downloaded from the UCSC genome browser (W. J. Kent *et al.* 2002)), using the Burrows-Wheeler Aligner (BWA) v.0.7.5 (H. Li and Durbin 2009). Aligned sam files were converted to bam format, sorted and indexed using samtools v.1.7 (H. Li *et al.* 2009). Picard v.0.7.5 (http://broadinstitute.github.io/picard/) was used for duplicate read removal, and to add read groups. Sample depth of coverage (DOC) was calculated using mosdepth v.0.2.9 (Pedersen and Quinlan 2018).

PCR genotyping

PCR genotyping was performed by Jennifer Hengeveld as part of ongoing research in the Complex Trait Genomics Laboratory, TCD, and was kindly provided for comparative purposes in this study. PCR genotypes of 23 genes for which *in silico* genotyping was available by at least one tool were provided for 338 samples who underwent PCR-free WGS. Fragment Length Analysis of multiplexed PCR products was performed by Eurofins, Germany and results were visualised and manually assessed using Peak Scanner v1.0.

C9orf72 genotyping

ALS patients were screened for the presence of the pathogenic *C9orf72* RE by rpPCR as described previously (Byrne *et al.* 2012). Amplified fragments were measured by capillary electrophoresis on an Applied Biosystems 3500 Series Genetic Analyzer and visualised using Gene Mapper v.4.0, screening for a decreasing sawtooth pattern which is indicative of a large RE. Patients with 30 hexanucleotide repeats or above and displaying a sawtooth rpPCR trace were deemed positive for the expansion.

STR genotyping

46 REs present in table 3.1 were studied in the five cohorts outlined above with a suite of STR genotyping tools including: ExpansionHunter v2.5.5 and v3.2.2, exSTRa v0.9.0, GangSTR v2.4.4 (in both targeted and genome-wide mode), HipSTR v0.6.2, RepeatSeq v0.8.2, STRetch v0.1.0, and TREDPARSE v0.6.6. Additionally the ALS WGS and controls were genotyped with ExpansionHunter Denovo v0.9.0. Certain loci were not genotyped across all software either because the repeat is complex, not present in the reference genome, or not present in the reference panel for the software (supplementary table S.3.1.). The full commands used to run each software are available at :
github.com/dohertymark/Thesis/Chapter3/Chapter3_Call_RE_Geno_Software.sh

ExpansionHunter3 is capable of reconstructing complex loci. The ATXN8OS locus harbours a complex $CTA_{N}CTG_{N}$ repeat. Reviewer v0.2.5 (Dolzhenko *et al.* 2021) was used to identify which haplotype the separately genotyped CTA and CTG alleles fall, in order to reconstruct the combined STR.

STRetch, exSTRa and ExpansionHunter Denovo require case and control cohorts. ExpansionHunter Denovo was run only in the ALS WGS case/control cohort as this requires PCR free WGS. When running STRetch and exSTRa in WES data and WGS with PCR, parental samples were treated as controls.

To reduce the possibility of erroneously excluding true positive expansions, the recommended minimal first-pass filtering was applied to genotyping results. GangSTR output was filtered using dumpSTR v.3.0.2 (Mousavi *et al.* 2019) (--gangstr-filter-spanbound-only --gangstr-filter-badCI --gangstr-max-call-DP 1000 --gangstr-min-call-DP 20 --filter-regions hg19_segmentalduplications.bed.gz --filter-regions-names SEGDUP).

HipSTR was filtered using the provided filtering script (--min-call-qual 0.9 --max-call-flank-indel 0.15 --max-call-stutter 0.15 --min-call-allele-bias -2 --min-call-strand-bias -2). For all other software, sites were retained if they were deemed a 'PASS' in the initial call and otherwise removed. Samples identified as positive expansions were further investigated to identify false positives.

Statistical analysis and plotting

The following formulae were used in analysis:

$$Sensitivity = \frac{Number\ of\ correctly\ genotyped\ true\ expansions}{Number\ of\ genotyped\ true\ expansions} \times 100$$

Equation 3.1

$$Specificity = \frac{Number\ of\ correctly\ genotyped\ true\ negatives}{Total\ number\ of\ genotyped\ true\ negatives} \times 100$$

Equation 3.2

Equation 3.3

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

where; RMSD= route-mean-square deviation; N= Number of data points; x = observed value; $\hat{x}_i$ = expected value

As *in silico* tools begin counting STR motifs at different starting positions, a correction was applied to *in silico* results when calculating RMSD between *in silico* predictions and PCR genotypes. To avoid division by zero errors when calculating odds ratios (ORs), a Haldane-Anscombe correction was applied (Lawson 2004). To identify significant ORs, Fisher exact tests were performed, applying Bonferroni corrections, accounting for the number of genes tested by each tool and the number of unique repeat units observed for each gene.

NIPA1 and *TNRC6A* were excluded from calculations of sensitivity and specificity as their pathogenic threshold of 8 repeats are risk factors rather than a strict pathogenic threshold (Blauw *et al.* 2012).

Unless otherwise stated, all statistical analyses and plotting for this chapter were performed using R v3.6.1 (Team 2014) utilising the packages stringr v1.4 (Hadley Wickham 2019) and qqman (S. D. Turner, n.d.) v0.1.8.

Results

Exome enrichment protocols

WES data was available for 66 samples prepared with the Agilent SureSelect v7 exome-enrichment probes and 326 samples sequenced with the SeqCap EZ Exome v3 exome-enrichment probes. To explore whether either enrichment panel provided beneficial coverage across exonic repeats, the observed coverage across 9 exonic repeats was compared between datasets (figure 3.5). SureSelect samples were on average sequenced to a higher DOC than SeqCap samples (figure 3.5.A), but this is unrelated to the chosen exome panel. A single repeat (*PPP2R2B*) was not present in the SeqCap exome panel. Of the remaining 8 repeats, 4 were sequenced to a higher DOC in the SeqCap panel than expected and four were sequenced to a higher DOC in the SureSelect panel than expected (figure 3.5.B-K), indicating that neither panel has an overall superiority for covering exonic STRs.

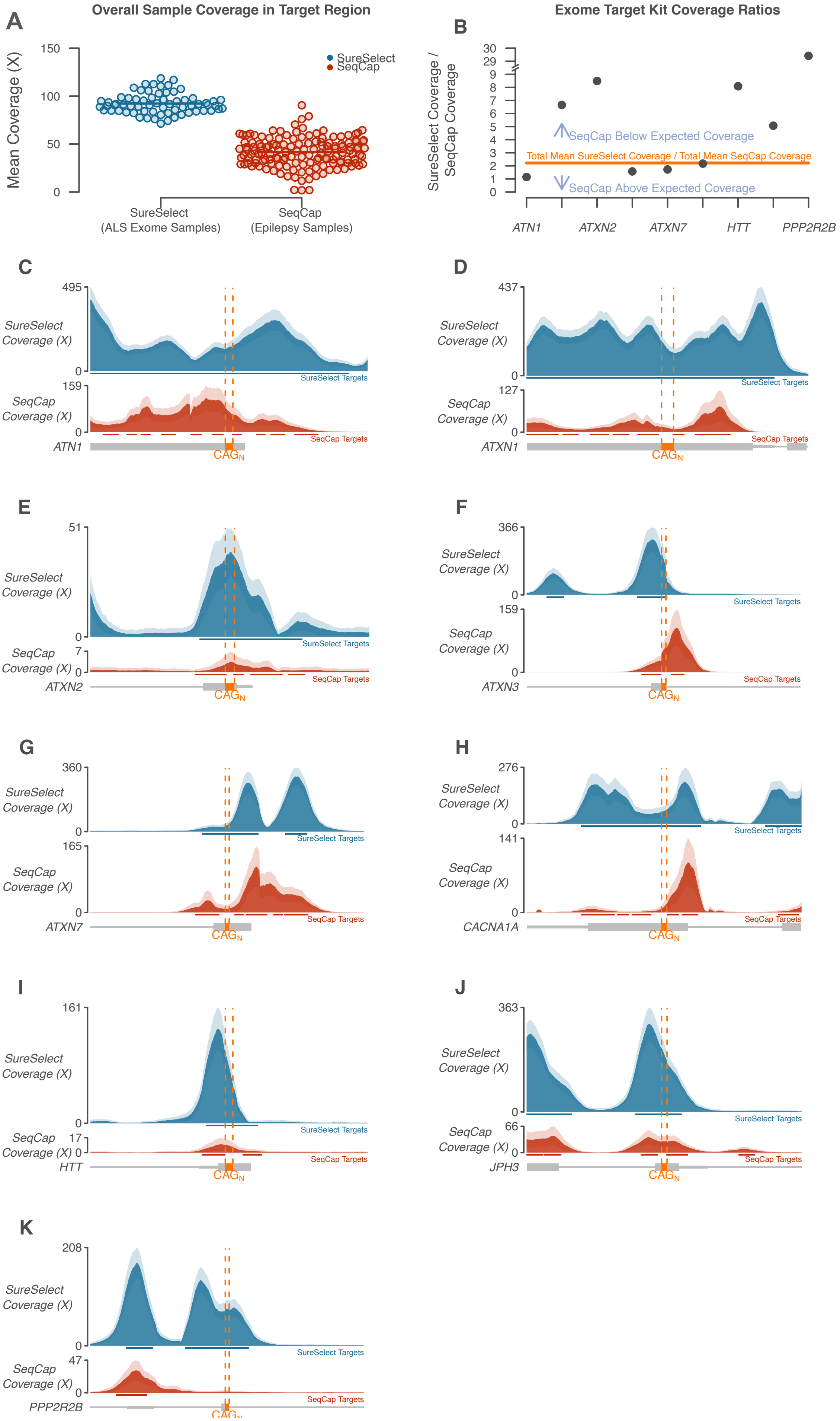
Benchmarking STR genotyping tools

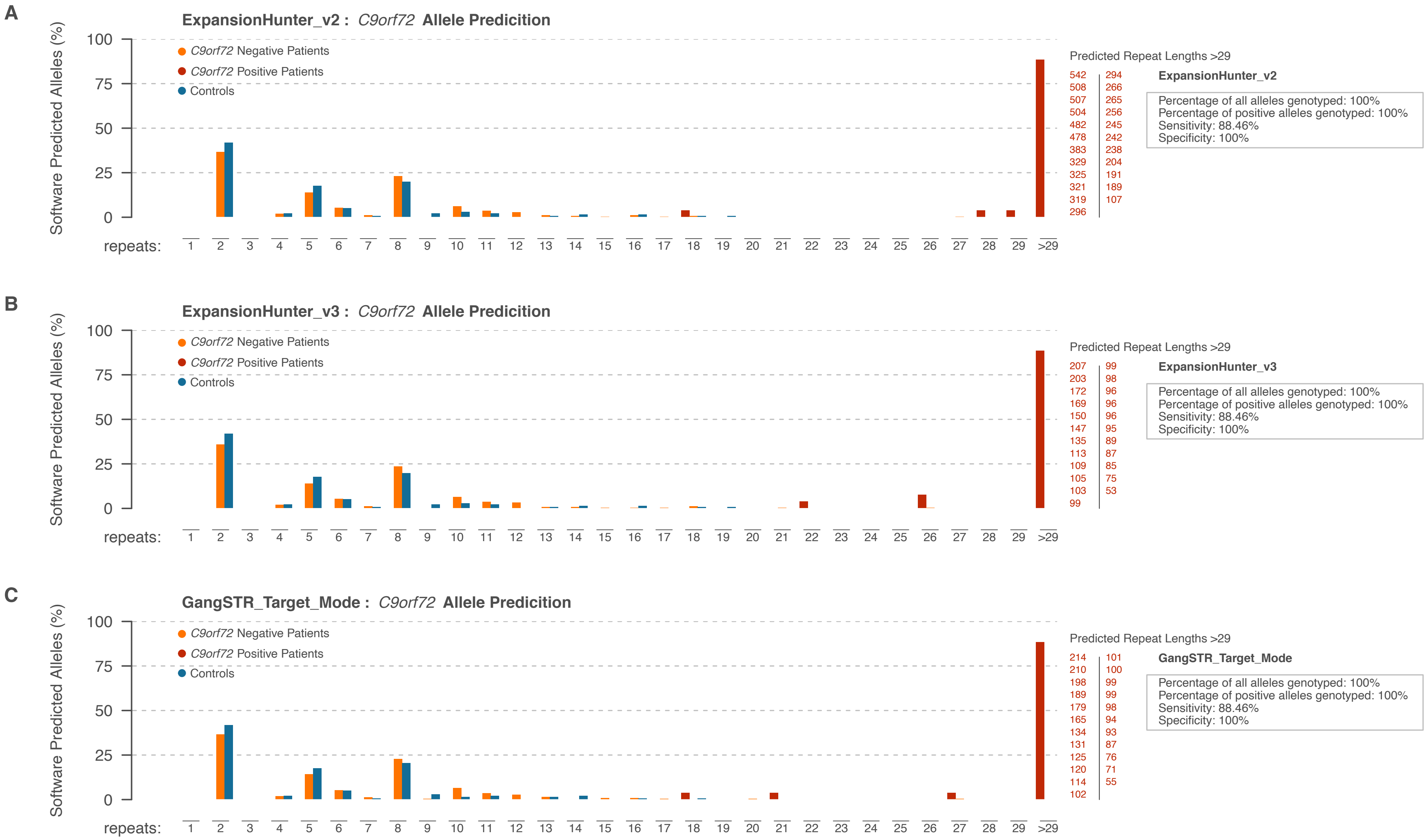
Identifying REs

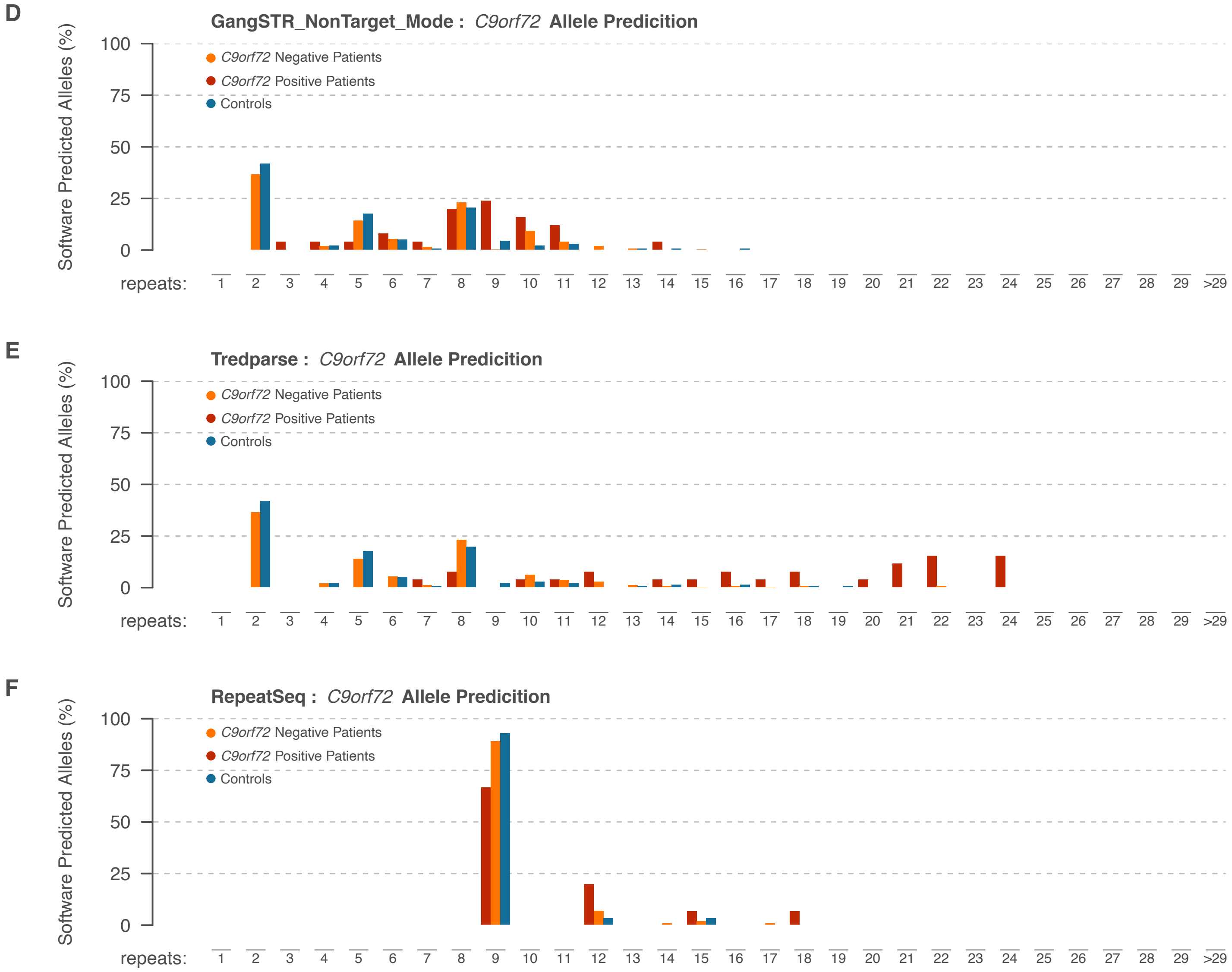
To examine each tool's ability to identify large REs, each software was used to genotype the *C9orf72* locus in 408 PCR-Free WGS samples for which rpPCR genotyping results were also available (supplementary table S3.1, figures 3.6 & 3.7). *C9orf72* rpPCR genotyping can only identify whether a sample is above or below 30 GGGGCC repeats, which is regarded as the pathogenic threshold; however, REs often extend to 10s of kilobases long (DeJesus-Hernandez *et al.* 2011; Renton *et al.* 2011). Figure 3.6 examines *in silico C9orf72* genotyping results. A strict pathogenic threshold of 30 repeats was used for all tools except STRetch and exSTRa wherein a significant p-value was required. Figure 3.7 displays sensitivity as a function of specificity; interrogating a tool's specificity when a certain percentage of expanded samples are correctly genotyped.

Figure 3.5: Differential repeat coverage with alternative exome targets (next page)

This figure examines whether either the SureSelect exome target kit or the SeqCap exome target kit provide better cover of exonic STRs. Figure 3.5.A describes the mean coverage of each sample prepared with either kit. The SureSelect samples are sequenced to a higher overall coverage but this is irrespective of the target panel. Figure 3.5.B demonstrates that whether an exonic STR is sequenced above or below the mean level of cover is gene specific, indicating that neither kit performs universally superior. Only one gene, PPP2R2B, is not targeted in the SeqCap panel (figure 3.5.K). Figures B-K depict the mean sample coverage at each base and the standard deviation.


A
Overall Sample Coverage in Target Region
Mean Coverage (X)
150
100
50
0
SureSelect
SeqCap
SureSelect
(ALS Exome Samples)
SeqCap
(Epilepsy Samples)
B
Exome Target Kit Coverage Ratios
SureSelect Coverage /
SeqCap Coverage
30
29
9
8
7
6
5
4
3
2
1
0
SeqCap Below Expected Coverage
Total Mean SureSelect Coverage / Total Mean SeqCap Coverage
SeqCap Above Expected Coverage
ATN1
ATXN2
ATXN7
HTT
PPP2R2B
C
SureSelect
Coverage (X)
495
0
SureSelect Targets
SeqCap
Coverage (X)
159
0
SeqCap Targets
ATN1
CAG$_N$
D
SureSelect
Coverage (X)
437
0
SureSelect Targets
SeqCap
Coverage (X)
127
0
SeqCap Targets
ATXN1
CAG$_N$
E
SureSelect
Coverage (X)
51
0
SureSelect Targets
SeqCap
Coverage (X)
7
0
SeqCap Targets
ATXN2
CAG$_N$
F
SureSelect
Coverage (X)
366
0
SureSelect Targets
SeqCap
Coverage (X)
159
0
SeqCap Targets
ATXN3
CAG$_N$
G
SureSelect
Coverage (X)
360
0
SureSelect Targets
SeqCap
Coverage (X)
165
0
SeqCap Targets
ATXN7
CAG$_N$
H
SureSelect
Coverage (X)
276
0
SureSelect Targets
SeqCap
Coverage (X)
141
0
SeqCap Targets
CACNA1A
CAG$_N$
I
SureSelect
Coverage (X)
161
0
SureSelect Targets
SeqCap
Coverage (X)
17
0
SeqCap Targets
HTT
CAG$_N$
J
SureSelect
Coverage (X)
363
0
SureSelect Targets
SeqCap
Coverage (X)
66
0
SeqCap Targets
JPH3
CAG$_N$
K
SureSelect
Coverage (X)
208
0
SureSelect Targets
SeqCap
Coverage (X)
47
0
SeqCap Targets
PPP2R2B
CAG$_N$


A
ExpansionHunter_v2 : C9orf72 Allele Predicition
Software Predicted Alleles (%)
100
75
50
25
0
C9orf72 Negative Patients
C9orf72 Positive Patients
Controls
repeats:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 >29
Predicted Repeat Lengths >29
542 508 507 504 482 478 383 329 325 321 319 296
294 266 265 256 245 242 238 204 191 189 107
ExpansionHunter_v2
Percentage of all alleles genotyped: 100%
Percentage of positive alleles genotyped: 100%
Sensitivity: 88.46%
Specificity: 100%
B
ExpansionHunter_v3 : C9orf72 Allele Predicition
Software Predicted Alleles (%)
100
75
50
25
0
C9orf72 Negative Patients
C9orf72 Positive Patients
Controls
repeats:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 >29
Predicted Repeat Lengths >29
207 203 172 169 150 147 135 113 109 105 103 99
99 98 96 96 96 95 89 87 85 75 53
ExpansionHunter_v3
Percentage of all alleles genotyped: 100%
Percentage of positive alleles genotyped: 100%
Sensitivity: 88.46%
Specificity: 100%
C
GangSTR_Target_Mode : C9orf72 Allele Predicition
Software Predicted Alleles (%)
100
75
50
25
0
C9orf72 Negative Patients
C9orf72 Positive Patients
Controls
repeats:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 >29
Predicted Repeat Lengths >29
214 210 198 189 179 165 134 131 125 120 114 102
101 100 99 99 98 94 93 87 76 71 55
GangSTR_Target_Mode
Percentage of all alleles genotyped: 100%
Percentage of positive alleles genotyped: 100%
Sensitivity: 88.46%
Specificity: 100%



GangSTR_NonTarget_Mode : *C9orf72* Allele Predicition

Tredparse : *C9orf72* Allele Predicition

RepeatSeq : *C9orf72* Allele Predicition

GangSTR_NonTarget_Mode

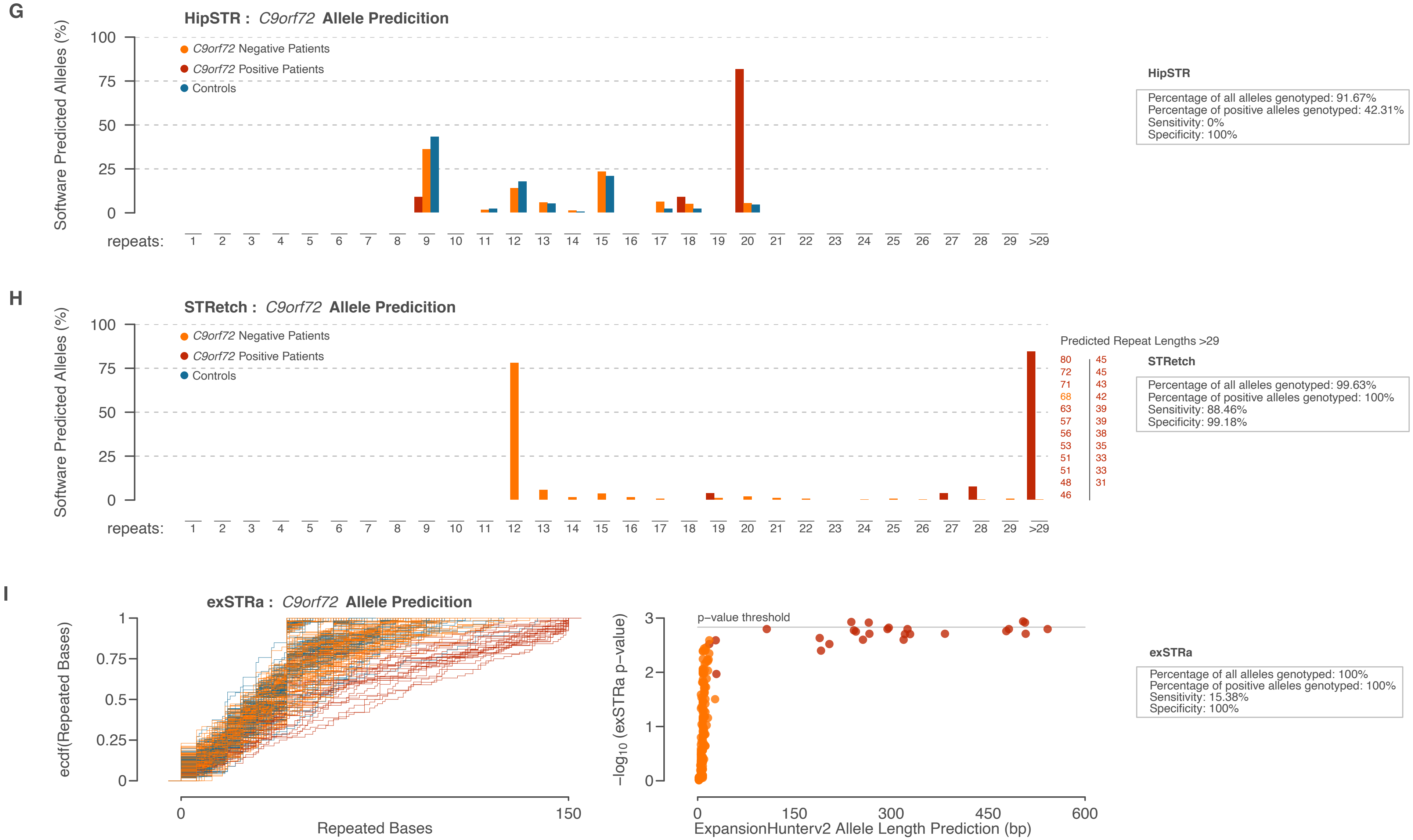
Percentage of all alleles genotyped: 99.75%
Percentage of positive alleles genotyped: 96.15%
Sensitivity: 0%
Specificity: 100%

Tredparse

Percentage of all alleles genotyped: 100%
Percentage of positive alleles genotyped: 100%
Sensitivity: 0%
Specificity: 100%

RepeatSeq

Percentage of all alleles genotyped: 42.4%
Percentage of positive alleles genotyped: 57.69%
Sensitivity: 0%
Specificity: 100%



HipSTR

Percentage of all alleles genotyped: 91.67%
Percentage of positive alleles genotyped: 42.31%
Sensitivity: 0%
Specificity: 100%

STRetch

Percentage of all alleles genotyped: 99.63%
Percentage of positive alleles genotyped: 100%
Sensitivity: 88.46%
Specificity: 99.18%

exSTRa

Percentage of all alleles genotyped: 100%
Percentage of positive alleles genotyped: 100%
Sensitivity: 15.38%
Specificity: 100%

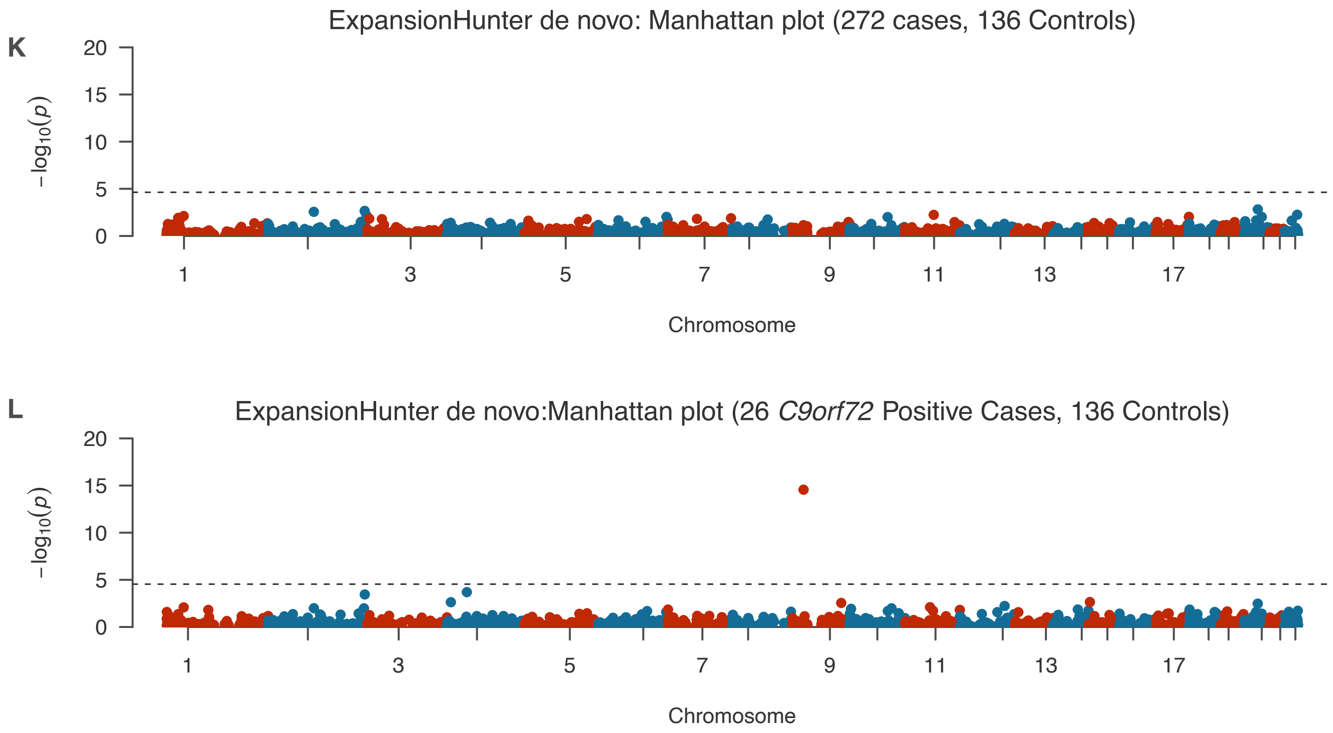


Figure 3.6: in silico genotyping of the *C9orf72* repeat expansion

Results are displayed for 272 ALS patients, 26 of whom carry a *C9orf72* repeat expansions, and 136 controls, sequenced with PCR-free WGS and genotyped with a range of in silico STR genotyping tools (A-L). For plots A-H the predicted alleles of positive and negative samples are directly compared. For figure I, the software exSTRa does not output allele predictions but instead gives p-values for predicted repeats (right) and values for the proportion of reads containing the repeat motif (left). ExpansionHunter Denovo compares genome-wide loci in cases and controls. When comparing 272 cases (including 26 *C9orf72* positive cases) to 136 controls, no significant loci are identified (K). Comparing 26 positive cases to 136 controls identifies the *C9orf72* RE.

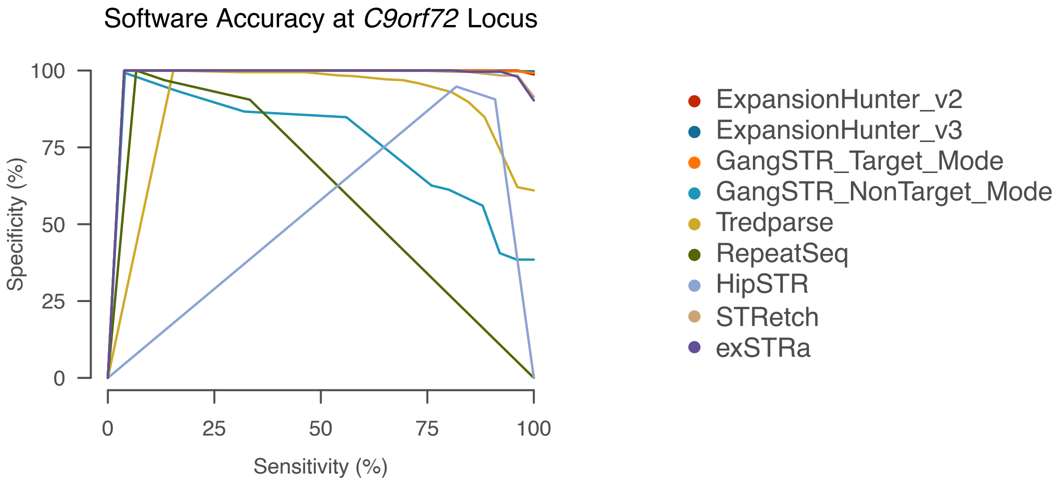


Figure 3.7: *C9orf72* in silico genotyping, specificity as a function of sensitivity

Rather than using a strict pathogenic threshold of 30 repeats, this figure examines how the percentage of *C9orf72* negative samples that are incorrectly genotyped changes are the number of positive cases are correctly genotyped. In figure 3.6 it is observed that TREDPARSE does not accurately classify any positive sample; however it is seen here that it significantly outperforms HipSTR, RepeatSeq and GangSTR in genome-wide mode as it maintains above 50% specificity while correctly identifying all positive samples.

All tools, with the exception of STRetch (which gives a single false-positive result), have 100% specificity when using a strict pathogenic threshold of 30 repeats (figure 3.6). ExpansionHunter v2 & v3, STRetch and GangSTR (targeted mode), provide the best RE discrimination; correctly identifying 88% of samples (figures 3.6.A, 3.6.B, 3.6.C & 3.6.H), this is followed by exSTRa (15%) and finally GangSTR-Genome-Wide, TREDPARSE, HipSTR and RepeatSeq, all of which fail to correctly genotype any positive samples. ExpansionHunter Denovo is trialled, firstly comparing 272 ALS cases to 136 controls, identifying zero loci (figure 3.6.K), and secondly comparing 26 RE positive patients to 136 controls, correctly identifying the *C9orf72* locus. ExpansionHunter Denovo is capable of identifying REs but requires sufficient power to do so.

ExpansionHunter, exSTRa, GangSTR (Targeted) and STRetch perform analogously when considering specificity as a function of sensitivity (figure 3.7). TREDPARSE is found to outperform HipSTR, GangSTR (Genome Wide) and RepeatSeq, as 100% sensitivity is achieved while maintaining above 50% specificity. This is likely as a result of TREDPARSE being limited to the fragment-length, rather than read-length.

Genotyping STRs in the broader population

Due to the phenotypic importance of STR variability in the general population (Gymrek 2017), each tool's capacity to accurately genotype STRs in a broad array of genes in a wide sample of individuals was assessed. PCR genotypes of 23 genes which were genotyped by at least one tool were available for 338 individuals who underwent PCR-free WGS. Applying the pathogenic thresholds in table 3.1, the overall sensitivity and specificity of each tool was measured (combining the results of the *C9orf72* locus with the additional loci), in addition to the overall observed RMSD and the RMSD observed per gene (table 3.3, figure 3.8, supplementary figure S3.1-S3.7).

Accounting for sensitivity, specificity and RMSD, both versions of ExpansionHunter outperform all other software. Other tools may match ExpansionHunter in one category but perform significantly worse in the other two. TREDPARSE, HipSTR, GangSTR (genome wide mode) and RepeatSeq have a similar or superior rate of false-positives; however, this is because they fail to predict longer alleles, and subsequently have very poor sensitivity.

STRetch manages to correctly identify 74% of alleles which are above the literature-reported pathogenic threshold. Its specificity is half a percent below that of ExpansionHunter. It is worth noting that with 23 genes and 338 samples a 0.5% decrease in specificity provides an additional 38 false positives. exSTRa identifies 46% of expanded loci as significant but has a high rate of false-positives.

GangSTR (targeted mode) correctly identifies a similar proportion of truly expanded loci as ExpansionHunter; however, it has a higher rate of false-positives. Examining the GangSTR output of the CTG repeat in *ATXN2* as an illustrative locus, and interrogating the 16 samples which are falsely predicted to have more than 50 repeats, reveals that GangSTR has misassigned reads to the *ATXN2* locus that are correctly aligned (with a MAPQ of 60) in the bam files to the CTG STR in *TCF4*. GangSTR does not account for the MAPQ of reads it assigns as off-target and so they are misincorporated.

In summary, considering a panel of 24 genes including several samples which have large repeats in the *C9orf72* locus; ExpansionHunter is found to outperform other software in terms of sensitivity, specificity and RMSD. GangSTR, STRetch and exSTRa are capable of identifying REs but with varying degrees of trade-off in terms of false-positives.

Table 3.3: RMSD, sensitivity & specificity of in *silico* genotyping tools relative to PCR data and between WES & WGS data

	C9orf72 Locus		All Loci		
Software	**Sensitivity (%)**	**Specificity**	**Sensitivity**	**Specificity**	**RMSD**
ExpansionHunter 2	88.46	100	81.25	99.86	3.12
ExpansionHunter 3	88.46	100	84.85	99.73	2.88
exSTRa	15.38	100	46.48	95.11	N/A
GangSTR (Targeted)	88.46	100	83.33	92.97	12
GangSTR (Genome Wide)	0	100	3.45	99.67	3.74
HipSTR	0	100	6.67	100	2.79
RepeatSeq	0	100	0	100	4.11
STRetch	88.46	99.18	72.73	99.26	8.66
TREDPARSE	0	100	12.12	99.74	3.52

	WES WGS Comparison				
	RMSD	**Longest WES Genotype**	**Mean WES Genotype**	**Longest WGS Genotype**	**Mean WGS Genotype**
ExpansionHunter 2	5.25	39	15.71 (SD: 7.31)	41	17.09 (SD: 7.76)
ExpansionHunter 3	8.05	85	19.83 (SD: 13.75)	58	18.83 (SD: 9.32)
exSTRa	N/A	N/A	N/A	N/A	N/A
GangSTR (Targeted)	16.79	75	16.67 (SD: 13.36)	113	20.56 (SD: 18.73)
GangSTR (Genome Wide)	8.12	71	16.13 (SD: 12.42)	39	15.53 (SD: 8.04)
HipSTR	1.13	36	14.9 (SD: 6.09)	33	14.86 (SD: 6.16)
RepeatSeq	1.22	21	13.03 (SD: 4.1)	23	13.05 (SD: 4.12)
STRetch	N/A	N/A	N/A	N/A	N/A
TREDPARSE	17.03	141	17.33 (SD: 15.73)	199	17.97 (SD: 15.01)

RMSD: Route Mean Square Deviation

RMSD is calculated at all loci excluding *C9orf72* as only positive or negative rpPCR genotypes were available at this locus

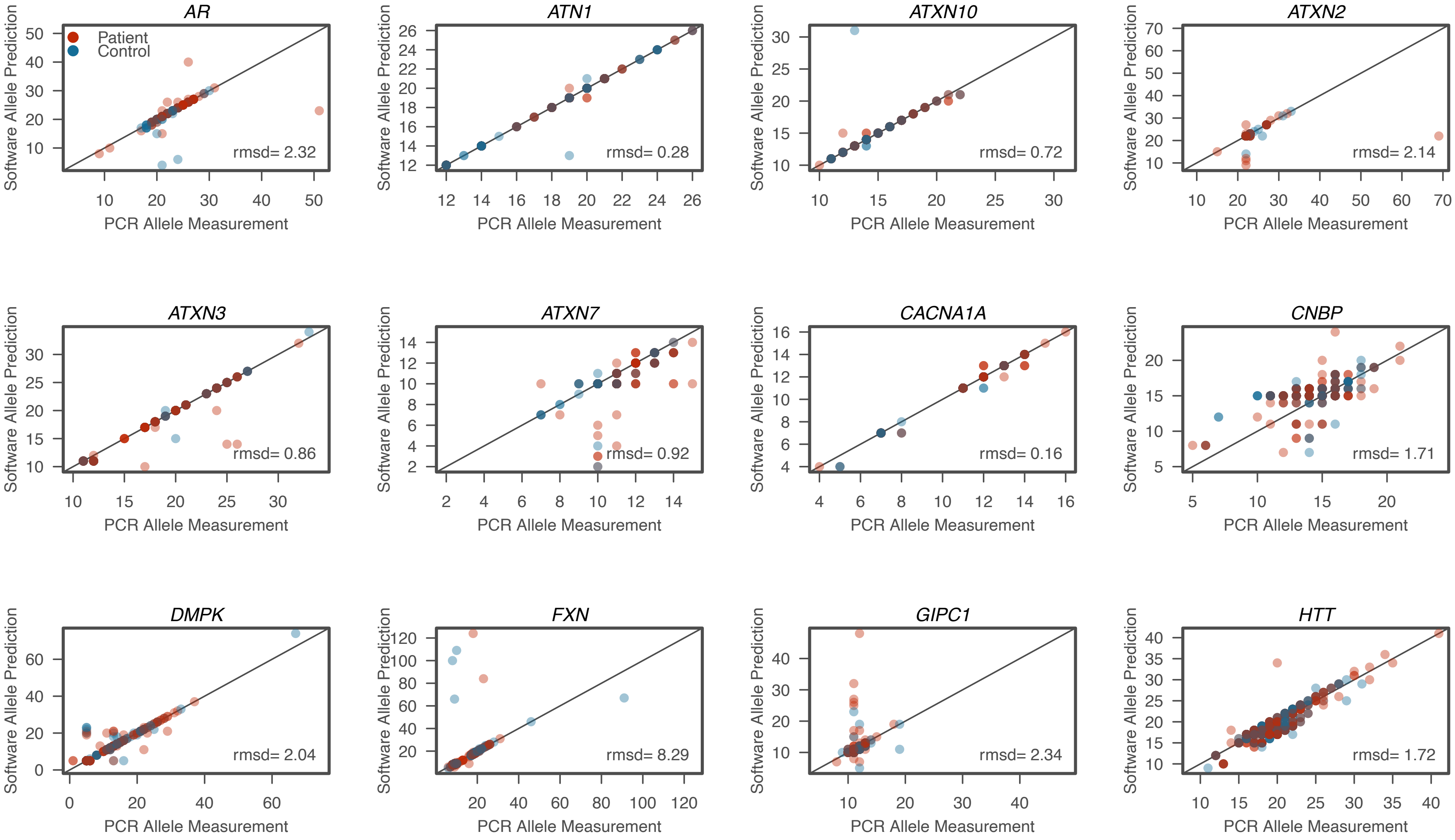
Reliability of WES *in silico* genotypes

It remains an open question which tools provide reliable genotyping results when used with exome data. To address this, *in silico* genotyping results from 29 epilepsy patients who had undergone both PCR-free WGS and WES were compared (table 3.3, figure3.9, supplementary figure S3.8-S3.13).

RMSD is used a measure of divergence between WGS and WES *in silico* genotyping calls. HipSTR and RepeatSeq are found to have best concordance between WES and WGS results; however, this is primarily due to a failure of both software to genotype longer alleles rather than an improvement in alleles which are genotyped (table 3.3, figure 3.9, supplementary figure S3.8-S3.13). TREDPARSE and GangSTR in targeted mode are found to be the worst performing software when comparing WES and WGS results. This is unsurprising for TREDPARSE which is only designed for WGS. As described above, the WGS results of GangSTR in targeted mode reveal that multiple samples at several loci are incorrectly called as expanded (figure S3.3,S3.10). This is found to be a result of GangSTR incorrectly attributing reads at potential off-target loci to the locus in question. These off-target loci are not present in the WES data so a large discrepancy in results appears.

ExpansionHunter2 (RMSD: 5.25) is found to slightly outperform ExpansionHunter3 (RMSD: 8.05) in genotyping exonic repeats, likely due to the ability to manually specify coverage in ExpansionHunter2. As both exSTRa and STRetch require sequencing and patient control data which was not available, these tools were not run on this data and this comparison is not calculated.

ExpansionHunter_v3 : Comparison of Gold Standard PCR Genotyping with Software Allele Prediction



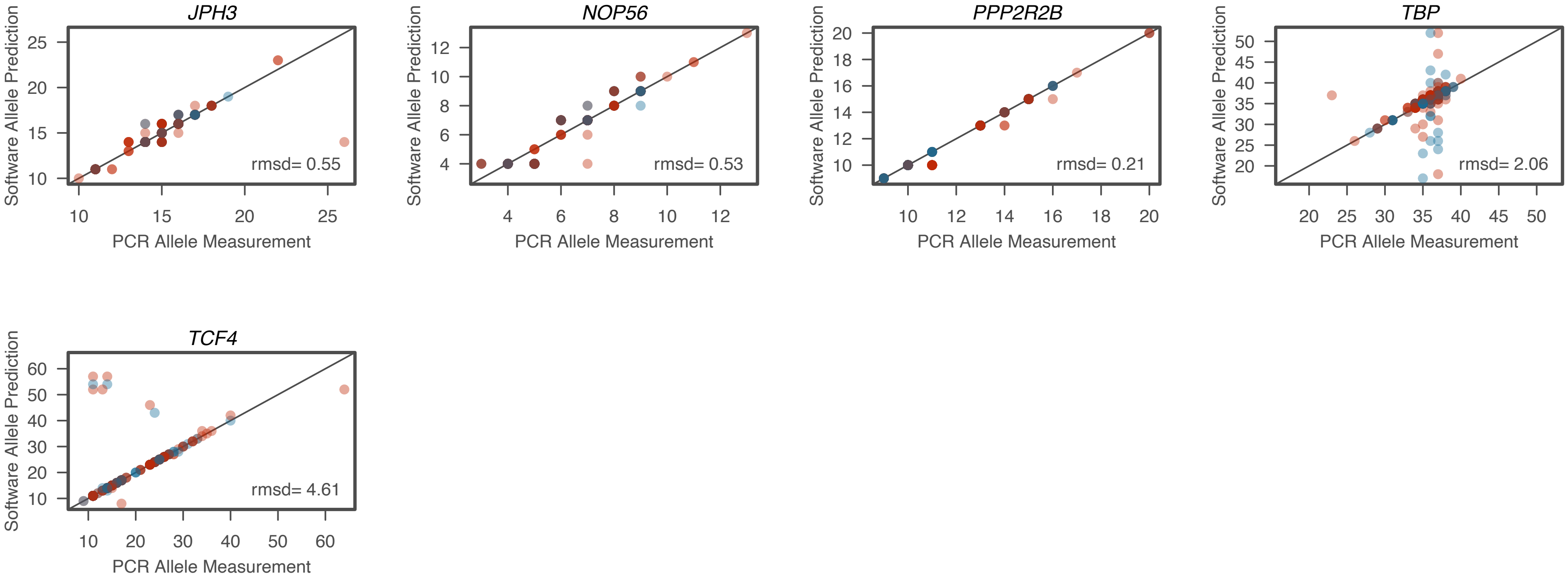
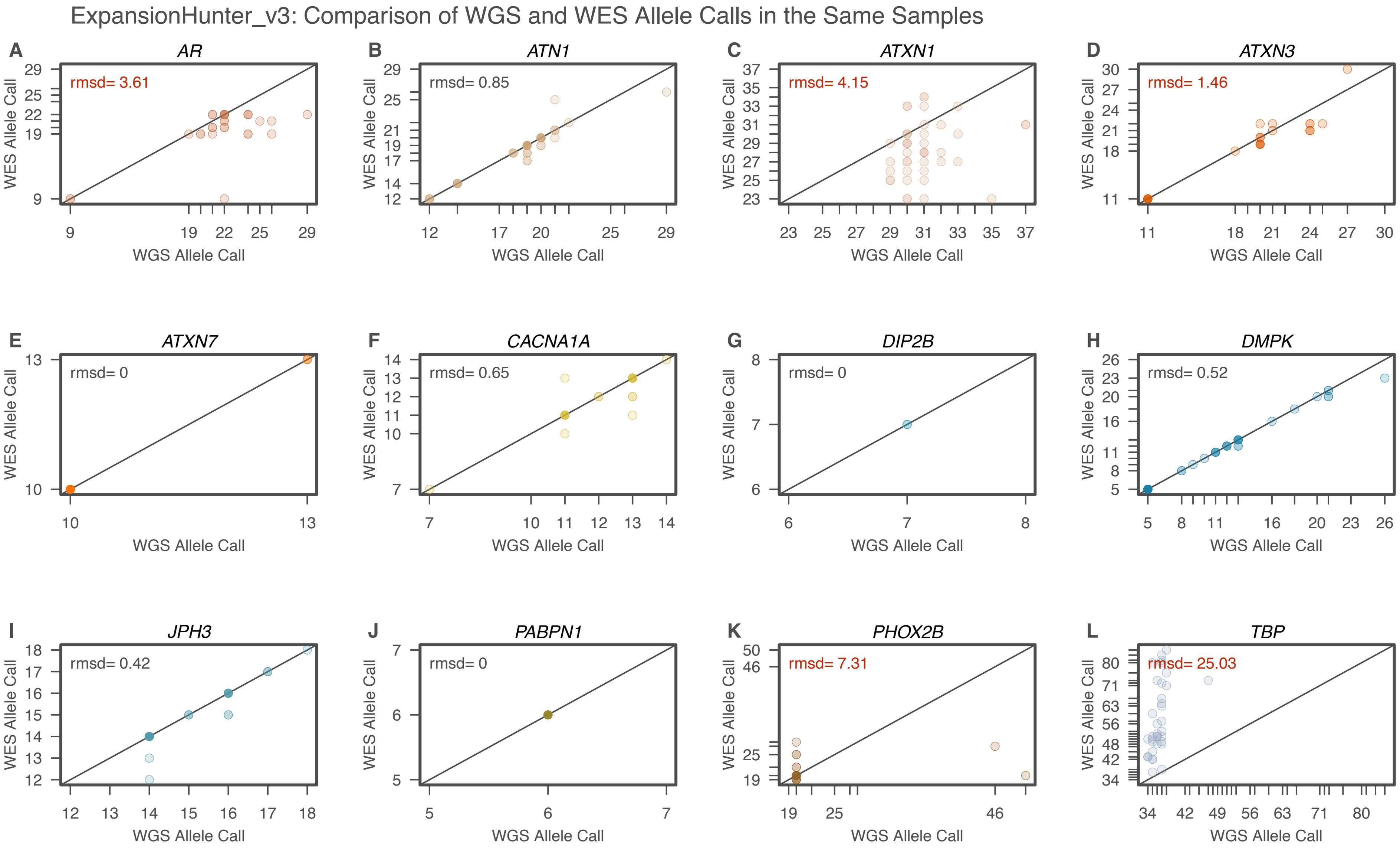


Figure 3.8: ExpansionHunter v3 comparison of gold standard PCR genotyping with in silico predictions

Gold standard PCR genotypes are compared to predicted alleles using the software ExpansionHunter 3.

Note: This is an example figure. Comparable figures are available for all tools (supplementary figures S.3.1-S.3.7)


ExpansionHunter_v3: Comparison of WGS and WES Allele Calls in the Same Samples
A
AR
rmsd= 3.61
WES Allele Call
29
25
22
19
9
WGS Allele Call
9
19
22
25
29
B
ATN1
rmsd= 0.85
WES Allele Call
29
25
21
19
17
14
12
WGS Allele Call
12
17
20
25
29
C
ATXN1
rmsd= 4.15
WES Allele Call
37
35
33
31
29
27
25
23
WGS Allele Call
23
25
27
29
31
33
35
37
D
ATXN3
rmsd= 1.46
WES Allele Call
30
27
24
21
18
11
WGS Allele Call
11
18
21
24
27
30
E
ATXN7
rmsd= 0
WES Allele Call
13
10
WGS Allele Call
10
13
F
CACNA1A
rmsd= 0.65
WES Allele Call
14
13
12
11
10
7
WGS Allele Call
7
10
11
12
13
14
G
DIP2B
rmsd= 0
WES Allele Call
8
7
6
WGS Allele Call
6
7
8
H
DMPK
rmsd= 0.52
WES Allele Call
26
23
20
16
11
8
5
WGS Allele Call
5
8
11
16
20
23
26
I
JPH3
rmsd= 0.42
WES Allele Call
18
17
16
15
14
13
12
WGS Allele Call
12
13
14
15
16
17
18
J
PABPN1
rmsd= 0
WES Allele Call
7
6
5
WGS Allele Call
5
6
7
K
PHOX2B
rmsd= 7.31
WES Allele Call
50
46
25
19
WGS Allele Call
19
25
46
L
TBP
rmsd= 25.03
WES Allele Call
80
71
63
56
48
42
34
WGS Allele Call
34
42
49
56
63
71
80

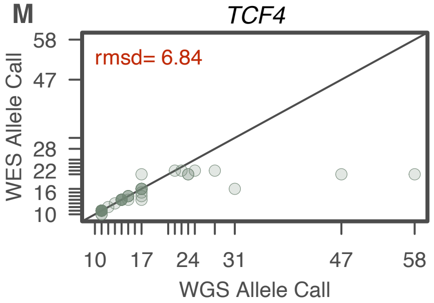


Figure 3.9: ExpansionHunter v3: Comparison of genotype calls from samples sequenced with WES and WGS

Allele calls are compared for 29 samples sequenced with both whole-exome sequencing and whole-genome sequencing, using route-mean-square deviation (RMSD) as a measure of conformation between the two.

Note: This is presented as a sample of the results obtained. Results for all tools are presented in supplementary figures S3.8-S3.13

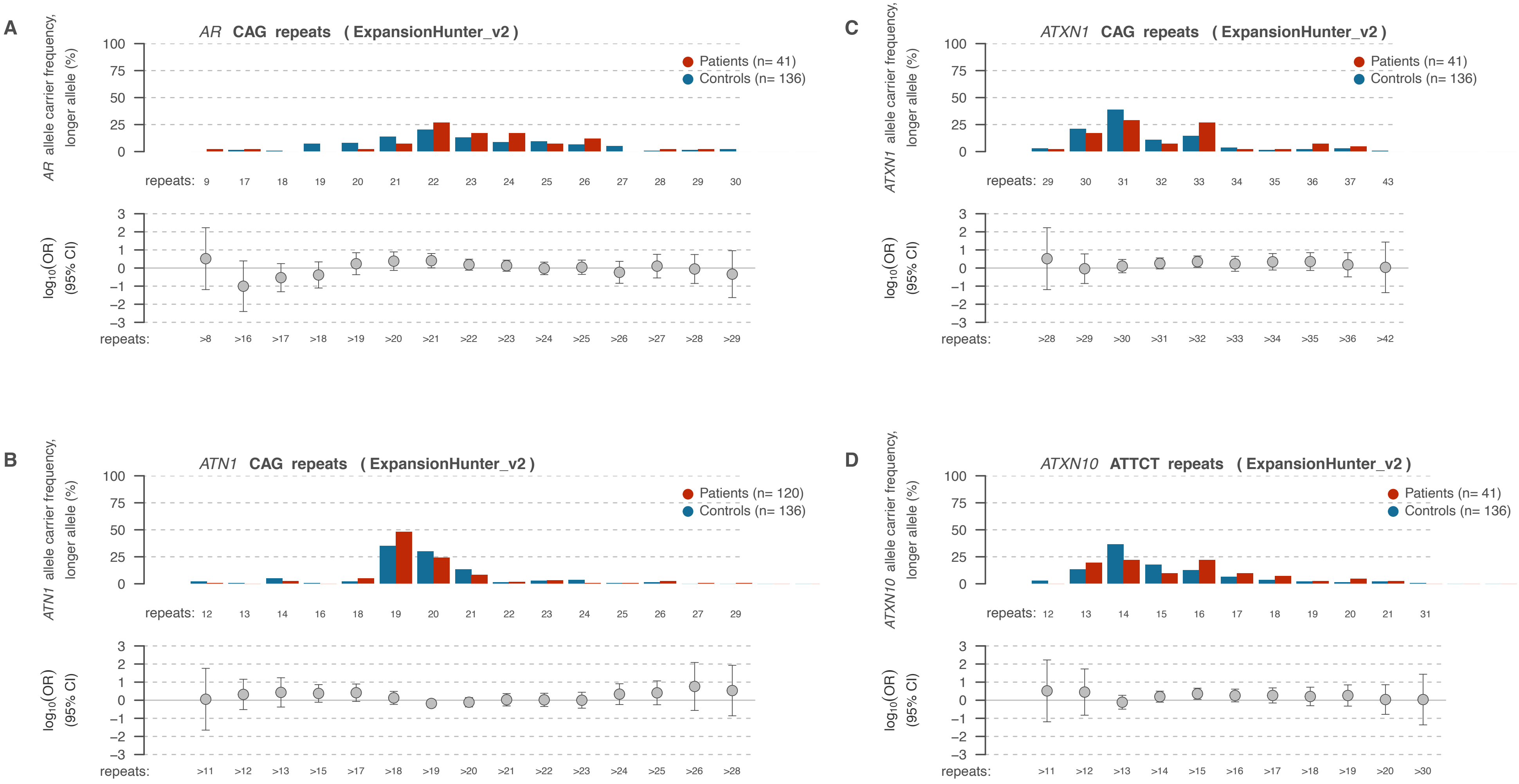
Repeat expansions in epilepsy

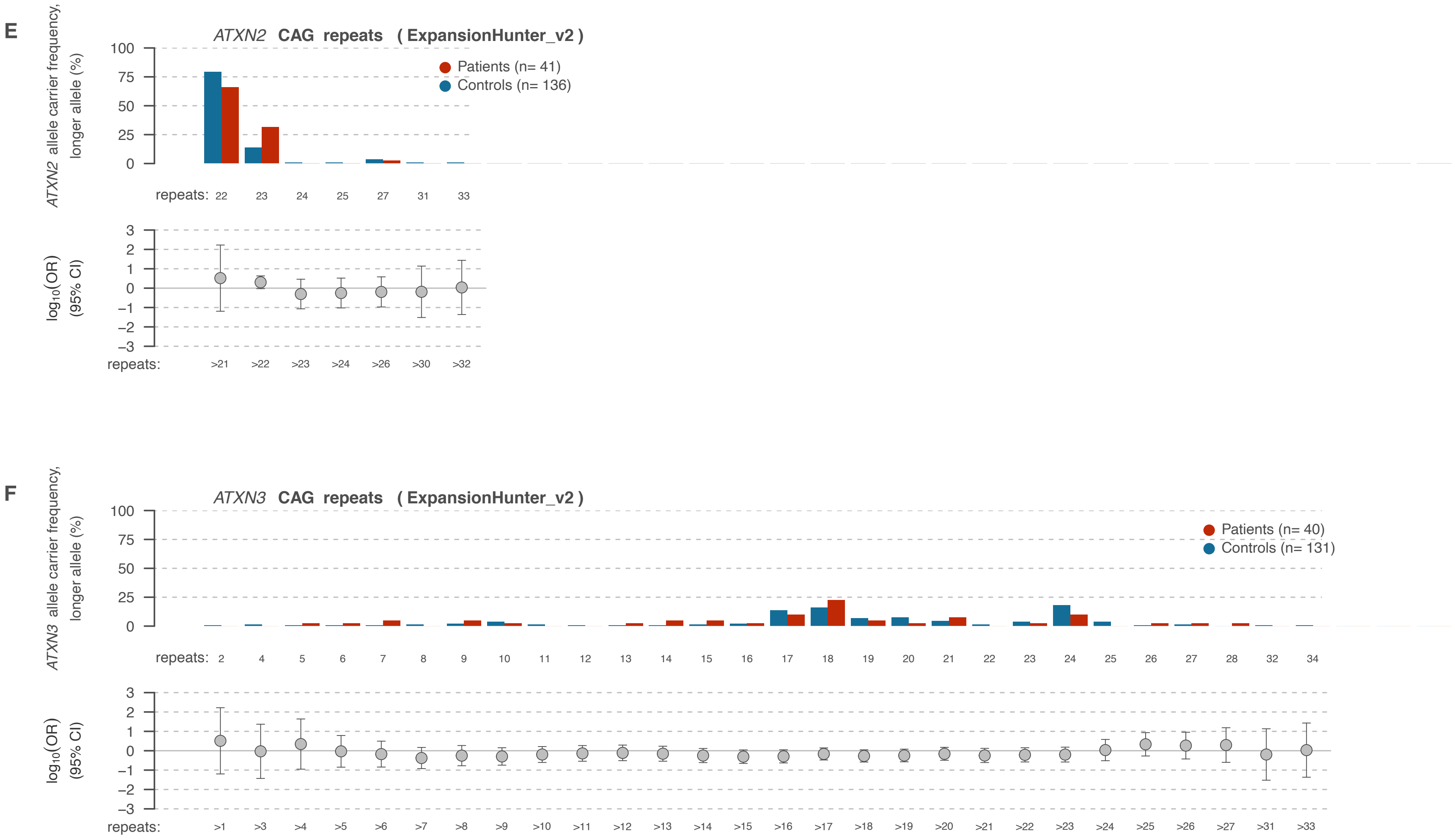
To study the potential pleiotropic effect that RE loci may have on cases of epilepsy in Ireland, 46 RE loci were studied with 7 *in silico* STR genotyping tools in: 114 patients (and most parents) for whom WES was available, 30 patients for whom PCR-free WGS was available and 11 patients (and most parents) for whom WGS with PCR was available. Not all tools were capable of genotyping each locus (supplementary table S3.1). Figures 3.10 & supplementary figures S3.14-S3.19 display the results for patients genotyped with five tools (excluding exSTRa and STRetch), compared to 136 Irish PCR-free WGS samples genotyped concurrently. For each gene, WES samples were included in this plot if a RMSD below one was observed when comparing WES and WGS genotypes, indicating reliable WES genotyping of the software in question.

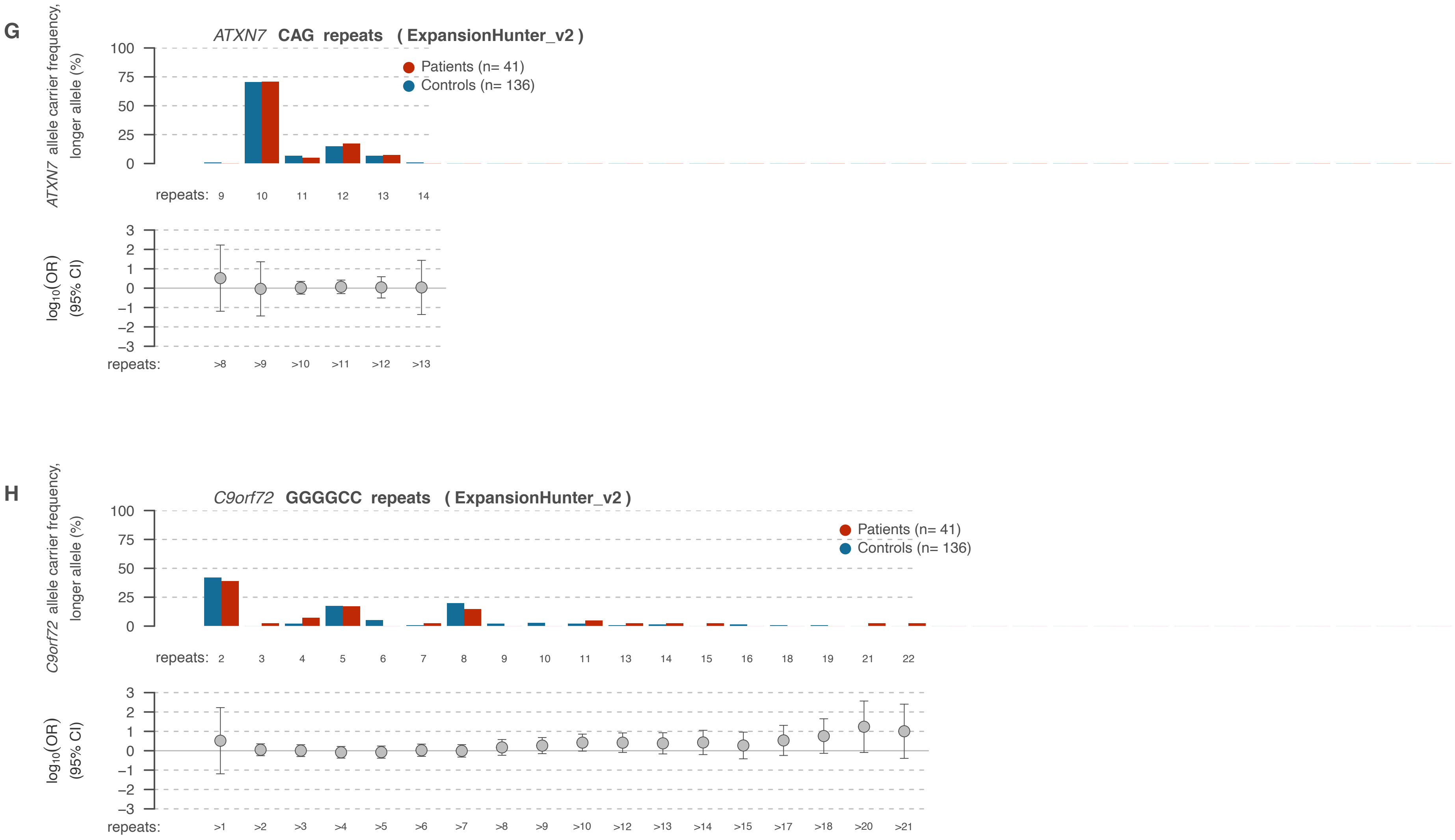
Statistically significant loci

Odds ratios

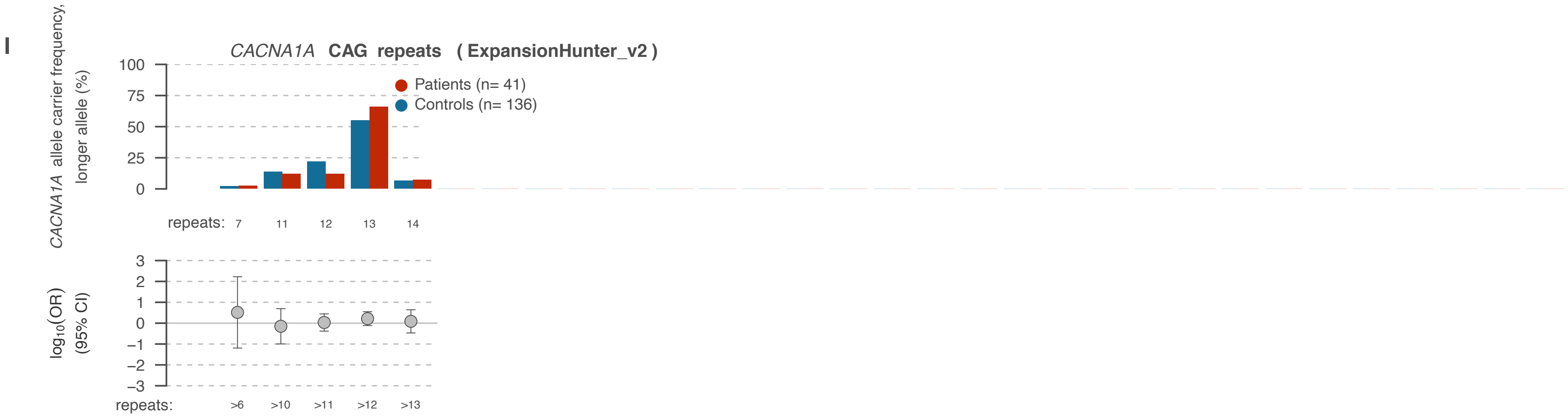
TREDPARSE identifies that more than 25 CTG repeats in the gene *TCF4* are a statistically significant risk factor for developing epilepsy (supplementary figure S.3.19.E.2). Several PCR-free WGS epilepsy samples are predicted to have more than 60 repeats, differing from the predicted distribution in controls. To investigate the validity of this finding, the results of ExpansionHunter v3, exSTRa and STRetch are interrogated. Comparing each tool to gold-standard PCR genotypes, TREDPARSE is found to be prone to false-positives at the *TCF4* locus and ExpansionHunter 3 is found to provide more reliable genotyping (figure 3.8). The same samples are not predicted to be expanded by ExpansionHunter 3 (figure 3.11.A). The CTG repeat in *TCF4* is not exonic; however, it is exon adjacent and subsequently, the WES epilepsy samples have sufficient cover for genotyping with both exSTRa and STRetch (figure 3.11.B); two tools which are capable of identifying significant expansions. STRetch does not identify any significant exome samples at the *TCF4* locus. None of the samples above 60 repeats in TREDPARSE are identified as significant by exSTRa (figure 3.11.C & 3.11.D). The results of ExpansionHunter 3, STRetch and exSTRa signify that the expanded alleles predicted by TREDPARSE are likely to represent false-positives, which TREDPARSE is prone to at this locus (supplementary figure S.3.7).

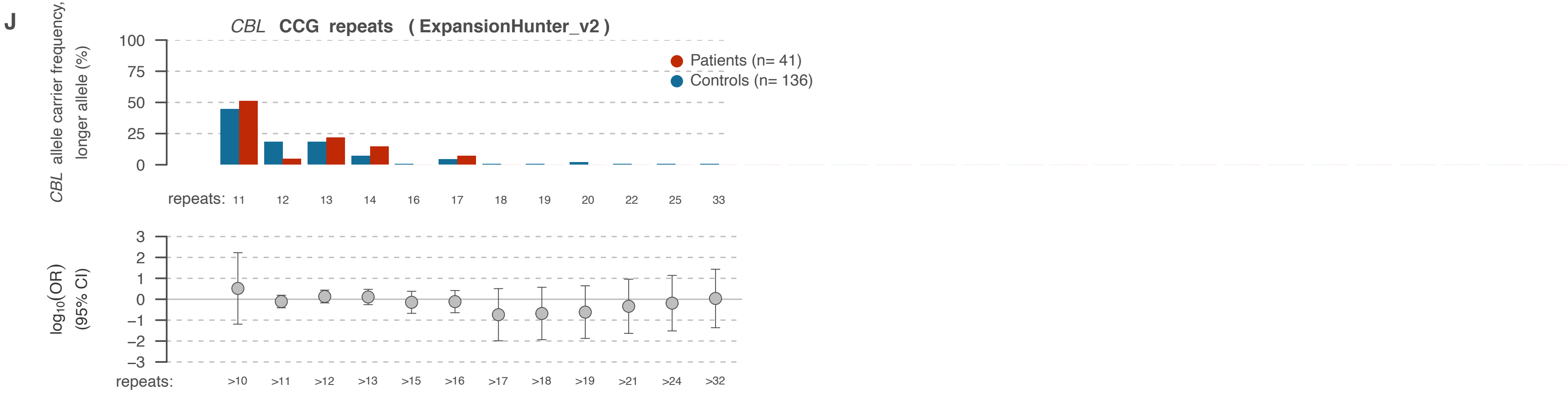

A
AR CAG repeats (ExpansionHunter_v2)
AR allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 136)
repeats: 9 17 18 19 20 21 22 23 24 25 26 27 28 29 30
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >29
B
ATN1 CAG repeats (ExpansionHunter_v2)
ATN1 allele carrier frequency, longer allele (%)
● Patients (n= 120)
● Controls (n= 136)
repeats: 12 13 14 16 18 19 20 21 22 23 24 25 26 27 29
repeats: >11 >12 >13 >15 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >28
C
ATXN1 CAG repeats (ExpansionHunter_v2)
ATXN1 allele carrier frequency, longer allele (%)
● Patients (n= 41)
● Controls (n= 136)
repeats: 29 30 31 32 33 34 35 36 37 43
repeats: >28 >29 >30 >31 >32 >33 >34 >35 >36 >42
D
ATXN10 ATTCT repeats (ExpansionHunter_v2)
ATXN10 allele carrier frequency, longer allele (%)
● Patients (n= 41)
● Controls (n= 136)
repeats: 12 13 14 15 16 17 18 19 20 21 31
repeats: >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >30

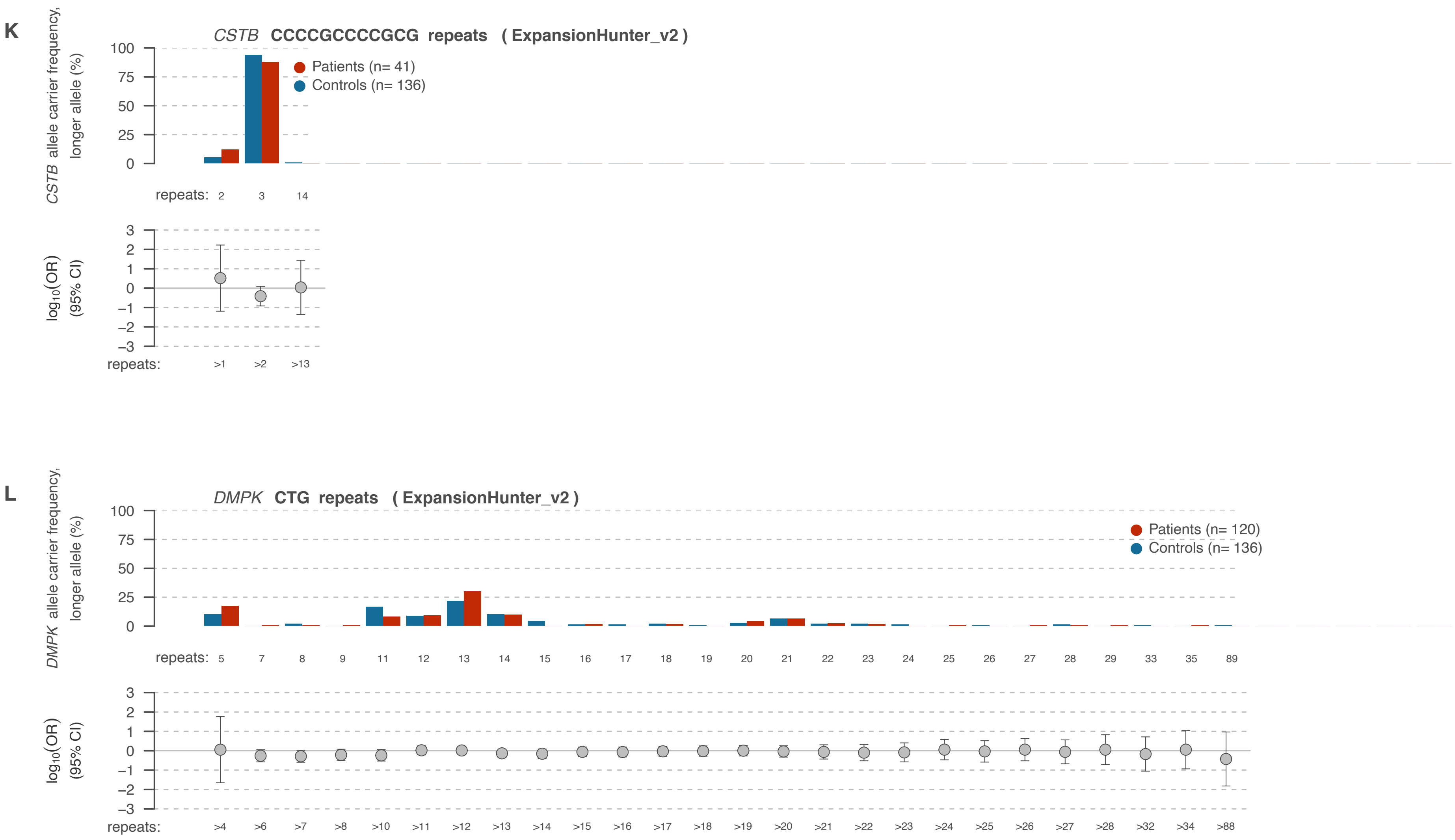

E
ATXN2 CAG repeats (ExpansionHunter_v2)
ATXN2 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 22 23 24 25 27 31 33
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >21 >22 >23 >24 >26 >30 >32
F
ATXN3 CAG repeats (ExpansionHunter_v2)
ATXN3 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 40)
Controls (n= 131)
repeats: 2 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 32 34
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >1 >3 >4 >5 >6 >7 >8 >9 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >31 >33

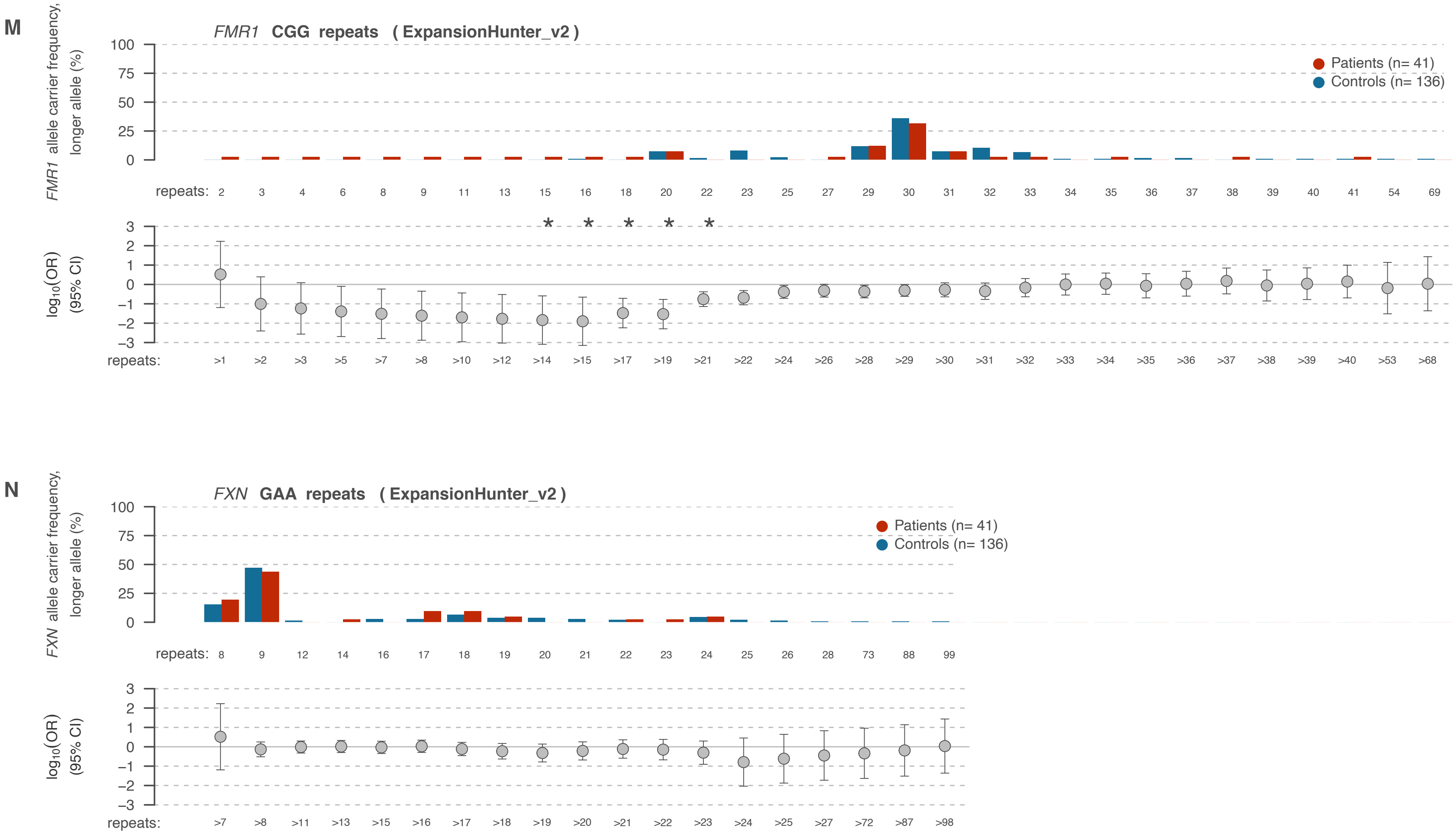


G
ATXN7 CAG repeats (ExpansionHunter_v2)
ATXN7 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 136)
repeats: 9 10 11 12 13 14
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >9 >10 >11 >12 >13
H
C9orf72 GGGGCC repeats (ExpansionHunter_v2)
C9orf72 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 136)
repeats: 2 3 4 5 6 7 8 9 10 11 13 14 15 16 18 19 21 22
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >1 >2 >3 >4 >5 >6 >7 >8 >9 >10 >12 >13 >14 >15 >17 >18 >20 >21


I
CACNA1A CAG repeats (ExpansionHunter_v2)
CACNA1A allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 7 11 12 13 14
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >10 >11 >12 >13


J
CBL CCG repeats (ExpansionHunter_v2)
CBL allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 11 12 13 14 16 17 18 19 20 22 25 33
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >10 >11 >12 >13 >15 >16 >17 >18 >19 >21 >24 >32


K
CSTB CCCCGCCCCGCG repeats (ExpansionHunter_v2)
CSTB allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 2 3 14
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >1 >2 >13
L
DMPK CTG repeats (ExpansionHunter_v2)
DMPK allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 120)
Controls (n= 136)
repeats: 5 7 8 9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 33 35 89
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >4 >6 >7 >8 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >32 >34 >88


M
FMR1 CGG repeats (ExpansionHunter_v2)
FMR1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 2 3 4 6 8 9 11 13 15 16 18 20 22 23 25 27 29 30 31 32 33 34 35 36 37 38 39 40 41 54 69
* * * * *
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >1 >2 >3 >5 >7 >8 >10 >12 >14 >15 >17 >19 >21 >22 >24 >26 >28 >29 >30 >31 >32 >33 >34 >35 >36 >37 >38 >39 >40 >53 >68
N
FXN GAA repeats (ExpansionHunter_v2)
FXN allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 8 9 12 14 16 17 18 19 20 21 22 23 24 25 26 28 73 88 99
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >11 >13 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >27 >72 >87 >98

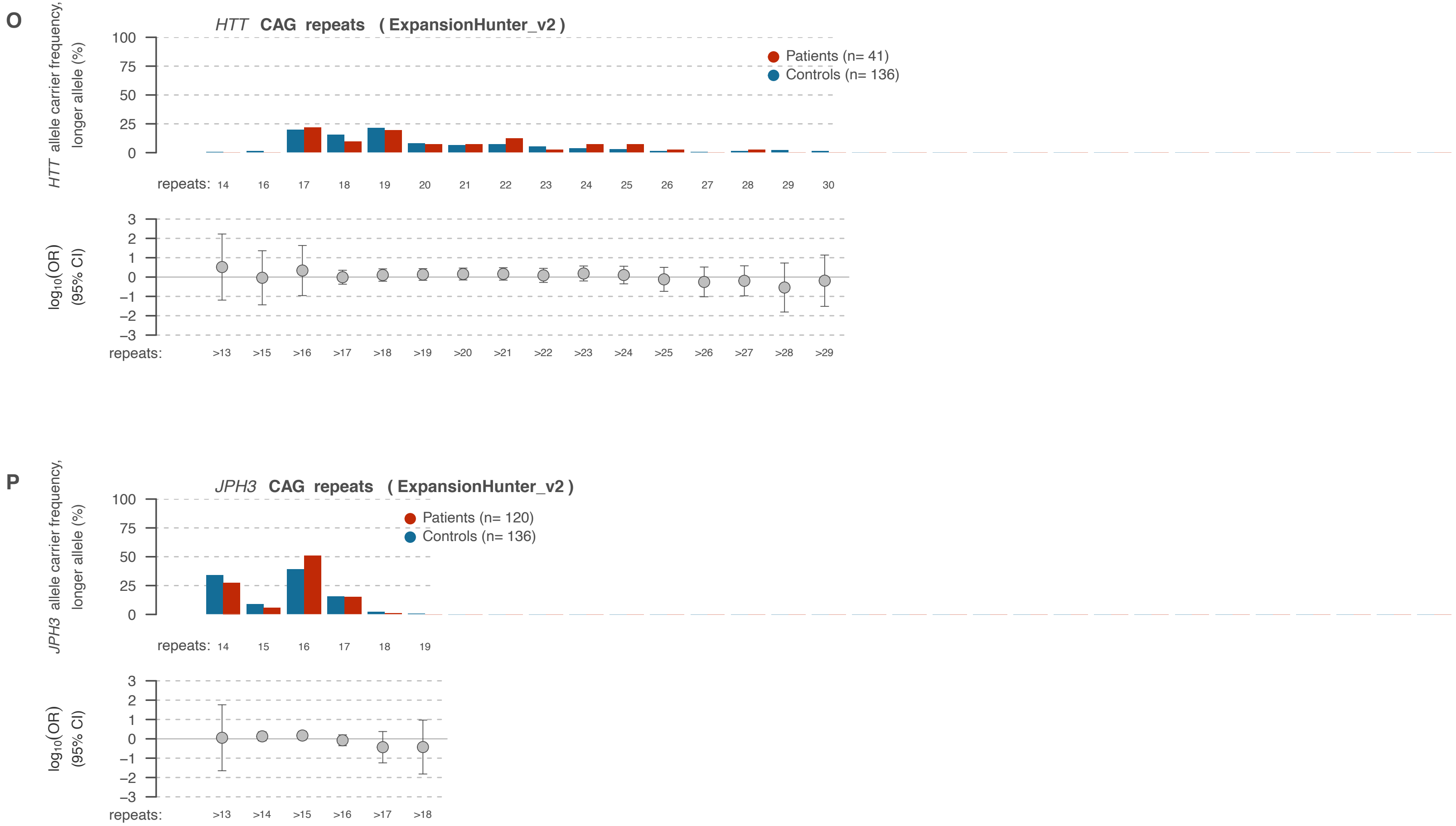


Figure 3.10: ExpansionHunter2 prediction of STR lengths in epilepsy patients

For each gene genotyped with ExpansionHunter2 the allele lengths in epilepsy patients are compared to 136 Irish controls. The upper plot shows the predicted allele lengths and the lower plot shows the OR. An asterisks indicate a significant OR. The epilepsy results include PCR-free WGS samples, PCR WGS samples and WES sample if an RMSD below one was observed when comparing WES results to WGS results for a given gene.

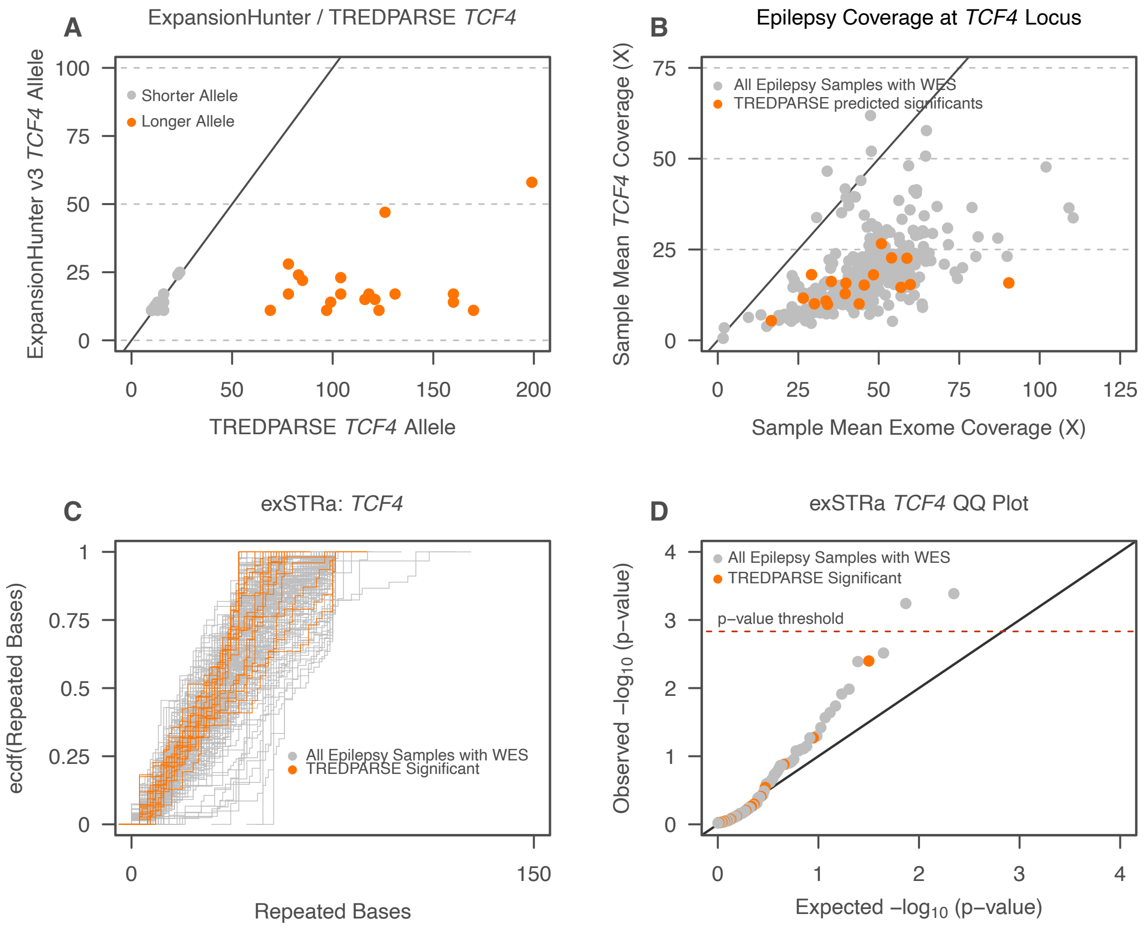


Figure 3.11: Exploration of TREDPARSE predicted *TCF4* expansions

From PCR-free WGS data, TREDPARSE identifies a statistically significant number of epilepsy patients with more than 60 repeats in the gene *TCF4*. To explore whether this is a true finding, the results are compared to other software. A) ExpansionHunter3 is identified to be more reliable at this locus (figure 3.9) and does not confirm these REs. B) These samples also have exome sequencing. The *TCF4* repeat is exon adjacent and these samples have good coverage in exome data and can thus be genotyped with exSTRa and STRetch. C & D) Using exSTRa these samples are not identified as having a statistically significant RE at this locus.

RepeatSeq identifies 18 CCTG repeats in *CNBP* as an epilepsy risk factor while 20 or more repeats is protective against epilepsy (supplementary figure S.3.18.L). This is not supported by other tools (figure 3.10, supplementary figure S.3.14-S.3.19) and RepeatSeq is shown to be unreliable at genotyping the *CNBP* repeat (figure S3.5).

There are instances where repeats above a certain number are identified as significantly protective against epilepsy (GangSTR (Target): *CNBP*, *RFC1*; GangSTR (NonTarget): *ATXN1*, *CNBP*; HipSTR: *DIP2B*; TREDPARSE: *CSTB*, *PABN1*,*HOXA13*; RepeatSeq: *ATXN3*, *DIP2B*, *PPP2R2B*). These findings do not replicate across datasets and are attributed to sequencing/ genotyping differences due to cases and controls originating from different datasets.

There are no instances where a tool identifies an expansion in a patient which is both larger than those observed in controls and in the pathogenic range for the locus in question.

STRetch

STRetch identifies no significant loci in the exome samples. In the WGS with PCR samples, four samples have a statistically significant expansion in *RFC1* and one sample has a statistically significant repeat in *ATXN3*. These samples are not identified as significant by exSTRa (table 3.4). ExpansionHunter 3 shows that the *ATXN3* sample that STRetch flags as significant has inherited the parental alleles. ExpansionHunter 3 shows some variability in *RFC1* genotyping; however, the proband alleles are within the 95% CI of the parental alleles (table 3.4) and are all below the range observed in controls (supplementary figure S3.5).

Table 3.4: Epilepsy patients with a predicted significant RE by STRetch

Gene	Patient	STRetch p-value	STRetch Allele	exSTRa p-value	EH3 Patient Alleles	EH3 Paternal Alleles	EH3 Maternal Alleles
RFC1	EP1A	4.8E-12	41	1	10 (10-10) / 48 (37-57)	10 (10-10) / 34 (30-38)	10 (10-10) / 45 (33-54)
RFC1	EP2A	8.9E-13	44	N/A	39 (30-48) / 48 (31-65)*	10 (10-10) / 56 (41-69)	10 (10-10) / 49 (38-58)
RFC1	EP3A	2E-10	37	1	38 (30-47) / 47 (30-65)*	9 (9-9) / 47 (34-57)	38 (30-46) / 46 (30-62)*
RFC1	EP4A	1.2E-4	23	1	10 (10-10) / 43 (30-52)	9 (9-9) / 46 (35-55)	10 (10-10) / 10 (10-10)
ATXN3	EP4A	7.2E-7	41	N/A	11 (11-11) / 27 (27-27)	11 (11-11) / 24 (24-24)	11 (11-11) / 27 (27-27)

*Flagged for Low Depth

EH: ExpansionHunter

EH alleles include the allele prediction and 95% CI

STRetch p-values are adjusted for multiple testing

exSTRa

exSTRa identifies 53 significantly expanded STRs in 15 patients (supplementary table S.3.2). Where genotypes are called, all putative expansions are examined with STRetch, ExpansionHunter v2 and V3, GangSTR (Targeted) and TREDPARSE. 44 of the 53 putative expansions are confirmed by at least two tools to either be not significant (STRetch), or to have two parental alleles or two alleles which are below the maximum allele length observed in controls, confirming that they are false-positives (supplementary table S3.2, figure 3.10, supplementary figure S3.14-S3.19).

There are insufficient genotyping calls to validate 9 putative REs identified by exSTRa, so these loci are manually inspected.

exSTRa identifies two samples with a significant expansion in *LRP12*. STRetch does not identify either sample as harbouring a significant expansion. Neither sample is genotyped by an additional tool; consequently, for further confirmation, reads were extracted from the repeat region in each sample and visually examined. One sample displays stutter error but appears to be heterozygous for 9 and 12 repeats and the second sample is heterozygous for 5 and 12 repeats (supplementary figure S3.20) . Both of these samples are in the normal range, well below the pathogenic cutoff (table 3.1).

exSTRa identifies a significant repeat in *SAMD12* in a single patient. Reads were manually extracted from the repeat region and visually inspected (supplementary figure S.3.21). The patient has two clearly identifiable unexpanded alleles of 12 and 20 repeats, both of which are in the normal range, and neither carry the pathogenic TTCAA interruptions (Ishiura *et al.* 2018).

A repeat in *YEATS2* in a single patient is identified as significant by exSTRa. This is an intronic repeat and only WES is available for the sample in question. The sample has only two reads in the repeat region, neither of which indicate a pathogenic expansion; and only four reads in the 500bp region surrounding the locus. There is insufficient evidence to support calling an RE at this locus.

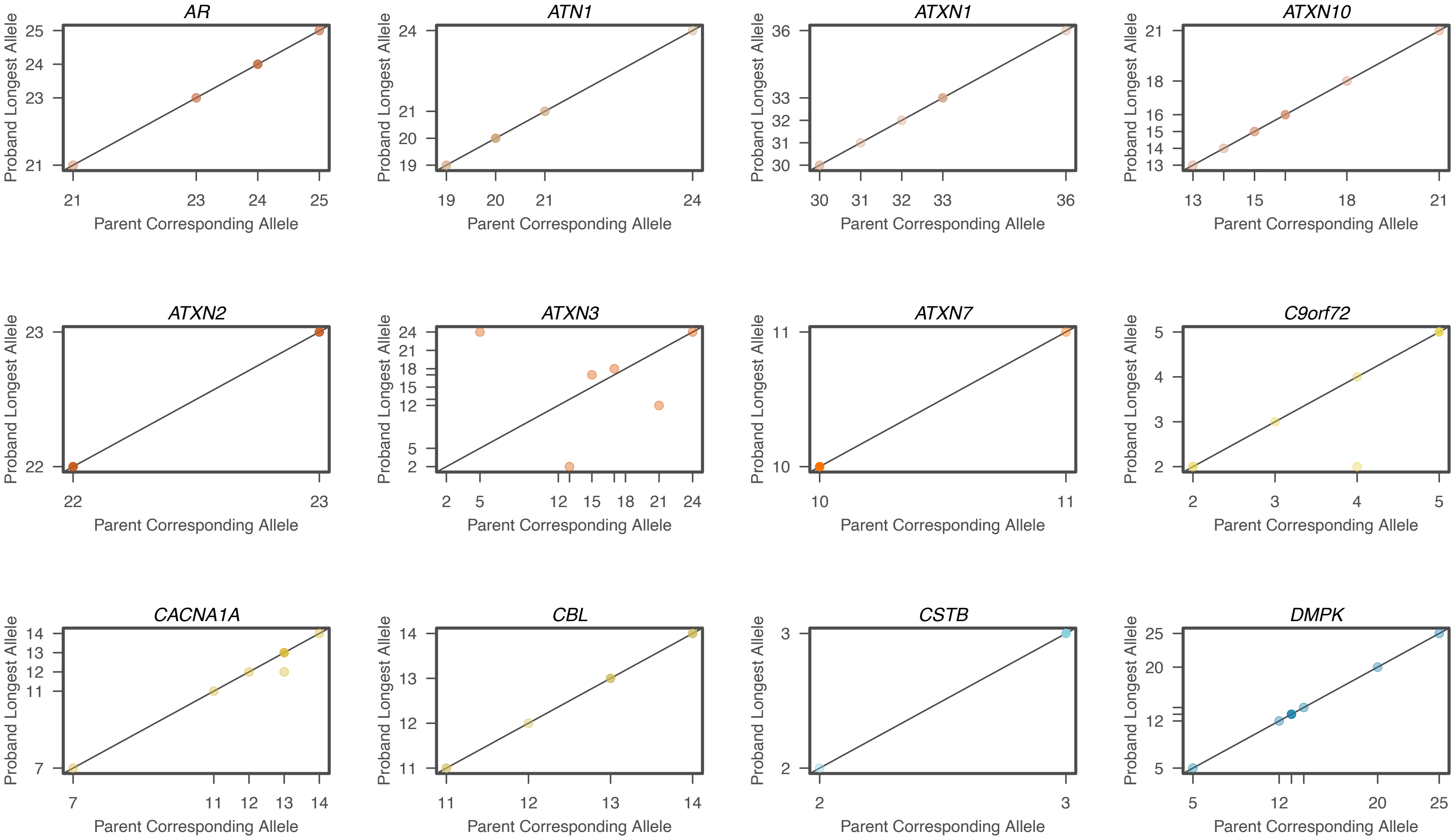
A putatively significant RE in two patients is observed in the gene *NUTM2B*. The two samples appear homozygous for 13 motif units (supplementary figure S.3.22), providing evidence that the predicted REs at this locus are false-positives.

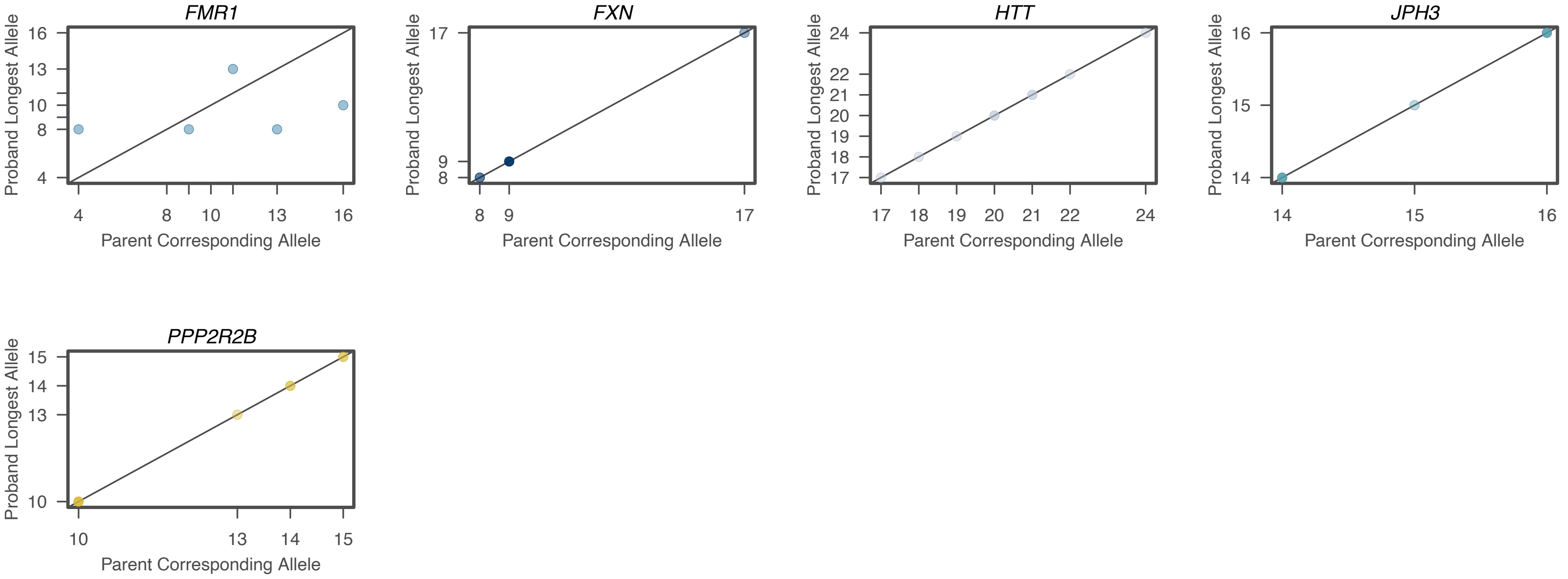
Three patients have predicted REs in *NOTCH2*. The samples are very deeply sequenced at this locus (supplementary figure S.3.23), which could indicate a true repeat; however, this is found to be proportional to their overall exome depth of sequencing (supplementary figure S.3.24). The observed reads at this locus do not support a prediction of an RE (supplementary figure S.3.23).

De novo STRs

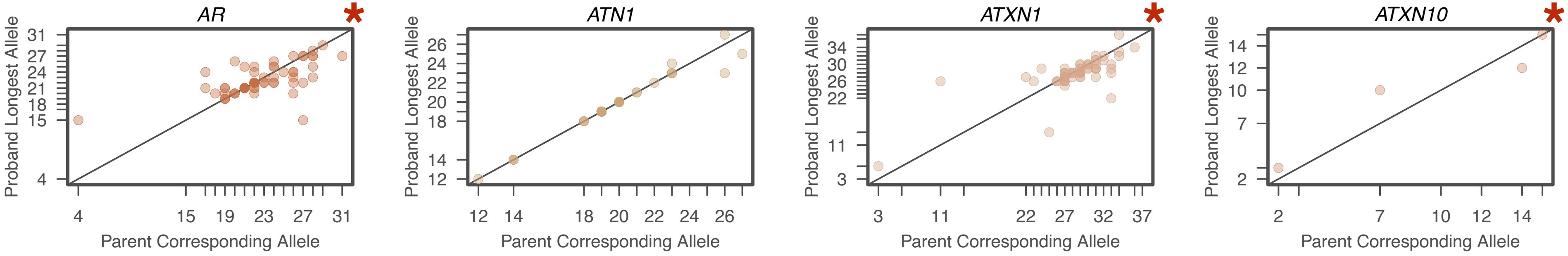
To directly study putative *de novo* mutations, across all tools, for both the WES samples and the WGS with PCR samples, the longest allele in each patient is compared to the longest allele in either of their parents (figure 3.12, supplementary figures S.3.25-S.3.30). While some variability is observed in the reporting of each locus (particularly for exome data), there are no instances where a *de novo* mutation is observed outside the range of alleles reported in parents.

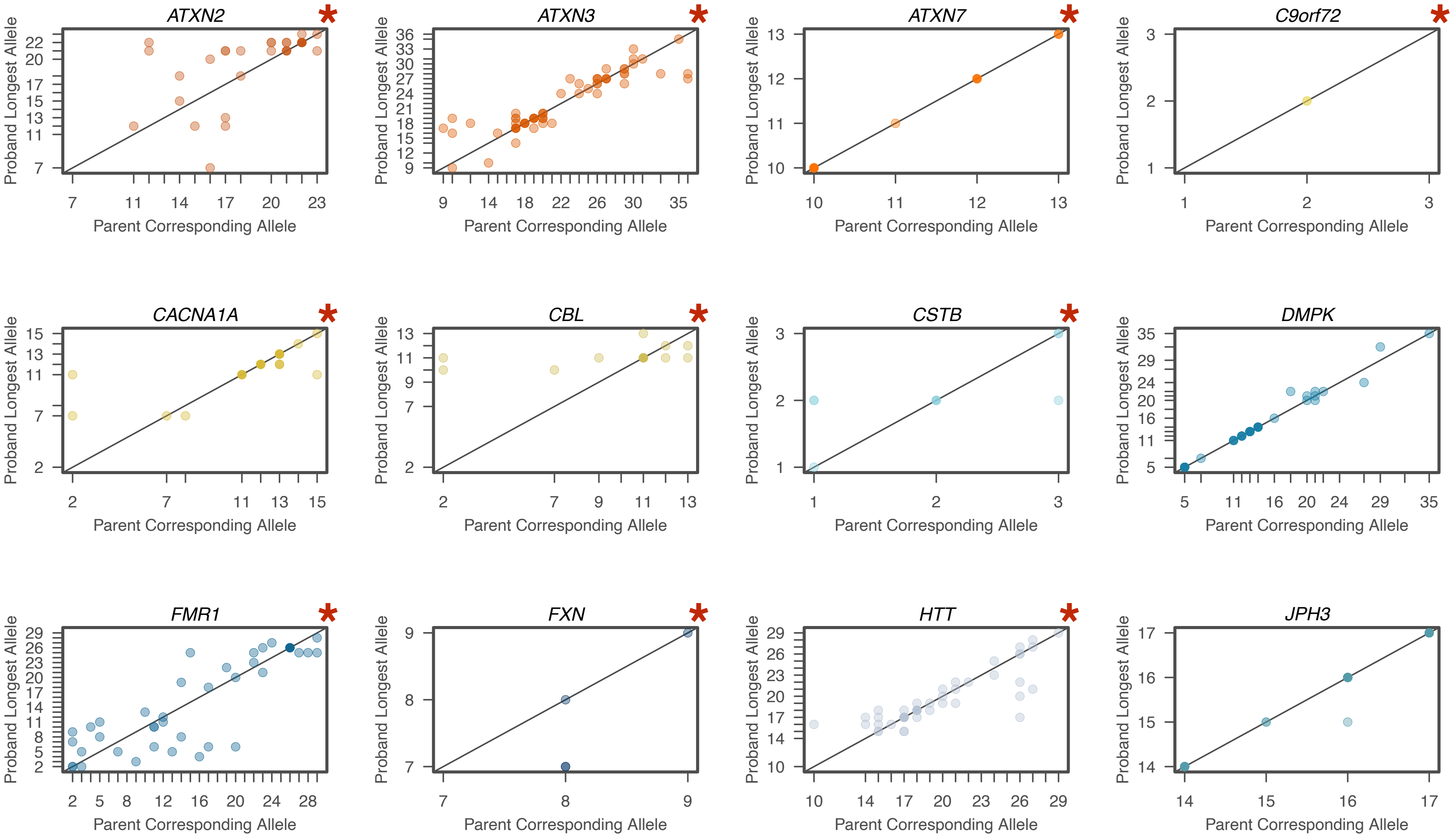
ExpansionHunter_v2: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WGS PCR Data





ExpansionHunter_v2: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WES Data




ATXN2 *
Proband Longest Allele
7 11 13 15 17 20 22
Parent Corresponding Allele
7 11 14 17 20 23
ATXN3 *
Proband Longest Allele
9 12 15 18 21 24 27 30 33 36
Parent Corresponding Allele
9 14 18 22 26 30 35
ATXN7 *
Proband Longest Allele
10 11 12 13
Parent Corresponding Allele
10 11 12 13
C9orf72 *
Proband Longest Allele
1 2 3
Parent Corresponding Allele
1 2 3
CACNA1A *
Proband Longest Allele
2 7 11 13 15
Parent Corresponding Allele
2 7 11 13 15
CBL *
Proband Longest Allele
2 7 9 11 13
Parent Corresponding Allele
2 7 9 11 13
CSTB *
Proband Longest Allele
1 2 3
Parent Corresponding Allele
1 2 3
DMPK
Proband Longest Allele
5 11 16 20 24 29 35
Parent Corresponding Allele
5 11 16 20 24 29 35
FMR1 *
Proband Longest Allele
2 5 8 11 14 17 20 23 26 29
Parent Corresponding Allele
2 5 8 12 16 20 24 28
FXN *
Proband Longest Allele
7 8 9
Parent Corresponding Allele
7 8 9
HTT *
Proband Longest Allele
10 14 17 20 23 26 29
Parent Corresponding Allele
10 14 17 20 23 26 29
JPH3
Proband Longest Allele
14 15 16 17
Parent Corresponding Allele
14 15 16 17

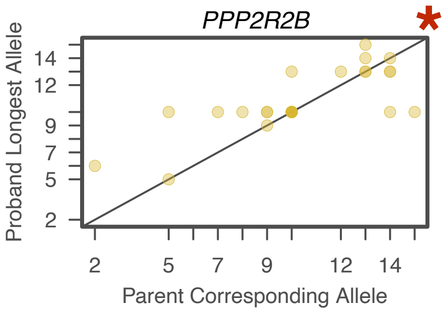


Figure 3.12: ExpansionHunter v2 exploration of potential de novo REs in epilepsy patients

For each patient the longest observed allele at a given locus is compared relative to the longest allele observed in parental samples. A red asterisks indicates that the gene in question had poor concordance when comparing WES and WGS genotypes for the same samples, consequently WES genotypes may not be reliable.

This figure is presented as an example. Results for all tools are available in supplementary figures S3.25-S3.30.

Discussion

STR expansions play a large role in human phenotypic variation and disease. *In silico* genotyping of STRs and identification of pathogenic REs holds great promise to broaden the understanding of neurological diseases and to revolutionise patient treatment, potentially greatly reducing a patient's time to diagnosis. However, it is essential to validate the accuracy of these tools and to interpret their results appropriately.

Here, three analyses are presented. Firstly, a useful benchmarking study is performed assessing the sensitivity, specificity and RMSD of seven STR genotyping tools, assessing their capabilities both for STRs in the normal range within the healthy population and for expanded repeats. Secondly, comparisons are drawn between identical samples sequenced with both WES and WGS to determine which tools are applicable to WES data and whether this is uniform across different loci. Finally, cohorts of patients with epilepsy were examined, to identify if any currently known pathogenic repeats have a pleiotropic effect, causing epilepsy, in the Irish population.

Benchmarking STR genotyping tools

In this study six STR genotyping tools are assessed in 272 Irish ALS patients and 136 age and population matched controls. Approximately 10% of these patients carry a large pathogenic GGGGCC repeat in the *C9orf72* gene which has been assayed by rpPCR. Additionally gold-standard PCR genotypes are available for 338 samples in a further 23 genes.

While each tool has presented its own supportive evidence upon publication, to date objective benchmarking studies have been limited in their number and scope. This study is unique in that PCR data is available for a larger array of genes and samples than has previously been included in benchmarking studies; providing novel insight into the generalisability of these tools and their ability to accurately genotype repeats in the normal range.

Overall both ExpansionHunter2 and 3 are found to outperform other tools when considering sensitivity, specificity and RMSD. ExpansionHunter is capable of accurately genotyping both large REs and short STRs. Other tools either do not (or have limited capacity to) identify large REs (HipSTR, GangSTR (genome-wide mode), RepeatSeq, TREDPARSE) or less successfully identify large REs while having a significantly higher number of false-positives (exSTRa, STRetch, GangSTR (target mode)).

A novel observation of this study is that, for all tools, including ExpansionHunter, the observed accuracy is gene dependent. For example, in figure 3.8 it is seen that ExpansionHunter is prone to false-positives in *FXN*, *GIPC1*, *TBP* and *TCF4*. Interestingly Rajan-Babu et.al. (2020) also identified a false-positive in *TBP*, indicating that these findings are not a unique feature of this dataset.

The findings of this study highlight the potential of *in silico* genotyping tools to accurately classify both REs and STRs in the general population but stress the importance of validating results as misdiagnoses could lead to poor patient treatment and outcome.

Genotyping REs in WES data

It remains an open question whether *in silico* STR genotyping tools can accurately classify STRs and REs in WES data. This study provides novel insight into this question for many tools. With the exception of TREDPARSE (which is not designed to work with WES data), the tools which do not provide accurate classification of long repeats (HipSTR and RepeatSeq) have good concordance between WES and WGS results; however, this is at the expense of failing to genotype long repeats (table 3.3). Consequently these tools may be useful to genotype short alleles in circumstances where no RE is suspected. ExpansionHunter provides the best results considering it is capable of genotyping both long and short alleles.

ExpansionHunter considers the 1kb region surrounding a repeat; as a consequence WES-WGS concordance is highly gene-dependent, in a similar manner to WGS-PCR concordance. For example in figure 3.9 it is seen that many large expansions are seen in the gene *TCF4* in the WES samples that are not seen in the WGS samples. Contrastingly, it is also possible that WES data may be more reliable at certain loci due the absence of misassigned off-target reads. For example the gene *TCF4* is prone to false positives in WGS data (figure 3.7): in figure 3.9 it is seen that some samples with predicted expansions in WGS data are not confirmed in WES data, so it is feasible that the unexpanded WES call is correct.

Repeat expansions in epilepsy

Repeat expansions have previously been shown to exhibit pleiotropic effects. For example the CAG repeat in *ATXN2* serves as both an ALS risk factor and a cause of SCA2 and the *C9orf72* RE can result in both ALS and FTD as well as rare cases of Parkinson disease, Huntington disease-like syndrome and Alzheimer's disease (Woollacott and Mead 2014).

Epilepsy is a highly heterogenous group of neurological diseases with many patients having an undiagnosed, underlying genetic basis. Repeat expansions have previously been linked to certain forms of adult-onset epilepsy (Ishiura *et al.* 2018). A study of epilepsy patients with ID and their parents in Ireland identified a genetic cause in 31% of patients (Benson *et al.* 2020). The current study utilises the benchmarking performed above in ALS data, and applies this to another neurological condition, investigating whether pleiotropy at currently known RE loci is a pathogenic cause of epilepsy in the Irish population.

Data was available from three cohorts: 114 patients with WES, 30 patients with PCR-free WGS and 11 patients with PCR WGS. Parental sequencing was also available for the many patients with WES and PCR WGS. REs in 46 genes were studied with 7 STR genotyping tools, screening each cohort for outliers, statistically significant numbers of repeats and finally screening for putative *de novo* repeats.

Statistically significant putative STRs were identified in 24 genes. All significant REs underwent further validation; combining information from the results of other tools, coverage and, where required, read-level data. Through a combination of these lines of evidence, it is shown that all significant STRs identified in these patients are false positives. This highlights the importance of thoroughly validating putative STRs, using *in silico* evidence as a first-pass and final PCR validation of remaining significant expansions.

This study does not find evidence supporting the pleiotropic role of known pathogenic REs in epilepsy in the Irish population.

Study limitations

There are limitations to the current study establishing the pleiotropic effects of established pathogenic STRs on epilepsy in Ireland. Firstly, care has been taken to validate which genes provide reliable results from WES data and thus can be reliably included in this analysis;

however, that validation did not include samples with known large expansions. Consequently, it is feasible that certain genes provide reliable WES genotyping in the normal range but do not successfully genotype large expansions, leading to false-negatives in WES samples. On the other hand, previous studies have successfully identified expansions from WES data, albeit from a small number of loci (Rajan-Babu *et al.* 2020).

WES is available for the majority of epilepsy patients in this study; and consequently, these samples are not genotyped for the majority of repeats which are intronic, except where sufficient off-target coverage is obtained. It is feasible that pathogenic non-exonic repeats may be present in patients for whom only WES sequencing is available.

By necessity this study combines different datasets. Specifically, it is not ideal that in the epilepsy study, cases and controls were not sequenced concurrently. However, it is likely that this has increased the rate of false-positives rather than false-negatives, as *in silico* genotyping of differentially sequenced samples can yield artifacts which appear as positive expansions. By implementing a strict downstream validation of putative positive results it is hoped that this is negated.

The final limitation of this study is that the expansions studied here are not exhaustive: new pathogenic expansions are regularly being reported. It is possible that revisiting these samples in future with an updated panel of expansions could lead to new insights.

Gene discovery

One area that is outside the scope of this thesis is the discovery of novel pathogenic STRs in either epilepsy or ALS. Regardless, it is worth discussing the challenges, highlighted by the current research, associated with gene discovery. While it is demonstrated here and in previous studies that *in silico* tools have good capacity to identify REs, the identification of novel STRs remains problematic.

ExpansionHunter Denovo is demonstrated here and in previous studies (Dolzhenko *et al.* 2020; Rafehi *et al.* 2019) to be capable of identifying novel loci, it also benefits from not requiring any *a priori* knowledge of target loci. However, it either requires large numbers of cases and controls or a highly homogenous disease cohort, wherein a single locus is responsible for a majority of cases. Many neurological conditions such as epilepsy and ALS are very heterogenous with patients having varied underlying genetic causes. Further, ExpansionHunter Denovo is only applicable to PCR-free WGS data.

This study demonstrates that many of the tools which are capable of performing genome-wide screens (HipSTR, GangSTR (genome-wide mode), TREDPARSE and RepeatSeq) do not successfully identify true REs. These tools also require a prespecified list of target loci, relying on *a priori* knowledge of likely repeat loci. This means that complex repeats or repeats which are not in the reference genome will be unknowingly excluded from analysis. While these tools do not accurately identify REs, they can show a shift in distribution from unexpanded samples (figure 3.6 & 3.7). A method has been developed utilising the minor discrimination of GangSTR or HipSTR to identify *de novo* STR expansions (Mitra *et al.* 2020). The poor sensitivity of both GangSTR and HipSTR makes this method unsuitable for identifying specific pathogenic STRs but useful for identifying patterns of variation. For example Mitra *et al.* (2020) demonstrated an excess of expanded STRs in promoters of fetal expressed genes in autism patients.

The majority of tools which are currently capable of identifying true REs (ExpansionHunter, GangSTR (target mode) and exSTRa) are not tractable at the genome-wide scale due to resource-costly screening of off-target loci and partner reads. These tools, especially GangSTR and exSTRa, have poor specificity, which if scaled genome-wide would provide thousands of false positives requiring verification.

STRetch is one tool which is scalable genome-wide, can identify true REs and functions in both exome and genome data. The disadvantages of STRetch are that, similar to other tools, it requires a target list of repeats so will not identify repeats that are absent from the reference genome and it requires a case/ control cohort. Further, it is shown here that STRetch has poor resolution for discerning repeat size (and is therefore only applicable to identifying large repeats), and is also shown here to be susceptible to false-positives which would result in thousands of false-positives at the genome-wide scale.

In summary, there are currently four approaches to identifying novel STR expansions. ExpansionHunter Denovo requires either a homogenous case cohort or a large number of cases and controls. HipSTR and GangSTR can be used to identify patterns of variation in pedigree data or STRetch can be used with further downstream validation of results to eliminate false-positives. Finally, these tools can be used in combination with linkage studies, first narrowing the location to a specific genomic region and applying genotyping tools to the isolated region (Bennett *et al.* 2020).

Summary

This is a useful benchmarking study that includes valuable data both in the form of accurate PCR genotypes from a broader array of genes than has previously been studied and a number of samples for whom both WES and PCR-free WGS was available. Considering all metrics ExpansionHunter is identified as the most accurate classifier of REs in both WES and WGS data; however, it has limitations. It is only available for a small number of loci and is identified here to have accuracy that is highly gene dependent.

While PCR-free WGS is ideal for *in silico* studies of REs, it is highlighted here that with appropriate filtering and stringent downstream validation it is possible to achieve meaningful insight from disparate datasets. This study does not identify the pleiotropic effect of established pathogenic REs as contributing to epilepsy in Ireland while acknowledging the limitations of the research conducted here.

Chapter 4

The genetic profile of ALS in Cuba

Introduction

Cuba is a Caribbean island with a population approaching 11.5 million people. The modern Cuban population is primarily a mix of Native Americans who first arrived between 4,500-4,000 BC, Spanish settlers who arrived in the 15th century and sub-Saharan Africans who arrived as slaves between the 16th and 19th centuries. For individuals, the ratio of ancestral origins differs across the country, however the average European ancestry for an individual is 71.1% (SD 23.4%), African ancestry accounts for 20.3% (SD 25.1%), Native American accounts for 6.9% (SD 4.6%) and East Asian ancestry accounts for 1.7% (SD 2.5%) (Fortes-Lima *et al.* 2018).

Chapter 2 highlighted the unfortunate fact that the majority of ALS patients still lack a genetic diagnosis. The true spectrum of ALS genetic variation cannot be understood if the majority of genetic studies are not representative of a diverse array of individuals and populations. A large proportion of confirmed Pathogenic or Likely Pathogenic variants exhibit significant geographic heterogeneity; they are present at an elevated rate in certain regions, countries or continents. Genetic research in understudied populations can identify variants and genes unseen in other populations and can aid in planning and stratifying human trials. Cuba is one such understudied population.

The ALS mortality rate in the Cuban population is similar to other Hispanic populations and slightly below the rates observed in Northern European populations (Zaldivar *et al.* 2009). Consistent with the Chapter 2 result and previous studies (Byrne *et al.* 2013) suggesting that countries with lower life expectancy exhibit an earlier age of onset on average, the mean age of onset for Cuban patients is earlier (53 (95% CI: 50.4-55.6)) than for Irish patients (61.6 (95% CI: 60.9 -62.4)) (Ryan, Zaldívar Vaillant, *et al.* 2019). The rate of FALS is higher in Cuba (15.8%) than in Ireland (11.8%). Previous research in Cuba has found that individuals with self-reported mixed ancestry have a lower risk of ALS than those who self-identify as black or white, indicating a protective effect of admixture (Zaldivar *et al.* 2009).

The work in this Chapter is the first study of the genetics of ALS in Cuba, a unique, understudied, admixed population. The journALS database outlined in Chapter 2 is utilised as an analysis and interpretation aid at all stages of the results process.

Research Aims

- Explore the profile of Cuban ALS genetics by performing targeted DNA sequencing of a panel of previously associated genes.
- Utilise the journALS database to interpret the sequencing results at the variant and population levels.

Methods

Study participants

All ALS patients participating in this study presented at the National Institute of Neurology, Havana, Cuba, which serves as a national tertiary referral centre for neurodegenerative conditions, between 1996 and 2017 (Ryan, Zaldívar Vaillant, *et al.* 2019). A specialist neurologist diagnosed all patients with definite, probable or possible ALS as defined by the El Escorial criteria (Brooks *et al.* 2000). Demographic and phenotypic information including age of onset, site of onset, disease duration, sex, family history and province of residency were recorded. In accordance with official Cuban guidance, self-reported skin colour ('white', 'black' or 'mestizo') was also reported. Cuban controls were neurologically normal at the time of sampling and included spouses of patients and volunteers. No further phenotype information is available for controls. DNA extraction from venous leucocytes was performed in Cuba.

DNA samples were divided into five batches, combining cases and controls within batches to prevent batch effects. The following sections describe the steps undertaken for each batch.

Targeted-sequencing library preparation

Dual-indexed sequencing libraries were prepared for each DNA sample following the KAPA HyperPlus KR1145-v3.16 protocol with minor modifications. DNA was quantified using either a Nanodrop ND-1000 spectrophotometer or a Qubit 2.0 fluorometer with dsDNA BR assay Kit. 300ng of DNA (or as much DNA as was available for low quality samples) was initially purified to remove any EDTA from the buffer using Agencourt Ampure XP beads and eluted in Tris-HCl. Resulting purified DNA samples were fragmented for 8 minutes to a target size of 400bp using Kapa HyperPlus fragmentation enzyme. DNA end-repair and A-tailing was performed using the Kapa HyperPlus library preparation kit. NEBNext hairpin adapters were ligated onto the resulting DNA fragments using a 60 minute ligation time. In order to remove uracil and thus open the adapters, the adapter-ligated libraries were treated with USER enzyme with a 60 minute incubation. The resulting libraries were PCR amplified (98°C:45 sec, 8x(98°C:15 sec, 60°C:30 sec, 72°C:30 sec), 72°C:1 min, 4°C:∞) using unique i5 and i7 adapters to index each sample with an individual identifier and to generate sequencer-ready libraries. Samples were assessed for quality (concentration, fragment size distribution) on an Agilent Tapestation.

DNA size selection

Size selection was carried out using gel extract size selection to obtain libraries of the optimum length for sequencing. A 1.5% low weight molecular agarose gel was prepared with the addition of SYBR to a final concentration of 1 in 5000. SYBR stained DNA was visualised with a UV screen and excised between 500bp-600bp. Size selected, library prepared DNA was extracted from the gel cut following the Qiagen MinElute Gel Extraction Protocol. DNA concentration and fragment size distribution were assessed using an Agilent Tapestation and Nanodrop ND-1000.

Design of target enrichment library

We designed an in-solution Integrated DNA technologies Ltd (IDT xGen Lockdown Probes) target enrichment kit to enrich the exons and surrounding 4 bps of 37 genes linked to either ALS or FTD (table 4.1) based on the GRCh37 build of the human genome. The kit was designed prior to the completion of Chapter 2; for this reason, genes were chosen based on their entry in the ALS Online Genetics Database (O. Abel *et al.* 2012) or the Alzheimer's Disease and FTD Mutation Database ("Center for Molecular Neurology" n.d.). *ERLIN1* (Tunca *et al.* 2018), *ERLIN2* (Muratet *et al.* 2019) and *PARK7* (Özoğuz *et al.* 2015; Hanagasi *et al.* 2016) as well as more recently linked genes such as *KIF5A* (Nicolas *et al.* 2018) are not included in the panel for this reason.

Target enrichment and next-generation sequencing

Samples were pooled to equal concentration. A pooled mass of 66ng of DNA was target enriched using the IDT Hybridization capture of DNA libraries using xGen Lockdown Probes protocol. Blocking oligos, Cot-1 DNA and the pooled library were combined and liquid was evaporated using a Savant DNA110 DNA SpeedVac Concentrator. Biotinylated capture probes were hybridised to the library with a 4 hour incubation at 65°C. Biotinylated probes and hybridised DNA were captured using streptavidin coated beads and a magnetic rack. For the first two batches, enriched DNA was PCR amplified with 15 cycles (98°C 45 sec, 15x(98°C 15 sec, 60°C 30 sec, 72°C 30 sec), 72°C 1 min, 4°C ∞), and this was reduced to 11 cycles for the final three batches (98°C 45 sec, 11x(98°C 15 sec, 60°C 30 sec, 72°C 30 sec), 72°C 1 min, 4°C ∞). The library was assessed for quality, concentration and fragment size distribution on an Agilent Tapestation, Nanodrop ND-1000 spectrophotometer and Qubit 2.0 fluorometer with dsDNA BR assay Kit.

Table 4.1: Genes included in targeted NGS panel

Symbol	Gene Name	First Reported Link to ALS, FTD or Dementia
ALS2	Alsin	(Hadano et al. 2001; Yang et al. 2001)
ANG	Angiogenin	(Greenway et al. 2004)
ATXN2	Ataxin 2	(Elden et al. 2010)
C21orf2	Cilia And Flagella Associated Protein 410	(van Rheenen et al. 2016)
CHCHD10	Coiled-Coil-Helix-Coiled-Coil-Helix Domain Containing 10	(Bannwarth et al. 2014)
CHMP2B	Charged multivesicular body protein 2b	(Skibinski et al. 2005; Parkinson et al. 2006)
DAO	D-Amino Acid Oxidase	(Mitchell et al. 2010)
DCTN1	Dynactin	(Münch et al. 2004)
ELP3	Elongator complex protein 3	(Simpson et al. 2009)
ERBB4	Erb-B2 Receptor Tyrosine Kinase 4	(Takahashi et al. 2013)
FIG4	Polyphosphoinositide phosphatase	(Chow et al. 2009)
FUS	Fused in sarcoma	(Kwiatkowski et al. 2009; Vance et al. 2009)
GRN	Progranulin	(Baker et al. 2006; Cruts et al. 2006)
HNRNPA1	Heterogeneous Nuclear Ribonucleoprotein A1	(Kim et al. 2013)
LMNB1	Lamin B1	(Johnson et al. 2014)
MAPT	Microtubule Associated Protein Tau	(Hutton et al. 1998)
MATR3	Matrin 3	(Johnson et al. 2014)
NEFH	Neurofilament, heavy polypeptide	(Figlewicz et al. 1994)
NEK1	NIMA Related Kinase 1	(Kenna et al. 2016)
OPTN	Optineurin	(Maruyama et al. 2010)
PFN1	Profilin 1	(Wu et al. 2012)
PRPH	Peripherin	(Beaulieu et al. 1999; Gros-Louis et al. 2004)
PSEN1	Presenilin 1	(Raux et al. 2000)
PSEN2	Presenilin 2	(Gallo et al. 2010)
SARM1	Sterile Alpha And TIR Motif Containing 1	(Fogh et al. 2014)
SETX	Senataxin	(Chen et al. 2004)
SIGMAR1	Sigma non-opioid intracellular receptor 1	(Luty et al. 2010)
SOD1	Superoxide dismutase 1	(Rosen et al. 1993)
SPAST	Spastin	(Meyer et al. 2005)
SPG11	Spatacsin	(Orlacchio et al. 2010)
SQSTM1	Sequestosome 1	(Fecto et al. 2011)
TAF15	TATA-Box Binding Protein Associated Factor 15	(Ticozzi et al. 2011)
TARDBP	TAR DNA-binding protein 43	(Arai et al. 2006; Neumann et al. 2006)
TBK1	TANK Binding Kinase 1	(Freischmidt et al. 2015)
UBQLN2	Ubiquilin 2	(Deng et al. 2011)
UNC13A	Unc-13 homolog A	(van Es et al. 2009)
VAPB	VAMP-associated protein B	(Nishimura et al. 2004)
VCP	Valosin-containing protein	(Kovach et al. 2001; Johnson et al. 2010)

Note: this panel was designed prior to the completion of chapter one and prior to the discovery of recently associated ALS genes. As such relevant genes including ERLIN1, ERLIN2, PARK7 and KIF5A are omitted.

The size-selected, pooled, target enriched libraries were diluted to 4nM in 5uL and sequenced on an Illumina MiSeq at the TrinSeq facility at St. James's Hospital with 300bp single end sequencing.

C9orf72 genotyping

Patients were screened for the presence of the pathogenic *C9orf72* RE by repeat-primed PCR (rpPCR) as described previously (Byrne *et al.* 2012). Amplified fragments were measured by capillary electrophoresis on an Applied Biosystems 3500 Series Genetic Analyzer and visualised using Gene Mapper v.4.0, screening for a decreasing sawtooth pattern which is indicative of a large RE. Patients with 30 hexanucleotide repeats or above and displaying a sawtooth rpPCR trace were deemed positive for the expansion.

Bioinformatic pre-processing

Next-generation sequencing generated single-end FASTQ files which were processed following the Genome Analysis Toolkit (GATK) best practices (as of 18/06/2018) (Van der Auwera *et al.* 2013). Sequences were adapter trimmed using cutadapt v.1.9.1 (M. Martin 2011). Reads were aligned to the GRCh37 version of the human reference genome (downloaded from the UCSC genome browser (W. J. Kent *et al.* 2002)), using the Burrows-Wheeler Aligner (BWA) v.0.7.5 (H. Li and Durbin 2009). Aligned sam files were converted to bam format, sorted, indexed and depth of coverage in targeted regions was calculated using samtools v.1.7 (H. Li *et al.* 2009). Picard v.0.7.5 (http://broadinstitute.github.io/picard/) was used for duplicate read removal, and to add read groups.

Base Quality Score Recalibration

Base Quality Score Recalibration (BQSR) was performed using GATK v.3.8 (McKenna *et al.* 2010). During next-generation sequencing, a quality score is assigned to each base. This quality score represents the likelihood that a base is sequenced incorrectly. BQSR detects, and corrects for, systematic errors made by the sequencing machine in assigning these quality scores. In the initial phase of BQSR, the genome is traversed to identify variant single-nucleotide sites and INDELs. SNPs and INDELs that are known to commonly vary (Sherry, Ward, and Sirotkin 1999; Mills *et al.* 2011), are masked to avoid counting truly variant sites as errors. For non-masked variant sites, the read group, reported quality score, machine cycle and previous dinucleotide are recorded. This is used to build a recalibration model that is subsequently utilised to adjust each base quality score according to the properties of the base.

Variant calling

Variant calling was performed in accordance with GATK best practices (as of 18/06/2018). Variant calling was performed using GATK's HaplotypeCaller and GenotypeGVCFs functions. Together these tools identify and assign likelihoods to SNPs and INDELs by performing local de-novo haplotype assembly in variant regions and assigning variant likelihoods based on the haplotypic context. Hard-filtering was applied to identify variants which fail QC based on various sequencing metrics (SNPs: QualByDepth (QD) < 2, FisherStrand (FS) > 60, StrandOddsRatio (SOR) > 3, RMSMappingQuality (MQ) < 40, MappingQualityRankSumTest (MQRankSum) < -12.5, ReadPosRankSumTest (ReadPosRankSum) < -8; INDELs: QD < 2, FS > 200 , ReadPosRankSum < -20).

Variant annotation

SNPs and INDELs were annotated to ensure compatibility with the journALS data from Chapter 2. Variants were normalised and annotated using Variant Tools v0.5772 (Tan, Abecasis, and Kang 2015), SnpEff v4.3s (Cingolani *et al.* 2012) and GEMINI v0.30.2 (Paila *et al.* 2013). As per Chapter 2, variants were annotated with variant AFs from Project MinE (van der Spek, van Rheenen, Pulit, Kenna, McLaughlin, *et al.* 2019), ALSVS ("ALS Variant Server, Worcester, MA" n.d.), ALSdb ("ALSdb, New York City, New York" n.d.; Cirulli *et al.* 2015), and the non-neuro subset of gnomAD (Karczewski *et al.* 2020). *In silico* annotations were added via dbNSFP 4.0a (X. Liu, Jian, and Boerwinkle 2011, 2013; X. Liu *et al.* 2016), spidex 1.0 (Xiong *et al.* 2015) and dbscSNV1.1 (Jian, Boerwinkle, and Liu 2014). Insertions and deletions (INDELs) were annotated using PROVEAN v1.1 (Choi *et al.* 2012; Choi 2012), SIFT (Sim *et al.* 2012) and VEST4 (Douville *et al.* 2016). *In silico* annotations were analysed as per Chapter 2.

Variant filtering and analysis

A bespoke analysis pipeline was applied to filter the observed variants to a set of putatively pathogenic variants. Variants failing variant calling QC filters were removed. Variants were then filtered to those present in cases, and either absent in controls, or, if the variant was homozygous in a case, was not homozygous in any control. Variants classified as Benign or Likely Benign in the journALS database were removed and only variants with a functional effect, as predicted by snpEff, were retained (conservative_inframe_deletion, conservative_inframe_insertion, disruptive_inframe_deletion, structural_interaction_variant, missense_variant, exon_loss_variant, disruptive_inframe_insertion, frameshift_variant, initiator_codon_variant, splice_acceptor_variant, splice_donor_variant, start_lost, stop_gained, stop_lost). Variants which were heterozygous in all cases were removed if they exceeded 1% in the gnomAD non neuro subset or if they exceeded 2% if any case was homozygous. The final filter was to remove variants with a control AF exceeding the case AF in the Project MinE data. Remaining variants were analysed in the context of the results of Chapter 2.

Exome sequencing

Following the initial targeted gene sequencing, sufficient DNA was available to perform exome sequencing for five of the six members of an affected pedigree to a mean target coverage of 35X on an Illumina NovaSeq with 2x150bp paired-end sequencing with Agilent SureSelect V7 target enrichment. Library preparation and sequencing were performed by Macrogen (Macrogen Inc.,1002, 254 Beotkkot-ro, Geumcheon-gu, Seoul, 153-781, Republic of Korea). Samples were sequenced concurrently with 44 Irish PLS samples and members of an Irish ALS pedigree including 4 affected family members and 13 currently unaffected relatives. These Irish samples are further described in Chapters 3 and 5. For the remainder of this chapter the Irish PLS samples are treated as sequencing controls on the basis that rare variants shared between a significant number of Irish PLS samples and a Cuban ALS pedigree exhibiting dominant inheritance, represent sequencing errors.

Exome alignment and variant calling

Data alignment, variant calling and annotation was performed as described above with the two exceptions that data was treated as paired-end rather than single-end and that, as per the GATK best-practices, there was sufficient data to perform Variant Quality Score Recalibration (VQSR) rather than variant hard-filtering.

With enough data, VQSR is preferable to the hard-filtering previously performed. VQSR constructs a model based on a training set of high-confidence variants in order to identify the manner in which various variant annotations of good and bad variants cluster and assign a new variant quality measure, the Variant Quality Score Log-Odds (VQSLOD); a continuous estimate of the probability that each variant is true. Each variant is now filtered or retained based on the profile of all of its quality scores rather than any individual score.

HapMap v3.3 (International HapMap Consortium 2003), 1000 Genomes phase 1 (1000 Genomes Project Consortium *et al.* 2015) and Mills INDELs (Mills *et al.* 2011) were used as training resources to identify true variant sites. The INDEL model accounted for the filters QD, FS, SOR, MQRankSum, ReadPosRankSum and the SNP model additionally accounted for MQ. A VQSLOD filter of 99.9% was assigned which retains 99.9% of the truth training sites.

Exome variant filtering

Potentially pathogenic variants were defined as those variants which were present in all family members, passed the VQSR threshold defined above, had an AF below 0.1% in gnomAD, had a functional impact (as defined above), and were present in no more than 10% of the Irish PLS samples which serve here as sequencing controls.

Exome sample relatedness

The relatedness of Cuban family members was confirmed using plink v.1.9 (Purcell *et al.* 2007). Variants called from exome sequencing data were restricted to SNPs and data was converted to plink format. To avoid artificial inflation of relatedness due to different ethnic backgrounds between Irish and Cuban samples, data was merged with 1,158 individuals from the 1000 Genomes phase 1 data, retaining the intersecting 68,973 SNPs. The plink command --genome was used to construct a relatedness matrix.

ATXN2 genotyping

The gene *ATXN2* contains a CAG RE that is known to cause spinocerebellar ataxia 2 when the repeat length exceeds 34 CAG motifs (Elden *et al.* 2010). Healthy individuals typically contain 22 or 23 repeats; however there is considerable variability in the population. Intermediate length repeats (between 27 and 34 repeats) have been shown to be an ALS risk factor (Elden *et al.* 2010; Van Damme *et al.* 2011; Gellera *et al.* 2012; M.-D. Wang *et al.* 2014; Sproviero *et al.* 2017). Typically PCR is used to determine the allele length at this locus (Pulst *et al.* 1996). Due to limited DNA availability in this study, *ATXN2* genotypes are inferred directly from sequencing data, as per Chapter 3.

The length of the *ATXN2* RE is inferred using both TREDPARSE v0.7.8 (Tang *et al.* 2017) and HipSTR v0.6.2 (Willems *et al.* 2017). These programs were chosen as they can both operate on single-end sequencing data to infer the length of REs that are below the length of a typical read (i.e. REs that are less than 300bp in this study).

The depth of coverage over the *ATXN2* CAG region was determined for each sample using bedtools v2.25.0 (Quinlan and Hall 2010). The root-mean-square deviation (RMSD) between genotype calls from TREDPARSE and HipSTR was calculated while removing genotype calls from samples which fell below a range of coverage thresholds. The optimum coverage threshold was identified which retained the maximum amount of samples whilst ensuring the reliability of genotype calls.

Burden analysis

A gene-based association analysis of rare variants was performed in order to ascertain if the exons of any of the genes identified as carriers of P or LP variants in Chapter 2 contain a statistically significant excess of either missense or LOF variants relative to controls in this study.

Efficient and Parallelizable Association Container Toolbox (EPACTS) v.3.3 ("EPACTS - Genome Analysis Wiki" n.d.) was used to assign both functional and gene annotation to all variants which passed sequencing filters and to perform Sequence Kernel Association Tests (SKAT) (M. C. Wu *et al.* 2011). Variants were grouped within genes and were filtered to variants with a MAF below 0.05. Two SKAT tests were performed; the first tested whether any gene harboured an excess of missense variants, the second tested whether any gene harboured an excess of LOF variants (StructuralVariation, Stop_Gain, Stop_Loss, Start_Gain, Start_Loss, Frameshift, CodonGain, CodonLoss, CodonRegion, Insertion, Deletion, Essential_Splice_Site, Nonsense). As phenotypes including age and sex were not available for controls, they were not included as covariates. 121 probands and 102 unrelated controls who passed the sequencing coverage filter were included in the analysis.

Oligogenic analysis

There is mounting evidence supporting the role of oligogenic inheritance in ALS, wherein ALS patients have been observed to harbour multiple variants in ALS-associated genes (van Blitterswijk *et al.* 2012; Nguyen *et al.* 2018; Kuuluvainen *et al.* 2019; McCann *et al.* 2020). Multiple patients are observed to carry multiple variants (table 4.4); to test if this is statistically significant, binomial tests were performed as per van Blitterswijk *et al.* (2012), wherein the expected frequency of two mutations occurring is the product of the frequency of variants in cases and the frequency of variants in controls.

Variants were first filtered to just those which passed sequencing filters and which have a functional effect (as predicted by snpEff: conservative_inframe_deletion, conservative_inframe_insertion, disruptive_inframe_deletion, structural_interaction_variant, missense_variant, exon_loss_variant, disruptive_inframe_insertion, frameshift_variant, initiator_codon_variant, splice_acceptor_variant, splice_donor_variant, start_lost, stop_gained, stop_lost). The filtered dataset was first tested retaining rare variants below a range of gnomAD AFs. As these results could be influenced by the inclusion of non-definitive ALS-associated genes, the same test was performed on variants falling within genes with P or LP variants identified in Chapter 2. A third group of tests was performed to take into account that samples with low coverage could affect results by the false exclusion of rare variants. 105 cases and 75 controls were found to have a mean coverage above 20X and a final group of tests was performed on variants within P or LP genes in these samples.

Statistical analysis and plotting

Unless otherwise stated all statistical analyses and plotting for this chapter were performed using R v3.6.1 (Team 2014) with a suite of packages including beeswarm v.0.4.0 (Eklund and Trimble 2021), berryFunctions v1.18.2 (Boessenkool 2019), binom v1.1-1 (Dorai-Raj 2014), data.table v1.14.0 (Dowle and Srinivasan 2019), ggplot2 v3.3.5 (Hadley Wickham 2016), kinship2 v 1.8.5 (Sinnwell, Therneau, and Schaid 2014), plyr v1.8.6 (H. Wickham 2011), raster v3.4.13 (Hijmans 2021), rcompanion v2.4.1 (Mangiafico 2021), rgdal v.1.5.23 (Bivand, Keitt, and Rowlingson 2021), scales v.1.1.1 (Hadley and Seidel 2019), stringr v1.4 (Hadley Wickham 2019) and tidyr v1.0.2 (Hadley Wickham and Henry 2020).

Results

Study participants

Targeted NGS sequencing was performed for 120 unrelated ALS patients, 6 members of a single pedigree (figure 4.1) and 111 unrelated healthy controls. Detailed phenotype information is available for 93 patients (table 4.4). The demographics of the cohort are found to closely resemble the global ALS population outlined in Chapter 2 (table 4.2). Patients with detailed phenotype information are present from all but one of Cuba's 16 provinces (figure 4.1). While only 19% of the Cuban population is resident in La Habana (ONEI 2021), 29% of patients are from this province, indicating that this region is overrepresented in our study population. The self-reported ethnicities of our cohort (white: 62.4%, mestizo: 24.7%, black: 12.9%) closely resemble the national figures (white: 64.1%, mestizo: 26.6%, black: 9.3%) (ONEI 2021); however, it is important to note that while this indicates that our dataset is not biased to a particular self-reported ethnicity, almost all individuals in Cuba are admixed to some degree regardless of self-reported ethnicity (Fortes-Lima *et al.* 2018). The presence of FTD was not a reported phenotype.

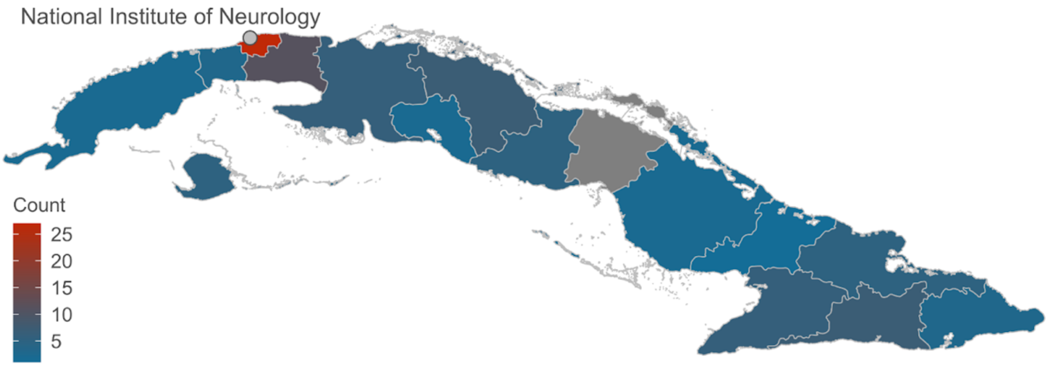


Figure 4.1: Birth provinces of Cuban ALS patients

Patients in this study are present from all but one of Cuba's 16 provinces.

Table 4.2: Summary of Cuban cohort demographics

	Cuba
Age of Onset (years)	54 (95% CI: 51-57)
Disease duration (months)	32 (95% CI: 26.5-43.47)
Sex (male)	55.90%
Site of Onset (spinal)	61.30%
Family History (familial)	12.36%

Note: age of onset and disease duration display the median time in years and months respectively. Disease duration only accounts for individuals who were deceased at the time of follow-up.

Bioinformatic pre-processing

Next-generation sequencing generated 106,981,857 reads across all samples. For each sample, an average of 99.75% (95% CI: 99.67-99.82) of reads aligned to the human genome. figure 4.2. A displays the results of successful adapter trimming for a demonstrative sample. 90% of reads require no trimming and the remaining reads display a range of sizes smaller than 300 bps. For samples that underwent 15 cycles of post-target enrichment PCR 58% (95% CI: 46-70%) of reads were found to be PCR duplicates, this was reduced to 35% (95% CI: 21-48%) by reducing the number of PCR cycles to 11 (figure 4.2 B). The mean sample coverage is 46X (95% CI: 39-52X) for cases and 28X (95% CI: 24-32X) for controls. Samples with a mean coverage below 5X were excluded from further analysis. Figure 4.2 D-F displays the successful application of BQSR for a demonstrative sample.

Variant calling

9 control samples had a target-wide coverage below 5X and were excluded from further analysis. Across all remaining samples, a total of 465 SNVs and 61 INDELs were identified. Following the variant filtering process (figure 4.3 A-I), 73 putatively pathogenic SNPs and 18 putatively pathogenic INDELs remained for further analysis (table 4.3, table 4.4, table 4.5, table 4.6), 39 of these 91 variants are located in genes identified as carriers of pathogenic or likely pathogenic variants in Chapter 2 and are investigated in further detail.

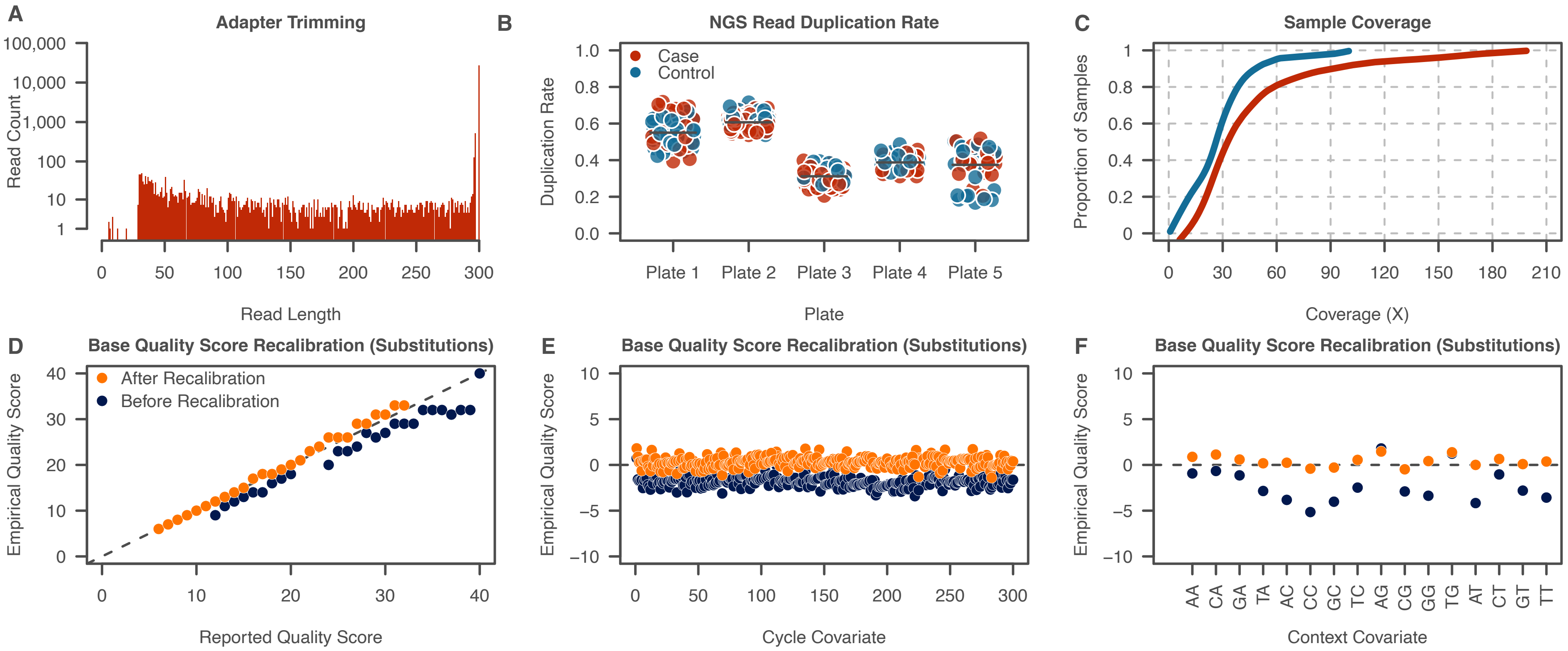


Figure 4.2: Bioinformatic pre-processing of NGS data

Following successful adapter trimming, 10% of reads were below 300bp indicating the successful removal of adapter sequence. B) Samples that underwent 15 cycles of post-target enrichment PCR had a mean duplication of 58% (95% CI: 46-70%). This was reduced to 35% (95% CI: 21-48%) by reducing the number of cycles to 11. C) A cumulative density function is displayed for the coverage in target regions for cases and controls. The mean coverage is 46X (95% CI: 39-52X) for cases and 28X (95% CI: 24-32X) for controls. D-F) For a demonstrative sample, the initial reported base quality score is compared to the recalibrated ‘empirical’ score. Scores are adjusted based on D) the reported quality score, E) read position and F) the preceding dinucleotide.

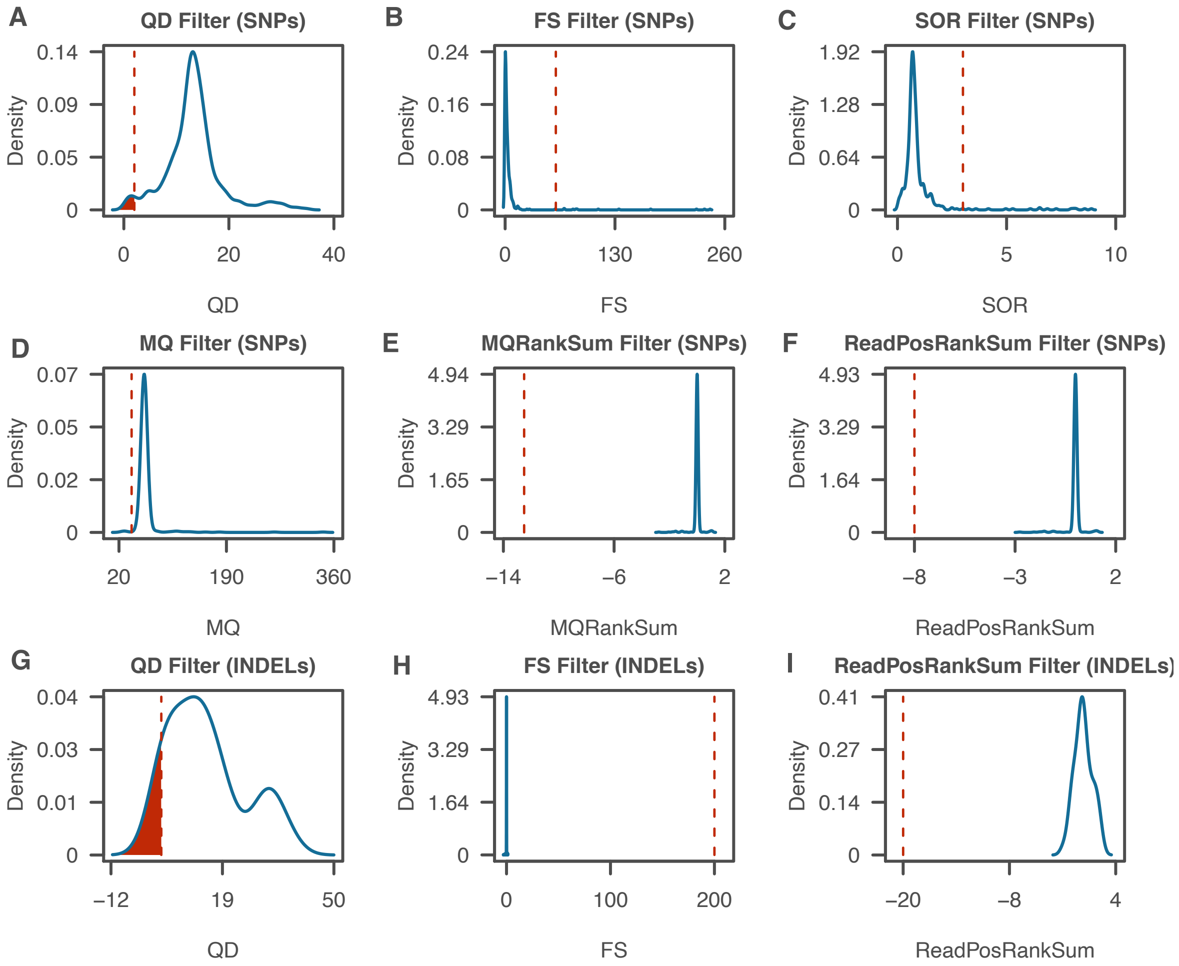


Figure 4.3: Hard-filtering of targeted sequencing variants

A-G show the distributions (blue) and cut-off thresholds (dashed red) of the various annotations used to assess the sequence quality of identified variants. Annotations and thresholds are further described in the text.

Table 4.3: Variant filtering

Filter Description	SNVs Remaining	INDELs Remaining	In journALS	In Literature
Initial variants	465	57	352	119
Variant Calling QC	442	54	344	114
Present in Cases	440	54	344	114
Absent in Controls *	192	25	144	38
Benign in journALS	152	21	100	27
Functional Filter	82	17	56	23
gnomAD Filter	82	16	56	23
ProjectMinE Filter	73	14	45	17
Putative Pathogenic Variants	73	14	45	17

* If homozygous in any case then not homozygous in any control, else if heterozygous in all cases then absent in all controls

Table 4.4 : Cuban ALS patients (1/5)

Pedigree	Sample_ID	AOO	Survival	Sex/Gender	Ethnicity	History	Onset	Condition	HGVS
NA	230101	NA	NA	NA	NA	NA	NA	NA	*ATXN2* :c.178_199dupCCCGGCCCCCCTCCCTCCCGGC(p.[Q67fs])
2302_01	230201	24	>212	F	White	Familial	Bulbar	Alive	*NEFH* :c.1104C>G(p.[D368E]) *hnRNPA1* :c.1018C>T(p.[P340S])
2302_02	230202	37	>11	M	White	Familial	Spinal	Alive	*hnRNPA1* :c.1018C>T(p.[P340S])
2302_03	230203	33	>79	M	White	Familial	Spinal	Alive	*hnRNPA1* :c.1018C>T(p.[P340S])
2302_04	230204	46	>29	M	White	Familial	Spinal	Alive	NA
2302_05	230205	27	490	M	White	Familial	Spinal	Deceased	*hnRNPA1* :c.1018C>T(p.[P340S])
2302_06	230206	NA	NA	F	NA	NA	Bulbar	NA	NA
NA	230301	59	24	M	White	Sporadic	Spinal	Deceased	*ALS2* :c.3167G>C(p.[G1056A])
NA	230401	61	11	M	Black	Sporadic	Bulbar	Deceased	*NEFH* :c.1138G>A(p.[A380T])
NA	230501	37	45	M	Mestizo	Sporadic	Spinal	Deceased	*ALS2* :c.3958A>T(p.[N1320Y]) *FUS* :c.1512_1513delAG(p.[G505fs])
NA	230601	47	24	F	Black	Sporadic	Spinal	Deceased	*MAPT* :c.1483G>A(p.[A495T])
NA	230701	49	59	M	Mestizo	Sporadic	Spinal	Deceased	*C21orf2* :c.505G>A(p.[E169K]) *FUS* :c.684_686dupCGG(p.[G229dup]) *NEFH* :c.410C>T(p.[A137V])
NA	230801	55	26	F	White	Sporadic	Bulbar	Deceased	NA
NA	230901	38	30	F	Mestizo	Sporadic	Bulbar	Deceased	*SETX* :c.6013G>A(p.[V2005M]) *SIGMAR1* :c.622C>T(p.[R208W])
NA	231001	53	21	F	White	Sporadic	Bulbar	Deceased	*GRN* :c.100C>T(p.[P34S])
NA	231101	34	90	M	Mestizo	Sporadic	Spinal	Deceased	NA
NA	231201	61	27	M	White	Sporadic	Spinal	Deceased	NA
NA	231301	62	78	M	White	Sporadic	Bulbar	Deceased	*ERBB4* :c.1669C>T(p.[P557S]) *TMEM199* :c.535C>T(p.[P179S])
NA	231401	52	63	M	White	Sporadic	Spinal	Deceased	*TBK1* :c.466dupA(p.[T156fs])
NA	231501	NA	NA	NA	NA	NA	NA	NA	NA
NA	231601	54	8	M	White	Sporadic	Bulbar	Deceased	*GRN* :c.1288C>G(p.[P430A])
NA	231701	44	86	F	White	Sporadic	Bulbar	Deceased	NA
NA	231801	NA	NA	NA	NA	NA	NA	NA	NA
NA	231901	NA	NA	NA	NA	NA	NA	NA	*ATXN2* :c.2806A>G(p.[T936A]) *PSEN1* :c.1109A>G(p.[N370S]) *TBK1* :c.539delT(p.[L180fs])
NA	232001	59	54	F	Black	Sporadic	Spinal	Deceased	*SPAST* :c.865C>T(p.[H289Y]) *SPG11* :c.6319G>A(p.[V2107I])
NA	232101	47	11	F	White	Sporadic	Spinal	Deceased	NA
NA	232201	56	20	M	White	Sporadic	Spinal	Deceased	NA
NA	232301	75	39	F	Mestizo	Sporadic	Bulbar	Deceased	*MAPT* :c.50C>T(p.[T17M])
NA	232401	46	58	M	White	Sporadic	Spinal	Deceased	NA
NA	232501	51	126	M	Mestizo	Sporadic	Spinal	Deceased	*CHMP2B* :c.560G>A(p.[S187N]) *MAPT* :c.1534C>T(p.[P512S])
NA	232601	66	27	F	White	Sporadic	Bulbar	Deceased	*FUS* :c.684_686dupCGG(p.[G229dup])

Table 4.4: Cuban ALS patients (2/5)

Pedigree	Sample_ID	AOO	Survival	Sex/Gender	Ethnicity	History	Onset	Condition	HGVS
NA	232701	53	19	M	White	Sporadic	Bulbar	Deceased	NA
NA	232801	47	>117	M	White	Sporadic	Spinal	Alive	NA
NA	232901	40	51	F	Mestizo	Sporadic	Spinal	Deceased	NA
NA	233001	58	68	M	White	Sporadic	Bulbar	Deceased	NA
NA	233101	34	45	F	White	Familial	Spinal	Deceased	*TBK1* :c.466dupA(p.[T156fs])
NA	233201	35	48	M	White	Sporadic	Spinal	Deceased	NA
NA	233301	73	19	M	White	Sporadic	Spinal	Deceased	NA
NA	233401	45	62	M	Black	Sporadic	Spinal	Deceased	*OPTN* :c.1457A>G(p.[H486R]) *SPG11* :c.7161A>T(p.[Q2387H])
NA	233501	47	32	M	Black	Sporadic	Spinal	Deceased	NA
NA	233601	53	25	M	White	Sporadic	Bulbar	Deceased	NA
NA	233701	41	55	F	White	Sporadic	Spinal	Deceased	*FIG4* :c.2459+1G>A
NA	233801	58	36	F	Mestizo	Familial	Spinal	Deceased	NA
NA	233901	50	20	M	White	Sporadic	Spinal	Deceased	*C9orf72* :c.-45+163GGGGCC[>24]
NA	234001	69	16	M	White	Sporadic	Spinal	Deceased	NA
NA	234101	40	32	F	Mestizo	Sporadic	Spinal	Deceased	*SPG11* :c.4687A>G(p.[R1563G])
NA	234201	46	29	M	Mestizo	Sporadic	Bulbar	Deceased	NA
NA	234301	54	115	F	White	Familial	Bulbar	Deceased	*SQSTM1* :c.714_716delGAA(p.[K238del])
NA	234401	58	>14	F	Black	Sporadic	Spinal	Alive	*SETX* :c.3663G>C(p.[K1221N])
NA	234501	56	>17	M	Black	Sporadic	Spinal	Alive	NA
NA	234601	NA	NA	NA	NA	NA	NA	NA	*SETX* :c.6122T>C(p.[I2041T])
NA	234701	57	36	M	Black	Sporadic	Spinal	Deceased	*ATXN2* :c.1769C>T(p.[S590L]) *SETX* :c.1807A>G(p.[N603D]) *SETX* :c.1957C>A(p.[Q653K]) *SQSTM1* :c.955G>A(p.[E319K])
NA	234801	55	18	F	White	Sporadic	Bulbar	Deceased	NA
NA	234901	57	72	M	White	Sporadic	Spinal	Deceased	NA
NA	235001	40	>21	F	White	Sporadic	Spinal	Alive	NA
NA	235101	59	38	M	White	Sporadic	Bulbar	Deceased	*NEK1* :c.2042delC(p.[S681fs])
NA	235201	66	15	F	Black	Sporadic	Bulbar	Deceased	*MAPT* :c.1483G>A(p.[A495T]) *UNC13A* :c.317-3_317-2delCA
NA	235301	35	>110	F	White	Sporadic	Spinal	Alive	NA
NA	235401	53	20	F	White	Sporadic	Spinal	Deceased	NA
NA	235501	69	24	M	White	Sporadic	Spinal	Deceased	*NEK1* :c.2190delC(p.[N731fs])
NA	235601	69	56	F	White	Sporadic	Bulbar	Deceased	NA
NA	235701	NA	NA	NA	NA	NA	NA	NA	NA
NA	235801	NA	NA	NA	NA	NA	NA	NA	NA
NA	235901	NA	NA	NA	NA	NA	NA	NA	*SETX* :c.4139C>T(p.[T1380I]) *MAPT*:c.1534C>T(p.[P512S])
NA	236001	61	20	M	White	Sporadic	Bulbar	Deceased	*ANG*:c.250A>G(p.[K84E])

Table 4.4: Cuban ALS patients (3/5)

Pedigree	Sample_ID	AOO	Survival	Sex/Gender	Ethnicity	History	Onset	Condition	HGVS
NA	236101	59	14	M	Black	Sporadic	Bulbar	Deceased	*SETX* :c.2113A>C(p.[I705L]) *SETX* :c.3965C>A(p.[T1322N]) *VCP* :c.79A>G(p.[I27V])
NA	236201	73	27	M	Black	Sporadic	Bulbar	Deceased	*hnRNPA1* :c.885_890delAGGCGG(p.[G296_G297del]) *MAPT*:c.1483G>A(p.[A495T]) *SETX* :c.3568A>G(p.[K1190E]) *SETX* :c.8078T>C(p.[L2693P])
NA	236501	53	51	M	Mestizo	Sporadic	Spinal	Deceased	*SETX* :c.1957C>A(p.[Q653K])
NA	236601	65	24	F	White	Familial	Bulbar	Deceased	NA
NA	236701	35	60	M	White	Familial	Spinal	Deceased	NA
NA	236801	75	17	F	White	Sporadic	Bulbar	Deceased	*ATXN2* :c.137C>A(p.[A46D])
NA	236901	60	>10	F	White	Sporadic	Bulbar	Alive	*VCP* :c.1147A>C(p.[I383L])
NA	237001	71	36	F	White	Sporadic	Bulbar	Deceased	*MAPT*:c.1535C>A(p.[P512H])
NA	237101	81	23	F	Black	Sporadic	Bulbar	Deceased	*NEFH*:c.985A>G(p.[T329A])
NA	237201	51	30	F	White	Familial	Spinal	Deceased	NA
NA	237301	49	73	F	White	Sporadic	Spinal	Deceased	*SPG11* :c.4687A>G(p.[R1563G])
NA	237401	60	96	M	Mestizo	Sporadic	Spinal	Deceased	NA
NA	237501	63	70	M	Mestizo	Sporadic	Spinal	Deceased	NA
NA	237601	63	37	M	Mestizo	Sporadic	Spinal	Deceased	NA
NA	237701	NA	NA	NA	NA	NA	NA	NA	NA
NA	237801	NA	NA	NA	NA	NA	NA	NA	*DAO* :c.430T>C(p.[Y144H])
NA	237901	NA	NA	NA	NA	NA	NA	NA	NA
NA	238001	NA	NA	NA	NA	NA	NA	NA	NA
NA	238101	NA	NA	NA	NA	NA	NA	NA	*SARM1* :c.1399T>C(p.[Y467H]) *SETX* :c.2401A>G(p.[K801E]) *SPAST*:c.865C>T(p.[H289Y]) *SPAST*:c.872G>T(p.[G291V])
NA	238201	65	28	F	Mestizo	Sporadic	Bulbar	Deceased	*SQSTM1* :c.955G>A(p.[E319K])
NA	238301	NA	NA	NA	NA	NA	NA	NA	*SIGMAR1* :c.622C>T(p.[R208W]) *SPG11* :c.4687A>G(p.[R1563G])
NA	238401	43	>47	M	White	Sporadic	Spinal	Alive	NA
NA	238501	74	26	F	White	Sporadic	Bulbar	Deceased	*PSEN2* :c.581A>G(p.[K194R]) *SARM1* :c.767C>G(p.[S256W])
NA	238601	54	71	F	White	Sporadic	Spinal	Deceased	*C21orf2* :c.505G>A(p.[E169K]) *NEK1* :c.3302G>A(p.[R1101H])
NA	238701	66	20	M	Mestizo	Sporadic	Spinal	Deceased	*SQSTM1* :c.955G>A(p.[E319K])
NA	238801	66	23	F	White	Sporadic	Bulbar	Deceased	NA
NA	238901	17	>70	M	Mestizo	Sporadic	Spinal	Alive	NA
NA	239001	NA	NA	NA	NA	NA	NA	NA	*CHCHD10* :c.100C>T(p.[P34S]) *SIGMAR1* :c.622C>T(p.[R208W]) *TBK1* :c.1954_1956delAAT

Table 4.4 : Cuban ALS patients (4/5)

Pedigree	Sample_ID	AOO	Survival	Sex/Gender	Ethnicity	History	Onset	Condition	HGVS
NA	239101	NA	NA	NA	NA	NA	NA	NA	*C21orf2* :c.1097G>A(p.[R366H])
NA	239201	64	>6	F	White	Sporadic	Bulbar	Alive	NA
NA	239301	64	18	F	White	Sporadic	Spinal	Deceased	NA
NA	239401	61	>87	M	White	Sporadic	Spinal	Alive	NA
NA	239501	33	NA	F	Mestizo	NA	Spinal	Alive	*SARM1* :c.1081C>T(p.[R361C])
NA	239601	NA	NA	NA	NA	NA	NA	NA	*NEFH* :c.1941G>C(p.[K647N])
NA	239701	51	>35	F	Mestizo	Sporadic	Bulbar	Alive	NA
NA	239801	NA	NA	NA	NA	NA	NA	NA	*ATXN2* :c.3228G>T(p.[M1076I]) *ELP3* :c.190G>A(p.[V64I])
NA	239901	72	>28	M	White	Sporadic	Bulbar	Alive	NA
NA	2310001	32	36	M	Mestizo	Sporadic	Spinal	Deceased	NA
NA	2310101	NA	NA	NA	NA	NA	NA	NA	*ATXN2* :c.563delA(p.[Q188fs]) *NEFH* :c.2512C>G(p.[P838A]) *SPG11* :c.4216G>T(p.[A1406S])
NA	2310201	58	>26	F	White	Familial	Bulbar	Alive	NA
NA	2310301	46	>28	M	White	Sporadic	Spinal	Alive	NA
NA	2310401	48	>17	M	White	Familial	Bulbar	Alive	NA
NA	2310501	75	4	F	White	Sporadic	Bulbar	Deceased	*C9orf72* :c.-45+163GGGGCC[>24]
NA	2310601	69	60	M	White	Sporadic	Spinal	Deceased	*C21orf2* :c.1097G>A(p.[R366H])
NA	2310701	38	>16	M	Mestizo	Sporadic	Spinal	Alive	NA
NA	2310901	NA	NA	NA	NA	NA	NA	NA	*SETX* :c.3935A>G(p.[D1312G]) *SETX* :c.4631T>C(p.[L1544S]) *NEFH* :c.1783C>T(p.[P595S]) *TBK1* :c.1522C>A(p.[L508I])
NA	2311001	NA	NA	NA	NA	NA	NA	NA	*C9orf72* :c.-45+163GGGGCC[>24]
NA	2311101	NA	NA	NA	NA	NA	NA	NA	*SPG11* :c.2656T>C(p.[Y886H]) *SQSTM1* :c.599T>C(p.[M200T])
NA	2311201	NA	NA	NA	NA	NA	NA	NA	*MATR3* :c.1735-2_1735-1insAA
NA	2311301	55	>14	F	White	Sporadic	Spinal	Alive	NA
NA	2311401	55	55	M	Mestizo	Sporadic	Spinal	Deceased	*FUS* :c.684_686dupCGG(p.[G229dup]) *SETX* :c.3310C>G(p.[Q1104E]) *TAF15* :c.1624G>A(p.[G542S])
NA	2311501	33	43	M	Mestizo	Familial	Spinal	Deceased	*FUS* :c.143C>T(p.[S48L]) *SQSTM1* :c.1201A>C(p.[M401L])
NA	2311601	44	>70	M	White	Sporadic	Spinal	Alive	*C21orf2* :c.404T>C(p.[L135P])
NA	2311701	NA	NA	NA	NA	NA	NA	NA	NA
NA	2311801	NA	NA	NA	NA	NA	NA	NA	*PRPH* :c.1303C>T(p.[R435W]) *TAF15* :c.1524_1544delCGGAGGAGATCGAGGAGGTTA(p.[G509_Y515del])
NA	2311901	NA	NA	NA	NA	NA	NA	NA	*NEFH* :c.1104C>G(p.[D368E])
NA	2312001	NA	NA	NA	NA	NA	NA	NA	*ERBB4* :c.1243A>G(p.[S415G])

Table 4.4: Cuban ALS patients (5/5)

Pedigree	Sample_ID	AOO	Survival	Sex/Gender	Ethnicity	History	Onset	Condition	HGVS
NA	2312101	NA	NA	NA	NA	NA	NA	NA	*DAO* :c.1028C>T(p.[P343L]) *ERBB4* :c.1243A>G(p.[S415G]) *SETX* :c.3568A>G(p.[K1190E])
NA	2312201	NA	NA	NA	NA	NA	NA	NA	*SPG11* :c.3121C>T(p.[R1041*])
NA	2312301	NA	NA	NA	NA	NA	NA	NA	*DCTN1* :c.2147A>G(p.[N716S])
NA	2312401	NA	NA	NA	NA	NA	NA	NA	NA

✶ : Homozygous

These variants presented here are those which are retained following the filtering process described in methods. Variants falling in genes with Pathogenic or Likely Pathogenic vairants identified in Chapter 1 are further discussed in the text.

Table 4.5: Putatively pathogenic variant properties (1/3)

Identifier	HGVS	Transcript	Impact	PM Case AF	PM Control AF	In Literature	gnomAD AF	*In silico* Prediction
2:202591402:C:G	*ALS2*:c.3167G>C(p.[G1056A])	ENST00000264276	missense	NA	NA	No	NA	P
2:202580441:T:A	*ALS2*:c.3958A>T(p.[N1320Y])	ENST00000264276	missense	NA	NA	Yes	NA	P
14:21161973:A:G	*ANG*:c.250A>G(p.[K84E])	ENST00000336811	missense	2.29e-4	0	Yes	1.54e-3	B
12:112037182:G:T	*ATXN2*:c.137C>A(p.[A46D])	ENST00000377617	missense	1.15e-4	0	No	7.47e-5	B
12:111954044:G:A	*ATXN2*:c.1769C>T(p.[S590L])	ENST00000377617	missense	NA	NA	No	6.83e-5	P
12:112037119:T:TGCCGGGAGGGAGGGGGGCCGGG	*ATXN2*:c.178_199dupCCCGGCCCCCCTCCCTCCCGGC (p.[Q67fs])	ENST00000377617	frameshift	NA	NA	No	NA	NA
12:111923648:T:C	*ATXN2*:c.2806A>G(p.[T936A])	ENST00000377617	missense	1.15e-4	0	No	4.80e-5	P
12:111908000:C:A	*ATXN2*:c.3228G>T(p.[M1076I])	ENST00000377617	missense	NA	NA	No	NA	P
12:112036755:CT:C	*ATXN2*:c.563delA(p.[Q188fs])	ENST00000377617	frameshift	NA	NA	No	1.49e-4	NA
21:45750112:C:T	*C21orf2*:c.1097G>A(p.[R366H])	ENST00000397956	missense	1.15e-4	0	No	9.30e-4	NA
21:45751867:A:G	*C21orf2*:c.404T>C(p.[L135P])	ENST00000397956	missense	NA	NA	No	7.71e-6	P
21:45751766:C:T	*C21orf2*:c.505G>A(p.[E169K])	ENST00000397956	missense	NA	NA	No	5.17e-4	NA
22:24109722:G:A	*CHCHD10*:c.100C>T(p.[P34S])	ENST00000401675	missense	4.24e-3	4.09e-3	Yes	1.53e-3	P
3:87302890:G:A	*CHMP2B*:c.560G>A(p.[S187N])	ENST00000263780	missense	NA	NA	Yes	1.92e-3	NA
12:109294295:C:T	*DAO*:c.1028C>T(p.[P343L])	ENST00000228476	missense	NA	NA	No	7.41e-5	B
12:109284027:T:C	*DAO*:c.430T>C(p.[Y144H])	ENST00000228476	missense	NA	NA	No	8.28e-4	P
2:74594860:T:C	*DCTN1*:c.2147A>G(p.[N716S])	ENST00000361874	missense	NA	NA	No	1.71e-5	NA
8:27957415:G:A	*ELP3*:c.190G>A(p.[V64I])	ENST00000256398	missense	NA	NA	No	2.89e-5	NA
2:212568875:T:C	*ERBB4*:c.1243A>G(p.[S415G])	ENST00000342788	missense	NA	NA	No	NA	B
2:212537936:G:A	*ERBB4*:c.1669C>T(p.[P557S])	ENST00000342788	missense	NA	NA	No	1.15e-4	B
6:110113868:G:A	*FIG4*:c.2459+1G>A	ENST00000230124	splice_donor	2.29e-4	0	No	1.99e-5	P
16:31193938:C:T	*FUS*:c.143C>T(p.[S48L])	ENST00000568685	missense	NA	NA	No	NA	P
16:31202396:CAG:C	*FUS*:c.1512 1513delAG(p.[G505fs])	ENST00000568685	frameshift	NA	NA	Yes	NA	NA
16:31196402:T:TGGC	*FUS*:c.684_686dupCGG(p.[G229dup])	ENST00000568685	inframe_insertion	8.02e-4	0	Yes	1.76e-3	NA
17:42426632:C:T	*GRN*:c.100C>T(p.[P34S])	ENST00000053867	missense	NA	NA	Yes	1.92e-5	B
17:42429491:C:G	*GRN*:c.1288C>G(p.[P430A])	ENST00000053867	missense	1.15e-4	0	No	8.83e-5	B
12:54677706:C:T	*hnRNPA1*:c.1018C>T(p.[P340S])	ENST00000340913	missense	NA	NA	No	NA	P
12:54676986:ACGGAGG:A	*hnRNPA1*:c.885_890delAGGCGG(p.[G296_G297del])	ENST00000340913	inframe_deletion	NA	NA	Yes	1.21e-3	NA
17:44068928:G:A	*MAPT*:c.1483G>A(p.[A495T])	ENST00000344290	missense	1.15e-4	0	No	4.75e-3	NA
17:44071316:C:T	*MAPT*:c.1534C>T(p.[P512S])	ENST00000344290	missense	1.15e-4	0	No	6.75e-4	NA
17:44071317:C:A	*MAPT*:c.1535C>A(p.[P512H])	ENST00000344290	missense	NA	NA	No	3.96e-5	B
17:44039753:C:T	*MAPT*:c.50C>T(p.[T17M])	ENST00000344290	missense	NA	NA	Yes	2.70e-4	B
5:138658149:T:TAA	*MATR3*:c.1735-2_1735-1insAA	ENST00000394800	splice_acceptor	NA	NA	No	4.81e-6	NA
22:29881732:C:G	*NEFH*:c.1104C>G(p.[D368E])	ENST00000310624	missense	NA	NA	No	4.81e-6	NA
22:29881766:G:A	*NEFH*:c.1138G>A(p.[A380T])	ENST00000310624	missense	NA	NA	Yes	4.22e-4	P
22:29885412:C:T	*NEFH*:c.1783C>T(p.[P595S])	ENST00000310624	missense	8.02e-4	2.73e-4	No	3.18e-3	NA
22:29885570:G:C	*NEFH*:c.1941G>C(p.[K647N])	ENST00000310624	missense	1.15e-4	0	No	1.01e-3	NA
22:29886141:C:G	*NEFH*:c.2512C>G(p.[P838A])	ENST00000310624	missense	NA	NA	No	5.23e-5	B

Table 4.5: Putatively pathogenic variant properties (2/3)

Identifier	HGVS	Transcript	Impact	PM Case AF	PM Control AF	In Literature	gnomAD AF	*In silico* Prediction
22:29876661:C:T	*NEFH* :c.410C>T(p.[A137V])	ENST00000310624	missense	NA	NA	No	4.72e-5	P
22:29879465:A:G	*NEFH* :c.985A>G(p.[T329A])	ENST00000310624	missense	NA	NA	No	NA	P
4:170400650:AG:A	*NEK1* :c.2042delC(p.[S681fs])	ENST00000507142	frameshift	NA	NA	No	4.81e-6	NA
4:170398597:TG:T	*NEK1* :c.2190delC(p.[N731fs])	ENST00000507142	frameshift	0	0	No	1.68e-5	NA
4:170327819:C:T	*NEK1* :c.3302G>A(p.[R1101H])	ENST00000507142	missense	NA	NA	No	1.35e-5	B
10:13174122:A:G	*OPTN* :c.1457A>G(p.[H486R])	ENST00000263036	missense	NA	NA	No	4.62E-05	P
12:49691776:C:T	*PRPH* :c.1303C>T(p.[R435W])	ENST00000257860	missense	5.72e-4	2.73e-4	No	3.13e-4	NA
14:73678538:A:G	*PSEN1* :c.1109A>G(p.[N370S])	ENST00000344094	missense	6.87e-4	2.73e-4	No	4.58e-4	NA
1:227073364:A:G	*PSEN2* :c.581A>G(p.[K194R])	ENST00000366782	missense	NA	NA	No	1.60e-5	P
17:26711469:C:T	*SARM1* :c.1081C>T(p.[R361C])	ENST00000457710	missense	NA	NA	No	1.01e-4	P
17:26712165:T:C	*SARM1* :c.1399T>C(p.[Y467H])	ENST00000457710	missense	NA	NA	No	1.88e-3	P
17:26708620:C:G	*SARM1* :c.767C>G(p.[S256W])	ENST00000457710	missense	NA	NA	No	6.14e-5	P
9:135205178:T:C	*SETX* :c.1807A>G(p.[N603D])	ENST00000372169	missense	0	0	No	1.36e-3	NA
9:135205028:G:T	*SETX* :c.1957C>A(p.[Q653K])	ENST00000372169	missense	1.15e-4	0	No	1.89e-3	NA
9:135204872:T:G	*SETX* :c.2113A>C(p.[I705L])	ENST00000372169	missense	1.15e-4	0	No	9.55e-4	NA
9:135204584:T:C	*SETX* :c.2401A>G(p.[K801E])	ENST00000372169	missense	1.15e-4	0	No	6.93e-4	B
9:135203675:G:C	*SETX* :c.3310C>G(p.[Q1104E])	ENST00000372169	missense	NA	NA	No	5.17e-4	B
9:135203417:T:C	*SETX* :c.3568A>G(p.[K1190E])	ENST00000372169	missense	NA	NA	No	9.91e-4	B
9:135203322:C:G	*SETX* :c.3663G>C(p.[K1221N])	ENST00000372169	missense	NA	NA	No	1.37e-3	NA
9:135203050:T:C	*SETX* :c.3935A>G(p.[D1312G])	ENST00000372169	missense	NA	NA	No	8.78e-6	B
9:135203020:G:T	*SETX* :c.3965C>A(p.[T1322N])	ENST00000372169	missense	NA	NA	No	1.49e-4	NA
9:135202846:G:A	*SETX* :c.4139C>T(p.[T1380I])	ENST00000372169	missense	NA	NA	No	NA	B
9:135202354:A:G	*SETX* :c.4631T>C(p.[L1544S])	ENST00000372169	missense	NA	NA	No	3.93e-5	B
9:135171352:C:T	*SETX* :c.6013G>A(p.[V2005M])	ENST00000372169	missense	NA	NA	No	1.79e-4	P
9:135164023:A:G	*SETX* :c.6122T>C(p.[I2041T])	ENST00000372169	missense	1.15e-4	0	No	1.29e-4	P
9:135139669:A:G	*SETX* :c.8078T>C(p.[L2693P])	ENST00000372169	missense	NA	NA	No	8.73e-5	B
9:34635679:G:A	*SIGMAR1* :c.622C>T(p.[R208W])	ENST00000277010	missense	2.06e-3	1.37e-3	Yes	8.15e-3	NA
2:32339889:C:T	*SPAST*:c.865C>T(p.[H289Y])	ENST00000315285	missense	0	0	No	6.50e-4	B
2:32339889:C:T	*SPAST*:c.865C>T(p.[H289Y])	ENST00000315285	missense	0	0	No	6.50e-4	B
2:32340772:G:T	*SPAST*:c.872G>T(p.[G291V])	ENST00000315285	missense	NA	NA	No	NA	B
15:44912566:A:G	*SPG11* :c.2656T>C(p.[Y886H])	ENST00000261866	missense	3.44e-4	2.73e-4	Yes	1.54e-3	B
15:44905652:G:A	*SPG11* :c.3121C>T(p.[R1041*])	ENST00000261866	stop_gained	NA	NA	No	8.72e-6	P
15:44888499:C:A	*SPG11* :c.4216G>T(p.[A1406S])	ENST00000261866	missense	1.15e-4	0	No	2.06e-5	P
15:44884585:T:C	*SPG11* :c.4687A>G(p.[R1563G])	ENST00000261866	missense	0	0	Yes	1.63e-3	NA
15:44864905:C:T	*SPG11* :c.6319G>A(p.[V2107I])	ENST00000261866	missense	0	0	Yes	2.89e-3	NA
15:44855490:T:A	*SPG11* :c.7161A>T(p.[Q2387H])	ENST00000261866	missense	NA	NA	No	3.05e-4	B
5:179263471:A:C	*SQSTM1* :c.1201A>C(p.[M401L])	ENST00000389805	missense	NA	NA	No	9.16e-5	P
5:179251249:T:C	*SQSTM1* :c.599T>C(p.[M200T])	ENST00000389805	missense	NA	NA	No	4.81e-6	B
5:179252182:TGAA:T	*SQSTM1* :c.714_716delGAA(p.[K238del])	ENST00000389805	inframe_deletion	NA	NA	Yes	9.61e-6	NA
5:179260232:G:A	*SQSTM1* :c.955G>A(p.[E319K])	ENST00000389805	missense	2.29e-4	0	Yes	3.67e-3	NA

Table 4.5: Putatively pathogenic variant properties (3/3)

Identifier	HGVS	Transcript	Impact	PM Case AF	PM Control AF	In Literature	gnomAD AF	*In silico* Prediction
17:34171806: TGGAGGAGATCGAGGAGGTTAC:T	*TAF15* :c.1524_1544delCGGAGGAGATCGAGGAGGTTA (p.[G509_Y515del])	ENST00000588240	inframe_deletion	5.76e-4	2.74e-4	No	9.00e-4	NA
17:34171927:G:A	*TAF15* :c.1624G>A(p.[G542S])	ENST00000588240	missense	0	0	No	1.85e-3	NA
12:64889263:C:A	*TBK1* :c.1522C>A(p.[L508I])	ENST00000331710	missense	NA	NA	No	7.59e-4	NA
12:64891032:CTAA:C	*TBK1* :c.1954_1956delAAT	ENST00000331710	structural_interaction	NA	NA	No	3.79e-4	NA
12:64860787:C:CA	*TBK1* :c.466dupA(p.[T156fs])	ENST00000331710	frameshift	NA	NA	No	0	NA
12:64860858:AT:A	*TBK1* :c.539delT(p.[L180fs])	ENST00000331710	frameshift	NA	NA	No	NA	NA
17:26708225:C:T	*TMEM199* :c.535C>T(p.[P179S])	ENST00000509083	missense	NA	NA	No	5.23e-5	NA
19:17785566:CTG:C	*UNC13A* :c.317-3_317-2delCA	ENST00000428389	splice_acceptor	NA	NA	No	1.00e-4	NA
9:35061621:T:G	*VCP* :c.1147A>C(p.[I383L])	ENST00000358901	missense	NA	NA	No	NA	P
9:35068298:T:C	*VCP* :c.79A>G(p.[I27V])	ENST00000358901	missense	0	0	Yes	5.95e-4	B

gnomAD AF represents the frequency in the non neuro subset as described in methods

Table 4.6: Putatively pathogenic variant *in silico* predictions (1/2)

HGVS	VEST_INDEL_CALL	SIFT_INDEL_CALL	cadd_CALL	ada_CALL	rf_CALL	VEST_CALL	REVEL_CALL	MetaSVM_CALL	MutationTaster_CALL	MCap_CALL	Combined Prediction
ALS2 :c.3167G>C(p.[G1056A])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic
ALS2 :c.3958A>T(p.[N1320Y])	B	B	P	NA	NA	B	B	B	B	P	Benign
ANG :c.250A>G(p.[K84E])	B	B	B	NA	NA	B	B	B	B	NA	Benign
ATXN2 :c.137C>A(p.[A46D])	B	P	B	NA	NA	B	B	B	B	P	Benign
ATXN2 :c.1769C>T(p.[S590L])	P	P	P	NA	NA	P	B	B	P	P	Pathogenic
ATXN2 :c.178_199dupCCCGGCCCCCCTCCCTCCCGGC (p.[Q67fs])	B	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
ATXN2 :c.2806A>G(p.[T936A])	P	B	P	NA	NA	P	B	P	P	P	Pathogenic
ATXN2 :c.3228G>T(p.[M1076I])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic
ATXN2 :c.563delA(p.[Q188fs])	B	NA	B	NA	NA	NA	NA	NA	NA	NA	NA
C21orf2 :c.1097G>A(p.[R366H])	B	B	P	NA	NA	B	B	B	B	NA	NA
C21orf2 :c.404T>C(p.[L135P])	P	P	P	NA	NA	P	B	B	P	P	Pathogenic
C21orf2 :c.505G>A(p.[E169K])	P	P	B	NA	NA	P	B	B	B	B	NA
CHCHD10 :c.100C>T(p.[P34S])	B	B	P	NA	NA	B	P	B	P	P	Pathogenic
CHMP2B :c.560G>A(p.[S187N])	B	B	B	NA	NA	B	B	B	P	NA	NA
DAO :c.1028C>T(p.[P343L])	B	P	B	NA	NA	B	B	B	P	B	Benign
DAO :c.430T>C(p.[Y144H])	P	P	P	NA	NA	P	P	P	P	NA	Pathogenic
DCTN1 :c.2147A>G(p.[N716S])	P	B	P	NA	NA	P	B	B	P	B	NA
ELP3 :c.190G>A(p.[V64I])	P	B	P	NA	NA	P	B	B	P	B	NA
ERBB4 :c.1243A>G(p.[S415G])	B	B	P	NA	NA	B	B	B	P	B	Benign
ERBB4 :c.1669C>T(p.[P557S])	B	B	B	NA	NA	B	B	B	B	B	Benign
FIG4 :c.2459+1G>A	P	NA	P	P	P	NA	NA	NA	P	NA	Pathogenic
FUS :c.143C>T(p.[S48L])	B	P	P	NA	NA	B	P	P	P	P	Pathogenic
FUS :c.1512_1513delAG(p.[G505fs])	B	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
FUS :c.684_686dupCGG(p.[G229dup])	P	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
GRN :c.100C>T(p.[P34S])	B	B	P	NA	NA	B	B	B	B	P	Benign
GRN :c.1288C>G(p.[P430A])	B	B	B	NA	NA	B	B	B	P	P	Benign
hnRNPA1 :c.1018C>T(p.[P340S])	P	P	P	NA	NA	P	P	B	P	P	Pathogenic
hnRNPA1 :c.885_890delAGGCGG(p.[G296_G297del])	B	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
MAPT :c.1483G>A(p.[A495T])	B	B	P	NA	NA	B	B	B	B	NA	NA
MAPT :c.1534C>T(p.[P512S])	B	P	P	NA	NA	B	B	B	P	NA	NA
MAPT :c.1535C>A(p.[P512H])	B	P	P	NA	NA	B	B	B	P	B	Benign
MAPT :c.50C>T(p.[T17M])	B	P	P	NA	NA	B	B	B	B	B	Benign
MATR3 :c.1735-2_1735-1insAA	P	NA	B	NA	NA	NA	NA	NA	NA	NA	NA
NEFH :c.1104C>G(p.[D368E])	P	B	B	NA	NA	P	P	B	B	P	NA
NEFH :c.1138G>A(p.[A380T])	B	B	P	NA	NA	B	P	P	P	P	Pathogenic
NEFH :c.1783C>T(p.[P595S])	B	B	B	NA	NA	B	B	P	B	NA	NA
NEFH :c.1941G>C(p.[K647N])	B	P	P	NA	NA	B	B	P	P	NA	NA
NEFH :c.2512C>G(p.[P838A])	B	P	B	NA	NA	B	B	B	P	P	Benign
NEFH :c.410C>T(p.[A137V])	B	P	P	NA	NA	B	B	P	P	P	Pathogenic
NEFH :c.985A>G(p.[T329A])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic
NEK1 :c.2042delC(p.[S681fs])	P	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
NEK1 :c.2190delC(p.[N731fs])	P	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
NEK1 :c.3302G>A(p.[R1101H])	B	B	P	NA	NA	B	B	B	B	B	Benign
OPTN :c.1457A>G(p.[H486R])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic
PRPH :c.1303C>T(p.[R435W])	B	P	P	NA	NA	B	P	P	B	P	NA
PSEN1 :c.1109A>G(p.[N370S])	NA	B	P	NA	NA	NA	NA	NA	NA	NA	NA
PSEN2 :c.581A>G(p.[K194R])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic
SARM1 :c.1081C>T(p.[R361C])	P	P	P	NA	NA	P	B	B	P	P	Pathogenic
SARM1 :c.1399T>C(p.[Y467H])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic
SARM1 :c.767C>G(p.[S256W])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic

Table 4.6: Putatively vathogenic variant *in silico* predictions (2/2)

HGVS	VEST_INDEL_CALL	SIFT_INDEL_CALL	cadd_CALL	ada_CALL	rf_CALL	VEST_CALL	REVEL_CALL	MetaSVM_CALL	MutationTaster_CALL	MCap_CALL	Combined Prediction
SETX :c.1807A>G(p.[N603D])	B	P	P	NA	NA	B	B	B	B	NA	NA
SETX :c.1957C>A(p.[Q653K])	B	P	B	NA	NA	B	B	B	B	NA	NA
SETX :c.2113A>C(p.[I705L])	B	P	B	NA	NA	B	B	B	B	NA	NA
SETX :c.2401A>G(p.[K801E])	B	P	P	NA	NA	B	B	B	B	P	Benign
SETX :c.3310C>G(p.[Q1104E])	B	P	P	NA	NA	B	B	B	B	P	Benign
SETX :c.3568A>G(p.[K1190E])	B	B	B	NA	NA	B	B	B	B	NA	Benign
SETX :c.3663G>C(p.[K1221N])	B	P	P	NA	NA	B	B	B	B	NA	NA
SETX :c.3935A>G(p.[D1312G])	B	P	P	NA	NA	B	B	B	B	P	Benign
SETX :c.3965C>A(p.[T1322N])	B	P	P	NA	NA	B	B	P	B	P	NA
SETX :c.4139C>T(p.[T1380I])	B	B	B	NA	NA	B	B	B	B	P	Benign
SETX :c.4631T>C(p.[L1544S])	B	P	P	NA	NA	B	B	B	B	P	Benign
SETX :c.6013G>A(p.[V2005M])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic
SETX :c.6122T>C(p.[I2041T])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic
SETX :c.8078T>C(p.[L2693P])	B	B	B	NA	NA	B	B	B	B	P	Benign
SIGMAR1 :c.622C>T(p.[R208W])	P	P	P	NA	NA	P	P	B	P	NA	NA
SPAST :c.865C>T(p.[H289Y])	B	B	P	NA	NA	B	B	B	P	B	Benign
SPAST :c.872G>T(p.[G291V])	B	B	P	B	B	B	B	B	P	B	Benign
SPG11 :c.2656T>C(p.[Y886H])	B	B	B	NA	NA	B	B	B	B	NA	Benign
SPG11 :c.3121C>T(p.[R1041*])	P	NA	P	NA	NA	P	NA	NA	P	NA	Pathogenic
SPG11 :c.4216G>T(p.[A1406S])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic
SPG11 :c.4687A>G(p.[R1563G])	B	P	P	NA	NA	B	B	B	B	NA	NA
SPG11 :c.6319G>A(p.[V2107I])	B	B	P	NA	NA	B	B	B	P	NA	NA
SPG11 :c.7161A>T(p.[Q2387H])	B	B	P	NA	NA	B	B	B	B	B	Benign
SQSTM1 :c.1201A>C(p.[M401L])	P	P	P	NA	NA	B	B	P	P	P	Pathogenic
SQSTM1 :c.599T>C(p.[M200T])	B	B	B	NA	NA	B	B	B	B	P	Benign
SQSTM1 :c.714_716delGAA(p.[K238del])	P	NA	B	NA	NA	NA	NA	NA	NA	NA	NA
SQSTM1 :c.955G>A(p.[E319K])	B	B	P	NA	NA	B	B	B	B	NA	NA
TAF15 :c.1524_1544delCGGAGGAGATCGAGGAGGTTA (p.[G509_Y515del])	B	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
TAF15 :c.1624G>A(p.[G542S])	B	B	P	NA	NA	B	B	B	P	NA	NA
TBK1 :c.1522C>A(p.[L508I])	B	B	P	B	B	B	B	B	P	NA	NA
TBK1 :c.1954_1956delAAT	P	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
TBK1 :c.466dupA(p.[T156fs])	P	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
TBK1 :c.539delT(p.[L180fs])	P	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
TMEM199 :c.535C>T(p.[P179S])	NA	B	P	NA	NA	NA	NA	NA	NA	NA	NA
UNC13A :c.317-3_317-2delCA	P	NA	P	NA	NA	NA	NA	NA	NA	NA	NA
VCP :c.1147A>C(p.[I383L])	P	P	P	NA	NA	P	P	P	P	P	Pathogenic
VCP :c.79A>G(p.[I27V])	B	B	P	NA	NA	B	B	B	P	B	Benign

Variant analysis

ATXN2

Intermediate length *ATXN2* CAG REs (27-34 repeats) are associated with increased risk for developing ALS. Sequence data was screened for intermediate length CAG REs using *in silico* programs HipSTR and TREDPARSE. 125 cases and 97 controls had *ATXN2* genotypes called by both HipSTR and TREDPARSE. As genotype calls may be less accurate for samples with low coverage across the *ATXN2* repeat, the concordance of both programs was assessed across a range of depth of coverage filters ranging from 1X to 15X (figure 4.4 A). Samples with coverage below 2X were removed from further *ATXN2* analysis. A more stringent coverage filter removed more samples from analysis but did not yield any improvement in RMSD (figure 4.4 A-C). 121 cases and 95 controls were retained after applying the 2X coverage filter. While the proportion of cases in the 27-34 repeat range was above the rate in controls, this did not yield a statistically significant result, with an OR of 1.34 (95% CI: 0.47-3.82) observed for both programs.

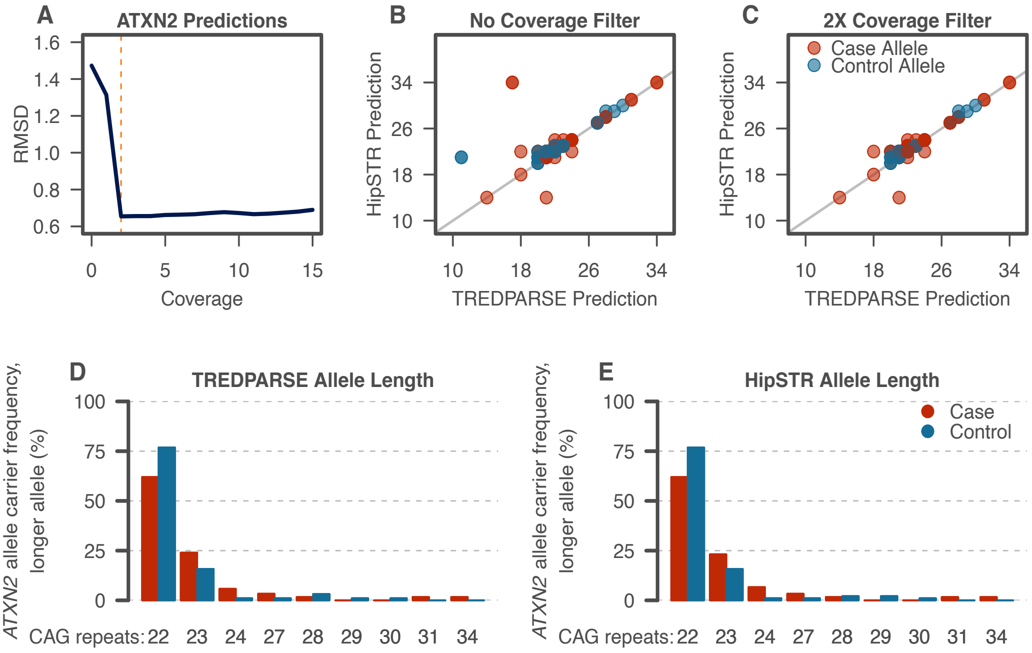


Figure 4.4: *ATXN2* genotyping

A) Removing samples with coverage below 2X significantly improves the concordance between genotype calls from the two programs. This is demonstrated by comparing B) program concordance before removing samples, and C) program concordance after applying the coverage filter. D-E) Considering an individual's longer *ATXN2* allele, both programs identify a higher proportion of carriers of 27-34 repeats in cases than in controls but it does not reach significance (OR of 1.34 (95% CI: 0.47-3.82))

C9orf72

Three samples displayed the characteristic sawtooth pattern displayed by carriers of the *C9orf72* hexanucleotide RE (figure 4.5 A-C). Phenotype information was available for two of these, both of whom had no family history of disease and onset at ages 50 and 75 and both of whom self-identify as white.

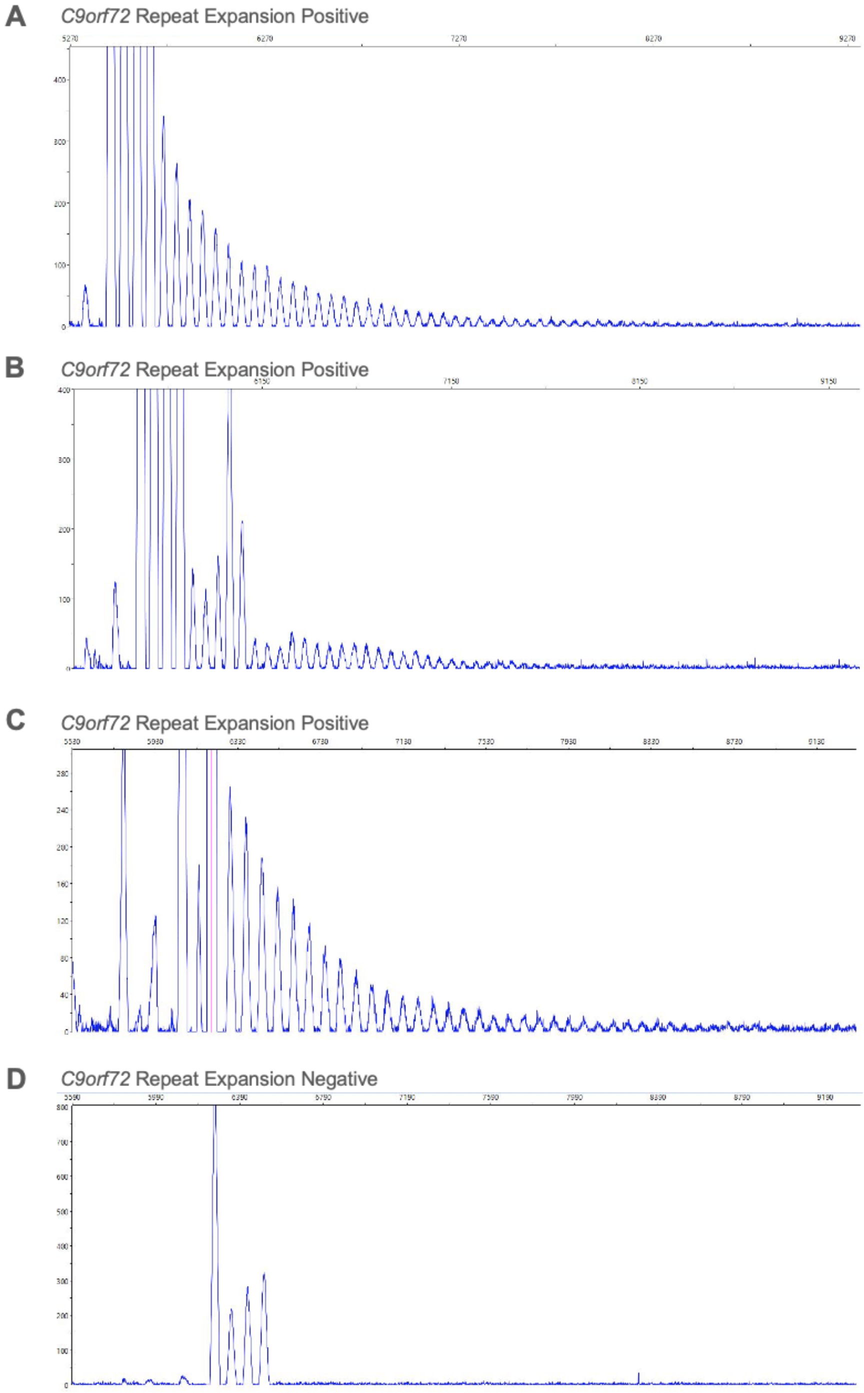


Figure 4.5: *C9orf72* rpPCR results

A-C) display the resultant traces from the three positive carriers of the *C9orf72* repeat expansion following rpPCR. All three display the characteristic sawtooth pattern. D) A demonstrative example of a *C9orf72* negative patient is shown. The initial peak represents 3 repeats. This individual carries 7 repeats.

ALS2

Two *ALS2* missense variants of uncertain significance are observed in two patients with ALS onset at ages 37 and 59. The variants are *ALS2*:c.3958A>T(p.[N1320Y]), which has previously been observed in an Italian ALS patient (Lamp *et al.* 2018) and *ALS2*:c.3958A>T(p.[N1320Y]) which is novel (absent in the ALS literature, gnomAD, Project MinE, ALSdb or ALSVS).

In Chapter 2, five typically homozygous, LOF variants were found to be Likely Pathogenic causes of extremely early onset ALS. Given that the variants observed here are missense, rather than homozygous LOF variants and that the two individuals in this study do not have extremely early onset, there is little reason to suspect the pathogenicity of these variants.

CHMP2B

In Chapter 2, a single C-terminal truncating *CHMP2B* splice acceptor variant (*CHMP2B*:c.532-1G>C) was identified as the LP cause of FTD in a well-characterised Danish pedigree (J. Brown *et al.* 1995; Skibinski *et al.* 2005; Holm *et al.* 2007; Urwin *et al.* 2010; Stokholm *et al.* 2013).

In this study, we observe a missense *CHMP2B* variant (*CHMP2B*:c.560G>A(p.[S187N]) in a sporadic ALS patient carrying a concurrent *MAPT* variant. This *CHMP2B* variant has previously been reported in an FTD patient from America (Ferrari *et al.* 2010). While missense variants all throughout *CHMP2B* have been observed in ALS and FTD, so too have benign missense variants. As such, little conclusion can be drawn as to the pathogenicity of this variant.

DCTN1

DCTN1:c.2147A>G(p.[N716S]) is observed in a single Cuban ALS patient lacking phenotypic information. This variant has not been previously reported in the literature, however is present at an AF of 1.71 × 10^{-5} in the gnomAD non neuro subset. JournALS identifies a single Likely Pathogenic variant at amino acid 59 and another variant at the same amino acid residue with weak supporting evidence. While rare missense variants have been reported in ALS patients all throughout *DCTN1*, so too have Benign missense variants, as such, the pathogenicity of *DCTN1*:c.2147A>G(p.[N716S]) should be interpreted with caution.

FUS

Three unique *FUS* variants are observed. A single patient carries *FUS*:c.1512_1513delAG(p.[G505fs]), a variant that has previously been observed in several patients (Kwon *et al.* 2012; Zou *et al.* 2013; L. Kent *et al.* 2014; Y.-E. Kim *et al.* 2015; Hirayanagi *et al.* 2016; Zou *et al.* 2016) but remains classified as VUS in journALS. The variant falls in the C-terminal domain that is identified as a hotspot for both missense and LOF variation. With the addition of this Cuban case to the journALS data, we now identify that, similarly to other *FUS* variants, this variant is associated with significantly early age of onset (figure 4.6 A) and is reclassified as Pathogenic. This expands the range of *FUS* pathogenic variants from missense and splice-site variants in the C-terminal domain to include frameshift variants in this region.

Three patients carry *FUS*:c.684_686dupCGG(p.[G229dup]), an inframe insertion that has previously been reported in four cases (Hewitt *et al.* 2010; Belzil *et al.* 2011; Rutherford, Finch, *et al.* 2012) and has a gnomAD AF of 1.8×10^{-3}. Unlike the G505fs variant, this variant does not fall in the C-terminal domain, and with the addition of data from this study does not display the characteristic early onset associated with *FUS* (figure 4.6 B). This variant remains classified as VUS with little evidence supporting its pathogenicity

1 patient with AOO of 33 carries a novel *FUS*:c.143C>T(p.[S48L]) variant in addition to the concurrent *SQSTM1*:c.1201A>C(p.[M401L]). While this patient does have early onset, the variant is outside the C-terminal domain and remains classified as a VUS with little supporting evidence.

GRN

Two patients carry two *GRN* variants (*GRN*:c.100C>T(p.[P34S]), *GRN*:c.1288C>G(p.[P430A])). All 10 identified Pathogenic or Likely Pathogenic *GRN* variants are LOF and are almost exclusively observed in FTD patients. These variants remain classified as VUS with little supporting evidence of pathogenicity.

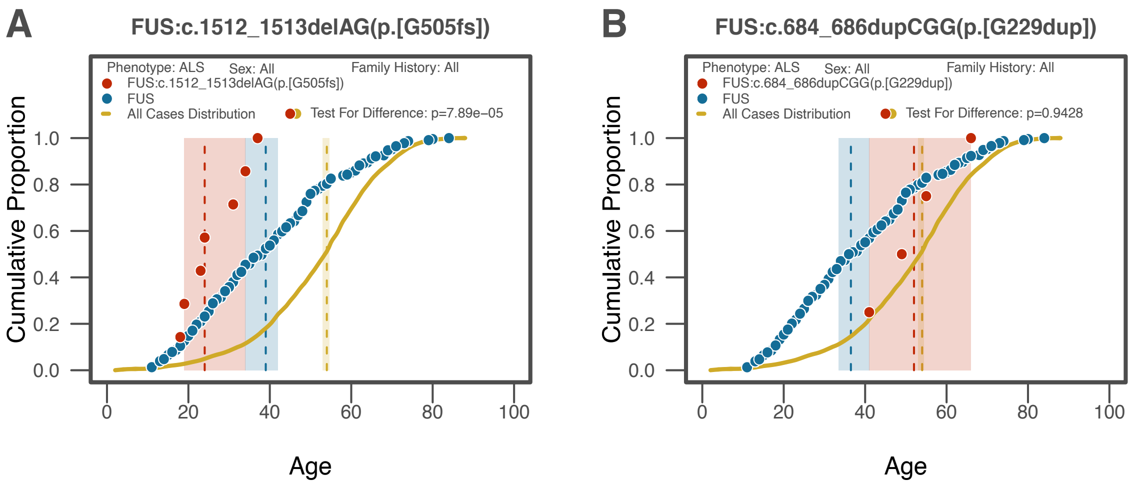


Figure 4.6: Age of onset for identified *FUS* variants

A) The age of onset of all *FUS*:c.1512_1513delAG(p.[G505fs]) carriers identified in Chapter 2 are combined with the individual from this study (red). These variants are found to exhibit a significantly earlier onset than the rest of the journALS cohort (gold). This is similar to other carriers of *FUS* variants (blue). The additional data point added here brings this to significance raising the status of *FUS*:c.1512_1513delAG(p.[G505fs]) from VUS to pathogenic variant. B) In contrast adding the three carriers of *FUS*:c.684_686dupCGG(p.[G229dup]) to this data reveals that carriers of this variant do not display the characteristic early onset associated with *FUS* variants.

MAPT

There are seven carriers of four unique missense *MAPT* variants (*MAPT*:c.50C>T(p.[T17M]); *MAPT*:c.1483G>A(p.[A495T]); *MAPT*:c.1534C>T(p.[P512S]); *MAPT*:c.1535C>A(p.[P512H])). Missense *MAPT* variants have been identified as causes of FTD, with rare reports of VUS identified in ALS patients. Of the four variants identified here, *MAPT*:c.50C>T(p.[T17M]) has previously been reported, interestingly in an ALS patient (Ghani *et al.* 2015). As these patients present with ALS rather than FTD and Benign missense variants have been observed in *MAPT* and indeed, even in this study rare missense variants are observed in controls, these variants are classified as VUS with little supporting evidence.

MATR3

A single carrier of the *MATR3* splice acceptor variant *MATR3*:c.1735-2_1735-1insAA is observed. This variant has not been previously reported in the literature. As the only previous reported *MATR3* variant that has been found to be pathogenic is a heterozygous missense variant (Johnson *et al.* 2014), there is little supporting the pathogenicity of this variant.

OPTN

Missense and homozygous frameshift variants in *OPTN* have been identified as pathogenic causes of ALS. *OPTN*:c.1457A>G(p.[H486R]) is identified in a patient who developed ALS at age 45 and who carries a concurrent *SPG11*:c.7161A>T(p.[Q2387H]) variant. There is supporting and conflicting support for this variant. Pathogenicity is supported by the fact that it is identified just 6 amino acids from the pathogenic *OPTN*:c.1433A>G(p.[E478G]) identified in Chapter 2. However, the four carriers of this variant in the gnomAD non-neuro subset were between 50 and 75 when they donated the blood, indicating that they were neurologically healthy at these ages, decreasing the likelihood that this is a highly pathogenic ALS variant.

SETX

14 putatively pathogenic *SETX* variants are identified in this study, none of which are previously reported in the literature. A single pathogenic heterozygous missense variant has previously been identified as a cause of early-onset, slowly progressing ALS. None of the *SETX* carriers identified in this study match this phenotype as the median AOO is 53 (95% CI: 53-59) survival 36 (14-55). There is not a statistically significant increase in missense variants in cases compared to controls. While these variants are VUS, the phenotype of the variant carriers and lack of statistical support does not support pathogenicity.

SIGMAR1

A homozygous *SIGMAR1* variant (*SIGMAR1*:c.304G>C(p.[E102Q])) has been identified as the LP cause of slowly progressive juvenile ALS in a Saudi Arabian family with a history of consanguinity (Al-Saif, Al-Mohanna, and Bohlega 2011). In this study an ALS patient presents with a heterozygous *SIGMAR1*:c.622C>T(p.[R208W]) variant. This variant has been reported in ten previous patients (Ghani *et al.* 2015; Zhang *et al.* 2018) however is found to be a similar frequency in gnomAD as in the Project MinE case cohort. Taking this into account, in addition to the variant's heterozygosity and the patient's later age of onset, there is little reason to suspect the pathogenicity of this variant.

TBK1

A missense variant (*TBK1*:c.1522C>A(p.[L508I])) in addition to three INDELs (*TBK1*:c.466dupA(p.[T156fs]),*TBK1*:c.539delT(p.[L180fs]),*TBK1*:c.1954_1956delAAT), two of which are frameshift variants and one of which is a structural interaction variant, are observed in *TBK1*. While missense *TBK1* variants are not known to be pathogenic, frameshift *TBK1* variants are a known causes of ALS. These variants have not previously been reported in the literature and both *TBK1*:c.466dupA(p.[T156fs]) and *TBK1*:c.539delT(p.[L180fs]) are novel. There are no reported INDELs in controls in this cohort. Lacking further evidence, these INDELs remain classified as VUS; however, evidence is suggestive of pathogenicity and they warrant further study.

VCP

In Chapter 2 seven heterozygous variants have been reported as Pathogenic or Likely Pathogenic causes of ALS, FTD, inclusion body myopathy, Paget disease of bone, or various combinations of these phenotypes. In this study two sporadic patients with bulbar onset carry *VCP*:c.1147A>C(p.[I383L]) and *VCP*:c.79A>G(p.[I27V]) which has previously been reported in a patient with ALS-FTD (Dols-Icardo *et al.* 2018). As these patients are not reported to present with PDB or IBM it is difficult to make any further ascertainment as to the pathogenicity of these variants.

Cuban ALS pedigree

Six affected members of a four generation ALS pedigree are described here (table 4.4, figure 4.7). The family present with early-onset slowly progressive ALS with some members presenting with bulbar onset and others spinal onset. The proband had onset at age 24 and was still alive after 17 years. All other patients for whom phenotype information is available had similar early onset and long duration.

Targeted sequencing in this family revealed that four family members carry *hnRNPA1*:c.1018C>T(p.[P340S]) and that this variant is absent in two family members. This variant has previously been reported in a Chinese pedigree with flail arm ALS (Q. Liu *et al.* 2016), an ALS variant associated with similar AOO and duration to this pedigree but with rare or delayed bulbar involvement.

Exome sequencing

Exome sequencing was performed on the five affected members of this family for whom sufficient DNA was available.

Data were merged with individuals from the 1000 genomes project to explore the relationships between individuals in this pedigree. All relationships in pedigree 2302 were confirmed to be that of siblings, half-siblings, or cousins, proving that this is a single family unit.

hnRNPA1:c.1018C>T(p.[P340S]) is confirmed to be absent in two family members (sequenced to depths of 155X and 180X) (figure 4.8), and as such is unlikely to be the sole causative variant in this family.

Exome variant identification

Having confirmed that *hnRNPA1*:c.1018C>T(p.[P340S]) did not segregate in this family, a screen was undertaken for potentially pathogenic variants. table 4.7 outlines the variants remaining through each stage of the filtering process. 60,957 variants were present in any family member and this was reduced to 14,712 when considering variants which passed VQSR filtering (figure 4.9) and were present in all family members. Applying the remaining filtering steps reduced this to two remaining potentially pathogenic variants: *RYBP*:c.633T>A(p.[S211R]) (ENST00000477973; chr3:113753890:C:A) and *KIAA1407*:c.700G>T(p.[Q234*]) (ENST00000295878; chr3:72428256:A:T).

Table 4.7: Exome variant filtering

Filter Description	Variants Remaining
In any Family Member	60957
Passing VQSR Filter	57993
Present in all Family Members	14712
0.1% gnomAD AF Filter	1402
Functional Filter	484
Sequencing Filter	2

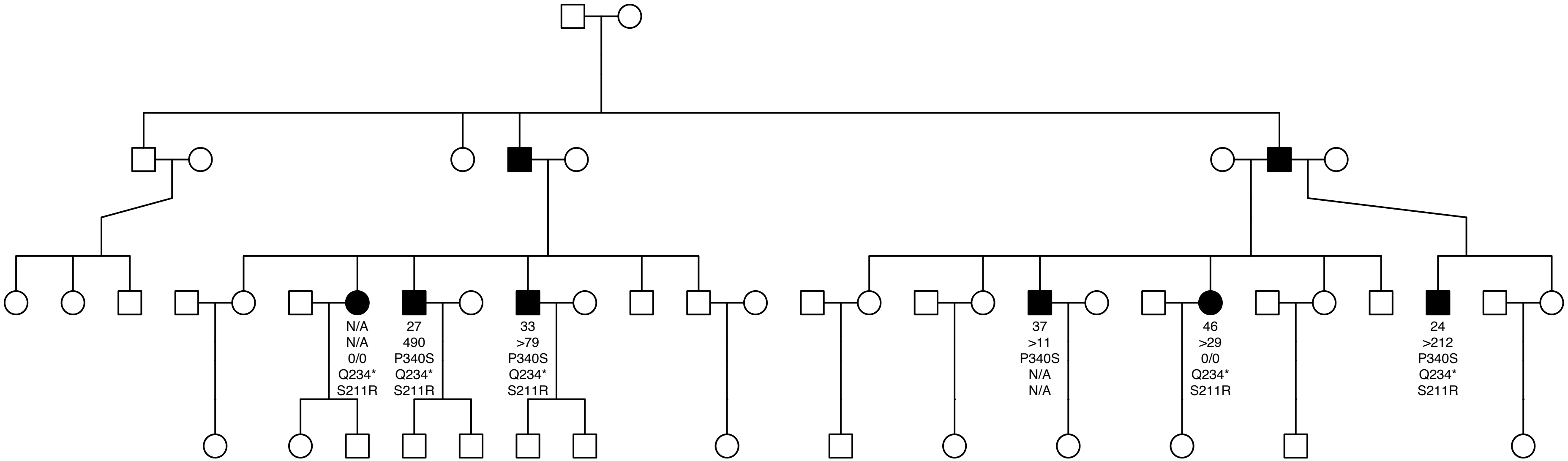


Figure 4.7: Pedigree 2302

The family present with early-onset slowly progressive ALS with some members presenting with bulbar onset and others spinal onset. Solid individuals are affected patients. The top line of information represents age of onset, the second line indicates the disease duration. The third line indicates carriers of *hnRNPA1*:c.1018C>T(p.[P340S]), this was confirmed to be absent in III.7 and III.22. The final two lines represent carriers of *KIAA1407*:c.700G>T(p.[Q234*]) and *RYBP*:c.633T>A(p.[S211R]) respectively; there was insufficient DNA to confirm these variants in patient III.19.

Genotype:	homozygous reference	heterozygous	heterozygous	homozygous reference	heterozygous
Pedigree Position:	III.7	III.8	III.10	III.22	III.26
Coverage:	155X	144X	158X	180X	131X

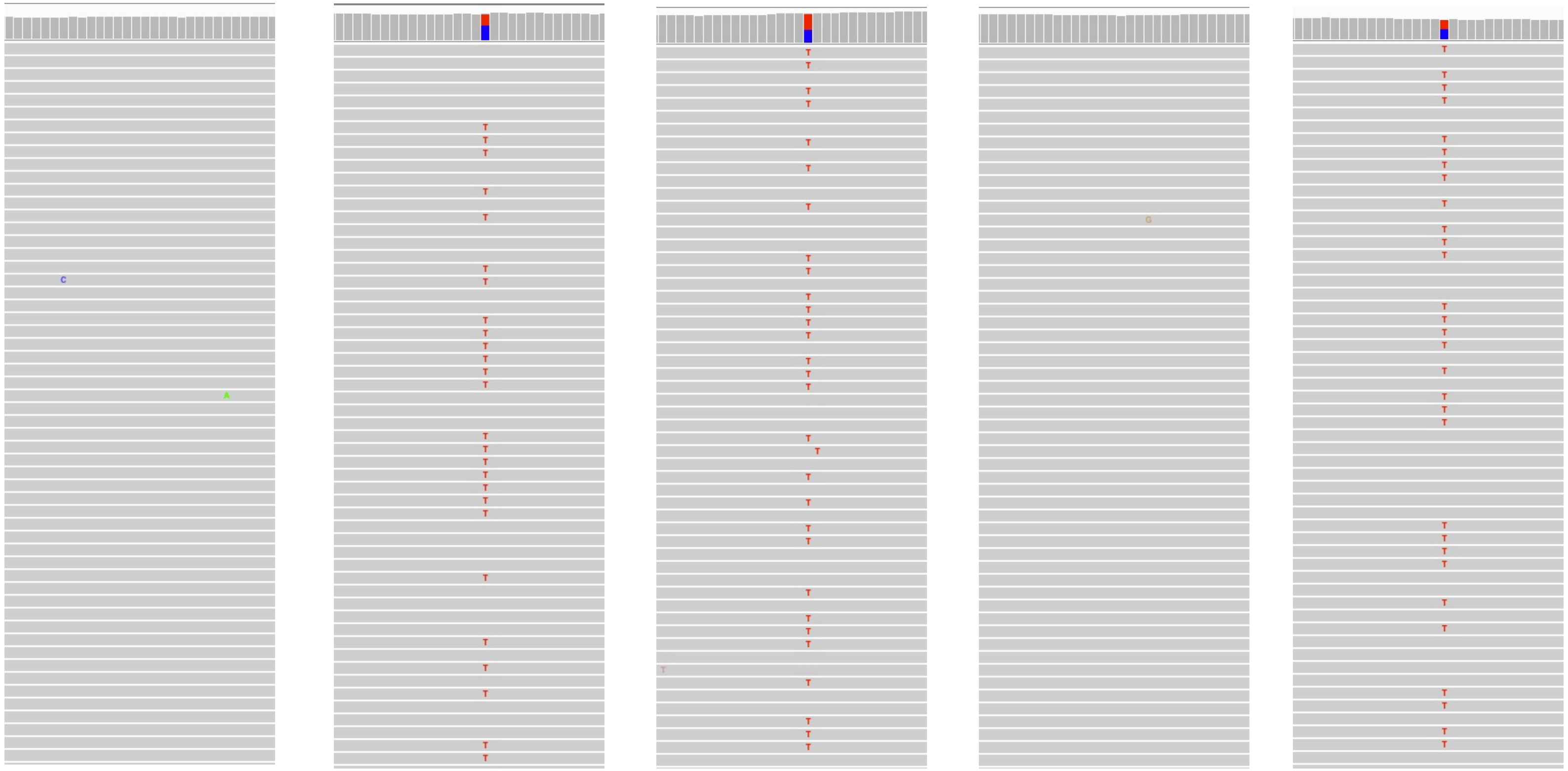


Figure 4.8: *hnRNPA1*:c.1018C>T(p.[P340S]) does not segregate in pedigree 2302

The figure is an Integrative Genomics Viewer (IGV) display of position 54,677,706 on chromosome 12 in the five family members who underwent exome sequencing. Three family members are confirmed to be heterozygous carriers of *hnRNPA1*:c.1018C>T(p.[P340S]) and two family members are homozygous for the reference allele.

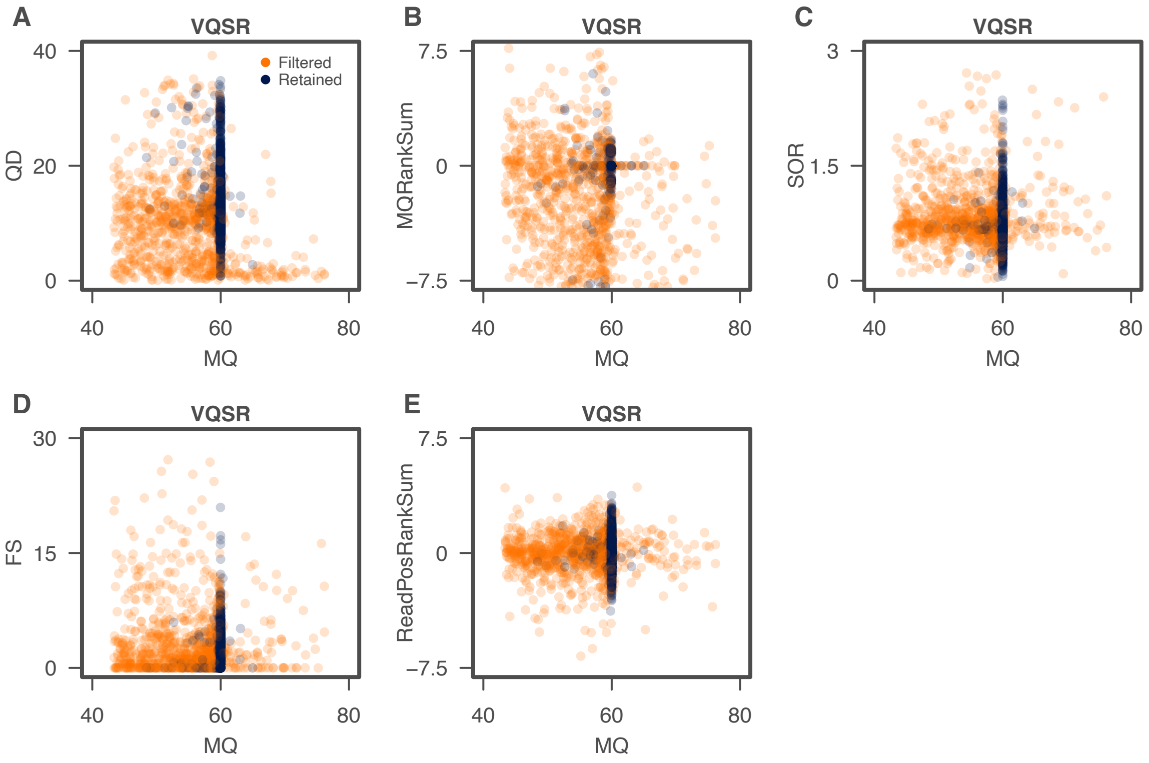


Figure 4.9: VQSR filtering

A-E show pairwise interactions between the MQ annotation and all other annotations. These are representative plots that demonstrate that VQSR does not apply a strict threshold to each annotation but rather is capable of considering all of a variant's annotations relative to the profile of 'good' variants.

Exome variant exploration

KIAA1407 is alternatively called coiled-coil domain containing 191 (*CCDC191*). The NCBI Gene Expression Omnibus database (Edgar, Domrachev, and Lash 2002; Barrett *et al.* 2013) shows broad expression which peaks in the testis and thyroid. This variant confers a LOF; however, gnomAD identifies this gene as being tolerant of LOF variants (observed/expected = 0.97 (95% CI: 0.77-1.23)). While this variant is absent in Project MinE controls, the frequency in the Project MinE cases (1.15x10^{-4}) is approximately equal to the gnomAD non-neuro subset (1.6x10^{-4}). Given the early ages of onset of these patients, carriers of a highly penetrant dominant ALS causing allele would be particularly unlikely to be present in the gnomAD non-neuro subset. Additionally several *KIAA1407* truncating variants are found to be more common in Project MinE cases than controls. The evidence suggests that *KIAA1407*:c.700G>T(p.[Q234*]) is not a pathogenic variant.

RYBP is a component of the Polycomb group (PcG) multiprotein PRC1-like complex. The Project MinE database shows no observed variants in this gene in cases or controls. While this variant is also absent in gnomAD, the gnomAD database shows this gene to be relatively tolerant of missense variants (observed/expected: 0.86 (95% CI: 0.75-0.98). Similar to other

PcG genes, previous studies have implicated *RYBP* in cancer aetiology (Novak and Phillips 2008; Zhu *et al.* 2017; Ali *et al.* 2018); and while it has been shown to be a requirement for central nervous system development (Pirity, Locker, and Schreiber-Agus 2005), it has not been associated with neurological disease. The confirmed segregation of the *RYBP*:c.633T>A(p.[S211R]) variant makes it a variant of interest but more study is required to ascertain its pathogenicity.

Burden analysis

Burden testing was performed in this study to identify if any gene had a significant excess of either missense (figure 4.10 C) or LOF variants (figure 4.10 D). No statistically significant results are observed, likely as a result of the small size and lower power of this study.

Oligogenic analysis

Following the observation of multiple cases carrying multiple variants, binomial testing was performed to explore whether there is statistically significant evidence of oligogenic inheritance in this study. Figure 4.10 E-G display that significance is not achieved when testing all rare, functional variants which pass sequencing filters (figure 4.10 E), when restricting these variants to just those present in P and LP genes (figure 4.10 F) or when restricting this further to just samples with mean coverage above 20X.

The lack of statistical evidence supporting oligogenic inheritance in this study indicates that either it is not a contributing factor in Cuban ALS genetics or, given the small size of the study and that oligogenic results have replicated in different studies, that our study is underpowered to observe this effect. It does however indicate that conclusions cannot be drawn as to the significance of carriers of multiple variants (table 4.4).

Cohort analysis and international comparison

In this study no pathogenic variants are observed in Cuban familial cases. 2.7% of sporadic cases carry the *C9orf72* RE and 0.9% of sporadic cases are explained by a reclassified pathogenic *FUS* variant. This profile stands in contrast to other North American countries and also to South American countries (figure 4.11). In other studied North American countries (primarily USA and Canada) *C9orf72* explains 33% of familial cases and 5% or sporadic cases. In South America these figures are 10.26% and 2.65% respectively. *SOD1* variants explain 12.4% of FALS in USA and Canada and 0.35% of SALS, Cuba much more closely resembles South America as both have an absence of *SOD1* variants. *VAPB*:c.166C>T(p.[P56S]) explains 33% of Brazilian FALS cases but is virtually absent in the rest of the world. We do not identify this variant in Cuba, further confirming that it is localised to Brazil.

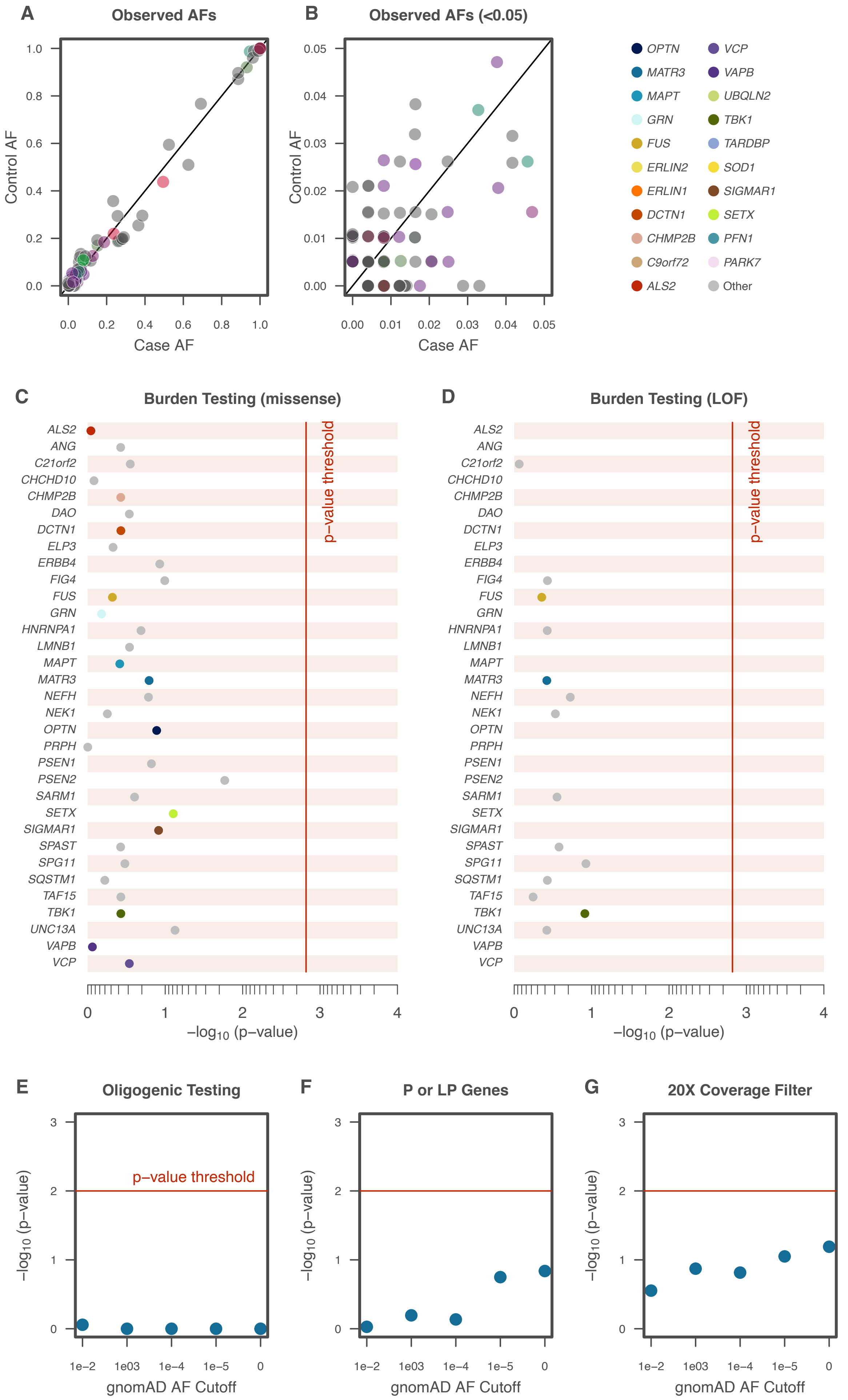
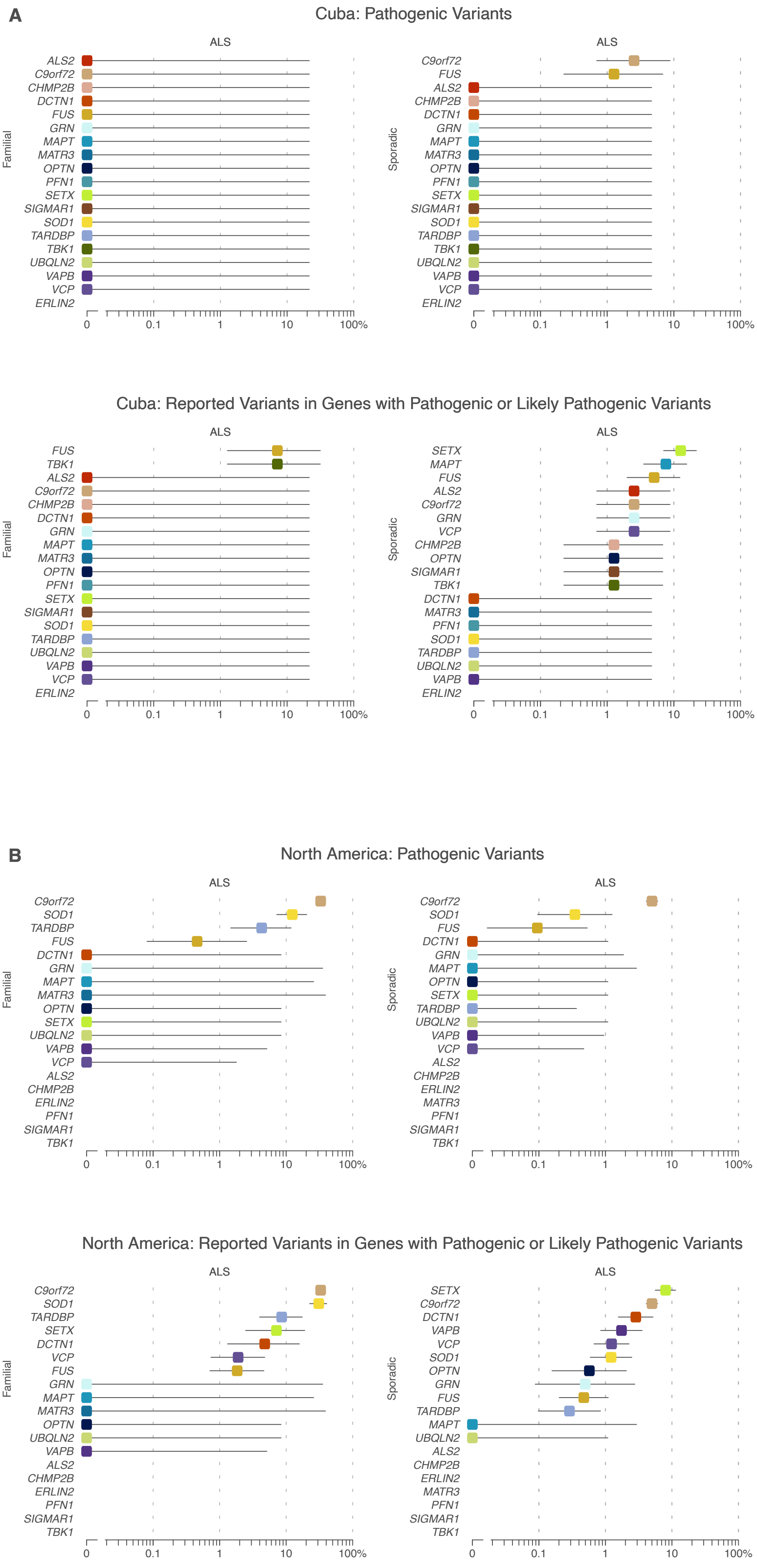

A
Observed AFs
Control AF
Case AF
0.0
0.2
0.4
0.6
0.8
1.0
B
Observed AFs (<0.05)
Control AF
Case AF
0.00
0.01
0.02
0.03
0.04
0.05
OPTN
MATR3
MAPT
GRN
FUS
ERLIN2
ERLIN1
DCTN1
CHMP2B
C9orf72
ALS2
VCP
VAPB
UBQLN2
TBK1
TARDBP
SOD1
SIGMAR1
SETX
PFN1
PARK7
Other
C
Burden Testing (missense)
D
Burden Testing (LOF)
ALS2
ANG
C21orf2
CHCHD10
CHMP2B
DAO
DCTN1
ELP3
ERBB4
FIG4
FUS
GRN
HNRNPA1
LMNB1
MAPT
MATR3
NEFH
NEK1
OPTN
PRPH
PSEN1
PSEN2
SARM1
SETX
SIGMAR1
SPAST
SPG11
SQSTM1
TAF15
TBK1
UNC13A
VAPB
VCP
p−value threshold
0
1
2
3
4
$-\log_{10}$ (p−value)
E
Oligogenic Testing
F
P or LP Genes
G
20X Coverage Filter
$-\log_{10}$ (p−value)
0
1
2
3
p−value threshold
1e−2
1e03
1e−4
1e−5
0
gnomAD AF Cutoff

Figure 4.10: Variant distribution in cases and controls

(figure on previous page)
A) and B) display the gnomAD AF of all variants observed in cases and controls (A) and rare variants in cases and controls (B). Both rare and common variants are observed in controls at similar rates to cases including the observation of novel variants in controls that are absent in cases. C-D) No gene is observed to carry a statistically significant excess of either missense (C) or LOF (D) variants. E-F) Cases are not found to be statistically more likely to carry multiple variants whether considering E) functional variants which pass sequencing filters, F) restricting this analysis to just genes with P or LP variants identified in Chapter 2 or G) restricting this further to individuals with average depth of coverage exceeding 20X.



A
Cuba: Pathogenic Variants
ALS
Familial
ALS2
C9orf72
CHMP2B
DCTN1
FUS
GRN
MAPT
MATR3
OPTN
PFN1
SETX
SIGMAR1
SOD1
TARDBP
TBK1
UBQLN2
VAPB
VCP
ERLIN2
0
0.1
1
10
100%
ALS
Sporadic
C9orf72
FUS
ALS2
CHMP2B
DCTN1
GRN
MAPT
MATR3
OPTN
PFN1
SETX
SIGMAR1
SOD1
TARDBP
TBK1
UBQLN2
VAPB
VCP
ERLIN2
0
0.1
1
10
100%
Cuba: Reported Variants in Genes with Pathogenic or Likely Pathogenic Variants
ALS
Familial
FUS
TBK1
ALS2
C9orf72
CHMP2B
DCTN1
GRN
MAPT
MATR3
OPTN
PFN1
SETX
SIGMAR1
SOD1
TARDBP
UBQLN2
VAPB
VCP
ERLIN2
0
0.1
1
10
100%
ALS
Sporadic
SETX
MAPT
FUS
ALS2
C9orf72
GRN
VCP
CHMP2B
OPTN
SIGMAR1
TBK1
DCTN1
MATR3
PFN1
SOD1
TARDBP
UBQLN2
VAPB
ERLIN2
0
0.1
1
10
100%
B
North America: Pathogenic Variants
ALS
Familial
C9orf72
SOD1
TARDBP
FUS
DCTN1
GRN
MAPT
MATR3
OPTN
SETX
UBQLN2
VAPB
VCP
ALS2
CHMP2B
ERLIN2
PFN1
SIGMAR1
TBK1
0
0.1
1
10
100%
ALS
Sporadic
C9orf72
SOD1
FUS
DCTN1
GRN
MAPT
OPTN
SETX
TARDBP
UBQLN2
VAPB
VCP
ALS2
CHMP2B
ERLIN2
MATR3
PFN1
SIGMAR1
TBK1
0
0.1
1
10
100%
North America: Reported Variants in Genes with Pathogenic or Likely Pathogenic Variants
ALS
Familial
C9orf72
SOD1
TARDBP
SETX
DCTN1
VCP
FUS
GRN
MAPT
MATR3
OPTN
UBQLN2
VAPB
ALS2
CHMP2B
ERLIN2
PFN1
SIGMAR1
TBK1
0
0.1
1
10
100%
ALS
Sporadic
SETX
C9orf72
DCTN1
VAPB
VCP
SOD1
OPTN
GRN
FUS
TARDBP
MAPT
UBQLN2
ALS2
CHMP2B
ERLIN2
MATR3
PFN1
SIGMAR1
TBK1
0
0.1
1
10
100%

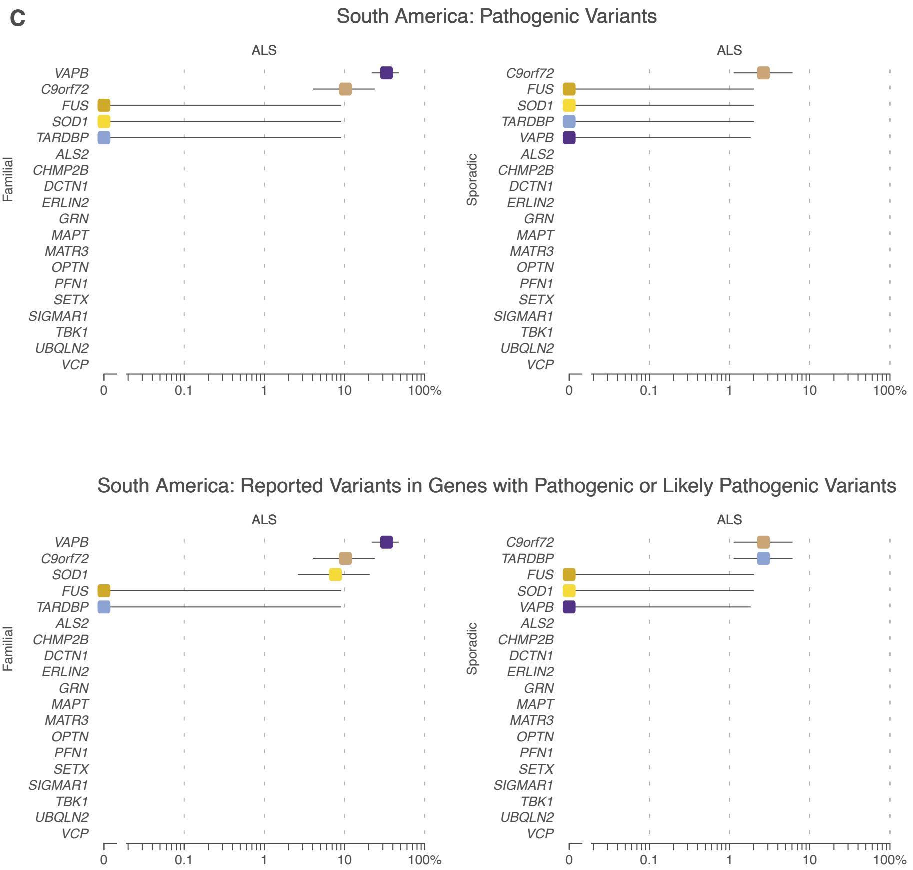


Figure 4.11: Cuba international comparisons

The proportions explained by pathogenic variants or any variants in pathogenic or likely pathogenic genes are displayed for A) Cuba, B) other North American countries and C) South America.

Discussion

Summary and significance

This work represents the first genetic screen of ALS patients from Cuba. NGS is performed for 120 unrelated ALS patients, 6 members of a single pedigree and 111 unrelated healthy controls. In summary, 2.7% of sporadic cases carry the *C9orf72* RE and 0.9% of sporadic cases are explained by a reclassified pathogenic *FUS* variant. No familial cases carry known pathogenic variants. Three previously unreported frameshift *TBK1* variants are identified that remain classified as VUS but have evidence that is supportive of pathogenicity.

The work in Chapter 2 highlighted that in order to both achieve global parity and to further increase our understanding of ALS genetics, it is vital that we begin to research understudied countries and regions. As a country with no previous ALS genetic screening and where most individuals have admixed ancestry including European, African, Native American and East Asian, Cuba fits this profile.

Both oligogenic and gene burden studies did not return statistically significant results. This is to be expected as these tests, particularly burden tests, typically require thousands of patients to find significant associations (Kenna *et al.* 2016; Nicolas *et al.* 2018). Rather, the purpose of these tests in this instance is to highlight that just because, for example, cases are found to harbour 14 rare *SETX* variants that are absent in controls, this is not indicative of a high rate of *SETX* variation in Cuban ALS cases, as similar numbers are observed in controls. Similarly, while several patients carrying multiple variants are identified, and similar results are often posited as further evidence of oligogenic inheritance in ALS; in fact a comparable rate of multiple variants is observed in controls.

Chapter 2 has provided a framework by which to analyse and interpret the research conducted here. The journALS database contains both novel analyses and variant annotation from several sources; as such, it is first utilised in this study as a tool in the variant filtering process. One variant that is present in our dataset but has not previously been discussed is *DCTN1*:c.2353C>T(p.[R785W]). This variant is identified in two patients in this study and has previously been identified in two affected siblings from Germany (Münch *et al.* 2004) and a case in the UK (Morgan *et al.* 2017). This variant has understandably pervaded the ALS literature since it was first published in 2004; however, in this analysis it is removed during filtering as it is at five times higher frequency in Project MinE controls than Project MinE cases and is at similar frequency in gnomAD. The journALS database facilitates a focused approach to variant filtration.

The second means by which journALS benefits this study, is as a means of interpreting variants that are retained after filtering. Having clarified the inheritance patterns and phenotypes associated with different genes, we are now able to interpret novel variants within this framework. For example, while many *ALS2* missense variants have previously been reported, we only find sufficient evidence supporting the pathogenicity of homozygous LOF variants which result in an early AOO phenotype. As such, the *ALS2* variants identified in this study are unlikely to be causative as they are not homozygous LOF and the patients are not early onset.

The final utility of journALS in this work is in understanding the broader context of the results. It is ascertained that Cuban ALS patients have a unique genetic profile quite distinct from the Northern American countries which are primarily of European ancestry and are characterised by high rates of *C9orf72* and *SOD1* variants and equally distinct from South American countries such as Brazil where variation in *VAPB* explains a large proportion of cases.

This work demonstrates the benefit and positive feedback loop that can be achieved through genetic screening. The understanding of two variants in particular has significantly increased due to this study. Firstly, *FUS*:c.1512_1513delAG(p.[G505fs]) is now reclassified as a pathogenic variant, and secondly *hnRNPA1*:c.1018C>T(p.[P340S]) is found to be is unlikely to be pathogenic as it does not segregate in the pedigree presented here. Increased understanding of these variants will help genetic counselling for future patients, increasing the rate of true positive diagnoses and decreasing the rate of false positive genetic diagnoses.

Two variants are identified as segregating in a large pedigree with early onset and long disease duration. These variants require additional follow up research to assess their frequency in both the general Cuban and ALS populations and further research to assess their functional impacts. It is possible that one of these variants is causative; however, it must also be acknowledged that there are several examples in journALS where a variant was initially found to segregate and now is either found to be more common in cases than controls (Mitchell *et al.* 2010), or sequencing additional affected family members revealed that the variant does not fully segregate (Johnson *et al.* 2014; Saez-Atienzar *et al.* 2020). There are other possible genetic explanations in this family that have not been explored. The true causative variant may be intronic or intergenic, or may be an unstudied RE.

Limitations

While targeted sequencing provides a means of identifying variants in previously reported genes, it does not facilitate the discovery of new genes or more complex variants such as genomic rearrangements. This is evident by the small proportion of cases with an identified pathogenic variant. As with the example of *VAPB* in Brazil, and the numerous other variants which exhibit geographic heterogeneity, it is possible that there is a single variant in an unknown gene affecting a large number of patients.

The second limitation of this study is that as exome sequencing was only performed for pedigree 2302, the frequency of the *RYBP* and *KIAA1407* variants in the remaining cases and controls is currently unknown.

Unfortunately no unaffected family members are available from pedigree 2302; this would greatly improve the power of segregation analysis. SNP genotyping is not performed but would enable the identification of a shared linkage region between patients.

The final limitation of this study is that we have not ascertained the ethnic background of the patients. While the Cuban population is very admixed, and these patients do represent the overall population well in terms of geographic distribution and self-reported ancestry, we do not know if any particular background is over- or under-represented at the genomic level, which could have interesting implications for identifying ALS risk-factors.

Future direction

Future work should focus on addressing the limitations outlined above. The frequency of *RYBP* and *KIAA1407* variants should be ascertained in the remainder of the cases and controls, this can be achieved either through Sanger sequencing of these variants or through the inclusion of Cuban cases and controls in whole-genome sequencing projects such as Project MinE. Prioritising these patients in a large WGS study would also aid in the identification of novel variants as instant international case/control comparisons would be available.

Patients should undergo SNP genotyping with the goal of answering the following questions:

1) Are there any shared genomic regions among Cuban cases who lack an explained genetic cause?

2) Is there any cryptic relatedness between samples that may help further interpret variants of uncertain significance either because segregation or lack of segregation is observed?

3) Are there any ethnic backgrounds which are over- or under-represented at the genomic level?

4) What is the shared linkage region between family members in pedigree 2302?

Conclusion

2.7% of sporadic cases in Cuba carry the *C9orf72* RE and 0.9% of sporadic cases are explained by a pathogenic *FUS* variant. Cuban ALS patients have a distinct genetic profile and, as a large proportion of cases lack an identified genetic cause, should be prioritised for further research.

This study demonstrates both the utility of the journALS database in conducting ALS research and how genetic screening can improve our knowledge of ALS genetics, providing benefit for the entire research community and ultimately for patients.

Chapter 5

The broader spectrum of motor neurone disease genetics in Ireland

Introduction

ALS sits within the ALS-FTD phenotypic continuum but also within a phenotypically and genetically heterogenous spectrum of MNDs. As discussed in Chapter 1, ALS is characterised by the loss of both lower motor neurons (LMNs) and upper motor neurons (UMNs). Loss of UMNs prevents signalling to the LMNs, resulting in muscle stiffness and weakness, while LMN degeneration prevents muscles from receiving signals, leading to weakness and muscular atrophy (Kent-Braun *et al.* 1998). While ALS patients lose both UMNs and LMNs, other MNDs are classified either by selective LMN degeneration (progressive muscular atrophy (PMA), spinal muscular atrophy (SMA)), or UMN degeneration (hereditary spastic paraplegia (HSP), primary lateral sclerosis (PLS)).

Much is still unknown about the genetic factors underpinning a patient's development of ALS or FTD and even more is unknown about the development of adult-onset selective motor neuron degeneration. This Chapter will examine genotypic variation in the Irish ALS population, but also will study the genetic basis of FTD and a PLS, a UMN disease, in Ireland.

Upper motor neurone disorders

HSP patients, as discussed in Chapter 1, typically experience stiffness and weakness of the lower extremities with gradual onset and slow progression, with onset occurring any time from childhood to adulthood. A hallmark of HSP is that it has distinct autosomal dominant, recessive or X-linked inheritance in pedigrees, and consequently variants in over 70 genes have been associated with HSP inheritance (table 5.1) (de Souza *et al.* 2017; Parodi *et al.* 2017; Klebe, Stevanin, and Depienne 2015; Lo Giudice *et al.* 2014).

PLS patients experience adult onset with stiffness in both arms and legs and often progress to difficulty in swallowing. While there is considerable overlap between the phenotypes of adult onset HSP and PLS, the upper body stiffness and bulbar symptoms often observed in PLS are rarely a feature of HSP (Frans Brugman *et al.* 2009). PLS is estimated to account for 7% of adult onset MNDs (W.-K. Kim *et al.* 2009). The rare nature of PLS and typical lack of familial segregation has historically made it difficult to study.

PLS genetics

PLS is often described as a sporadic condition. This is a label that distinguishes it from HSP which distinctly segregates in families. However, just as in ALS, the term sporadic does not necessarily mean there is no genetic basis to the condition; indeed, the term sporadic is not even always an accurate description for PLS patients as many have family members affected by PLS or other MNDs (F. Brugman *et al.* 2005; Dupré *et al.* 2007; Valdmanis, Dupré, and Rouleau 2008; Praline *et al.* 2010). This shared aetiology within pedigrees and shared phenotypic overlap with other MNDs highlights the likely contribution of genetic factors in PLS pathogenesis.

Due to the rarity of PLS cases, genetic screening has been scarce and with small numbers. Combining the results of 8 studies, *C9orf72* RE is observed in 2.1% (95% CI: 0.9-4.8%) of patients diagnosed with PLS (table 5.2) (Stewart *et al.* 2012; García-Redondo *et al.* 2013; Mitsumoto *et al.* 2015; van Rheenen *et al.* 2012; Rutherford, DeJesus-Hernandez, *et al.* 2012; Hübers *et al.* 2014; Nicola Ticozzi *et al.* 2014; de Vries *et al.* 2017), although it is important to acknowledge the possibility that some of these patients may have subsequently developed LMN symptoms after the period of reporting.

Table 5.1: Genes associated with HSP (1/2)

Gene	Clinical_Phenotype	Inheritance	Exons	TOM	Original Study
ACP33	SPG21	AR	9	PM,ins	Simpson et al. (2003)
ALDH18A1	SPG9A	AD		PM	Panza et al. (2016)
AMPD2	SPG63	AR	18	Del	Novarino et al. (2014)
AP4B1	SPG47	AR	10	Ins, del	Abou Jamra et al. (2011)
AP4E1	SPG51	AR	21	Ss	Abou Jamra et al. (2011)
AP4M1	SPG50	AR	15	Ss	Abou Jamra et al. (2011)
AP4S1	SPG52	AR	6	PM	Abou Jamra et al. (2011)
ARL6IP1	SPG61	AR	6	Del	Novarino et al. (2014)
ARSI	SPG66	AR	2	Ins	Novarino et al. (2014)
ATAD3A	NA	AR	1	PM	Harel et al. (2016)
ATL1	SPG3A	AD	14	PM	Zhao et al. (2001)
ATP2B4	NA	AD		PM	Li et al . (2014)
B4GALNT1	SPG26	AR	11	PM,del,dupl	Boukhris et al. (2013)
BICD2	NA	AR	7	PM	Novarino et al. (2014)
BSCL2	SPG17	AD	12	PM	Windpassinger et al. (2004)
C12orf65	SPG55	AR	3	PM	Shimazaki et al. (2012)
C19orf12	SPG43	AR	3	PM,del	Landouré et al. (2013)
CCT5	NA	AR	11	PM	Bouhouche et al. (2006)
CPT1C	PG73	AD		PM	Carrasco et al. (2013)
CYP2U1	SPG56	AR	5	PM,del	Tesson et al. (2012)
CYP7B1	SPG5A	AR	6	PM	Tsaousidou et al. (2008)
DDHD1	SPG28	AR	13	PM,del	Tesson et al. (2012)
DDHD2	SPG54	AR	18	PM,ins,ss	Schuurs-Hoeijmakers et al. (2012)
DNM2	NA	AD		PM	Sambuughin et al. 2015
ENTPD1	SPG64	AR	10	PM	Novarino et al. (2014)
ERLIN1	SPG62	AR	11	PM	Novarino et al. (2014)
ERLIN2	SPG18	AR	12	Del	Alazami et al. (2011)
EXOSC3	NA	AR		PM	Halevy et al. (2014)
FA2H	SPG35	AR	7	PM,del	Dick et al. (2010)
FLRT1	SPG68	AR	2	PM	Novarino et al. (2014)
GAD1	NA	AR	17	PM	Lynex et al. (2004)
GBA2	SPG46	AR	18	PM,del	Martin et al. (2013)
GJC2	SPG44	AR	2	PM	Orthmann-Murphy et al. (2009)
HSPD1	SPG13	AD	12	PM	Hansen et al. (2002)
KIAA0196	SPG8	AD	29	PM,del	Valdmanis et al. (2007)
KIAA0415	SPG48	AR	17	Indel	Slabicki et al. (2010)
KIAA1840	SPG11	AR	40	PM,dupl,ins,del,ss	Stevanin et al. (2007)
KIF1A	SPG30	AR	50	PM	Erlich et al. (2011)
KIF1C	SPG58	AR	23	PM,del	Novarino et al. (2014)
KIF5A	SPG10	AD	29	PM	Reid et al. (2002)
L1CAM	SPG1	X-linked	29	PM	Jouet et al. (1994)
LYST	NA	AR	53	PM	Shimazaki et al. (2014)
MAG	NA	AR	12	PM	Novarino et al. (2014)
MARS	SPG70	AR	21	PM	Novarino et al. (2014)
NIPA1	SPG6	AD	5	PM	Rainier et al. (2003)

Table 5.1: Genes associated with HSP (2/2)

Gene	Clinical_Phenotype	Inheritance	Exons	TOM	Original Study
NT5C2	SPG56	AR	18	PM,ss	Novarino et al. (2014)
OPA3	NA	AR	3	PM	Arif et al. (2013)
PGAP1	SPG67	AR	27	Ss	Novarino et al. (2014)
PLP1	SPG2	X-linked	8	PM,del,dupl	Saugier-Veber et al. (1994)
PNPLA6	SPG39	AR	34	PM,ins	Rainier et al. (2003)
RAB3GAP2	SPG69	AR	35	PM	Novarino et al. (2014)
REEP1	SPG31	AD	7	PM,del,ss,ins	Züchner et al. (2006)
REEP2	SPG72	AD/AR	8	PM	Esteves et al. (2014)
RTN2	SPG12	AD	11	PM,ins,del	Montenegro et al. (2012)
SLC16A2	SPG22	X-linked	6	PM,del,ins	Schwartz et al. (2005)
SLC33A1	SPG42	AD	6	PM	Lin et al. (2008)
SPAST	SPG4	AD	17	PM,ss,del,dupl	Hazan et al. (1999)
SPG20	SPG20	AR	9	PM,del	Patel et al. (2002)
SPG7	SPG7	AR	17	PM,del,ins	Casari et al. (1998)
TECPR2	SPG49	AR	20	NA	Oz-Levi et al. (2012)
TFG	SPG57	AR	8	PM	Beetz et al. (2013)
TUBB4A	NA	AD		PM	Kancheva et al. (2015)
USP8	SPG59	AR	21	PM	Novarino et al. (2014)
VPS37A	SPG53	AR	12	PM	Zivony-Elboum et al. (2012)
WDR48	SPG60	AR	19	Del	Novarino et al. (2014)
ZFR	SPG71	AR	20	PM	Novarino et al. (2014)
ZFYVE26	SPG15	AR	42	PM,del,ss,ins	Hanein et al. (2008)
ZFYVE27	SPG33	AD	12,13	PM	Mannan et al. (2006)

Table compiled from de Souza et al. (2017), Parodi et al. (2017), Klebe, Stevanin, and Depienne (2015) and Lo Giudice et al. (2014).

AD= autosomal dominant; AR= autosomal recessive; Del= deletion; Ins= insertion; PM= point mutation; Ss= splice site; TOM= type of mutation;

Table 5.2: *C9orf72* repeat expansion screening studies in PLS

Study	Country	PLS Patients Screened	C9orf72 RE Positive	Proportion of Carriers
van Rheenen et al. (2012)	Netherlands	110	1	0.9% (95% CI 0.2-0.5%)
Mitsumoto et al. (2015)	USA	41	1	2.4% (95% CI: 0.4-12.6%)
Hübers et al. (2014)	Germany	30	0	0% (95% CI: 0-11.4%)
Stewart et al. (2012)	Canada	23	2	8.7% (95% CI: 2.4-26.8%)
García-Redondo et al. (2013)	Spain	22	1	4.6% (95% CI: 0.8-21.8%)
de Vries et al. (2017)	Netherlands	4	0	0% (95% CI: 0-49.0%)
Ticozzi et al. (2014)	UK	2	0	0% (95% CI: 0-65.8%)
Rutherford et al. (2012)	USA	2	0	0% (95% CI: 0-65.8%)
Total		**234**	**5**	**2.1% (95% CI: 0.9-4.8%)**

Due to the strong clinical overlap with both adult-onset HSP and ALS, the few genetic studies that have been performed have focused on identifying potentially pleiotropic pathogenic variants in these genes. Yang *et al.* (2016) identified a PLS pedigree with five affected members that carried compound heterozygous *SPG7* variants. Compound heterozygosity in *SPG7* would typically be a hallmark of HSP; however, the patients had bulbar and upper limb involvement which are signatures of PLS rather than HSP. McDermott *et al.* (2003) examined the genes *SPAST* and *SPG7* in 7 PLS patients and did not identify any variants. Mitsumoto *et al.* (2015) performed the only broad NGS screen of a PLS patient cohort to date. WES was performed for 41 PLS patients and identified heterozygous variants in *SPG7*, *DCTN1* and *PARK2* that are previously reported as pathogenic in HSP, ALS and Parkinson's respectively. It is worth noting that the *DCTN1* variant (c.3746C>T(p.[T1249I])) is observed at a higher frequency in Project MinE controls than cases so is unlikely to be truly pathogenic, the *SPG7* variant (c.1529C>T(p.[A510V])) is associated with HSP in either recessive or compound heterozygous state and the *PARK2* variant (c.823C>T(p.[R275W]) is a well-established Parkinson's variant but only in homozygosity (Abbas *et al.* 1999). This renders the pathogenicity of these variants in PLS uncertain. There have been other rare reports of PLS patients carrying variants in ALS-associated genes including *OPTN* (Del Bo *et al.* 2011), *UBQLN2* (H.-X. Deng *et al.* 2011) and two variants in *FIG4* (Chow *et al.* 2009).

Given the rarity of PLS it is difficult to gather sufficient patients to conduct in-depth genetic analyses; however, to further our understanding of the genetic basis underpinning the MND phenotypic spectrum it is vital to further investigate the potential causes of this rare phenotype.

The Irish ALS register

The Irish ALS register was first established in 1995 and continues to the present day. The register records and monitors progress and disease progression for all consenting patients who not only present with ALS but also PLS, PMA and other rare forms of adult-onset MND. The register forms the backbone to research spanning genetics, epidemiology, neuroimaging, neurophysiology, neuropsychology as well as patient and carer support and wellbeing.

Genetic screening of ALS patients in Ireland has been published previously (Kenna, McLaughlin, Byrne, *et al.* 2013; McLaughlin, Kenna, Vajda, Heverin, *et al.* 2015; Byrne *et al.* 2012); however, genetic screening of FTD and PLS patients has not. Additionally, as gene-based ALS clinical trials are now underway in Ireland, it is vital to continually re-evaluate genetic results as new patients and new contextual information becomes available.

Recent work utilising the Irish ALS register has identified Irish families with multiple affected family members who have discordant *C9orf72* genotyping (Ryan *et al.* 2018), the basis of this discordance requires further investigation.

Research Aims

1. Perform the first comprehensive genetic screen of an FTD patient cohort in Ireland.
2. Perform the first genetic screen of a PLS patient cohort in Ireland.
3. Analyse the largest Irish ALS cohort to date in the context of the journALS study of Chapter 2 and in the context of information available from the Irish ALS register.
4. Analyse Irish ALS pedigree with discordant *C9orf72* genotyping.

Methods

Participants

ALS and PLS patients attended the national specialist MND clinic at Beaumont Hospital Dublin. All ALS patients were diagnosed as definite, probable or possible ALS by specialist neurologists in accordance with the El Escorial criteria (Brooks *et al.* 2000). A PLS diagnosis was made based on the consensus diagnostic criteria (Turner *et al.* 2020). FTD patients were recruited from the cognitive clinic at St James's Hospital and the neurodegenerative clinic at Beaumont Hospital and patients were diagnosed by a specialist neurologist based on the Rascvosky criteria (Rascovsky *et al.* 2011).

FTD DNA sequencing, processing and analysis

The targeted-sequencing pipeline developed for the Cuban patients in Chapter 4 was applied here to 51 patients with FTD. Briefly; the exons and surrounding 4bps of 37 genes previously linked to ALS, FTD or dementia (table 4.1) underwent target enrichment with 11 cycles of PCR. Samples were pooled and sequenced on an Illumina MiSeq at the TrinSeq facility at St. James's Hospital with 300bp single end sequencing.

The resultant FASTQ files were adapter-trimmed, aligned to the GRCh37 version of the human reference genome, duplicate reads were removed and samples underwent BQSR. GATK best practices with hard-filtering were followed for variant calling. Variants were annotated with a suite of *in silico* prediction tools, population datasets and disease specific databases to ensure compatibility with the journALS study in Chapter 2.

Observed variants were analysed jointly with variants present in 136 PCR-free Irish control samples; following the variant analysis pipeline developed in Chapter 4, filtering variants classified as benign or likely benign in the journALS database and retaining rare variants with a functional effect.

PLS DNA sequencing, processing and analysis

DNA from 44 PLS patients underwent Agilent SureSelect WES enrichment and 150bp PE sequencing to a target depth of 90X on an Illumina NovaSeq (table 3.2). These samples were sequenced concurrently with the Cuban family described in Chapter 4 and a large Irish pedigree containing affected and unaffected individuals (further described in this chapter).

Read alignment and variant calling with VQSR for these samples has been described in Chapter 4.

Multidimensional scaling (MDS) with Plink v1.9 (Purcell *et al.* 2007) was used to confirm whether 136 Irish PCR-free WGS controls were suitably comparable to PLS patients (figure 5.1). SNPs sharing linkage were pruned by removing SNPs within a 50bp range with an R^2 exceeding 0.2 (--indep-pairwise 50 5 0.2). MDS was performed after calculating pairwise identity by state (IBS) between samples. 236,019 SNPs and INDELs which were retained after merging WGS and WES data, filtering for variants which pass sequencing filters and pruning SNPs in high linkage disequilibrium. MDS analysis showed no significant bias between WGS and WES samples (figure 5.1).

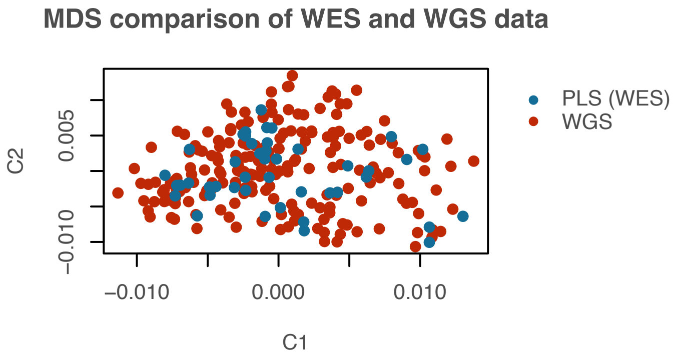


Figure 5.1: Validation of use of WGS controls with WES PLS data

Variants called from PLS data (blue) do not show significant bias when compared to exomic variants extracted from WGS data, indicating that controls from the WGS data are suitably comparable to WES samples.

PLS patients were screened for variants in genes linked to ALS or FTD using the previously described pipeline. Additionally, 70 genes linked to HSP (table 5.1) and the gene *PARK2* were included in variant screening. To account for the fact that the HSP literature was not screened for inclusion in the journALS data browser, variants identified in PLS patients were compared to HGMD v.2021.4 (public) and ClinVar (GRCh37_clinvar_20220313) to determine if they were previously reported in the literature.

A single SKAT test was performed to identify if PLS patients carry an excess of functional variants in HSP associated genes relative to controls. EPACTS v.3.3 was used to assign functional and gene annotations to all variants which passed sequencing filters. The exons of all genes within table 5.1 were treated as a single group and all missense and LOF variants (Missense, Nonsynonymous, StructuralVariation, Stop_Gain, Stop_Loss, Start_Gain, Start_Loss, Frameshift, CodonGain, CodonLoss, CodonRegion, Insertion, Deletion, Essential_Splice_Site, Nonsense) with a MAF below 0.05 were included in analysis.

In Chapter 3 ExpansionHunter v3 was found to be capable of accurately classifying STRs in the normal range and REs while also genotyping more exonic sites than ExpansionHunter v2. PLS samples are genotyped with ExpansionHunter v3 as described in Chapter 3 and the results are compared to PCR-free WGS control samples.

ALS data curation, processing and analysis

To perform the largest study of variation in the Irish ALS population to date, NGS data were collated from two sources. PCR-free WGS data were available for 272 ALS patients and 136 controls, these data were sequenced as part of Project MinE (van der Spek, van Rheenen, Pulit, Kenna, van den Berg, *et al.* 2019; Project MinE ALS Sequencing Consortium 2018) and have been previously described in Chapter 4. Targeted sequencing data were available for a further 404 patients and 311 controls from a previous study (Kenna, McLaughlin, Byrne, *et al.* 2013).

Targeted sequencing data were processed from FASTQ to variant calling via the previously described GATK best practices pipeline. Variants identified in the WGS and targeted data were filtered through the variant filtering pipeline developed in Chapter 4. Patients with only targeted sequencing data were screened for the previously described list of genes (table 4.1), and patients with WGS data were additionally screened for variants in *ERLIN1*, *ERLIN2*, *PARK7* and *KIF5A*; the first three of these genes are identified as significant in the journALS study of chapter 2 and *KIF5A* has reliably been linked to ALS pathogenesis through exome burden studies (Nicolas *et al.* 2018).

15 members of an Irish family that has discordant *C9orf72* genotyping underwent exome sequencing and variant calling in conjunction with the previously described PLS patients and Cuban family.

C9orf72 genotyping

C9orf72 genotyping was performed for all FTD and PLS patients as described in Chapter 4. *C9orf72* genotyping of Irish ALS patients has routinely been performed for Irish ALS patients since 2011 (including retrospectively where possible) and was continued for this study.

Discordant families

Recent work by Dr. Marie Ryan in the Academic Unit of Neurology TCD has utilised the Irish ALS register to identify Irish families with multiple affected family members who have discordant *C9orf72* genotyping (Ryan *et al.* 2018). The basis and confirmation of this discordance is further investigated here in 3 pedigrees for which sufficient information is available.

Discordant pedigrees are investigated using all available genotyping data (SNP, WGS, WES, targeted sequencing data and rpPCR genotyping). Where sufficient DNA is available, discordant patients have undergone a secondary rpPCR to confirm the initial result. Where SNP genotyping data are available, relatedness is confirmed using identity-by-descent (IBD) and haplotype analysis is performed to determine whether *C9orf72* RE positive and negative patients carry the *C9orf72* haplotype (described below). The observed haplotype is compared to the established *C9orf72* haplotypes observed in Europe (Smith *et al.* 2013), Finland (Laaksovirta *et al.* 2010), Sweden (Chiang *et al.* 2017) and the UK (Mok *et al.* 2012). Where WGS data is available the presence or absence of the RE is confirmed using ExpansionHunter v2 and ExpansionHunter v3 as described in Chapter 3. Finally where targeted NGS, WES or WGS is available patients are screened for any segregating variants.

SNP genotyping

SNP genotyping data from five Irish cohorts (table 5.3) were analysed to study the haplotype surrounding the *C9orf72* RE in families with discordant RE genotyping and to confirm relatedness. Plink v1.9 (Purcell *et al.* 2007) was used to perform data QC separately for each dataset (table 5.4, table 5.5).

To verify relationships between family members, IBD matrices were calculated by first filtering to common SNPs (MAF >0.35), removing SNPs that were absent in more than 5%

of samples and removing SNPs that were significantly out of HWE in controls. SNPs sharing linkage were pruned by removing SNPs within a 50bp range with an R^2 exceeding 0.2.

To assess the haplotypes surrounding the *C9orf72* locus, each dataset was phased with Beagle v4.1 (Browning and Browning 2007) using 1,000 Genomes Project Phase 3 (1000 Genomes Project Consortium *et al.* 2015) reference data.

Table 5.3: Source for SNP datasets

Identifier	Platform	Source
GWA1	Illumina HumanHap550v3.0	McLaughlin *et al.* 2015 a
GWA2	Illumina Human610-Quadv1.0	McLaughlin *et al.* 2015 b
GWA3	HumanOmniExpressExome-8v1	van Rheenen *et al.* 2016
GWA4	Illumina GSA	van Rheenen *et al.* 2021
LP	Illumina Infinium HumanOmni2.5-8 SNP array v1.2	Project MinE ALS Sequencing Consortium 2018

Table 5.4: SNP QC filters

Step	Name	Description
QC1	Sample Missingness	Remove individuals with SNP missingness greater than three SDs from the mean
QC2	Sample Heterozygosity	Remove Individuals with heterozygoisity greater than three SDs from the mean
QC3	Duplicate Individuals	Remove individuals with greater than 85% IBD to another individual
QC4	SNP Missingness	Remove SNPs absent in more than 3% of samples
QC5	SNP AF	Remove SNPs with a MAF below 0.01
QC6	SNP Hardy-Weinberg	Remove SNPs with HWE below 1e-6 in controls
QC7	Duplicate SNPs	Remove duplicate SNPs

Table 5.5: SNP dataset filtering

	GWA1			GWA2			GWA3			GWA4			LP		
	Cases	Controls	SNPs	Cases	Controls	SNPs	Cases	Controls	SNPs	Cases	Controls	SNPs	Cases	Controls	SNPs
Prefilter	221	216	521396	131	139	501516	323	3	629862	311	357	500399	269	136	1954769
QC1	216	214	521396	128	138	501516	321	3	629862	310	356	500399	269	133	1954769
QC2	215	211	521396	128	137	501516	321	3	629862	310	356	500399	265	133	1954769
QC3	215	211	521396	126	135	501516	315	3	629862	310	356	500399	265	133	1954769
QC4	215	211	521396	126	135	498841	315	3	625186	310	356	498347	265	133	1912362
QC5	215	211	512606	126	135	490935	315	3	609411	310	356	454551	265	133	1312555
QC6	215	211	512587	126	135	490883	315	3	609411	310	356	454551	265	133	1312540
QC7	215	211	512587	126	135	490883	315	3	609411	310	356	437524	265	133	1307016

Statistical analysis and plotting

Unless otherwise stated statistical analysis was conducted in R v3.6.1 (Team 2014).

Results

FTD

51 FTD patients underwent targeted DNA sequencing. Patients had a mean AOO of 64.8 (SD 8.25). 13 patients have subsequently developed ALS and the remainder are diagnosed with a variety of FTD subphenotypes (table 5.6).

Table 5.6: Phenotypes of FTD patients

Subphenotype	Count
FTD-CBS	3
FTD-MND	15
FTD-PSP	3
PNFA	11
PNFA-CBS	1
SD	2
bvFTD	16
Total	**51**

Samples were sequenced to an average target DOC of 45.85 (SD 31.41). 302 SNPs and 39 INDELs were initially observed in FTD patients and controls. After applying the novel variant filtration pipeline developed in Chapter 4, this was reduced to 25 putatively pathogenic variants (table 5.7, table 5.8). 10 of these variants have been previously reported in the literature with all classified as VUS.

Table 5.7: Variant filtering in FTD samples

Filter Description	SNVs Remaining	INDELs Remaining
Initial variants	302	39
Variants calling QC	285	31
Present in cases	191	21
Absent in controls *	58	4
Benign in journALS	47	4
Functional filter	26	2
gnomAD filter	26	2
ProjectMinE filter	23	2
Putative pathogenic variants	23	2

* If homozygous in any case then not homozygous in any control, else if heterozygous in all cases then absent in all controls

A single bvFTD patient and two FTD-MND patients were found to harbour the *C9orf72* RE, resulting in a frequency of 5.9% (95% CI 2.0-15.9%) in the entire cohort and 2.6% (95% CI: 0.5-13.5%) in "pure" FTD.

Table 5.8: Variants observed in FTD patients

Identifier	HGVS	Transcript	Impact	Carrier Subphenotype	PM Case AF	PM Control AF	In Literature	gnomAD AF	in silico Prediction	Case count
2:202575717:T:C	*ALS2*:c.4119A>G(p.[I1373M])	ENST00000264276	missense_variant	PNFA,FTD-PSP	5.5e-3	5.5e-3	No	2.6e-3	NA	2
12:112037104:G:A	*ATXN2*:c.215C>T(p.[S72F])	ENST00000377617	missense_variant	FTD-MND	2.8e-3	1.9e-3	No	1.2e-4	P	1
12:112037095:T:C	*ATXN2*:c.224A>G(p.[D75G])	ENST00000377617	missense_variant	FTD-MND	1.1e-4	0	No	5.4e-5	B	1
21:45750089:C:T	*C21orf2*:c.1120G>A(p.[A374T])	ENST00000397956	missense_variant	FTD-MND	2.6e-3	8.2e-4	No	5.1e-4	B	1
21:45753117:C:A	*C21orf2*:c.172G>T(p.[V58L])	ENST00000397956	missense_variant	PNFA	2.1e-2	1.2e-2	Yes	7.9e-3	NA	1
21:45750608:A:C	*C21orf2*:c.737T>G(p.[V246G])	ENST00000397956	missense_variant	bvFTD	1.1e-4	5.5e-4	No	1.7e-4	B	1
2:74593663:G:C	*DCTN1*:c.2551C>G(p.[L851V])	ENST00000361874	missense_variant	bvFTD	3.4e-4	0	No	1.0e-4	NA	1
2:74597797:G:C	*DCTN1*:c.999C>G(p.[D333E])	ENST00000361874	missense_variant	bvFTD	NA	NA	No	3.3e-4	B	1
6:110037748:C:T	*FIG4*:c.266C>T(p.[A89V])	ENST00000230124	missense_variant	FTD-PSP	1.1e-4	0	No	2.2e-5	NA	1
5:138661006:G:C	*MATR3*:c.2170G>C(p.[E724Q])	ENST00000394800	missense_variant	PNFA	NA	NA	No	2.2e-5	P	1
22:29876522:G:C	*NEFH*:c.271G>C(p.[V91L])	ENST00000310624	missense_variant	PNFA	NA	NA	No	NA	B	1
10:13168035:AAG:A	*OPTN*:c.1241_1242delAG(p.[E414fs])	ENST00000263036	frameshift_variant	FTD-PSP	1.1e-4	0	No	4.4e-6	NA	1
10:13166053:A:T	*OPTN*:c.941A>T(p.[Q314L])	ENST00000263036	missense_variant	bvFTD	1.1e-4	0	Yes	1.5e-4	P	1
9:135206522:T:G	*SETX*:c.1015A>C(p.[K339Q])	ENST00000372169	missense_variant	FTD-CBS	NA	NA	No	2.4e-5	NA	1
9:135202325:A:C	*SETX*:c.4660T>G(p.[C1554G])	ENST00000372169	missense_variant	bvFTD	3.2e-3	2.2e-3	Yes	5.7e-3	NA	1
9:34635679:G:A	*SIGMAR1*:c.622C>T(p.[R208W])	ENST00000277010	missense_variant	bvFTD	2.1e-3	1.4e-3	Yes	8.1e-3	NA	1
2:32289031:C:T	*SPAST*:c.131C>T(p.[S44L])	ENST00000315285	missense_variant	PNFA	1.2e-3	9.6e-3	Yes	4.3e-3	P	1
15:44856827:G:A	*SPG11*:c.7069C>T(p.[L2357F])	ENST00000261866	missense_variant	bvFTD	1.8e-3	1.6e-3	No	1.2e-3	NA	1
15:44944406:G:C	*SPG11*:c.928C>G(p.[P310A])	ENST00000261866	missense_variant	bvFTD	NA	NA	No	4.7e-5	NA	1
5:179263447:C:T	*SQSTM1*:c.1177C>T(p.[R393W])	ENST00000389805	missense_variant	FTD-MND	NA	NA	No	5.7e-5	P	1
5:179250905:G:A	*SQSTM1*:c.349G>A(p.[A117T])	ENST00000389805	missense_variant	bvFTD	NA	NA	No	NA	B	1
17:34171623: TAGAAGTGGGGGCGGCTATGGTGGAGAC:T	*TAF15*:c.1332_1358del CGGCTATGGTGGAGACAGAAGTGGGGG(p.[G445_G453del])	ENST00000588240	disruptive_inframe_deletion	FTD-MND	NA	NA	No	2.9e-4	NA	1
1:11082347:G:A	*TARDBP*:c.881G>A(p.[G294E])	ENST00000240185	missense_variant	bvFTD	NA	NA	No	8.5e-6	P	1
12:64854098:A:G	*TBK1*:c.217A>G(p.[I73V])	ENST00000331710	missense_variant	FTD-MND	1.1e-4	0	Yes	4.4e-5	B	1
19:17768944:G:A	*UNC13A*:c.958C>T(p.[P320S])	ENST00000428389	missense_variant	bvFTD	1.1e-4	0	No	4.4e-6	B	1

All observed variants were heterozygous

Variants reported here are those that are retained after filtering

In addition to the *C9orf72* RE, 13 VUS are observed in genes identified in the journALS study of Chapter 2 to harbour pathogenic or likely pathogenic ALS or FTD variants. There is little reason to suspect the pathogenicity of the observed heterozygous missense variants in *ALS2*, *DCTN1*, *MATR3*, *SETX* or *SIGMAR1*. Disruptive homozygous variants in *ALS2* are associated with early-onset ALS and there is little evidence supporting the role of heterozygous missense variants in this gene. *DCTN1*, *MATR3* and *SETX* have previously only been linked to ALS rather than FTD and rare missense variants are frequently observed in these genes. A homozygous *SIGMAR1* variant is associated with ALS in journALS, but there is little evidence supporting the role of heterozygous variants such as identified here.

A missense and frameshift variant are observed in *OPTN*. While frameshift *OPTN* variants have only been linked to ALS in homozygosity, certain heterozygous missense variants have previously been shown to cause ALS. *OPTN*:c.941A>T(p.[Q314L]), observed here, is a VUS that has previously been observed in 12 cases of ALS and never previously in FTD. It is a rare variant that is absent in Project MinE controls; however, it is at a similar frequency in Project MinE cases and in the gnomAD non-neuro subset. Its pathogenicity is uncertain.

An observed *TBK1* missense variant (*TBK1*:c.217A>G(p.[I73V])) has previously been reported in two FTD patients (van der Zee *et al.* 2017); however only disruptive *TBK1* variants have definitively been shown to be pathogenic and the potential role of missense variants remains uncertain.

TARDBP:c.881G>A(p.[G294E]) is observed here in a single FTD patient. This variant has not previously been reported in the ALS and FTD literature, is absent in Project MinE and rare in gnomAD. While this variant lacks sufficient evidence to be classified as pathogenic, there is evidence supporting its pathogenicity. It is in the C-terminal domain of *TARDBP*, where pathogenic variants aggregate. Also, there are two previously reported changes of the same amino acid in patients with both ALS and FTD in the journALS database. *TARDBP*:c.881G>C(p.[G294A]) is reported in 3 patients and *TARDBP*:c.881G>T(p.[G294V]) is reported 15 times.

Comparing the frequency of FTD variants in Ireland to other European and global cohorts (figure 5.2), it is notable that Irish FTD patients do not have any *MAPT* or *GRN* variants. Globally these genes account for 2-4% of patients if considering strictly pathogenic and likely pathogenic variants and almost 20% of patients when VUS variants are included.

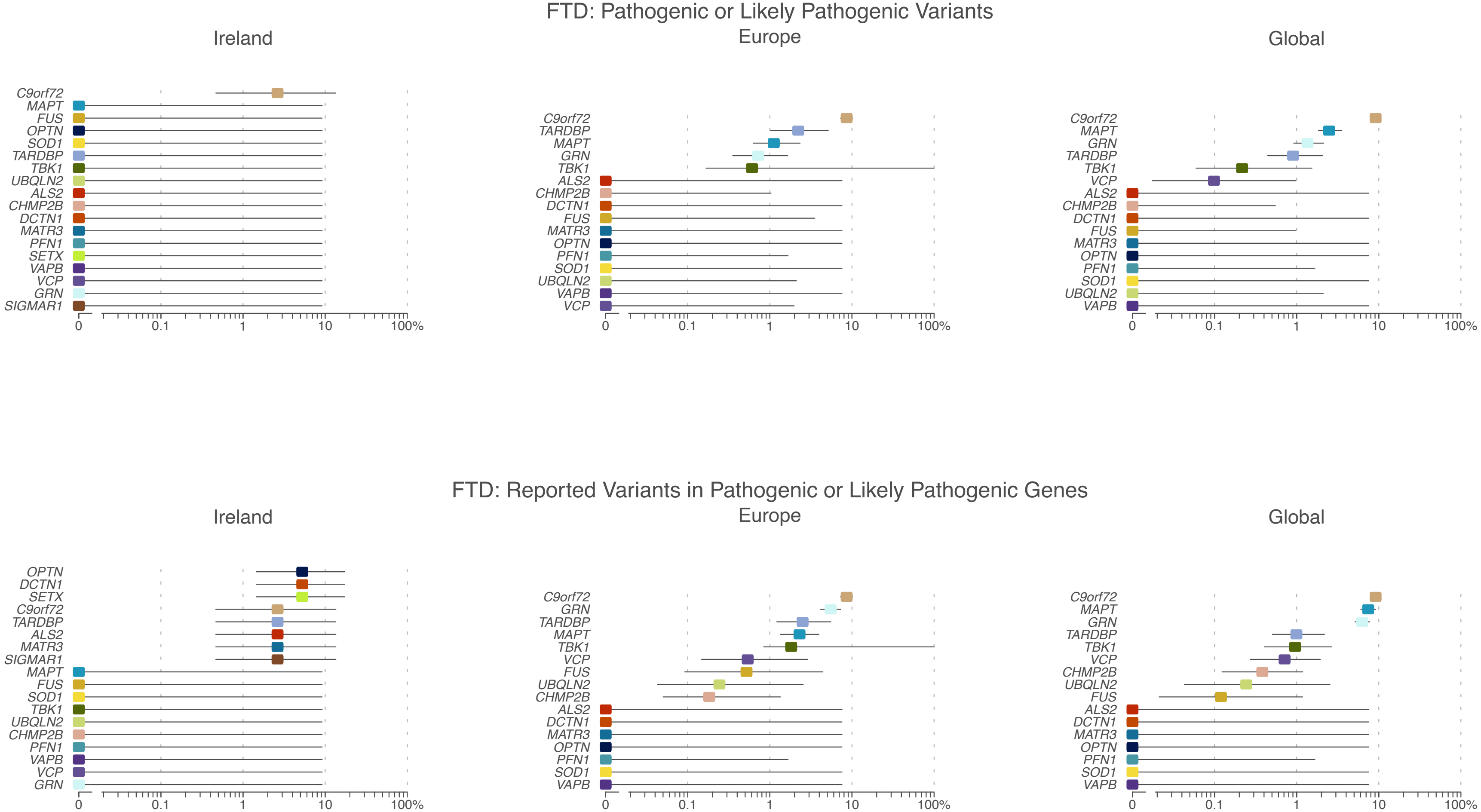


Figure 5.2: Proportion of FTD cases carrying genetic variants

The upper panel displays the proportion of FTD cases in Ireland, Europe and globally that carry a pathogenic or likely pathogenic variant. The lower panel shows the proportion of patients who carry a variant that is not benign or likely benign in the same set of genes. European and global proportions are calculated based on an estimated familial proportion of 40%. The displayed percentages are calculated for cases of pure FTD and exclude individuals and studies of FTD-ALS.

PLS

WES and *C9orf72* RE genotyping were performed for 44 PLS patients. Detailed phenotype information was available for 43 patients (table 5.9). Notably, while no patients were related, 9.8% of patients had a positive family history for MND, further calling into question the notion that PLS is a 'sporadic' disorder.

Table 5.9: Summary of PLS patients included in this study

Age of Onset (years)	52.1 (95% CI: 49.3-54.9)
Alive	91.70%
Disease duration (months)	186 (95% CI: 161-212)
Sex (male)	53.7% (95% CI: 38.7-67.9%)
Site of Onset (Spinal/Bulbar/Other)	83.7% / 16.7% / 2.4%
Family History (familial)	9.8% (95% CI: 3.9-22.5%)
Concomittant FTD	0% (95% CI: 0-8.6%)

Note: disease duration is for patient who are currently alive. For decesead patients the time to death was 100.7 months (95% CI: 90-131.4)

All PLS patients were found to be negative for the *C9orf72* RE. Combining the current research with previous PLS *C9orf72* studies (table 5.2), a revised PLS *C9orf72* carrier frequency of 1.8% (95% CI: 0.8-4.1%) is observed, with the repeat expansion present in 5 out of 280 patients.

SKAT analysis was performed to identify if PLS patients carry an excess of rare SNVs and INDELs in genes previously linked to HSP. A single SKAT test was performed treating all exons of 70 genes linked to HSP as a single unit. 437 functional variants with allele frequency below 5% were observed in cases and controls and 323 of these passed sequencing filters. No significant excess of variants was observed in PLS patients (p=0.38).

To examine whether pathogenic variants in ALS, FTD or HSP associated genes contribute to PLS pathogenesis in Ireland, these genes were screened through a novel pipeline which has been developed to prioritise putatively pathogenic variants. 3,225 variants were observed in cases and controls (table 5.10). After filtering variants which fail sequencing filters, were present in controls, non-functional, benign in journALS or at a higher frequency in Project MinE cases than controls, 45 SNVs and 4 INDELs remained (table 5.11). Following an interrogation of HGMD and ClinVar, 7 variants were identified as being previously reported in the literature.

Table 5.10: Variant filtering in PLS samples

Filter Description	SNVs Remaining	INDELs Remaining
Initial variants	2731	494
Variants calling QC	2610	470
Present in cases	2068	420
Absent in controls *	249	39
Benign in journALS	223	35
Functional filter	50	4
gnomAD filter	49	4
ProjectMinE filter	45	4
Putative pathogenic variants	45	4

* If homozygous in any case then not homozygous in any control, else if heterozygous in all cases then absent in all controls

Previously reported variants in PLS patients

AP4E1:c.613C>A(p.[H205N]) is identified here in 2 PLS patients. A North American individual affected with a persistent stutter was previously identified to carry this variant but lacked further phenotypic or familial information (Raza *et al.* 2015). *In silico* prediction tools form a consensus agreement that this variant is likely to affect protein function. There is insufficient evidence to classify this variant as either pathogenic or benign; however the frequency of the variant in gnomAD (discussed below), is suggestive of a benign variant.

WES sequencing previously identified *MARS1*:c.403T>C(p.[F135L]) in a patient with a fatal case of H1N1 influenza; however, there was little other evidence supporting the pathogenicity of the variant in that instance (Schulert *et al.* 2016).

A heterozygous *PNPLA6*:c.2389G>A(p.[V797M]) was previously reported in a compound heterozygote HSP patient who also carried a second heterozygous *PNPLA6* variant (c.3585C>G[D1195Q]) (D'Amore *et al.* 2018). The PLS patient here did not carry any further *PNPLA6* variants. *In silico* tools form a consensus prediction that this is likely to be a benign variant that will not significantly effect protein structure or function. There is insufficient evidence to ascertain the pathogenic effect of this variant in heterozygosity.

PSEN2:c.811C>T(p.[L271F]) has been previously reported as a possible risk factor in patients with AZD. Blauwendraat *et al.* (2016) and Sala Frigerio *et al.* (2015) each identified the variant in a single sporadic AZD patient with no further supporting evidence for either patient. The PLS patient in this study experienced PLS onset at age 56 and had no reported dementia by age 69. The pathogenicity of this variant is uncertain.

Table 5.11: Putative variants in PLS patients (1/2)

Identifier	HGVS	Transcript	Impact	PM Case AF	PM Control AF	In Literature	gnomAD AF	in silico Prediction	Case count
1:110168011:G:A	*AMPD2*:c.340G>A(p.[D114N])	ENST00000256578	missense	NA	NA	N	8.0e-6	NA	1
15:51221276:C:A	*AP4E1*:c.613C>A(p.[H205N])	ENST00000261842	missense	1.0e-3	2.7e-4	Y	3.6e-4	P	2
14:31553964:G:GT	*AP4S1*:c.367-3dupT	ENST00000216366	splice acceptor	9.2e-4	8.2e-4	N	1.3e-3	NA	2
7:4830747:G:A	*AP5Z1*:c.2155G>A(p.[A719T])	ENST00000348624	missense	NA	NA	N	1.4e-5	NA	1
1:1455523:A:T	*ATAD3A*:c.661A>T(p.[T221S])	ENST00000378755	missense	NA	NA	N	NA	P	1
9:95481748:A:T	*BICD2*:c.1179T>A(p.[N393K])	ENST00000356884	missense	8.0e-4	5.5e-4	N	7.3e-4	P	1
21:45751726:G:A	*C21orf2*:c.545C>T(p.[T182I])	ENST00000397956	missense	1.1e-3	5.5e-4	N	3.5e-4	B	1
19:50212047:A:AC	*CPT1C*:c.1521dupC(p.[T508fs])	ENST00000323446	frameshift	NA	NA	N	NA	NA	1
4:108866485:T:C	*CYP2U1*:c.850T>C(p.[F284L])	ENST00000332884	missense	1.0e-3	5.5e-4	N	1.8e-3	NA	1
4:108868556:G:T	*CYP2U1*:c.1151G>T(p.[R384I])	ENST00000332884	missense	4.8e-3	3.6e-3	N	2.5e-3	P	1
12:109294181:T:C	*DAO*:c.914T>C(p.[V305A])	ENST00000228476	missense	NA	NA	N	1.2e-5	P	1
8:27957364:G:A	*ELP3*:c.139G>A(p.[A47T])	ENST00000256398	missense	NA	NA	N	7.6e-5	NA	2
10:101911898:C:T	*ERLIN1*:c.1037G>A(p.[S346N])	ENST00000407654	missense	NA	NA	N	6.0e-5	B	1
8:37602227:C:T	*ERLIN2*:c.437C>T(p.[S146F])	ENST00000523887	missense	1.2e-4	0	N	NA	B	1
6:5613405:C:T	*FARS2*:c.1069C>T(p.[L357F])	ENST00000274680	missense	1.2e-4	0	N	8.4e-5	B	1
11:63883777:C:G	*FLRT1*:c.38C>G(p.[T13R])	ENST00000246841	missense	1.8e-3	1.4e-3	N	8.8e-4	NA	1
11:63885451:G:T	*FLRT1*:c.1712G>T(p.[G571V])	ENST00000246841	missense	1.0e-3	5.5e-4	N	6.4e-4	NA	1
11:63885582:C:T	*FLRT1*:c.1843C>T(p.[R615C])	ENST00000246841	missense	1.1e-3	8.2e-4	N	1.3e-3	P	1
9:35737341:C:T	*GBA2*:c.2627G>A(p.[R876Q])	ENST00000545786	missense	NA	NA	N	2.6e-4	B	1
9:35740222:C:T	*GBA2*:c.1285G>A(p.[G429S])	ENST00000545786	missense	NA	NA	N	3.4e-3	NA	1
1:228353776:G:A	*IBA57*:c.259G>A(p.[G87R])	ENST00000366711	missense	NA	NA	N	1.6e-5	B	1
2:163144694:T:C	*IFIH1*:c.1046A>G(p.[K349R])	ENST00000263642	missense	1.7e-3	1.4e-3	N	3.1e-3	NA	1
2:163174589:G:A	*IFIH1*:c.229C>T(p.[R77W])	ENST00000263642	missense	1.0e-3	5.5e-4	N	7.0e-4	B	1
2:241689933:G:C	*KIF1A*:c.2890C>G(p.[P964A])	ENST00000498729	missense	NA	NA	N	NA	P	1
12:57975670:C:T	*KIF5A*:c.2927C>T(p.[T976I])	ENST00000455537	missense	NA	NA	N	2.7e-4	B	1
1:235827874:C:T	*LYST*:c.11086G>A(p.[V3696I])	ENST00000389793	missense	5.7e-4	2.7e-4	N	6.1e-4	B	1
12:57883330:T:C	*MARS1*:c.403T>C(p.[F135L])	ENST00000262027	missense	NA	NA	Y	1.2e-5	P	1
12:57884160:G:A	*MARS1*:c.661G>A(p.[E221K])	ENST00000262027	missense	3.4e-4	0	N	1.3e-4	P	1
5:138653337:G:A	*MATR3*:c.1235G>A(p.[R412K])	ENST00000394800	missense	NA	NA	N	NA	P	1
2:197710636:T:C	*PGAP1*:c.2256A>G(p.[I752M])	ENST00000354764	missense	1.2e-4	0	N	2.0e-5	B	1
19:7600891:A:G	*PNPLA6*:c.244A>G(p.[R82G])	ENST00000414982	missense	1.2e-4	0	N	7.0e-5	NA	1
19:7606451:C:T	*PNPLA6*:c.1076C>T(p.[T359I])	ENST00000414982	missense	1.2e-4	0	N	6.8e-5	NA	1
19:7618859:G:A	*PNPLA6*:c.2389G>A(p.[V797M])	ENST00000414982	missense	2.7e-3	2.7e-4	Y	1.8e-3	B	1
19:7626428:G:A	*PNPLA6*:c.4108G>A(p.[G1370S])	ENST00000414982	missense	9.3e-3	6.3e-3	N	5.1e-3	NA	1
19:7623736:C:T	*PNPLA6*:c.3428C>T(p.[A1143V])	ENST00000414982	missense	NA	NA	N	8.3e-6	P	1
1:227076675:C:T	*PSEN2*:c.811C>T(p.[L271F])	ENST00000366782	missense	1.2e-4	0	N	1.6e-5	P	1
2:86444181:G:A	*REEP1*:c.413C>T(p.[S138L])	ENST00000541910	missense	NA	NA	N	2.5e-5	P	1
9:135204703:G:C	*SETX*:c.2282C>G(p.[S761W])	ENST00000372169	missense	NA	NA	N	8.0e-6	NA	1
17:26726628:G:A	*SLC46A1*:c.307C>T(p.[H103Y])	ENST00000582735	missense	NA	NA	N	8.1e-6	NA	1
2:32370064:G:A	*SPAST*:c.1675G>A(p.[G559S])	ENST00000315285	missense	NA	NA	N	4.0e-6	P	1
16:89616965:C:G	*SPG7*:c.1727C>G(p.[S576W])	ENST00000268704	missense	NA	NA	Y	8.0e-6	P	1
5:179249949:ACAAT:A	*SQSTM1*:c.268_271delAATC(p.[N90fs])	ENST00000504627	frameshift	1.2e-4	0	N	2.8e-5	NA	1

Table 5.11: Putative variants in PLS patients (2/2)

Identifier	HGVS	Transcript	Impact	PM Case AF	PM Control AF	In Literature	gnomAD AF	in silico Prediction	Case count
17:34171623: TAGAAGTGGGGGCGGCTATGGTGGAGAC:T	*TAF15*:c.1332_1358del CGGCTATGGTGGAGACAGAAGTGGGGG(p.[G445_G453del])	ENST00000588240	disruptive inframe deletion	NA	NA	N	2.7e-4	NA	1
15:50774096:A:G	*USP8*:c.1637A>G(p.[K546R])	ENST00000307179	missense	NA	NA	N	1.4e-5	B	1
1:101198111:C:G	*VCAM1*:c.1663C>G(p.[L555V])	ENST00000294728	missense	3.8e-3	8.2e-4	Y	1.5e-3	B	2
8:17125873:G:T	*VPS37A*:c.307G>T(p.[V103L])	ENST00000324849	missense	NA	NA	N	NA	B	1
3:39108050:T:G	*WDR48*:c.280T>G(p.[S94A])	ENST00000302313	missense	4.0e-3	3.3e-3	Y	3.0e-3	P	1
14:68250088:C:T	*ZFYVE26*:c.3781G>A(p.[A1261T])	ENST00000347230	missense	NA	NA	N	1.6e-5	B	1
14:68272021:C:A	*ZFYVE26*:c.1184G>T(p.[G395V])	ENST00000347230	missense	3.4e-3	2.7e-4	N	3.4e-3	B	2

SPG7:c.1727C>G(p.[S576W]) is observed here in a single PLS patient. A patient with a family history of autosomal recessive HSP was previously reported to be a compound heterozygote, carrying both S576W and *SPG7*:c.1529C>T(p.[A510V]) (Kumar *et al.* 2013; Wali *et al.* 2020). Compound heterozygosity, particularly with the A510V variant is an established method of pathogenicity in *SPG7* (Kumar *et al.* 2013). There is insufficient evidence to ascertain the pathogenic effect of this variant in heterozygosity.

VCAM1:c.1663C>G(p.[L555V]) has previously been reported in an Irish patient with atherosclerosis (Parra *et al.* 1992; Schmitz *et al.* 2013). No other familial or phenotypic information is available for the patient and it remains classified as VUS.

WDR48:c.280T>G(p.[S94A]) was previously observed in a HSP patient with an autosomal recessive family history; however, segregation of the variant with disease could not be confirmed (Morais *et al.* 2017).

In summary, seven previously reported variants are observed in heterozygosity in this PLS cohort. No variants are definitively shown to be pathogenic variants associated with ALS, FTD or HSP; however there are three variants of interest. *PNPLA6*:c.2389G>A(p.[V797M]), *SPG7*:c.1727C>G(p.[S576W]) and *WDR48*:c.280T>G(p.[S94A]) have all previously been observed in either compound heterozygosity or homozygosity in HSP patients. Although none of these variants have been confirmed to segregate with disease, all match the expected pattern of inheritance for their respective genes. The effect of these variants in heterozygosity remains uncertain; however, they are variants of interest.

Variants observed in multiple PLS patients

Of the 49 putatively pathogenic variants observed in our cohort, 5 are observed in two patients. The 5 variants all have the maximum genotype quality score (99) and pairs of carriers show relatedness that is in line with the background rate in the population (2% (95% CI: 0-4.2%)). None of the ten individuals have a reported familial history for MND.

AP4E1:c.613C>A(p.[H205N]) is observed in two PLS patients and has been discussed previously as a variant that is present in the literature. This variant is present in 0.04% of individuals in the gnomAD non-neuro subset and this rises to 0.07% in individuals of European descent. There is no available estimate for the lifetime risk of developing PLS;

however, applying equation 2.1, where P(A) is the frequency of the allele in the general population (0.0007), P(A|D) is the frequency of the allele in patients (0.045) and P(D|A) is the assumed penetrance of the variant (1); this results in an estimated lifetime risk for PLS of 1/64 (95% CI: 1/18-1/216). This far exceeds reasonable estimates for the lifetime risk for this rare condition. The lifetime risk of ALS is approximately 1/400 (McGuire *et al.* 1996; Traynor *et al.* 1999; E. Beghi *et al.* 2007; Vázquez *et al.* 2008; Ryan, Heverin, *et al.* 2019). As ALS is both more common than PLS and is associated with a higher mortality rate, it should be expected to have a higher lifetime risk. It is possible that this variant is at a higher frequency in the Irish population than the 0.07% observed in gnomAD, but not at a high enough frequency to appear in our control cohort of 136 individuals.

ELP3:c.139G>A(p.[A47T]) is present in 2 PLS patients. This variant is not observed in Project MinE and has an AF of 7.6x10^{-5} in gnomAD, with all carriers being of European descent. Applying the same criteria as above this provides a PLS lifetime risk estimate of 1/599 (95% CI: 1/165-1/1991). *In silico* tools do not form a consensus as to the pathogenicity of this variant. While it is possible this variant is at a higher population in Ireland than elsewhere in Europe, it cannot be excluded as a variant of interest that may be associated with PLS pathogenesis.

AP4S1:c.367-3dupT , *VCAM1*:c.1663C>G(p.[L555V]) and *ZFYVE26*:c.1184G>T(p.[G395V]) are each present in 2 PLS patients. All are at similar frequencies in Project MinE cases and controls and are present in more than 0.1% of individuals in gnomAD, which is too high a frequency for these to be highly penetrant pathogenic PLS variants. *In silico* tools predict all three variants to be benign. These variants are likely to represent benign variants that are possibly at higher frequency in the Irish population than elsewhere in the world.

In summary, for the 5 variants which are observed in more than one PLS patient, 4 variants appear at too high a frequency in population datasets to be pathogenic variants and the final variant is a variant of uncertain significance that cannot be excluded as a variant of interest. However, even for the four variants that are unlikely to be pathogenic, it is still beneficial to catalogue the observation of these variants. As more PLS cohorts are studied in future years and the underlying genetics are further elucidated, it may transpire that these variants are modifiers of disease.

Other PLS variant of interest

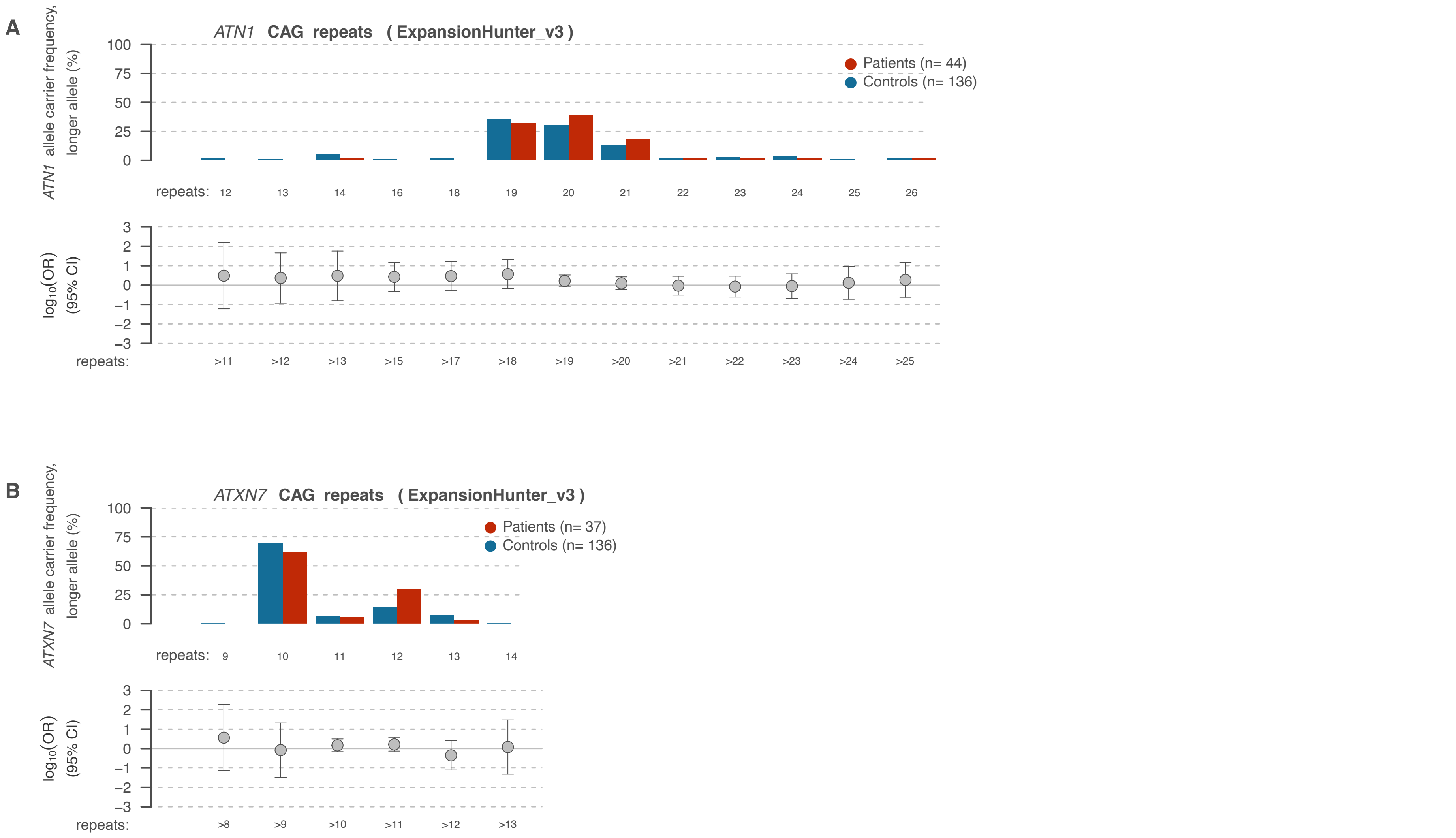
SPAST:c.1675G>A(p.[G559S]) is observed in a single PLS patient. It is rare in gnomAD (AF: $4.0x10^{-6}$) and is predicted to be pathogenic by *in silico* tools, reflecting the conserved nature of this amino acid residue. This variant has not previously been reported in the literature; however, two variants at the same amino acid have previously been observed in three HSP families.

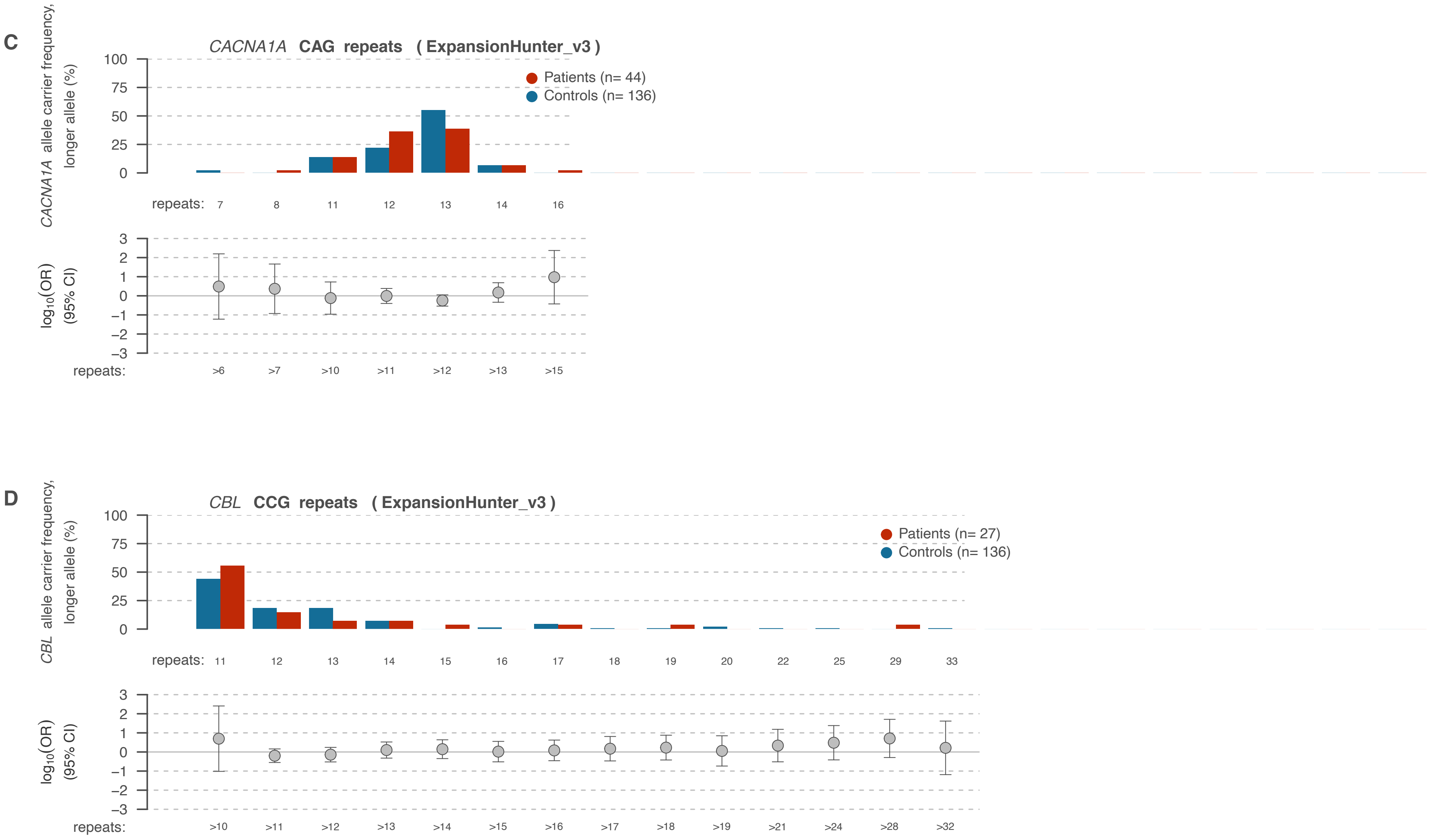
Nanetti *et al.* (2012) reported two related HSP patients with autosomal dominant inheritance and onset in their 40s and 50s who both carry a *SPAST*:c.1675G>C(p.[G559R]) variant, both patients had lower limb stiffness with no upper limb stiffness. Hentati *et al.* (2000) reported a pedigree with age of onset between 38 and 42 carrying *SPAST*:c.1676G>A(p.[G559D]). The patients are described as pure HSP and no reference to upper limb symptoms is noted. The segregation of the variant in the pedigree is not described. McCorquodale *et al.* (2011) report another heterozygous c.1676G>A(p.[G559D]) variant in a second family with unspecified segregation. The family is again described as pure HSP with no reference to upper limb symptoms, with mean age of onset of 40.

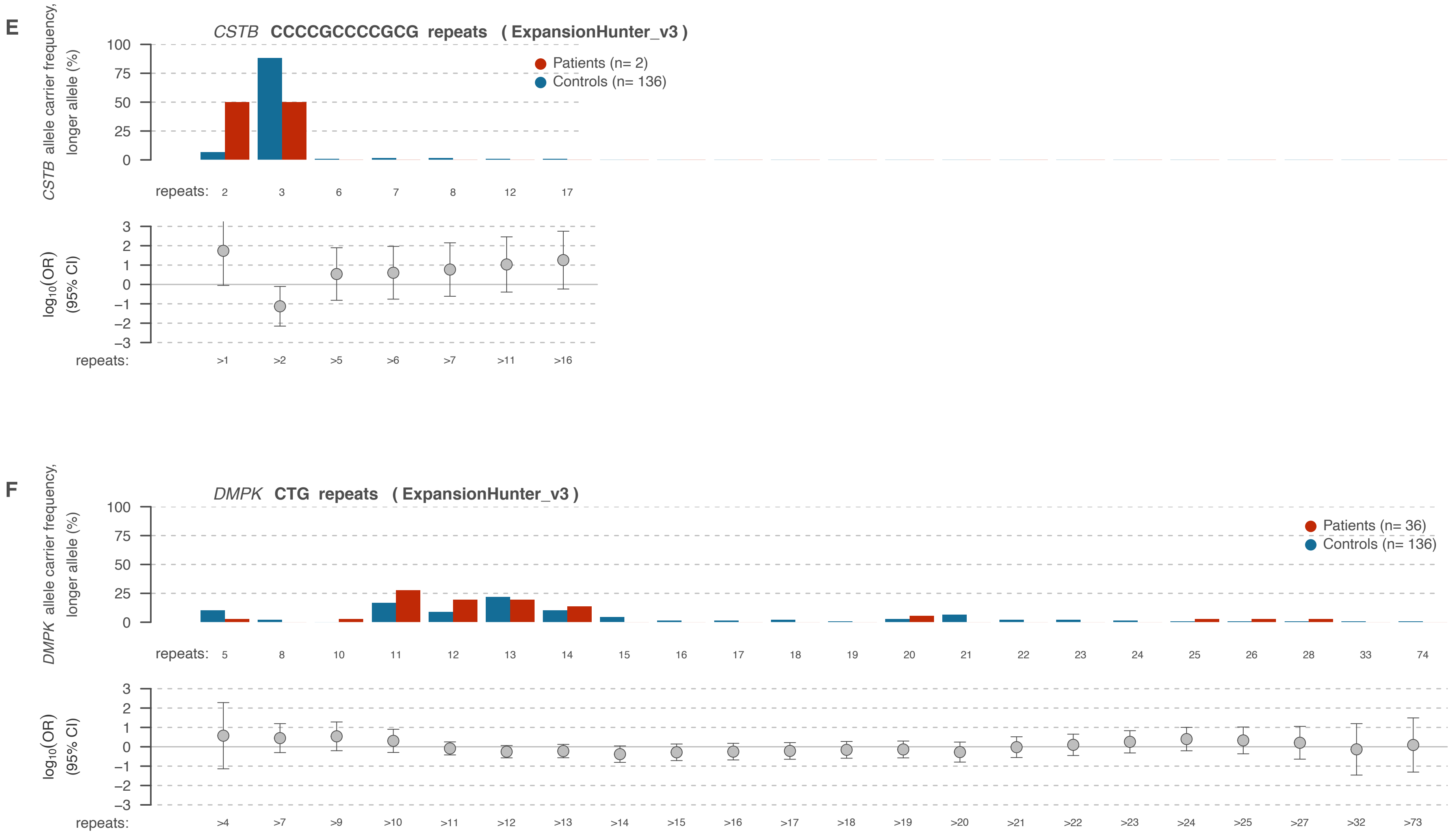
The PLS patient in this study had lower limb onset at age 62 and survived for 72 months, which indicates fast decline for a PLS patient. The patient was measured on the ALS functional rating scale (ALS-FRS) in four clinic visits up to 46 months from symptom onset. Over this period the patient did not exhibit bulbar symptoms but did show a decline in upper limb fine motor skills. Evidence indicates that missense changes in this amino acid are associated with UMN pathogenesis, although due to the clinical overlap between adult onset PLS and HSP it is unclear whether the patient reported here had a different aetiology to previously reported HSP patients with variants in the same amino acid.

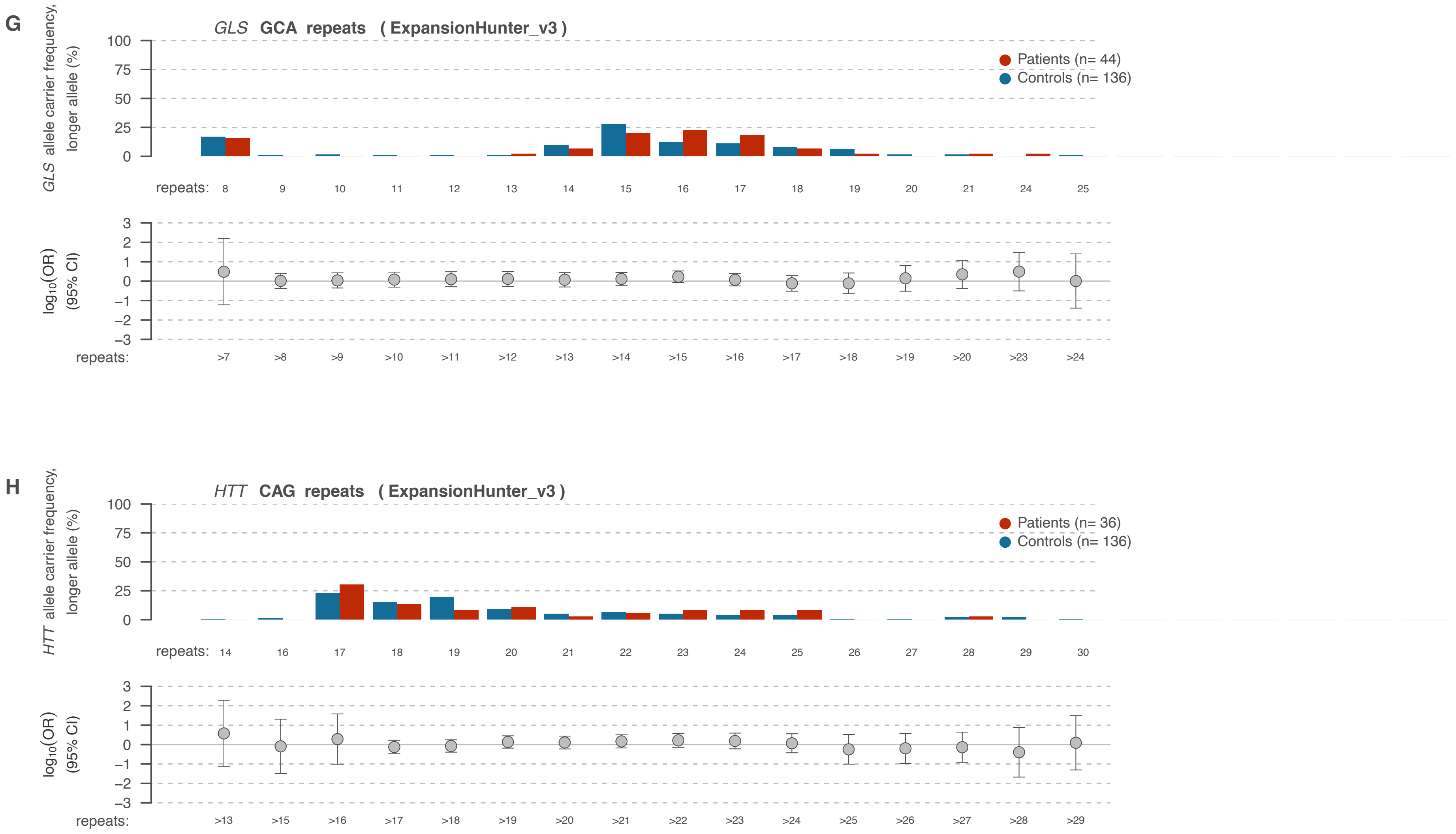
Repeat expansions in PLS

12 exonic STR loci are genotyped in 44 PLS patients and 136 controls using ExpansionHunter v3 (figure 5.3). No significantly enriched expansions are observed.

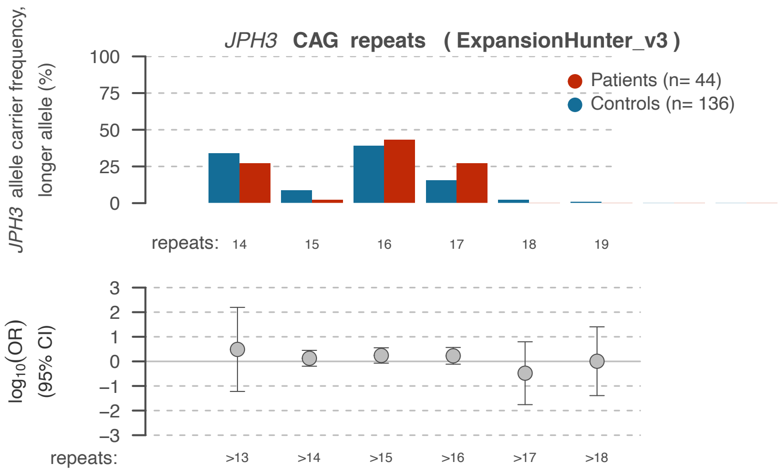

A
ATN1 CAG repeats (ExpansionHunter_v3)
ATN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 44)
● Controls (n= 136)
repeats: 12 13 14 16 18 19 20 21 22 23 24 25 26
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >11 >12 >13 >15 >17 >18 >19 >20 >21 >22 >23 >24 >25
B
ATXN7 CAG repeats (ExpansionHunter_v3)
ATXN7 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 37)
● Controls (n= 136)
repeats: 9 10 11 12 13 14
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >9 >10 >11 >12 >13


C
CACNA1A CAG repeats (ExpansionHunter_v3)
CACNA1A allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 44)
Controls (n= 136)
repeats: 7 8 11 12 13 14 16
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >7 >10 >11 >12 >13 >15
D
CBL CCG repeats (ExpansionHunter_v3)
CBL allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 27)
Controls (n= 136)
repeats: 11 12 13 14 15 16 17 18 19 20 22 25 29 33
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >21 >24 >28 >32

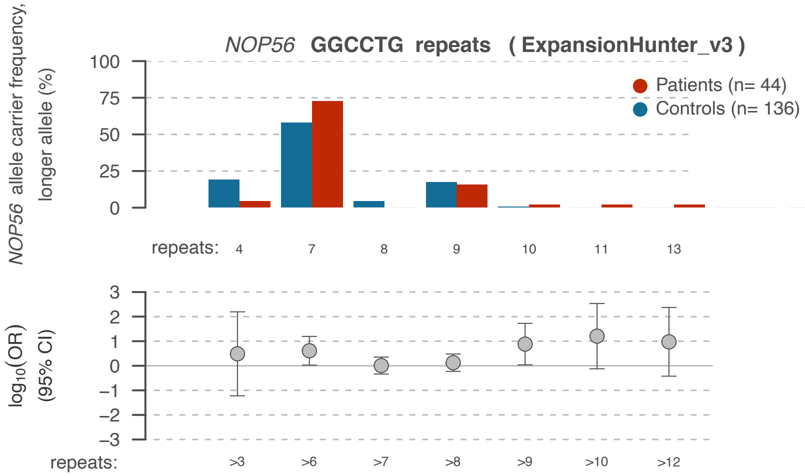

E
CSTB CCCCGCCCCGCG repeats (ExpansionHunter_v3)
CSTB allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 2)
● Controls (n= 136)
repeats: 2 3 6 7 8 12 17
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >1 >2 >5 >6 >7 >11 >16
F
DMPK CTG repeats (ExpansionHunter_v3)
DMPK allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 36)
● Controls (n= 136)
repeats: 5 8 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 33 74
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >4 >7 >9 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >27 >32 >73


G
GLS GCA repeats (ExpansionHunter_v3)
GLS allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 44)
Controls (n= 136)
repeats: 8 9 10 11 12 13 14 15 16 17 18 19 20 21 24 25
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >9 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >23 >24
H
HTT CAG repeats (ExpansionHunter_v3)
HTT allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 36)
Controls (n= 136)
repeats: 14 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >13 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >29

I


JPH3 CAG repeats (ExpansionHunter_v3)
JPH3 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 44)
● Controls (n= 136)
repeats: 14 15 16 17 18 19
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >13 >14 >15 >16 >17 >18

J


NOP56 GGCCTG repeats (ExpansionHunter_v3)
NOP56 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 44)
● Controls (n= 136)
repeats: 4 7 8 9 10 11 13
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >3 >6 >7 >8 >9 >10 >12

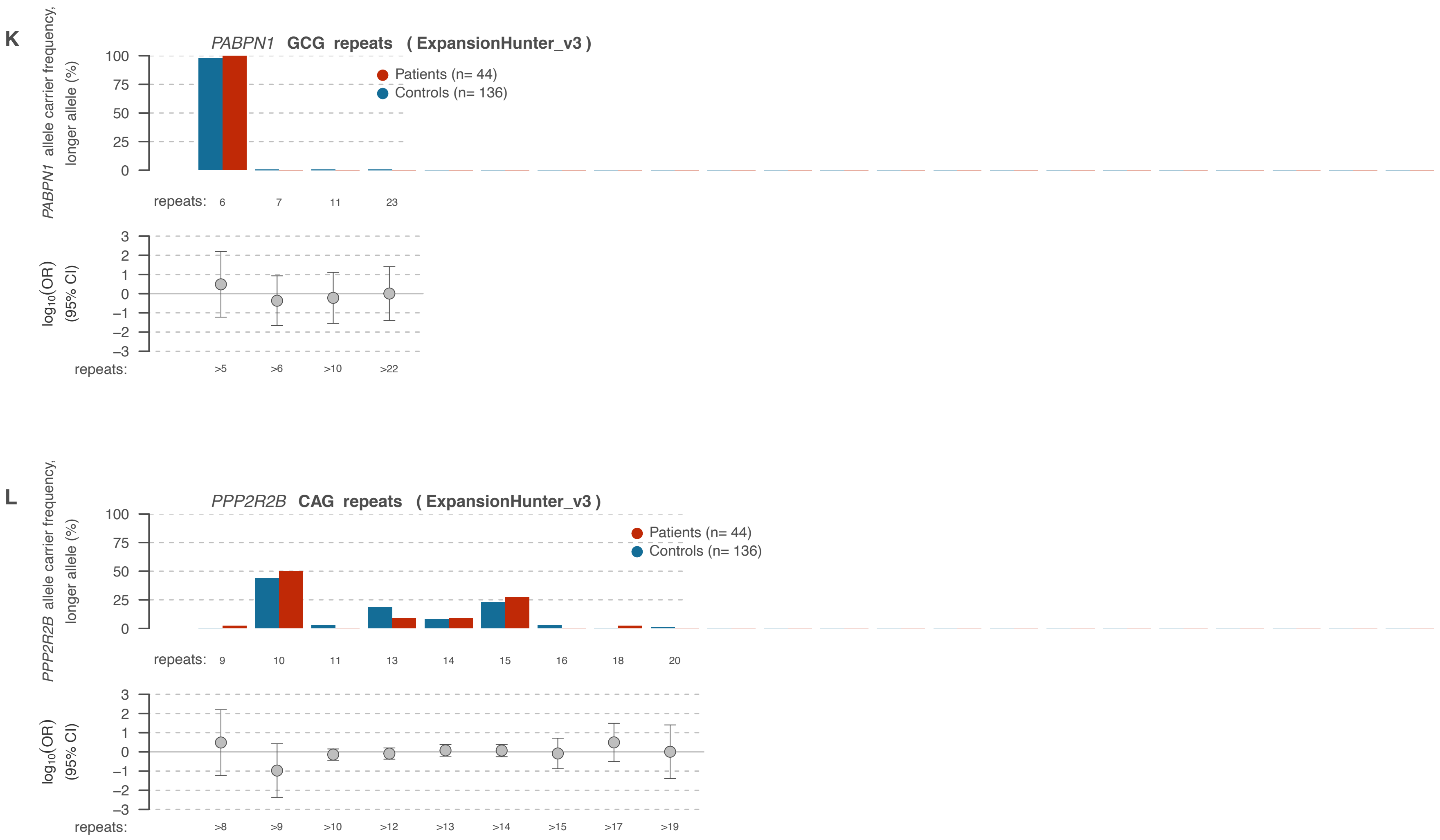


Figure 5.3: STR genotyping in PLS patients

12 exonic STR loci were genotyped in PLS patients using ExpansionHunter v3. No significant expansions are observed.

ALS

Including DNA sequencing and *C9orf72* RE genotyping, 1,549 ALS patients are included in this study. The phenotypes of these patients are summarised in table 5.12.

Table 5.12: Summary of Irish ALS patients included in this study

	Irish ALS Register	Sequencing	*C9orf72* Testing
Age of Onset (years)	63 (95%CI: 63-64)	62 (95% CI: 61-62)	62 (95% CI: 61-62)
Disease duration (months)	38 (95% CI: 37-40)	44 (95% CI: 41-48)	40 (95% CI: 38-42)
Sex (male)	57% (95% CI: 56-59)	59% (95% CI: 55-63)	60% (95% CI: 57-62)
Site of Onset (Spinal/Bulbar/Other)	69% / 33% / 8%	65% / 30% / 15%	65% / 28% /17%
Family History (familial)	11% (95% CI: 10-13)	16% (95% CI: 13-19)	15% (95% CI: 13-17)
Concomittant FTD	6% (95% CI: 5-6)	7% (95% CI: 6-10)	7% (95% CI: 6-8)

"Sequencing" indicates the phenotypes of patients for whom targeted or whole genome DNA sequencing was available.
"*C9orf72* testing" indicates the phenotypes of patients for whom C9orf72 RE genotyping was available.
This overall phenotypes for the Irish ALS register are for comparitive purposes

Of the 1,526 patients tested for the *C9orf72* RE, 9.7% were found to be carriers of the expansion. 36% of patients with a positive family history for ALS are carriers. The DNA of 676 Irish ALS patients are screened here for other causative ALS variants. 120 of the initially observed 1,721 variants present in the combined dataset remain following the variant filtration process (table 5.13, table 5.14). 37 of these 120 variants are present in the literature. Individual variants will not be discussed in depth here as Irish ALS genetics have been described previously (Kenna, McLaughlin, Byrne, *et al.* 2013; McLaughlin, Kenna, Vajda, Heverin, *et al.* 2015; Byrne *et al.* 2012); however, similar to FTD, it is of note that the profile of ALS genetics in Ireland is distinct from the rest of the Europe by its absences (figure 5.4). With the exception of a single patient carrying *SOD1*:c.317C>T(p.[S106L]), no *SOD1* variants are observed in Ireland. Similarly a very low rate of *FUS*, *TARDBP* and *TBK1* variation relative to the rest of Europe are observed.

Table 5.13: Variant filtering in ALS samples

Filter Description	SNVs Remaining	INDELs Remaining
Initial variants	1481	240
Variants calling QC	1189	219
Present in cases	1052	189
Absent in controls *	472	67
Benign in journALS	414	66
Functional filter	109	18
gnomAD filter	109	18
ProjectMinE filter	102	18
Putative pathogenic variants	**102**	**18**

Table 5.14: Variants Observed in Irish ALS Patients (1/3)

				PM	PM		In	gnomAD	in silico	Het	Hom	Het	Hom
Identifier	HGVS	Transcript	Impact	Case AF	Control AF	ACMG	Literature	AF	Pred	Patients	Patients	Controls	Controls
2:202622313:G:T	ALS2:c.1283C>A(p.[T428N])	ENST00000264276	missense_variant	2.3e-4	0	VUS	No	3.0e-5	B	2	0	0	0
2:202603402:T:C	ALS2:c.2408A>G(p.[K803R])	ENST00000264276	missense_variant	NA	NA	VUS	Yes	1.2e-5	NA	1	0	0	0
2:202598013:T:C	ALS2:c.2566A>G(p.[T856A])	ENST00000264276	missense_variant	NA	NA	VUS	Yes	1.0e-5	B	1	0	0	0
2:202589070:G:T	ALS2:c.3460C>A(p.[Q1154K])	ENST00000264276	missense_variant	1.1e-4	0	VUS	No	NA	NA	1	0	0	0
2:202575717:T:C	ALS2:c.4119A>G(p.[I1373M])	ENST00000264276	missense_variant	5.5e-3	5.5e-3	VUS	Yes	2.6e-3	NA	5	0	0	0
2:202572650:C:T	ALS2:c.4345G>A(p.[E1449K])	ENST00000264276	missense_variant	1.1e-4	0	VUS	No	1.7e-5	B	1	0	0	0
2:202626232:G:C	ALS2:c.485C>G(p.[T162S])	ENST00000264276	missense_variant	1.1e-4	0	VUS	No	NA	P	1	0	0	0
12:112037107:G:A	ATXN2:c.212C>T(p.[P71L])	ENST00000377617	missense_variant	9.2e-4	8.2e-4	VUS	Yes	7.3e-5	NA	3	0	0	0
12:112037104:G:A	ATXN2:c.215C>T(p.[S72F])	ENST00000377617	missense_variant	2.8e3	1.9e-3	VUS	Yes	1.9e-4	P	3	0	0	0
12:112037095:T:C	ATXN2:c.224A>G(p.[D75G])	ENST00000377617	missense_variant	1.1e-4	0	VUS	No	5.4e-5	B	1	0	0	0
12:111895043:C:T	ATXN2:c.3491G>A(p.[S1164N])	ENST00000377617	missense_variant	NA	NA	NA	No	1.5e-6	NA	1	0	0	0
12:112036879:C:G	ATXN2:c.440G>C(p.[C147S])	ENST00000377617	missense_variant	1.1e-4	0	VUS	No	9.6e-6	NA	1	0	0	0
21:45750127:C:T	C21orf2:c.1082G>A(p.[R361Q])	ENST00000397956	missense_variant	1.1e-4	0	VUS	No	4.9e-6	B	1	0	0	0
21:45750089:C:T	C21orf2:c.1120G>A(p.[A374T])	ENST00000397956	missense_variant	2.6e-3	8.2e-4	VUS	No	5.1e-4	B	4	0	0	0
21:45753117:C:A	C21orf2:c.172G>T(p.[V58L])	ENST00000397956	missense_variant	2.1e-2	1.2e-2	VUS	Yes	7.8e-3	NA	9	0	0	0
21:45753085:G:T	C21orf2:c.204C>A(p.[Y68*])	ENST00000397956	stop_gained	2.3e-4	0	VUS	No	NA	NA	1	0	0	0
21:45753071:C:G	C21orf2:c.218G>C(p.[R73P])	ENST00000397956	missense_variant	6.9e-4	2.7e-4	VUS	No	2.9e-4	NA	1	0	0	0
21:45751772:C:T	C21orf2:c.499G>A(p.[A167T])	ENST00000397956	missense_variant	1.1e-4	0	VUS	No	5.9e-4	B	1	0	0	0
21:45751726:G:A	C21orf2:c.545C>T(p.[T182I])	ENST00000397956	missense_variant	1.1e-3	5.5e-4	VUS	No	3.4e-4	B	1	0	0	0
21:45750608:A:C	C21orf2:c.737T>G(p.[V246G])	ENST00000397956	missense_variant	1.1e-3	5.5e-4	VUS	No	1.7e-4	B	3	0	0	0
3:87302948:A:C	CHMP2B:c.618A>C(p.[Q206H])	ENST00000263780	missense_variant	NA	NA	VUS	Yes	8.0e-6	P	1	0	0	0
2:74597660:C:T	DCTN1:c.1060G>A(p.[A354T])	ENST00000361874	missense_variant	NA	NA	NA	No	1.3e-5	P	1	0	0	0
2:74596006:G:C	DCTN1:c.1703C>G(p.[A568G])	ENST00000361874	missense_variant	1.1e-4	0	VUS	No	NA	P	1	0	0	0
2:74594037:A:G	DCTN1:c.2339T>C(p.[I780T])	ENST00000361874	missense_variant	2.3e-4	0	VUS	No	4.8e-5	P	1	0	0	0
2:74590135:G:A	DCTN1:c.3515C>T(p.[T1172I])	ENST00000361874	missense_variant	1.1e-4	0	VUS	No	NA	NA	1	0	0	0
8:27957364:G:A	ELP3:c.139G>A(p.[A47T])	ENST00000256398	missense_variant	NA	NA	VUS	No	7.9e-5	NA	3	0	0	0
8:27957431:G:T	ELP3:c.206G>T(p.[R69L])	ENST00000256398	missense_variant	NA	NA	VUS	Yes	4.8e-6	NA	1	0	0	0
8:27995228:AC:A	ELP3:c.923delC(p.[P308fs])	ENST00000256398	frameshift_variant	2.3e-4	0	VUS	No	7.2e-5	NA	1	0	0	0
2:212570063:C:T	ERBB4:c.1178G>A(p.[R393Q])	ENST00000342788	missense_variant	1.1e-4	0	VUS	No	2.9e-5	B	1	0	0	0
2:212530084:C:T	ERBB4:c.1835G>A(p.[R612Q])	ENST00000342788	missense_variant	NA	NA	VUS	No	3.9e-5	B	1	0	0	0
2:212251859:C:T	ERBB4:c.3200G>A(p.[R1067Q])	ENST00000342788	missense_variant	1.1e-4	0	VUS	No	5.9e-5	B	1	0	0	0
10:101914682:C:A	ERLIN1:c.760G>T(p.[A254S])	ENST00000407654	missense_variant	1.1e-4	0	VUS	No	1.3e-4	P	1	0	0	0
8:37599308:G:A	ERLIN2:c.208G>A(p.[E70K])	ENST00000276461	missense_variant	1.1e-4	0	VUS	No	NA	P	1	0	0	0
8:37601893:T:C	ERLIN2:c.257T>C(p.[F86S])	ENST00000276461	missense_variant	1.1e-4	0	VUS	No	NA	P	1	0	0	0
6:110081535:C:T	FIG4:c.1220C>T(p.[P407L])	ENST00000230124	missense_variant	1.1e-4	0	VUS	No	3.1e-5	P	1	0	0	0
6:110085177:C:T	FIG4:c.1426C>T(p.[R476C])	ENST00000230124	missense_variant	NA	NA	VUS	No	1.4e-5	P	1	0	0	0
6:110107592:G:A	FIG4:c.2036G>A(p.[R679Q])	ENST00000230124	missense_variant	1.1e-4	0	VUS	No	9.6e-6	B	1	0	0	0
6:110112668:G:C	FIG4:c.2270G>C(p.[S757T])	ENST00000230124	missense_variant	1.1e-4	0	VUS	No	NA	B	1	0	0	0
6:110113868:G:A	FIG4:c.2459+1G>A	ENST00000230124	splice_donor_variant	2.3e-4	0	VUS	No	1.5e-5	P	2	0	0	0
6:110117972:G:C	FIG4:c.2464G>C(p.[V822L])	ENST00000230124	missense_variant	1.1e-4	0	VUS	No	NA	B	1	0	0	0
6:110062665:G:A	FIG4:c.794G>A(p.[R265Q])	ENST00000230124	missense_variant	1.1e-4	0	VUS	No	1.4e-5	NA	1	0	0	0
16:31201719:C:T	FUS:c.1295C>T(p.[P432L])	ENST00000568685	missense_variant	3.4e-4	0	VUS	Yes	1.2e-4	P	1	0	0	0
16:31202740:G:T	FUS:c.1565G>T(p.[R522L])	ENST00000568685	missense_variant	2.3e-4	0	P	Yes	NA	P	1	0	0	0
16:31202752:C:T	FUS:c.1577C>T(p.[P526L])	ENST00000568685	missense_variant	2.3e-4	0	P	Yes	0	P	2	0	0	0
16:31196402:T:TGGC	FUS:c.684_686dupCGG(p.[G229dup])	ENST00000568685	disruptive_inframe_insertion	8.0e-4	0	VUS	Yes	2.5e-3	NA	2	0	0	0
12:57965910:A:G	KIF5A:c.1429A>G(p.[N477D])	ENST00000455537	missense_variant	1.1e-4	0	VUS	No	NA	P	1	0	0	0
12:57975281:A:G	KIF5A:c.2839A>G(p.[T947A])	ENST00000455537	missense_variant	1.1e-4	1.1e-3	VUS	No	1.0e-3	B	1	0	0	0
12:57975696:G:A	KIF5A:c.2953G>A(p.[G985S])	ENST00000455537	missense_variant	2.3e-4	0	VUS	No	5.3e-5	B	1	0	0	0
5:126154711:T:C	LMNB1:c.1037T>C(p.[M346T])	ENST00000261366	missense_variant	NA	NA	NA	No	NA	P	1	0	0	0
5:126158516:A:G	LMNB1:c.1430A>G(p.[D477G])	ENST00000261366	missense_variant	NA	NA	NA	No	NA	P	1	0	0	0

Table 5.14: Variants Observed in Irish ALS Patients (2/3)

				PM	PM		In	gnomAD	in silico	Het	Hom	Het	Hom
Identifier	HGVS	Transcript	Impact	Case AF	Control AF	ACMG	Literature	AF	Pred	Patients	Patients	Controls	Controls
5:126141379:G:C	LMNB1:c.633G>C(p.[M211I])	ENST00000261366	missense_variant	NA	NA	NA	No	1.9e-5	B	1	0	0	0
17:44073923:G:A	MAPT:c.1720G>A(p.[A574T])	ENST00000344290	missense_variant	1.1e-3	5.5e-4	VUS	Yes	1.0e-3	B	2	0	0	0
17:44055794:C:A	MAPT:c.361C>A(p.[H121N])	ENST00000344290	missense_variant	NA	NA	NA	No	NA	B	1	0	0	0
17:44060672:C:T	MAPT:c.502C>T(p.[R168C])	ENST00000344290	missense_variant	2.3e-4	0	VUS	No	6.8e-5	B	1	0	0	0
5:138658499:A:C	MATR3:c.1991A>C(p.[E664A])	ENST00000394800	missense_variant	5.7e-4	2.7e-4	VUS	Yes	3.0e-3	P	1	0	0	0
5:138651409:A:G	MATR3:c.998A>G(p.[N333S])	ENST00000394800	missense_variant	1.1e-4	0	VUS	No	NA	B	1	0	0	0
22:29885458:A:C	NEFH:c.1829A>C(p.[E610A])	ENST00000310624	missense_variant	NA	NA	NA	No	NA	B	1	0	0	0
22:29885638:T:A	NEFH:c.2009T>A(p.[V670E])	ENST00000310624	missense_variant	NA	NA	NA	No	7.4e-3	NA	17	1	7	0
22:29885644:C:A	NEFH:c.2015C>A(p.[A672E])	ENST00000310624	missense_variant	NA	NA	NA	No	6.2e-3	B	16	1	8	0
4:170482986:A:T	NEK1:c.1137T>A(p.[D379E])	ENST00000507142	missense_variant	5.7e-4	0	VUS	No	1.4e-4	B	2	0	0	0
4:170482633:T:C	NEK1:c.1264A>G(p.[K422E])	ENST00000507142	missense_variant	1.1e-4	0	VUS	No	NA	NA	1	0	0	0
4:170477185:T:C	NEK1:c.1328A>G(p.[Y443C])	ENST00000507142	missense_variant	1.1e-4	0	VUS	No	9.6e-6	P	1	0	0	0
4:170359294:T:C	NEK1:c.2704A>G(p.[S902G])	ENST00000507142	missense_variant	1.1e-4	0	VUS	No	NA	B	1	0	0	0
4:170523753:A:G	NEK1:c.29T>C(p.[I10T])	ENST00000507142	missense_variant	1.1e-4	0	VUS	No	1.3e-5	P	1	0	0	0
4:170345819:G:C	NEK1:c.3107C>G(p.[S1036*])	ENST00000507142	stop_gained	2.5e-3	0	VUS	Yes	8.8e-5	P	1	0	0	0
4:170345733:T:C	NEK1:c.3193A>G(p.[T1065A])	ENST00000507142	missense_variant	6.9e-4	2.7e-4	VUS	No	2.4e-4	NA	2	0	0	0
4:170327847:TA:T	NEK1:c.3273delT(p.[M1092fs])	ENST00000507142	frameshift_variant	1.1e-4	0	VUS	Yes	NA	NA	1	0	0	0
4:170315672:T:C	NEK1:c.3850A>G(p.[N1284D])	ENST00000507142	missense_variant	2.3e-4	0	VUS	No	8.7e-6	P	2	0	0	0
4:170498110:TG:T	NEK1:c.988delC(p.[H330fs])	ENST00000507142	frameshift_variant	1.1e-4	0	VUS	Yes	9.6e-6	NA	1	0	0	0
10:13167989:C:G	OPTN:c.1192C>G(p.[Q398E])	ENST00000263036	missense_variant	1.1e-4	0	VUS	Yes	9.6e-6	B	1	0	0	0
10:13178784:C:T	OPTN:c.1652C>T(p.[P551L])	ENST00000263036	missense_variant	1.1e-4	0	VUS	No	9.6e-6	P	1	0	0	0
1:8030955:C:G	PARK7:c.254C>G(p.[S85C])	ENST00000338639	missense_variant	NA	NA	NA	No	4.8e-6	P	1	0	0	0
12:49689305:T:C	PRPH:c.322T>C(p.[F108L])	ENST00000257860	missense_variant	1.1e-4	0	VUS	No	2.0e-5	P	1	0	0	0
12:49689399:G:A	PRPH:c.416G>A(p.[R139H])	ENST00000257860	missense_variant	1.1e-4	0	VUS	No	4.8e-6	P	1	0	0	0
14:73678599:G:A	PSEN1:c.1078G>A(p.[A360T])	ENST00000324501	missense_variant	3.4e-4	0	VUS	No	4.0e-5	P	1	0	0	0
14:73637721:T:G	PSEN1:c.304T>G(p.[S102A])	ENST00000324501	missense_variant	1.1e-4	0	VUS	No	NA	P	1	0	0	0
1:227073271:C:T	PSEN2:c.488C>T(p.[S163L])	ENST00000366782	missense_variant	9.2e-4	5.5e-4	VUS	Yes	7.0e-4	P	3	0	0	0
17:26708773:C:T	SARM1:c.920C>T(p.[A307V])	ENST00000457710	missense_variant	NA	NA	NA	No	5.0e-5	NA	1	0	0	0
9:135205781:G:A	SETX:c.1204C>T(p.[R402C])	ENST00000372169	missense_variant	1.1e-4	0	VUS	No	9.6e-6	P	1	0	0	0
9:135205694:G:A	SETX:c.1291C>T(p.[Q431*])	ENST00000372169	stop_gained	1.1e-4	0	VUS	No	NA	P	1	0	0	0
9:135204431:TATC:T	SETX:c.2551_2553delGAT(p.[D851del])	ENST00000372169	conservative_inframe_deletion	1.1e-4	0	VUS	No	NA	NA	1	0	0	0
9:135204143:G:T	SETX:c.2842C>A(p.[P948T])	ENST00000372169	missense_variant	NA	NA	VUS	Yes	4.7e-5	B	1	0	0	0
9:135203422:G:C	SETX:c.3563C>G(p.[T1188S])	ENST00000372169	missense_variant	1.1e-4	0	VUS	No	NA	B	1	0	0	0
9:135202897:C:T	SETX:c.4088G>A(p.[R1363Q])	ENST00000372169	missense_variant	1.1e-4	0	VUS	No	4.7e-5	B	1	0	0	0
9:135202226:G:A	SETX:c.4759C>T(p.[P1587S])	ENST00000372169	missense_variant	1.1e-4	0	VUS	No	8.7e-6	B	1	0	0	0
9:135173661:T:C	SETX:c.5587A>G(p.[T1863A])	ENST00000372169	missense_variant	NA	NA	VUS	Yes	4.9e-6	P	1	0	0	0
9:135172294:G:A	SETX:c.5929C>T(p.[L1977F])	ENST00000372169	missense_variant	1.1e-4	0	VUS	No	5.6e-6	P	1	0	0	0
9:135140316:A:C	SETX:c.7431T>G(p.[I2477M])	ENST00000372169	missense_variant	1.1e-4	0	VUS	No	NA	NA	1	0	0	0
9:135140221:G:A	SETX:c.7526C>T(p.[A2509V])	ENST00000372169	missense_variant	3.4e-4	0	VUS	No	1.2e-4	B	1	0	0	0
9:135210019:G:C	SETX:c.814C>G(p.[H272D])	ENST00000372169	missense_variant	NA	NA	VUS	Yes	4.8e-6	P	1	0	0	0
9:135210013:T:C	SETX:c.820A>G(p.[M274V])	ENST00000372169	missense_variant	1.1e-4	0	VUS	Yes	1.9e-5	P	1	0	0	0
21:33039648:C:T	SOD1:c.317C>T(p.[S106L])	ENST00000270142	missense_variant	1.1e-4	0	VUS	Yes	0	NA	1	0	0	0
2:32289031:C:T	SPAST:c.131C>T(p.[S44L])	ENST00000315285	missense_variant	1.2e-2	9.6e-3	VUS	Yes	4.3e-3	P	2	0	0	0
2:32340778:C:T	SPAST:c.878C>T(p.[P293L])	ENST00000315285	missense_variant	1.1e-4	0	VUS	No	4.8e-5	B	1	0	0	0
15:44921004:T:A	SPG11:c.1930A>T(p.[T644S])	ENST00000261866	missense_variant	1.1e-4	0	VUS	Yes	2.9e-5	P	1	0	0	0
15:44907696:C:G	SPG11:c.2903G>C(p.[G968A])	ENST00000261866	missense_variant	2.3e-4	0	VUS	No	2.2e-5	P	1	0	0	0
15:44892671:T:C	SPG11:c.3680A>G(p.[K1227R])	ENST00000261866	missense_variant	NA	NA	VUS	Yes	9.6e-6	B	1	0	0	0
15:44952678:T:C	SPG11:c.394A>G(p.[S132G])	ENST00000261866	missense_variant	NA	NA	VUS	Yes	NA	B	1	0	0	0
15:44888372:C:T	SPG11:c.4343G>A(p.[C1448Y])	ENST00000261866	missense_variant	NA	NA	VUS	Yes	4.8e-6	P	1	0	0	0

Table 5.14: Variants Observed in Irish ALS Patients (3/3)

				PM	PM		In	gnomAD	in silico	Het	Hom	Het	Hom
Identifier	**HGVS**	**Transcript**	**Impact**	**Case AF**	**Control AF**	**ACMG**	**Literature**	**AF**	**Pred**	**Patients**	**Patients**	**Controls**	**Controls**
15:44876119:CCT:C	SPG11:c.5757_5758delAG(p.[E1921fs])	ENST00000261866	frameshift_variant	1.1e-4	0	VUS	No	1.5e-5	NA	2	0	0	0
15:44861684:A:G	SPG11:c.6497T>C(p.[I2166T])	ENST00000261866	missense_variant	NA	NA	NA	No	8.2e-5	P	1	0	0	0
15:44856827:G:A	SPG11:c.7069C>T(p.[L2357F])	ENST00000261866	missense_variant	1.8e-3	1.6e-3	VUS	Yes	1.2e-3	NA	4	0	0	0
15:44855327:C:G	SPG11:c.7324G>C(p.[A2442P])	ENST00000261866	missense_variant	NA	NA	VUS	Yes	4.9e-5	NA	1	0	0	0
5:179263548:G:A	SQSTM1:c.1028G>A(p.[R343Q])	ENST00000510187	missense_variant	3.4e-4	0	VUS	No	2.8e-4	B	1	0	0	0
5:179263501:G:A	SQSTM1:c.1231G>A(p.[G411S])	ENST00000389805	missense_variant	1.1e-4	0	VUS	Yes	5.3e-5	P	1	0	0	0
5:179247940:G:A	SQSTM1:c.4G>A(p.[A2T])	ENST00000389805	missense_variant	1.1e-4	0	VUS	No	NA	NA	1	0	0	0
5:179248021:C:T	SQSTM1:c.85C>T(p.[P29S])	ENST00000389805	missense_variant	6.9e-4	0	VUS	No	9.2e-5	NA	1	0	0	0
5:179260200:C:T	SQSTM1:c.923C>T(p.[A308V])	ENST00000389805	missense_variant	1.1e-4	0	VUS	No	3.2e-5	B	1	0	0	0
17:34171551:C:CGGCTATGGTGGAGACAGAAGTGGGGGT	TAF15:c.1269_1295dupTGGGGGTGGCTATGGTGGAGACAGAAG(p.[S432_S433insGGGYGGDRS])	ENST00000588240	disruptive_inframe_insertion	NA	NA	NA	No	3.9e-4	NA	1	0	0	0
17:34147214:A:G	TAF15:c.146A>G(p.[N49S])	ENST00000588240	missense_variant	1.1e-4	0	VUS	No	1.5e-5	B	1	0	0	0
17:34171806:TGGAGGAGATCGAGGAGGTTAC:T	TAF15:c.1524_1544delCGGAGGAGATCGAGGAGGTTA(p.[G509_Y515del])	ENST00000588240	disruptive_inframe_deletion	5.8e-4	2.7e-4	VUS	No	9.0e-4	NA	3	0	0	0
17:34149742:A:C	TAF15:c.389A>C(p.[D130A])	ENST00000588240	missense_variant	1.1e-4	0	VUS	No	3.5e-5	NA	1	0	0	0
1:11082325:G:A	TARDBP:c.859G>A(p.[G287S])	ENST00000240185	missense_variant	2.3e-4	0	VUS	Yes	9.7e-6	P	2	0	0	0
12:64879749:C:T	TBK1:c.1292C>T(p.[T431I])	ENST00000331710	missense_variant	1.1e-4	0	VUS	No	4.7e-5	NA	1	0	0	0
12:64854098:A:G	TBK1:c.217A>G(p.[I73V])	ENST00000331710	missense_variant	1.1e-4	0	VUS	Yes	4.4e-5	B	1	0	0	0
12:64895152:C:CTT	TBK1:c.2182_2183insTT(p.[C728fs])	ENST00000331710	frameshift_variant	1.1e-4	0	VUS	No	NA	NA	1	0	0	0
12:64875638:C:G	TBK1:c.829C>G(p.[L277V])	ENST00000331710	missense_variant	1.1e-4	0	VUS	Yes	4.8e-6	NA	1	0	0	0
12:64875683:T:C	TBK1:c.874T>C(p.[C292R])	ENST00000331710	missense_variant	NA	NA	NA	No	NA	P	1	0	0	0
19:17746950:A:T	UNC13A:c.3362T>A(p.[V1121D])	ENST00000428389	missense_variant	1.1e-4	0	VUS	Yes	4.9e-6	NA	1	1	0	0
20:57016039:GTTC:G	VAPB:c.479_481delCTT(p.[S160del])	ENST00000475243	disruptive_inframe_deletion	3.9e-3	3.8e-3	VUS	Yes	1.6e-3	NA	2	0	0	0

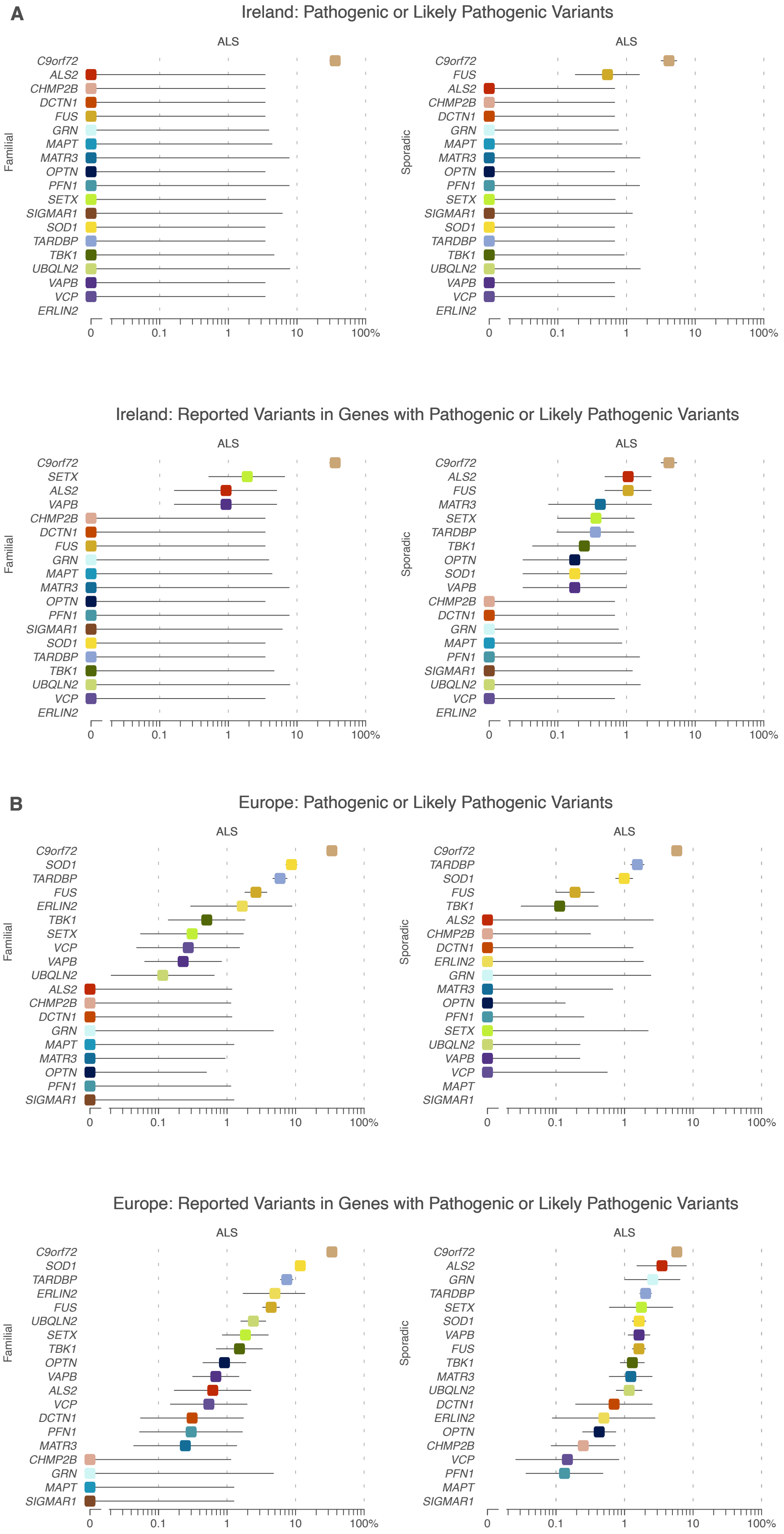


Figure 5.4: ALS genetic variation in Ireland and Europe

Discordant families

Recent work in the Academic Unit of Neurology at TCD has identified pedigrees wherein members of the same family are affected by ALS but have different *C9orf72* genotyping results. Combining analysis of rpPCR, targeted NGS, WES, WGS and SNP data, sufficient information is available to investigate the basis of this discordance in three pedigrees.

Pedigree 3

Family three has 6 siblings affected by either ALS or FTD (figure 5.5). *C9orf72* genotyping of 4 available affected siblings identified that three siblings are positive for the repeat expansion and one sibling is negative. Sufficient DNA for two samples was available to repeat the PCR, confirming the result in one positive sibling and the negative sibling. The RE is also observed in a currently unaffected sibling. Both ExpansionHunter v2 and ExpansionHunter v3 provide further confirmation that the negative patient is heterozygous for 2 and 5 GGGGCC repeat motifs.

SNP genotyping was available for one positive sibling and the negative sibling. A sibling relatedness was confirmed (pi-hat=0.5383), verifying both that the negative sibling is truly related to the family and that the result is not attributable to sample mix-up. SNP genotyping confirms that the positive sibling carries the elongated *C9orf72* haplotype (figure 5.6). The negative sample is homozygous for the non-risk allele at two critical SNPs (rs3849942 and rs10812605). Rare recombination is known to occur in this haplotype, with Smith *et al.* (2013) identifying that 1.43% and 2.86% of expansion carriers have the non-risk allele at each of these SNPs respectively, so it cannot be confirmed whether the *C9orf72* RE negative patient did not inherit the risk allele or whether recombination occurred in the inheritance of the haplotype.

Targeted NGS was available for two positive RE carriers and WGS was available for the negative sibling. The only putative variant (table 5.14) observed in the negative patient was *ATXN2*:c.224A>G(p.[D75G]). This variant is predicted to be benign by *in silico* tools and to date only an intermediate CAG repeat expansion in *ATXN2* has been linked to ALS pathogenesis. This evidence suggests that this variant is not pathogenic, although it still remains a VUS. *ATXN2* was not included in the ALS target NGS panel so cannot be confirmed in the two positive siblings.

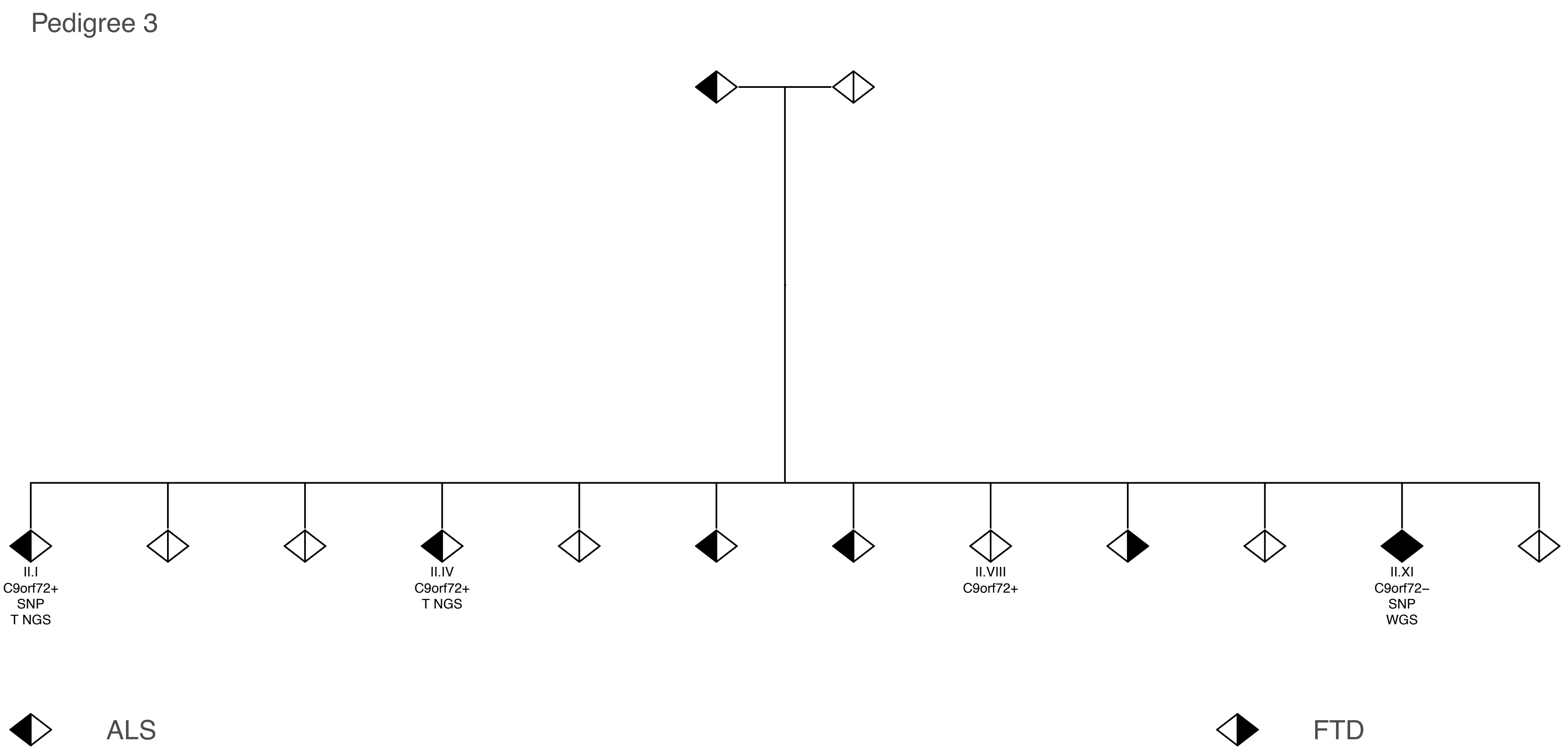
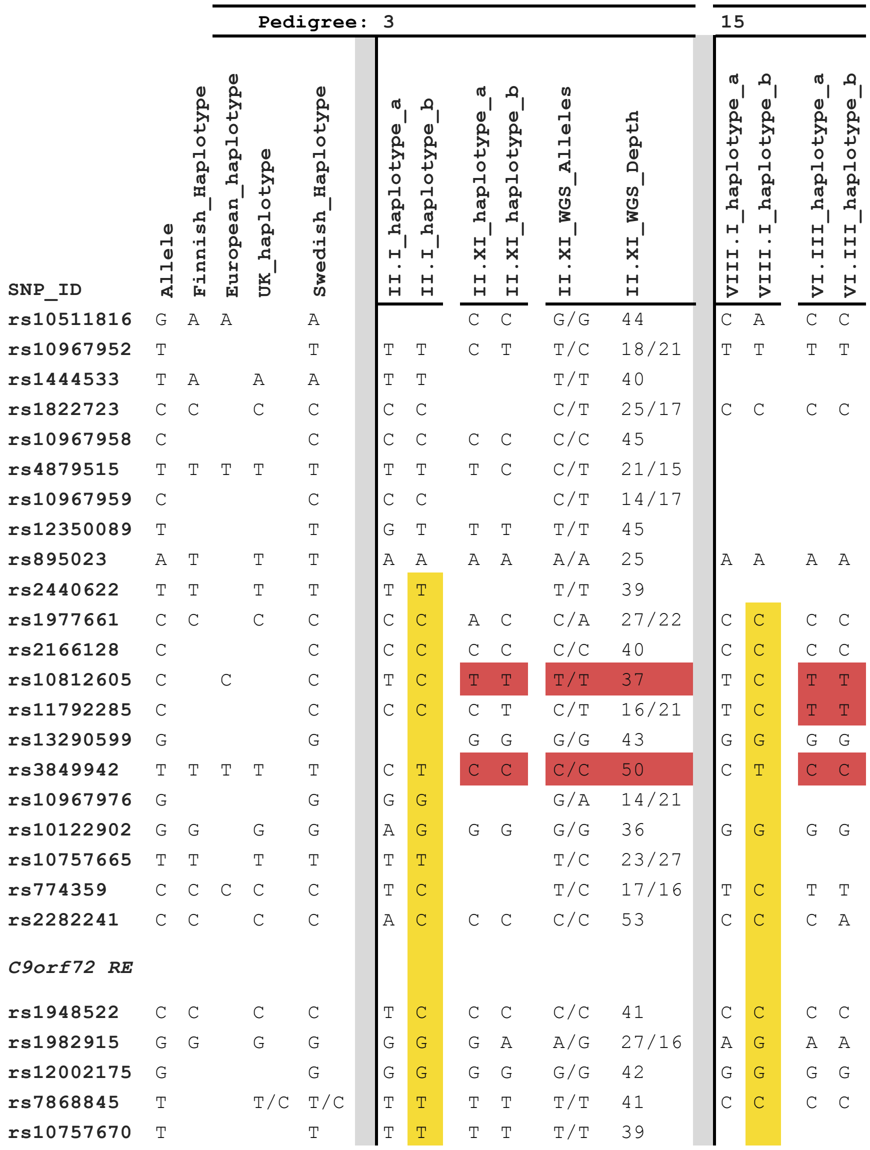


Figure 5.5: Pedigree 3 - discordant *C9orf72* genotyping in affected siblings

C9orf72+ indicates a carrier of the repeat expansion. *C9orf72-* indicates the individual does not carry the expansion. T NGS indicates that there is targeted sequencing data available. WGS indicates that there is whole genome sequencing available. SNP indicates that there is SNP genotyping available.



						Pedigree: 3						15			
SNP_ID	Allele	Finnish_Haplotype	European_haplotype	UK_haplotype	Swedish_Haplotype	II.I_haplotype_a	II.I_haplotype_b	II.XI_haplotype_a	II.XI_haplotype_b	II.XI_WGS_Alleles	II.XI_WGS_Depth	VIII.I_haplotype_a	VIII.I_haplotype_b	VI.III_haplotype_a	VI.III_haplotype_b
rs10511816	G	A	A		A			C	C	G/G	44	C	A	C	C
rs10967952	T				T	T	T	C	T	T/C	18/21	T	T	T	T
rs1444533	T	A		A	A	T	T			T/T	40				
rs1822723	C	C		C	C	C	C			C/T	25/17	C	C	C	C
rs10967958	C				C	C	C	C	C	C/C	45				
rs4879515	T	T	T	T	T	T	T	T	C	C/T	21/15				
rs10967959	C				C	C	C			C/T	14/17				
rs12350089	T				T	G	T	T	T	T/T	45				
rs895023	A	T		T	T	A	A	A	A	A/A	25	A	A	A	A
rs2440622	T	T		T	T	T	T			T/T	39				
rs1977661	C	C		C	C	C	C	A	C	C/A	27/22	C	C	C	C
rs2166128	C				C	C	C	C	C	C/C	40	C	C	C	C
rs10812605	C		C		C	T	C	T	T	T/T	37	T	C	T	T
rs11792285	C				C	C	C	C	T	C/T	16/21	T	C	T	T
rs13290599	G				G			G	G	G/G	43	G	G	G	G
rs3849942	T	T	T	T	T	C	T	C	C	C/C	50	C	T	C	C
rs10967976	G				G	G	G			G/A	14/21				
rs10122902	G	G		G	G	A	G	G	G	G/G	36	G	G	G	G
rs10757665	T	T		T	T	T	T			T/C	23/27				
rs774359	C	C	C	C	C	T	C			T/C	17/16	T	C	T	T
rs2282241	C	C		C	C	A	C	C	C	C/C	53	C	C	C	A
C9orf72 RE															
rs1948522	C	C		C	C	T	C	C	C	C/C	41	C	C	C	C
rs1982915	G	G		G	G	G	G	G	A	A/G	27/16	A	G	A	A
rs12002175	G				G	G	G	G	G	G/G	42	G	G	G	G
rs7868845	T			T/C	T/C	T	T	T	T	T/T	41	C	C	C	C
rs10757670	T				T	T	T	T	T	T/T	39				

Figure 5.6: *C9orf72* haplotype analysis for pedigrees 3 and 15

The yellow highlight indicates that the two positive samples carry the established elongated *C9orf72* haplotype. The red highlight indicates two loci where the *C9orf72* negative patients in pedigrees 3 and 15 are homozygous for the non-risk allele indicating that they either did not inherit the haplotype or recombination occurred in the inherited haplotype.

Pedigree 15

Pedigree 15 contains two affected siblings who are confirmed to carry the *C9orf72* RE and a fourth cousin who is also affected but does not carry the expansion (figure 5.7). SNP genotyping for an affected patient and the distant cousin confirms, as for pedigree 3, that the *C9orf72* RE negative individual either did not inherit the haplotype or that recombination occurred in the inherited haplotype. Relatedness is observed to be 3.5%, which is at the background level of the population but is not unexpected for distant cousins. Targeted sequencing was available for the negative distant cousin and no putative variants were observed.

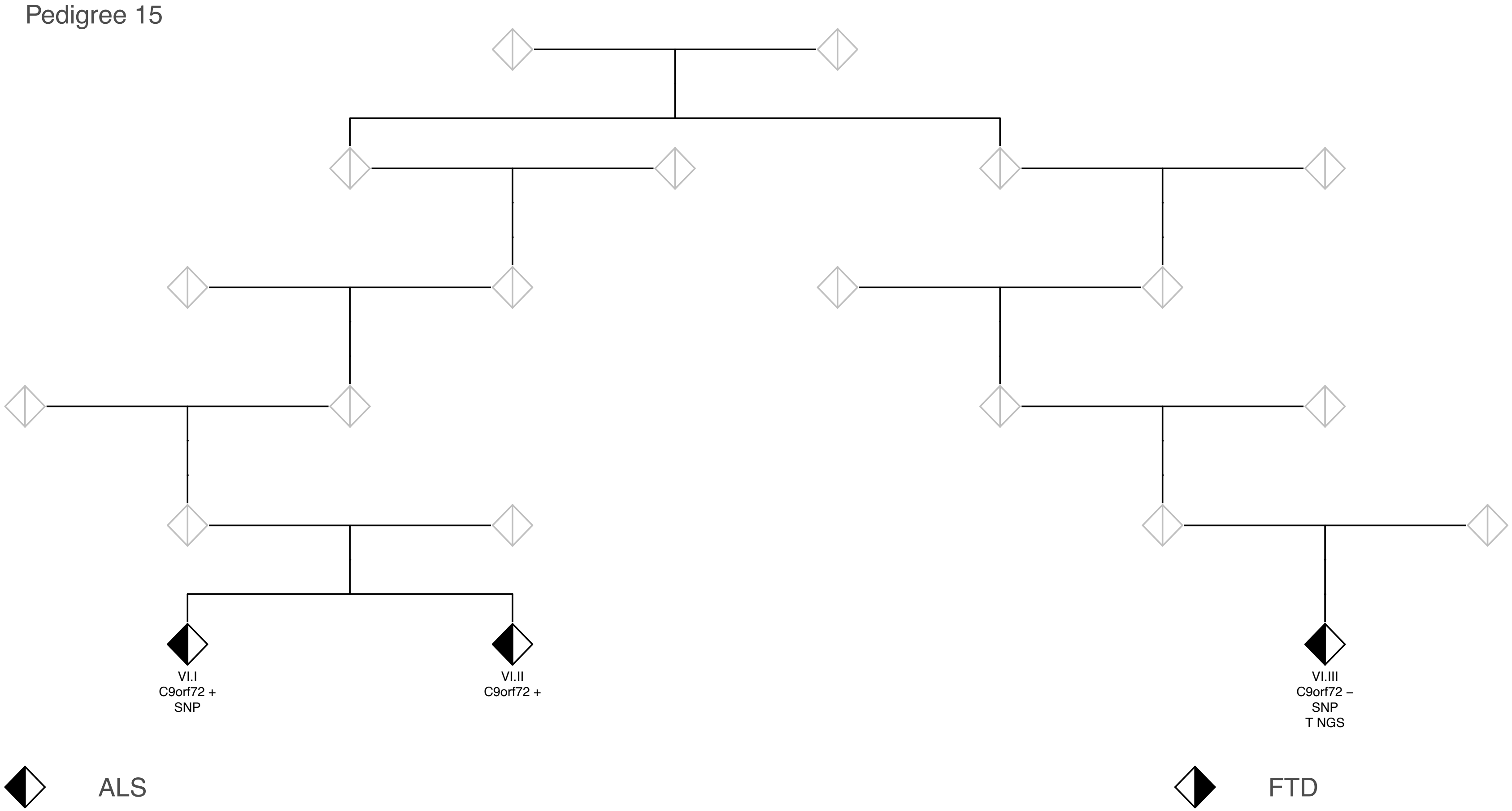


Figure 5.7: Pedigree 15 - discordant *C9orf72* genotyping in distant cousins

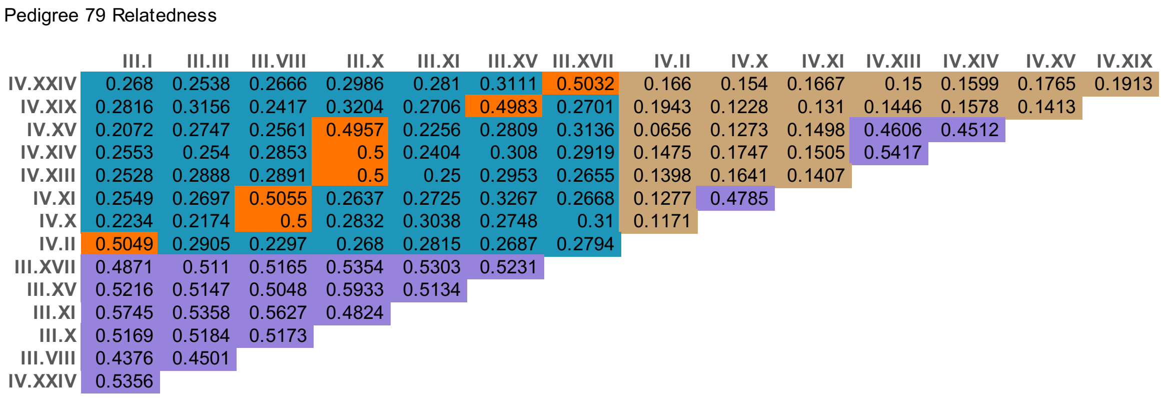
C9orf72+ indicates a carrier of the repeat expansion. *C9orf72*- indicates the individual does not carry the expansion. T NGS indicates that there is targeted sequencing data available. WGS indicates that there is whole genome sequencing available. SNP indicates that there is SNP genotyping available.

Pedigree 79

Pedigree 79 is a four generation family with 6 recorded cases of ALS (figure 5.9). In the fourth generation an ALS patient is negative for the *C9orf72* expansion but their affected parent is positive. Sufficient DNA was available to repeat the rpPCR for the negative sample and their parent's affected sibling (individual III.I) with both PCRs confirming the initial result.

WES was performed for 15 members of the pedigree. The expected relatedness percentages were confirmed with the *C9orf72* negative affected individual having 50% relatedness to their affected *C9orf72* positive parent (figure 5.8). The presence of the *C9orf72* RE was further confirmed in patient III.I with ExpansionHunter v2 and v3 with allele predictions of 2/238 and 2/99 respectively.

Individual III.I carries *KIF5A*:c.2953G>A(p.[G985S]) in addition to the *C9orf72* RE. This is also identified in their unaffected child (IV.II), sibling (III.XVII) and nibling (IV.XXIV). This variant is predicted to be benign by *in silico* tools and crucially, is also absent in the *C9orf72* negative patient (IV.X) and their affected parent (III.VIII), so is unlikely to be contributing to the observed discordance.



Pedigree 79 Relatedness

	III.I	III.III	III.VIII	III.X	III.XI	III.XV	III.XVII	IV.II	IV.X	IV.XI	IV.XIII	IV.XIV	IV.XV	IV.XIX
IV.XXIV	0.268	0.2538	0.2666	0.2986	0.281	0.3111	0.5032	0.166	0.154	0.1667	0.15	0.1599	0.1765	0.1913
IV.XIX	0.2816	0.3156	0.2417	0.3204	0.2706	0.4983	0.2701	0.1943	0.1228	0.131	0.1446	0.1578	0.1413	
IV.XV	0.2072	0.2747	0.2561	0.4957	0.2256	0.2809	0.3136	0.0656	0.1273	0.1498	0.4606	0.4512		
IV.XIV	0.2553	0.254	0.2853	0.5	0.2404	0.308	0.2919	0.1475	0.1747	0.1505	0.5417			
IV.XIII	0.2528	0.2888	0.2891	0.5	0.25	0.2953	0.2655	0.1398	0.1641	0.1407				
IV.XI	0.2549	0.2697	0.5055	0.2637	0.2725	0.3267	0.2668	0.1277	0.4785					
IV.X	0.2234	0.2174	0.5	0.2832	0.3038	0.2748	0.31	0.1171						
IV.II	0.5049	0.2905	0.2297	0.268	0.2815	0.2687	0.2794							
III.XVII	0.4871	0.511	0.5165	0.5354	0.5303	0.5231								
III.XV	0.5216	0.5147	0.5048	0.5933	0.5134									
III.XI	0.5745	0.5358	0.5627	0.4824										
III.X	0.5169	0.5184	0.5173											
III.VIII	0.4376	0.4501												
IV.XXIV	0.5356													

Figure 5.8: Relatedness in pedigree 79

The observed relatedness for all individuals matches the expected relatedness based on the reported pedigree. Parent-offspring (orange), have a mean relatedness of 50.1% (95% CI 50-50.2%). Siblings (purple) have a mean relatedness of 51% (95% CI: 50-51.9%). Aunt/uncle-nibling (green) have a mean relatedness of 27.3% (95% CI: 26.9-27.7%). Cousins (brown) have a mean relatedness of 14.8% (95% CI: 14.3-15.4%).

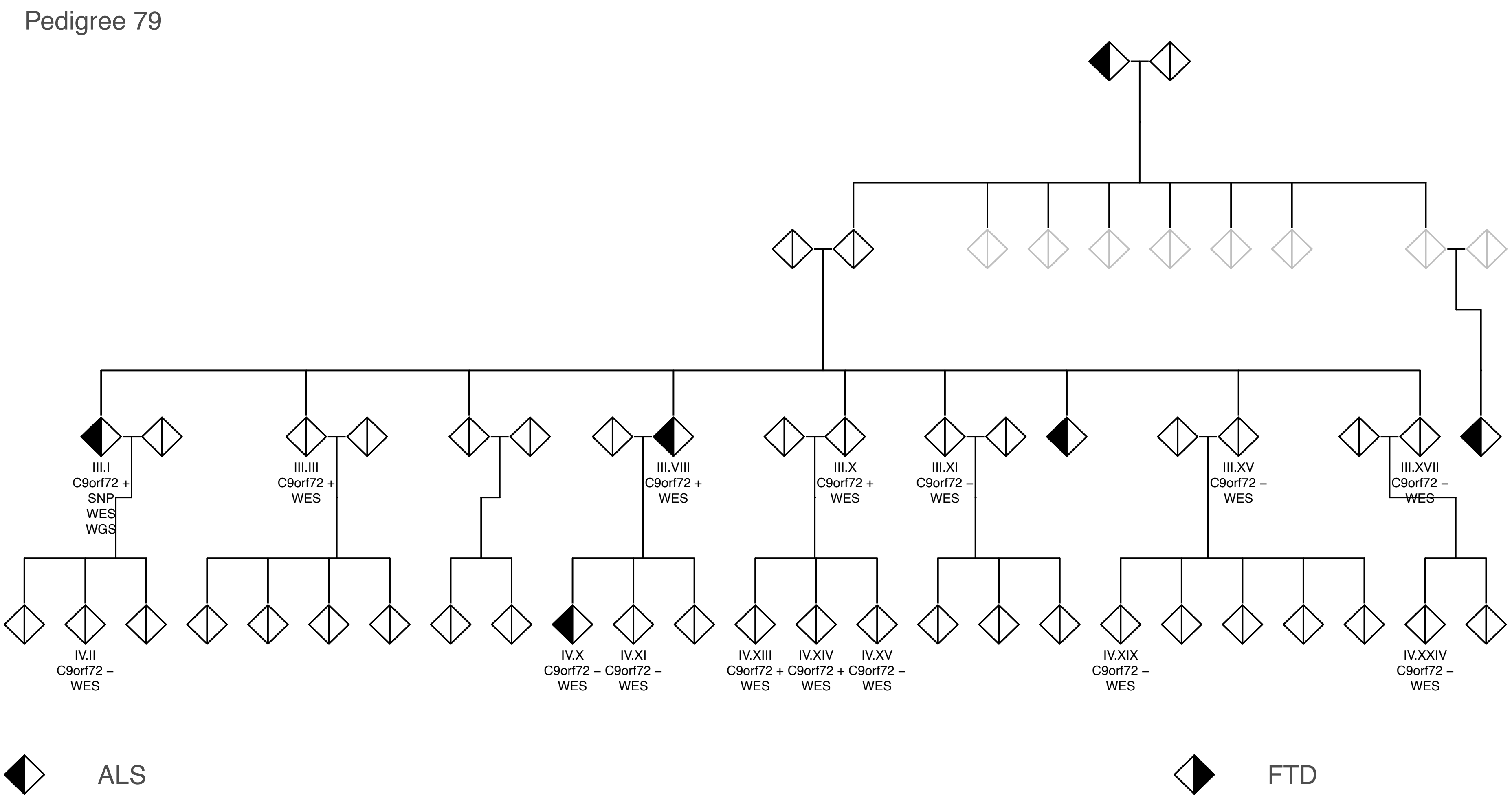


Figure 5.9: Pedigree 79 - discordant *C9orf72* genotyping in parent-offspring

C9orf72+ indicates a carrier of the repeat expansion. *C9orf72-* indicates the individual does not carry the expansion. T NGS indicates that there is targeted sequencing data available. WGS indicates that there is whole genome sequencing available. SNP indicates that there is SNP genotyping available.

Discussion

PLS genetics in Ireland

The largest NGS study of a PLS cohort to date is presented here. *C9orf72* genotyping and WES was performed for 44 patients. Due to the phenotypic and clinical overlap with ALS and HSP the study focused on identifying pathogenic variants in these genes. The *C9orf72* RE was not observed in any patient. The first NGS analysis performed, tested whether PLS patients carry a statistical excess of rare variants in HSP genes. The result was not statistically significant and the null hypothesis that there is no difference in variation in these genes cannot be rejected; however, the small size of this study cohort would only be powered to detect a very large effect so this question is still uncertain.

No previously reported, definitively pathogenic, ALS, FTD or HSP variants were observed. There is little reason to suspect the pathogenicity of 4 out of 7 observed previously reported variants. The remaining three previously reported variants have either been observed in homozygosity or compound heterozygosity in HSP patients and their pathogenicity in heterozygosity here remains uncertain.

A single patient was found to carry *SPAST*:c.1675G>A(p.[G559S]). While this variant has not previously been reported, variants in this amino acid have been found in three families with pure HSP. The evidence here indicates that missense variants in this amino acid are responsible for UMN degeneration.

ALS and FTD genetics in Ireland

This is the first NGS screen of FTD patients in Ireland and the largest Irish ALS genetics study to date.

51 FTD patients underwent *C9orf72* genotyping and NGS screening for pathogenic variants. A low rate of the *C9orf72* RE was observed (below 3% in “pure” FTD cases) relative to the rest of the world where approximately 10% of cases carry the expansion. No other definitively pathogenic variants are observed in this Irish FTD cohort. In Europe and globally, variants in *MAPT* and *GRN* are the second most common cause of FTD; however, no rare variation in these genes is observed in Ireland.

Similar to FTD, the landscape of ALS genetics in Ireland is notable by its absences. While our rate of *C9orf72* is on par with other European countries we observe a small amount of *FUS* and *TARDBP* variation with virtually no *SOD1* or *TBK1* variants, while these are major causative genes in other countries.

Unfortunately the fact that we observe fewer pathogenic variants in Irish ALS cases does not translate to fewer cases in the Irish population, as a similar incidence rate is observed in Ireland to the rest of the world (O'Toole *et al.* 2008). Nor does it mean that there is less genetic contribution to ALS cases in Ireland, as heritability and the number of observed familial cases is again similar to the rest of the world (Ryan, Heverin, *et al.* 2019; Ryan *et al.* 2018). Rather, the result suggests that there are as yet, undiscovered pathogenic variants in the Irish ALS and FTD populations that are potentially identifiable with increased genome sequencing of both patients and controls.

Discordant pedigrees

In Chapter 2 it was identified that the *C9orf72* RE displays reduced penetrance for the risk of developing ALS, suggesting that other developmental, environmental or genetic factors may contribute to pathogenesis. However, the prevailing expectation within pedigrees with *C9orf72* expansions is that only individuals who demonstrate a RE in the pathologic range should develop ALS or FTD. The observation of numerous discordant affected relatives in the Irish population challenges this orthodoxy.

Three pedigrees are examined to further explore the basis of this discordance and to study potential explanations. The possibility that the result is attributable to laboratory error is removed by firstly replicating the rpPCR results where possible, and secondly confirming a sibling relatedness in one family and a parent-offspring relatedness in a second family. Affected individuals in the third family are distant cousins so excess relatedness is not expected. The presence or absence of the RE is further confirmed in two samples using *in silico* RE genotyping from WGS data.

A second potential explanation is that these pedigrees may have a second pathogenic variant circulating in the family. For two previously reported families with discordant *C9orf72* family members this has been the case (van Blitterswijk *et al.* 2012; Ismail *et al.* 2013); however, analysis of targeted NGS, WES and WGS does not reveal any other pathogenic variants. It is still possible that there are other circulating pathogenic variants that are as yet unknown.

It is possible that the discordance is attributable to somatic mosaicism. REs have previously been found to exhibit somatic instability during development (Sharma *et al.* 2002; McMurray 2010). As DNA for this project is extracted from blood, it is possible that a patient could carry a RE in their motor neurones, which are derived from the ectoderm during embryogenesis, but not in their blood, which is derived from the mesoderm. To examine this possibility, haplotype analysis was performed in *C9orf72* RE positive and negative patients, where SNP data was available.

Laaksovirta *et al.* (2010) first identified a 232kb block of linkage disequilibrium on chromosome 9 that was significantly associated with familial ALS in Finland (OR=21.0 (95% CI: 11.2-39.1); p=4.24x10-33). This haplotype was subsequently found to tag the *C9orf72* RE (DeJesus-Hernandez *et al.* 2011; Renton *et al.* 2011). Subsequent research has found that the *C9orf72* RE only arose once on this haplotype (Smith *et al.* 2013). However, while all carriers of the RE also have the haplotype, Laaksovirta *et al.* (2010) identified that it is also present in the healthy population at a rate of 3.6%. Rare recombination has been observed in the haplotype, particularly as distance from the *C9orf72* RE increases (Smith *et al.* 2013).

Evidence here indicates, but is inconclusive, that the two negative patients for whom SNP genotyping is available, did not inherit the *C9orf72* haplotype. This is indicated by two SNPs where each sample is homozygous for the non-risk allele. However, this is inconclusive as both samples do potentially carry a short version of the *C9orf72* haplotype directly surrounding the location of RE. It is possible that recombination has occurred in the inheritance of the haplotype in these patients, as Smith *et al.* identified that 1.43% and 2.86% of *C9orf72* RE positive patients carry each of these non-risk alleles.

This is a hugely important question to be addressed in future studies. If these patients are indeed exhibiting somatic mosaicism, this indicates that the rate of the *C9orf72* RE in ALS patients is being underrepresented by testing patient’s blood. Unfortunately, the discordant patients in this study are either no longer alive or no longer consenting to research. A future study in which DNA is not only extracted from patient’s blood but also cheek epithelial cells, which derive from the same germ layer as motor neurones, could identify if somatic mosaicism is causing an underreporting of the true rate of the *C9orf72* RE.

If these patients have not inherited the *C9orf72* haplotype, this indicates that somatic mosaicism is not a factor and that there is an alternative, as yet unexplained, cause of ALS in these pedigrees. It also cannot be ruled out that there are unknown pathogenic variants also circulating in these families. It may be possible that there are circulating variants that promote genomic instability, which could manifest primarily at the *C9orf72* locus but also at currently unknown loci in the absence of the *C9orf72* haplotype.

Study limitations

Studies of rare diseases in a small population will always be limited by the size of the available patient cohort. This is slightly ameliorated in Ireland by the quality and duration of the Irish ALS register; however the studies of FTD and PLS presented here are still of a relatively small size. Despite being the largest NGS study of PLS to date, this is still too small a cohort to examine the effect of variants that may not be fully penetrant.

While cohort studies may be limited by the availability of patients, they could also be greatly improved by increasing the availability of controls. Of five variants observed in two PLS patients, there is evidence suggesting that four of these may be more common in Ireland than elsewhere in the world. A large publicly available Irish genomics resource would greatly improve analysis of all rare diseases in Ireland.

Future direction

Future efforts should focus on creating an all-Ireland genomics resource that would benefit the study of all rare diseases. For PLS and FTD, international collaborations should focus on pooling patient cohorts to improve studies.

Monitoring of the development of currently unaffected individuals in ALS pedigrees should be made a priority, to further explore the extent of discordant *C9orf72* inheritance. Additionally, a study of potential somatic mosaicism in future discordant families should be undertaken, by testing for the *C9orf72* RE in DNA extracted not only from blood but also from cheek epithelial cells.

Chapter 6

Discussion, limitations & future direction

The overarching aim of this thesis is to clarify and further our understanding of the genetic causes of ALS and related diseases. It is hoped that achieving this can help bring clarity to patients, relatives and carers by improving genetic counselling and aiding in the design of clinical trials by improving patient stratification based on genetic background.

In Chapter 2 the extant body of genetics literature in ALS and FTD was screened to uniformly and objectively assess the evidence supporting each variant and to provide an accessible web application for patients, clinicians and researchers (available at alsftd.tcd.ie). 2,914 articles were screened, of which, 1,028 were found to be relevant ALS or FTD genetic studies. 3,114 previously reported variants were identified in 356 genes and all reported phenotype and segregation data was recorded. Ultimately, 112 variants in 21 genes were found to cross the evidence threshold to be classified as pathogenic or likely pathogenic. This is less than 1% of variants which have previously been reported in ALS or FTD patients. A further 10% of reported variants are classified as benign or likely benign and the vast majority are variants of uncertain significance.

Globally, it is found that reported variants in the 21 genes with observed ALS or FTD pathogenic or likely pathogenic variants can currently explain at most 68.7% of fALS, 51.2% of fFTD, 21.4% of sALS and 9.6% of sFTD; however, these figures are considerably lower when considering strictly pathogenic or likely pathogenic variants (48.67%, 28.6%, 6.51%

and 4.59% respectively). Considering that most cases of both ALS and FTD are sporadic, a clear picture emerges that despite the high heritability of ALS and FTD, the majority of cases still lack a clear genetic diagnosis. 11% of the identified pathogenic or likely pathogenic variants were found to exhibit geographic heterogeneity, highlighting the often population-specific genetic basis underlying ALS. It was also observed that the majority of studies have been confined to a small number of regions. In order to both improve global parity and to further our understanding of ALS and FTD genetics, it is essential to broaden the areas in which genetic screening occurs.

It is not surprising that just below 90% of previously reported variants in Chapter 2 receive a VUS classification, after all, in the absence of significant evidence in either direction, this is the default status of any observed variant. However, there is a large degree of nuance within this category. By definition, a categorisation of VUS means that there is insufficient evidence to infer whether a variant is pathogenic or benign; however, evidence can be supportive of benignity, supportive of pathogenicity, have conflicting support or have little support in either direction. Chapters 4 and 5 aim to capture this nuance when discussing variants observed in ALS in Cuba and FTD in Ireland respectively.

It is confirmed that variant penetrance plays a significant role in ALS and FTD pathogenesis, with several variants of intermediate penetrance identified in the research of Chapter 2. Reduced lifetime penetrance of the *C9orf72* RE has been observed previously (Spargo *et al.* 2021); however, by combining analysis of both ALS and FTD, this study identifies that the likelihood of developing disease along the ALS-FTD spectrum ranges from 0.76 to 1 for carriers of the *C9orf72* RE. While this is an unfortunate finding, it can hopefully provide clarity to patients and relatives carrying this variant. Improvements in the size and availability of national and international genome biobanks continues to improve will see a corresponding improvement in the confidence with which penetrance estimates can be calculated.

With the exception of the *C9orf72* RE, it was necessary to omit the analysis of REs from Chapter 2. Due to their nature, it has traditionally not been possible to measure REs from NGS data and they have therefore not been uniformly reported across previous studies. There have been several tools developed in recent years that purport to facilitate this research, however objective benchmarking studies of these tools have been limited.

Chapter 3 aims to objectively benchmark 7 *in silico* STR / RE genotyping tools through three analyses. Firstly, each tool's ability to accurately identity large REs is tested by screening 408 samples, 26 of which are known to carry large REs at the *C9orf72* locus. Secondly, the accuracy of each tool each tool is assessed by comparing gold-standard PCR genotyping and *in silico* predictions for 23 genes in 338 samples. Finally, the results of *in silico* genotyping are compared between 23 samples for which WGS and WES was available. While ExpansionHunter is found to perform best overall across the three metrics, no one tool provides perfect discrimination and accuracy, with results being highly gene dependent, and several genes being prone to false positives. The presence of false positives indicates that either a consensus approach should be taken between tools or all predicted expansions require further validation.

In Chapter 3 it is demonstrated that methods developed and validated for one neurological condition can have broader impact in the field of neurological disease research. An analysis of STR loci in 132 epilepsy patients was performed, utilising the results of the benchmarking study, which was performed primarily in data derived from ALS patients. Data from epilepsy patients was comprised of PCR-free WGS, WGS with PCR and WES data. While PCR-free WGS is ideal data for *in silico* RE genotyping, useful insight can be gleaned from analysis of WES and WGS data with PCR, providing significant or interesting results are interpreted cautiously, as is done here. Statistically significant putative STRs were identified in 24 genes, however after inspection of reads and comparison with other tools all positive results were found to be false positives. This study does not find evidence supporting the pleiotropic role of known pathogenic REs in epilepsy in the Irish population.

Findings from Chapter 2 revealed that the majority of ALS and FTD genetics research has been concentrated in a small number of countries, and also that there are regions such as Brazil and Sardinia that exhibit significant geographic heterogeneity with a single variant explaining a large proportion of cases. It is therefore worthwhile to broaden the scope of where ALS genetics research is conducted in order to improve global parity and to further improve our knowledge of the underlying causes of ALS. Chapter 4 attempts to redress this by studying the genetics of ALS in Cuba. 126 Cuban ALS patients and 111 controls underwent targeted NGS and rpPCR genotyping of the *C9orf72* RE. 6 of these patients were from a single pedigree and also underwent WES. A low rate of the *C9orf72* RE is observed in Cuba (2.7%), this is likely reflective of the partial European ancestry in the population. The profile of ALS genetics in Cuba is unique from other North and South American countries with none of the prevalent *SOD1*, *TARDBP* or *VAPB* variants that are observed in those regions.

A *FUS* variant (*FUS*:c.1512_1513delAG(p.[G505fs])) is observed in a single sporadic patient with relatively early onset. The observation of this variant in an early onset patient is found to be sufficient evidence, when combined with journALS data, to reclassify this variant as a pathogenic. This reclassification will hopefully bring clarity to current and future patients who hold this variant and highlights the importance of continuous phenotyping and genotyping of ALS patients and the benefits this can have for the broader community.

Several studies have previously been published reporting an oligogenic basis to ALS, wherein ALS patients are found to carry multiple variants in associated genes. While statistical evidence supporting this finding has been provided (van Blitterswijk *et al.* 2012; Pang *et al.* 2017; Morgan *et al.* 2017), the majority of publications on the topic describe all cases that carry more than one variant in ALS-associated genes as demonstrative examples of oligogenic inheritance without determining if there is a statistical difference between cases and controls (Zhang *et al.* 2018; McCann *et al.* 2020; Kuuluvainen *et al.* 2019; Giannoccaro *et al.* 2017; Bury *et al.* 2016). In this study of 126 patients and 111 controls, no statistical difference is observed between the number of cases and the number of controls that carry multiple variants across a range of tested variables. This does not disprove that oligogenic inheritance is relevant in ALS, it may just be the case that oligogenic inheritance is not a feature in Cuba or that this study is underpowered to detect the effect; however, this does demonstrate that an observation of two or more variants should not be assumed to be an oligogenic cause of ALS as this is also frequently observed in controls.

While Chapter 4 studies ALS in a population which has not previously undergone ALS genetic screening, Chapter 5 studies a well characterised cohort, enabling in depth analysis of related conditions and anomalies. Chapter 5 presents the first comprehensive screen of FTD and PLS in Ireland, the largest analysis of ALS in Ireland to date and explores the genetic basis of multiple families with affected individuals who are found to be discordant for the *C9orf72* RE. Results are analysed through the framework developed in the journALS study, demonstrating the utility of this research.

In Chapter 5 the profile of genetic variation in ALS and FTD in Ireland is found to be distinct from the rest of the world by its absences. While rates of the *C9orf72* RE are similar to other

European countries, Irish patients lack variants that are commonly observed elsewhere: in FTD no *MAPT* or *GRN* variation is observed, while in ALS no *TBK1* variation and little *SOD1* or *TARDBP* variation is observed. Unfortunately, as discussed in Chapter 5, this does not mean there is a lower rate of these diseases in Ireland, rather it suggests either that there are as-yet undiscovered genetic causes of ALS and FTD in the Irish population or that these causes are individually so rare elsewhere that their absence does not have a notable effect on disease incidence in Ireland.

The largest NGS screening study of PLS patients to date is performed here and reveals that PLS does not appear to be largely driven by pathogenic variants in HSP or ALS genes. A patient is observed to carry a previously unreported *SPAST* variant (c.1675G>A(p.[G559S])). Variants in the same amino acid have previously been observed to cause cases of adult onset familial HSP with similar ages of onset to the patient here. While the phenotypic overlap between PLS and HSP makes it difficult to determine whether these patients had a different aetiology it is clear that heterozygous missense variants in this amino acid are responsible for UMN degeneration.

Recent work has identified Irish pedigrees wherein members of the same family are affected by ALS or FTD but have different *C9orf72* genotyping results. Combining analysis of rpPCR, targeted NGS, WES, WGS and SNP data, sufficient information is available to investigate the basis of this discordance in three pedigrees. The possibility that the discordant results are attributable to lab error (either sample mix up or false positives / negatives) was eliminated by repeating rpPCR genotyping in at least one positive and one negative sample in each pedigree and by confirming the expected relatedness of individuals using SNP genotyping. This work also demonstrates the utility of the benchmarking study carried out in Chapter 3 as the presence and absence of the RE is also confirmed in two samples for whom WGS was available using ExpansionHunter.

It is found that two of the patients who are negative for the *C9orf72* RE have either not have inherited the associated haplotype or that recombination has occurred in the inheritance of the haplotype. Identifying which of these is the case is a hugely important topic for future studies. If the patients have indeed inherited the *C9orf72* haplotype but are testing negatively both by rpPCR and ExpansionHunter, one possible explanation for this is that they may be exhibiting somatic mosaicism wherein they carry they RE in their motor neurones both not in their blood. This would indicate that the rate of cases that is attributable to the *C9orf72*

RE is higher than reported. This hypothesis can be tested in future studies by extracting DNA both from blood and cheek epithelial cells. If the patients are not inheriting the haplotype, this indicates that there is an entirely different cause of ALS in these families. This can be tested in future studies by further SNP genotyping of discordant trios or larger pedigrees to determine if the inherited haplotype is a recombined haplotype or a separate haplotype segregating in the family.

The case / control studies of Chapters 4 and 5 are primarily limited by sample size, both for cases and controls. Due to limited statistical power in these chapters, analysis has been restricted to the study of highly penetrant pathogenic variants and has not extended to the potential association of low-penetrance variants. This problem has long plagued ALS research and is also true of FTD and PLS. The problem is exacerbated in these three conditions due to the highly heterogenous patient populations, wherein, with the exception of the *C9orf72*, individual variants are present in a very small percentage of patients globally.

Despite the fact that ALS was first described over 150 years ago and that the first ALS gene was first identified 30 years ago, less than 50% of familial patients and less than 7% of sporadic patients can currently receive a confident genetic diagnosis. This situation needs to be addressed as a priority. The future of research in ALS, FTD and PLS has to be global and collaborative. There are of course considerable financial, organisational and infrastructural challenges to such approaches; however, there are four primary and immediate benefits to taking a global, collaborative approach to these conditions. Firstly, patients from historically underserved areas will be able to receive genetic counselling and may be found to be eligible for clinical trials. Secondly, as seen in Chapter 2, there are regions where a large number of cases are explained by a single variant; identifying these regions could greatly improve enrolment and power in clinical trials in addition to furthering our understanding of biology. Thirdly, increasing the sample size of studies will increase the statistical power to detect both high and low penetrance variants. Finally, the sharing of intellectual and physical resources supports research in under-funded and under-resourced regions, empowering ALS research on a global scale while combining expertise to address the many challenges facing the field.

ALS clinical trials targeting specific genetic variants are now underway. It is a source of hope in the community that treatment for some patients may be possible in the near future.

There is much that can and should be done in the coming years to improve these trials and increased and refined genetic screening with should be at the forefront of these improvements. It is hoped that the research presented in this thesis is a positive step in this direction.

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Abbas, N., C. B. Lücking, S. Ricard, A. Dürr, V. Bonifati, G. De Michele, S. Bouley, et al. 1999. “A Wide Variety of Mutations in the Parkin Gene Are Responsible for Autosomal Recessive Parkinsonism in Europe. French Parkinson’s Disease Genetics Study Group and the European Consortium on Genetic Susceptibility in Parkinson’s Disease.” *Human Molecular Genetics* 8 (4): 567–74.

Abel, Ernest L. 2007. “Football Increases the Risk for Lou Gehrig’s Disease, Amyotrophic Lateral Sclerosis.” *Perceptual and Motor Skills* 104 (3 Pt 2): 1251–54.

Abel, Olubunmi, John F. Powell, Peter M. Andersen, and Ammar Al-Chalabi. 2012. “ALSoD: A User-Friendly Online Bioinformatics Tool for Amyotrophic Lateral Sclerosis Genetics.” *Human Mutation*. https://doi.org/10.1002/humu.22157.

Abou Jamra, Rami, Orianne Philippe, Annick Raas-Rothschild, Sebastian H. Eck, Elisabeth Graf, Rebecca Buchert, Guntram Borck, et al. 2011. “Adaptor Protein Complex 4 Deficiency Causes Severe Autosomal-Recessive Intellectual Disability, Progressive Spastic Paraplegia, Shy Character, and Short Stature.” *American Journal of Human Genetics* 88 (6): 788–95.

Abou Tayoun, Ahmad N., Tina Pesaran, Marina T. DiStefano, Andrea Oza, Heidi L. Rehm, Leslie G. Biesecker, Steven M. Harrison, and ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI). 2018. “Recommendations for Interpreting the Loss of Function PVS1 ACMG/AMP Variant Criterion.” *Human Mutation* 39 (11): 1517–24.

Akarsu, A. N., I. Stoilov, E. Yilmaz, B. S. Sayli, and M. Sarfarazi. 1996. “Genomic Structure of HOXD13 Gene: A Nine Polyalanine Duplication Causes Synpolydactyly in Two Unrelated Families.” *Human Molecular Genetics* 5 (7): 945–52.

Alazami, Anas M., Nouran Adly, Hisham Al Dhalaan, and Fowzan S. Alkuraya. 2011. “A Nullimorphic ERLIN2 Mutation Defines a Complicated Hereditary Spastic Paraplegia Locus (SPG18).” *Neurogenetics* 12 (4): 333–36.

Al-Chalabi, A., F. Fang, M. F. Hanby, P. N. Leigh, C. E. Shaw, W. Ye, and F. Rijsdijk. 2010. “An Estimate of Amyotrophic Lateral Sclerosis Heritability Using Twin Data.” *Journal of Neurology, Neurosurgery, and Psychiatry* 81 (12): 1324–26.

Al-Chalabi, Ammar, Leonard H. van den Berg, and Jan Veldink. 2017. “Gene Discovery in Amyotrophic Lateral Sclerosis: Implications for Clinical Management.” *Nature Reviews. Neurology* 13 (2): 96–104.

Al-Chalabi, Ammar, Andrea Calvo, Adriano Chio, Shuna Colville, Cathy M. Ellis, Orla Hardiman, Mark Heverin, et al. 2014. “Analysis of Amyotrophic Lateral Sclerosis as a Multistep Process: A Population-Based Modelling Study.” *Lancet Neurology* 13 (11): 1108–13.

Ali, Mohammad A. M., Hilmar Strickfaden, Brian L. Lee, Leo Spyracopoulos, and Michael J. Hendzel. 2018. “RYBP Is a K63-Ubiquitin-Chain-Binding Protein That Inhibits Homologous Recombination Repair.” *Cell Reports* 22 (2): 383–95.

Alonso, A., G. Logroscino, S. S. Jick, and M. A. Hernán. 2009. “Incidence and Lifetime Risk of Motor Neuron Disease in the United Kingdom: A Population-Based Study.” *European Journal of Neurology: The Official Journal of the European Federation of Neurological Societies* 16 (6): 745–51.

"ALS Signal Dashboard." n.d. Iamals.Org. Accessed May 21, 2021. http://iamals.org/alssignal.

"ALS Variant Server, Worcester, MA." n.d. Als.Umassmed.Edu. Accessed January 9, 2020. http://als.umassmed.edu/.

Al-Saif, Amr, Futwan Al-Mohanna, and Saeed Bohlega. 2011. "A Mutation in Sigma-1 Receptor Causes Juvenile Amyotrophic Lateral Sclerosis." *Annals of Neurology* 70 (6): 913–19.

"ALSdb, New York City, New York." n.d. Alsdb.Org. Accessed January 9, 2020. http://alsdb.org.

Amiel, Jeanne, Béatrice Laudier, Tania Attié-Bitach, Ha Trang, Loïc de Pontual, Blanca Gener, Delphine Trochet, et al. 2003. "Polyalanine Expansion and Frameshift Mutations of the Paired-like Homeobox Gene PHOX2B in Congenital Central Hypoventilation Syndrome." *Nature Genetics* 33 (4): 459–61.

Andersen, Peter M., and Ammar Al-Chalabi. 2011. "Clinical Genetics of Amyotrophic Lateral Sclerosis: What Do We Really Know?" *Nature Reviews. Neurology* 7 (11): 603–15.

Annesi, Grazia, Giovanni Savettieri, Pierfrancesco Pugliese, Marco D'Amelio, Patrizia Tarantino, Paolo Ragonese, Vincenzo La Bella, et al. 2005. "DJ-1 Mutations and Parkinsonism-Dementia-Amyotrophic Lateral Sclerosis Complex." *Annals of Neurology* 58 (5): 803–7.

Arai, Tetsuaki, Masato Hasegawa, Haruhiko Akiyama, Kenji Ikeda, Takashi Nonaka, Hiroshi Mori, David Mann, et al. 2006. "TDP-43 Is a Component of Ubiquitin-Positive Tau-Negative Inclusions in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis." *Biochemical and Biophysical Research Communications* 351 (3): 602–11.

Arif, Beenish, Kishore R. Kumar, Philip Seibler, Franca Vulinovic, Amara Fatima, Susen Winkler, Gudrun Nürnberg, et al. 2013. "A Novel OPA3 Mutation Revealed by Exome Sequencing: An Example of Reverse Phenotyping." *JAMA Neurology* 70 (6): 783–87.

Arkblad, Eva, Már Tulinius, Anna-Karin Kroksmark, Mirja Henricsson, and Niklas Darin. 2009. "A Population-Based Study of Genotypic and Phenotypic Variability in Children with Spinal Muscular Atrophy." *Acta Paediatrica* 98 (5): 865–72.

Ash, Peter E. A., Kevin F. Bieniek, Tania F. Gendron, Thomas Caulfield, Wen-Lang Lin, Mariely Dejesus-Hernandez, Marka M. van Blitterswijk, et al. 2013. "Unconventional Translation of C9ORF72 GGGGCC Expansion Generates Insoluble Polypeptides Specific to C9FTD/ALS." *Neuron* 77 (4): 639–46.

Auguie, Baptiste. 2017. *GridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Baker, Matt, Ian R. Mackenzie, Stuart M. Pickering-Brown, Jennifer Gass, Rosa Rademakers, Caroline Lindholm, Julie Snowden, et al. 2006. "Mutations in Progranulin Cause Tau-Negative Frontotemporal Dementia Linked to Chromosome 17." *Nature* 442 (7105): 916–19.

Balduzzi, Sara, Gerta Rücker, and Guido Schwarzer. 2019. "How to Perform a Meta-Analysis with R: A Practical Tutorial." *Evidence-Based Mental Health* 22 (4): 153–60.

Banack, Sandra Anne, and Paul Alan Cox. 2003. "Biomagnification of Cycad Neurotoxins in Flying Foxes:" *Neurology* 61 (3): 387–89.

Bandres-Ciga, Sara, Alastair J. Noyce, Gibran Hemani, Aude Nicolas, Andrea Calvo, Gabriele Mora, ITALSGEN Consortium, et al. 2019. "Shared Polygenic Risk and Causal Inferences in Amyotrophic Lateral Sclerosis." *Annals of Neurology* 85 (4): 470–81.

Bañez-Coronel, Monica, Fatma Ayhan, Alex D. Tarabochia, Tao Zu, Barbara A. Perez, Solaleh Khoramian Tusi, Olga Pletnikova, et al. 2015. "RAN Translation in Huntington Disease." *Neuron* 88 (4): 667–77.

Bannwarth, Sylvie, Samira Ait-El-Mkadem, Annabelle Chaussenot, Emmanuelle C. Genin, Sandra Lacas-Gervais, Konstantina Fragaki, Laetitia Berg-Alonso, et al. 2014. "A Mitochondrial Origin for Frontotemporal Dementia and Amyotrophic Lateral Sclerosis through CHCHD10 Involvement." *Brain: A Journal of Neurology* 137 (Pt 8): 2329–45.

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2013. "NCBI GEO: Archive for Functional Genomics Data Sets--Update." *Nucleic Acids Research* 41 (Database issue): D991-5.

Beaulieu, Jean-Martin, Minh Dang Nguyen, and Jean-Pierre Julien. 1999. "Late Onset Death of Motor Neurons in Mice Overexpressing Wild-Type Peripherin." *The Journal of Cell Biology* 147 (3): 531–44.

Beck, Jon, Mark Poulter, Davina Hensman, Jonathan D. Rohrer, Colin J. Mahoney, Gary Adamson, Tracy Campbell, et al. 2013. "Large C9orf72 Hexanucleotide Repeat Expansions Are Seen in Multiple Neurodegenerative Syndromes and Are More Frequent than Expected in the UK Population." *American Journal of Human Genetics* 92 (3): 345–53.

Beetz, Christian, Adam Johnson, Amber L. Schuh, Seema Thakur, Rita-Eva Varga, Thomas Fothergill, Nicole Hertel, et al. 2013. "Inhibition of TFG Function Causes Hereditary Axon Degeneration by Impairing Endoplasmic Reticulum Structure." *Proceedings of the National Academy of Sciences of the United States of America* 110 (13): 5091–96.

Beghi, E., A. Millul, A. Micheli, E. Vitelli, G. Logroscino, and SLALOM Group. 2007. "Incidence of ALS in Lombardy, Italy." *Neurology* 68 (2): 141–45.

Beghi, Ettore, Giancarlo Logroscino, Adriano Chiò, Orla Hardiman, Andrea Millul, Douglas Mitchell, Robert Swingler, and Bryan J. Traynor. 2010. "Amyotrophic Lateral Sclerosis, Physical Exercise, Trauma and Sports: Results of a Population-Based Pilot Case-Control Study." *Amyotrophic Lateral Sclerosis: Official Publication of the World Federation of Neurology Research Group on Motor Neuron Diseases* 11 (3): 289–92.

Belli, Stefano, and Nicola Vanacore. 2005. "Proportionate Mortality of Italian Soccer Players: Is Amyotrophic Lateral Sclerosis an Occupational Disease?" *European Journal of Epidemiology* 20 (3): 237–42.

Belzil, Veronique V., Peter O. Bauer, Mercedes Prudencio, Tania F. Gendron, Caroline T. Stetler, Irene K. Yan, Luc Pregent, et al. 2013. "Reduced C9orf72 Gene Expression in C9FTD/ALS Is Caused by Histone Trimethylation, an Epigenetic Event Detectable in Blood." *Acta Neuropathologica* 126 (6): 895–905.

Belzil, Veronique V., Hussein Daoud, Judith St-Onge, Anne Desjarlais, Jean-Pierre Bouchard, Nicolas Dupre, Lucette Lacomblez, et al. 2011. "Identification of Novel FUS Mutations in Sporadic Cases of Amyotrophic Lateral Sclerosis." *Amyotrophic Lateral Sclerosis: Official Publication of the World Federation of Neurology Research Group on Motor Neuron Diseases* 12 (2): 113–17.

Bengtsson, Henrik. 2020. *R.Utils: Various Programming Utilities*. https://CRAN.R-project.org/package=R.utils.

Bennett, Mark F., Karen L. Oliver, Brigid M. Regan, Susannah T. Bellows, Amy L. Schneider, Haloom Rafehi, Neblina Sikta, et al. 2020. "Familial Adult Myoclonic Epilepsy Type 1 SAMD12 TTTCA Repeat Expansion Arose 17,000 Years Ago and Is Present in Sri Lankan and Indian Families." *European Journal of Human Genetics: EJHG* 28 (7): 973–78.

Benson, Katherine A., Maire White, Nicholas M. Allen, Susan Byrne, Robert Carton, Elizabeth Comerford, Daniel Costello, et al. 2020. “A Comparison of Genomic Diagnostics in Adults and Children with Epilepsy and Comorbid Intellectual Disability.” *European Journal of Human Genetics: EJHG* 28 (8): 1066–77.

Bivand, Roger, Tim Keitt, and Barry Rowlingson. 2021. “Rgdal: Bindings for the ‘Geospatial’ Data Abstraction Library.” https://CRAN.R-project.org/package=rgdal.

Blauw, Hylke M., Wouter van Rheenen, Max Koppers, Philip Van Damme, Stefan Waibel, Robin Lemmens, Paul W. J. van Vught, et al. 2012. “NIPA1 Polyalanine Repeat Expansions Are Associated with Amyotrophic Lateral Sclerosis.” *Human Molecular Genetics* 21 (11): 2497–2502.

Blauwendraat, Cornelis, Carlo Wilke, Iris E. Jansen, Claudia Schulte, Javier Simón-Sánchez, Florian G. Metzger, Benjamin Bender, et al. 2016. “Pilot Whole-Exome Sequencing of a German Early-Onset Alzheimer’s Disease Cohort Reveals a Substantial Frequency of PSEN2 Variants.” *Neurobiology of Aging* 37 (January): 208.e11-208.e17.

Blitterswijk, Marka van, Mariely DeJesus-Hernandez, Ellis Niemantsverdriet, Melissa E. Murray, Michael G. Heckman, Nancy N. Diehl, Patricia H. Brown, et al. 2013. “Association between Repeat Sizes and Clinical and Pathological Characteristics in Carriers of C9ORF72 Repeat Expansions (Xpansize-72): A Cross-Sectional Cohort Study.” *Lancet Neurology* 12 (10): 978–88.

Blitterswijk, Marka van, Michael A. van Es, Eric A. M. Hennekam, Dennis Dooijes, Wouter van Rheenen, Jelena Medic, Pierre R. Bourque, et al. 2012. “Evidence for an Oligogenic Basis of Amyotrophic Lateral Sclerosis.” *Human Molecular Genetics* 21 (17): 3776–84.

Blitterswijk, Marka van, Tania F. Gendron, Matthew C. Baker, Mariely DeJesus-Hernandez, Nicole A. Finch, Patricia H. Brown, Lillian M. Daughrity, et al. 2015. “Novel Clinical Associations with Specific C9ORF72 Transcripts in Patients with Repeat Expansions in C9ORF72.” *Acta Neuropathologica* 130 (6): 863–76.

Boessenkool, Berry. 2019. *BerryFunctions: Function Collection Related to Plotting and Hydrology*. https://CRAN.R-project.org/package=berryFunctions.

Bouhouche, A., A. Benomar, N. Bouslam, T. Chkili, and M. Yahyaoui. 2006. “Mutation in the Epsilon Subunit of the Cytosolic Chaperonin-Containing t-Complex Peptide-1 (Cct5) Gene Causes Autosomal Recessive Mutilating Sensory Neuropathy with Spastic Paraplegia.” *Journal of Medical Genetics* 43 (5): 441–43.

Boukhris, A., G. Stevanin, I. Feki, P. Denora, N. Elleuch, M. I. Miladi, C. Goizet, et al. 2009. “Tunisian Hereditary Spastic Paraplegias: Clinical Variability Supported by Genetic Heterogeneity.” *Clinical Genetics* 75 (6): 527–36.

Brais, B., J. P. Bouchard, Y. G. Xie, D. L. Rochefort, N. Chrétien, F. M. Tomé, R. G. Lafrenière, et al. 1998. “Short GCG Expansions in the PABP2 Gene Cause Oculopharyngeal Muscular Dystrophy.” *Nature Genetics* 18 (2): 164–67.

Brandt, Tracy, Laura M. Sack, Dolores Arjona, Duanjun Tan, Hui Mei, Hong Cui, Hua Gao, et al. 2020. “Adapting ACMG/AMP Sequence Variant Classification Guidelines Forsingle-Gene Copy Number Variants.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 22 (2): 336–44.

Brnich, Sarah E., Ahmad N. Abou Tayoun, Fergus J. Couch, Garry R. Cutting, Marc S. Greenblatt, Christopher D. Heinen, Dona M. Kanavy, et al. 2019. “Recommendations for Application of the Functional Evidence PS3/BS3 Criterion Using the ACMG/AMP Sequence Variant Interpretation Framework.” *BioRxiv*. bioRxiv. https://doi.org/10.1101/709428.

Broce, Iris J., Chun C. Fan, Nicholas T. Olney, Catherine Lomen-Hoerth, Steve Finkbeiner, Nazem Atassi, Merit E. Cudkowicz, et al. 2018. “Partitioning the Genetic

Architecture of Amyotrophic Lateral Sclerosis." *BioRxiv*. https://doi.org/10.1101/505693.

Brooks, B. R., R. G. Miller, M. Swash, T. L. Munsat, and World Federation of Neurology Research Group on Motor Neuron Diseases. 2000. "El Escorial Revisited: Revised Criteria for the Diagnosis of Amyotrophic Lateral Sclerosis." *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders: Official Publication of the World Federation of Neurology, Research Group on Motor Neuron Diseases* 1 (5): 293–99.

Brown, J., A. Ashworth, S. Gydesen, A. Sorensen, M. Rossor, J. Hardy, and J. Collinge. 1995. "Familial Non-Specific Dementia Maps to Chromosome 3." *Human Molecular Genetics* 4 (9): 1625–28.

Brown, L. Y. 2001. "Holoprosencephaly Due to Mutations in ZIC2: Alanine Tract Expansion Mutations May Be Caused by Parental Somatic Recombination." *Human Molecular Genetics*. https://doi.org/10.1093/hmg/10.8.791.

Browning, Sharon R., and Brian L. Browning. 2007. "Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering." *American Journal of Human Genetics* 81 (5): 1084–97.

Brugman, F., J. H. J. Wokke, J. M. B. Vianney de Jong, H. Franssen, C. G. Faber, and L. H. Van den Berg. 2005. "Primary Lateral Sclerosis as a Phenotypic Manifestation of Familial ALS." *Neurology* 64 (10): 1778–79.

Brugman, Frans, Jan H. Veldink, Hessel Franssen, Marianne de Visser, J. M. B. Vianney de Jong, Carin G. Faber, Berry H. P. Kremer, et al. 2009. "Differentiation of Hereditary Spastic Paraparesis from Primary Lateral Sclerosis in Sporadic Adult-Onset Upper Motor Neuron Syndromes." *Archives of Neurology* 66 (4): 509–14.

Bury, Joanna J., J. Robin Highley, Johnathan Cooper-Knock, Emily F. Goodall, Adrian Higginbottom, Christopher J. McDermott, Paul G. Ince, Pamela J. Shaw, and Janine Kirby. 2016. "Oligogenic Inheritance of Optineurin (OPTN) and C9ORF72 Mutations in ALS Highlights Localisation of OPTN in the TDP-43-Negative Inclusions of C9ORF72-ALS." *Neuropathology: Official Journal of the Japanese Society of Neuropathology* 36 (2): 125–34.

Byrne, Susan, Marwa Elamin, Peter Bede, Aleksey Shatunov, Cathal Walsh, Bernie Corr, Mark Heverin, et al. 2012. "Cognitive and Clinical Characteristics of Patients with Amyotrophic Lateral Sclerosis Carrying a C9orf72 Repeat Expansion: A Population-Based Cohort Study." *Lancet Neurology* 11 (3): 232–40.

Byrne, Susan, Iain Jordan, Marwa Elamin, and Orla Hardiman. 2013. "Age at Onset of Amyotrophic Lateral Sclerosis Is Proportional to Life Expectancy." *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration* 14 (7–8): 604–7.

Byrne, Susan, Cathal Walsh, Catherine Lynch, Peter Bede, Marwa Elamin, Kevin Kenna, Russell McLaughlin, and Orla Hardiman. 2011. "Rate of Familial Amyotrophic Lateral Sclerosis: A Systematic Review and Meta-Analysis." *Journal of Neurology, Neurosurgery, and Psychiatry* 82 (6): 623–27.

Campuzano, Oscar, Georgia Sarquella-Brugada, Anna Fernandez-Falgueras, Mónica Coll, Anna Iglesias, Carles Ferrer-Costa, Sergi Cesar, et al. 2020. "Reanalysis and Reclassification of Rare Genetic Variants Associated with Inherited Arrhythmogenic Syndromes." *EBioMedicine* 54 (April): 102732.

Campuzano, V., L. Montermini, M. D. Moltò, L. Pianese, M. Cossée, F. Cavalcanti, E. Monros, et al. 1996. "Friedreich's Ataxia: Autosomal Recessive Disease Caused by an Intronic GAA Triplet Repeat Expansion." *Science* 271 (5254): 1423–27.

Cannon, Ashley, Shinsuke Fujioka, Nicola J. Rutherford, Tanis J. Ferman, Daniel F. Broderick, Kevin B. Boylan, Neill R. Graff-Radford, et al. 2013. "Clinicopathologic

Variability of the GRN A9D Mutation, Including Amyotrophic Lateral Sclerosis." *Neurology* 80 (19): 1771–77.

Carrasco, Patricia, Jordi Jacas, Ignasi Sahún, Helena Muley, Sara Ramírez, Beatriz Puisac, Pau Mezquita, Juan Pié, Mara Dierssen, and Núria Casals. 2013. "Carnitine Palmitoyltransferase 1C Deficiency Causes Motor Impairment and Hypoactivity." *Behavioural Brain Research* 256 (November): 291–97.

Casari, G., M. De Fusco, S. Ciarmatori, M. Zeviani, M. Mora, P. Fernandez, G. De Michele, et al. 1998. "Spastic Paraplegia and OXPHOS Impairment Caused by Mutations in Paraplegin, a Nuclear-Encoded Mitochondrial Metalloprotease." *Cell* 93 (6): 973–83.

"Center for Molecular Neurology." n.d. Accessed July 17, 2021. http://www.molgen.vib-ua.be/ADMutations/.

Cervenakova, L., I. I. Protas, A. Hirano, V. I. Votiakov, M. K. Nedzved, N. D. Kolomiets, I. Taller, et al. 2000. "Progressive Muscular Atrophy Variant of Familial Amyotrophic Lateral Sclerosis (PMA/ALS)." *Journal of the Neurological Sciences* 177 (2): 124–30.

Chang, Winston, Joe Cheng, J. J. Allaire, Yihui Xie, and Jonathan McPherson. 2019. *Shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny.

Chang, Winston, and Hadley Wickham. 2019. *Ggvis: Interactive Grammar of Graphics*. https://CRAN.R-project.org/package=ggvis.

Charcot, Jean-Martin, and Alix Joffroy. 1869. *Deux Cas d'atrophie Musculaire Progressive : Avec Lésions de La Substance Grise et Des Faisceaux Antérolatéraux de La Moelle Épinière*. Paris: Masson.

Chen, Ying-Zhang, Craig L. Bennett, Huy M. Huynh, Ian P. Blair, Imke Puls, Joy Irobi, Ines Dierick, et al. 2004. "DNA/RNA Helicase Gene Mutations in a Form of Juvenile Amyotrophic Lateral Sclerosis (ALS4)." *American Journal of Human Genetics* 74 (6): 1128–35.

Chernoff, N., D. J. Hill, D. L. Diggs, B. D. Faison, B. M. Francis, J. R. Lang, M. M. Larue, et al. 2017. "A Critical Review of the Postulated Role of the Non-Essential Amino Acid, β-N-Methylamino-L-Alanine, in Neurodegenerative Disease in Humans." *Journal of Toxicology and Environmental Health. Part B, Critical Reviews* 20 (4): 183–229.

Chew, Jeannie, Tania F. Gendron, Mercedes Prudencio, Hiroki Sasaguri, Yong-Jie Zhang, Monica Castanedes-Casey, Chris W. Lee, et al. 2015. "Neurodegeneration. C9ORF72 Repeat Expansions in Mice Cause TDP-43 Pathology, Neuronal Loss, and Behavioral Deficits." *Science* 348 (6239): 1151–54.

Chia, Ruth, Adriano Chiò, and Bryan J. Traynor. 2018. "Novel Genes Associated with Amyotrophic Lateral Sclerosis: Diagnostic and Clinical Implications." *Lancet Neurology* 17 (1): 94–102.

Chiang, Huei-Hsin, Charlotte Forsell, Anna-Karin Lindström, Lena Lilius, Håkan Thonberg, Inger Nennesmo, and Caroline Graff. 2017. "No Common Founder for C9orf72 Expansion Mutation in Sweden." *Journal of Human Genetics* 62 (2): 321–24.

Chintalaphani, Sanjog R., Sandy S. Pineda, Ira W. Deveson, and Kishore R. Kumar. 2021. "An Update on the Neurological Short Tandem Repeat Expansion Disorders and the Emergence of Long-Read Sequencing Diagnostics." *Acta Neuropathologica Communications* 9 (1): 98.

Chiò, A., G. Logroscino, B. J. Traynor, J. Collins, J. C. Simeone, L. A. Goldstein, and L. A. White. 2013. "Global Epidemiology of Amyotrophic Lateral Sclerosis: A Systematic Review of the Published Literature." *Neuroepidemiology* 41 (2): 118–30.

Chiò, A., G. Mora, A. Calvo, L. Mazzini, E. Bottacchi, R. Mutani, and PARALS. 2009. "Epidemiology of ALS in Italy: A 10-Year Prospective Population-Based Study." *Neurology* 72 (8): 725–31.

Chio, Adriano, Andrea Calvo, Maurizia Dossena, Paolo Ghiglione, Roberto Mutani, and Gabriele Mora. 2009. "ALS in Italian Professional Soccer Players: The Risk Is Still Present and Could Be Soccer-Specific." *Amyotrophic Lateral Sclerosis: Official Publication of the World Federation of Neurology Research Group on Motor Neuron Diseases* 10 (4): 205–9.

Chiò, Adriano, Andrea Calvo, Cristina Moglia, Antonio Canosa, Maura Brunetti, Marco Barberis, Gabriella Restagno, et al. 2015. "ATXN2 PolyQ Intermediate Repeats Are a Modifier of ALS Survival." *Neurology* 84 (3): 251–58.

Chiò, Adriano, Giancarlo Logroscino, Orla Hardiman, Robert Swingler, Douglas Mitchell, Ettore Beghi, Bryan G. Traynor, and Eurals Consortium. 2009. "Prognostic Factors in ALS: A Critical Review." *Amyotrophic Lateral Sclerosis: Official Publication of the World Federation of Neurology Research Group on Motor Neuron Diseases* 10 (5–6): 310–23.

Chiò, Adriano, Letizia Mazzini, Sandra D'Alfonso, Lucia Corrado, Antonio Canosa, Cristina Moglia, Umberto Manera, et al. 2018. "The Multistep Hypothesis of ALS Revisited: The Role of Genetic Mutations." *Neurology* 91 (7): e635–42.

Cho, You Kyung, Dhong-Gun Won, Changwon Keum, Beom Hee Lee, Go Hun Seo, and Byung-Chul Lee. 2020. "A Novel PS4 Criterion Approach Based on Symptoms of Rare Diseases and In-House Frequency Data in a Bayesian Framework." *BioRxiv*. bioRxiv. https://doi.org/10.1101/2020.07.22.215426.

Choi, Yongwook. 2012. "A Fast Computation of Pairwise Sequence Alignment Scores between a Protein and a Set of Single-Locus Variants of Another Protein." In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 414–17. BCB '12. New York, NY, USA: Association for Computing Machinery.

Choi, Yongwook, Gregory E. Sims, Sean Murphy, Jason R. Miller, and Agnes P. Chan. 2012. "Predicting the Functional Effect of Amino Acid Substitutions and Indels." *PloS One* 7 (10): e46688.

Chow, Clement Y., John E. Landers, Sarah K. Bergren, Peter C. Sapp, Adrienne E. Grant, Julie M. Jones, Lesley Everett, et al. 2009. "Deleterious Variants of FIG4, a Phosphoinositide Phosphatase, in Patients with ALS." *American Journal of Human Genetics* 84 (1): 85–88.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3." *Fly* 6 (2): 80–92.

Cirulli, Elizabeth T., Brittany N. Lasseigne, Slavé Petrovski, Peter C. Sapp, Patrick A. Dion, Claire S. Leblond, Julien Couthouis, et al. 2015. "Exome Sequencing in Amyotrophic Lateral Sclerosis Identifies Risk Genes and Pathways." *Science* 347 (6229): 1436–41.

Ciura, Sorana, Serena Lattante, Isabelle Le Ber, Morwena Latouche, Hervé Tostivint, Alexis Brice, and Edor Kabashi. 2013. "Loss of Function of C9orf72 Causes Motor Deficits in a Zebrafish Model of Amyotrophic Lateral Sclerosis." *Annals of Neurology* 74 (2): 180–87.

Cooper-Knock, Johnathan, Pamela J. Shaw, and Janine Kirby. 2014. "The Widening Spectrum of C9ORF72-Related Disease; Genotype/Phenotype Correlations and Potential Modifiers of Clinical Phenotype." *Acta Neuropathologica* 127 (3): 333–45.

Cooper-Knock, Johnathan, Matthew J. Walsh, Adrian Higginbottom, J. Robin Highley, Mark J. Dickman, Dieter Edbauer, Paul G. Ince, et al. 2014. "Sequestration of Multiple RNA Recognition Motif-Containing Proteins by C9orf72 Repeat Expansions." *Brain: A Journal of Neurology* 137 (Pt 7): 2040–51.

Corbett, Mark A., Thessa Kroes, Liana Veneziano, Mark F. Bennett, Rahel Florian, Amy L. Schneider, Antonietta Coppola, et al. 2019. “Intronic ATTTC Repeat Expansions in STARD7 in Familial Adult Myoclonic Epilepsy Linked to Chromosome 2.” *Nature Communications* 10 (1): 4920.

Cortese, Andrea, Roberto Simone, Roisin Sullivan, Jana Vandrovcova, Huma Tariq, Wai Yan Yau, Jack Humphrey, et al. 2019. “Biallelic Expansion of an Intronic Repeat in RFC1 Is a Common Cause of Late-Onset Ataxia.” *Nature Genetics* 51 (4): 649–58.

Covanis, Athanasios, Alla Guekht, Shichuo Li, Mary Secco, Raad Shakir, and Emilio Perucca. 2015. “From Global Campaign to Global Commitment: The World Health Assembly’s Resolution on Epilepsy.” *Epilepsia* 56 (11): 1651–57.

Cox, Laura E., Laura Ferraiuolo, Emily F. Goodall, Paul R. Heath, Adrian Higginbottom, Heather Mortiboys, Hannah C. Hollinger, et al. 2010. “Mutations in CHMP2B in Lower Motor Neuron Predominant Amyotrophic Lateral Sclerosis (ALS).” *PloS One* 5 (3): e9872.

Cox, Paul Alan, and Oliver W. Sacks. 2002. “Cycad Neurotoxins, Consumption of Flying Foxes, and ALS-PDC Disease in Guam.” *Neurology* 58 (6): 956–59.

Coyle-Gilchrist, Ian T. S., Katrina M. Dick, Karalyn Patterson, Patricia Vázquez Rodríquez, Eileen Wehmann, Alicia Wilcox, Claire J. Lansdall, et al. 2016. “Prevalence, Characteristics, and Survival of Frontotemporal Lobar Degeneration Syndromes.” *Neurology* 86 (18): 1736–43.

Cruts, Marc, Ilse Gijselinck, Julie van der Zee, Sebastiaan Engelborghs, Hans Wils, Daniel Pirici, Rosa Rademakers, et al. 2006. “Null Mutations in Progranulin Cause Ubiquitin-Positive Frontotemporal Dementia Linked to Chromosome 17q21.” *Nature* 442 (7105): 920–24.

D’Amore, Angelica, Alessandra Tessa, Carlo Casali, Maria Teresa Dotti, Alessandro Filla, Gabriella Silvestri, Antonella Antenora, et al. 2018. “Next Generation Molecular Diagnosis of Hereditary Spastic Paraplegias: An Italian Cross-Sectional Study.” *Frontiers in Neurology* 9 (December): 981.

Daoud, Hussein, Véronique Belzil, Sandra Martins, Mike Sabbagh, Pierre Provencher, Lucette Lacomblez, Vincent Meininger, et al. 2011. “Association of Long ATXN2 CAG Repeat Sizes with Increased Risk of Amyotrophic Lateral Sclerosis.” *Archives of Neurology* 68 (6): 739–42.

Dashnow, Harriet, Monkol Lek, Belinda Phipson, Andreas Halman, Simon Sadedin, Andrew Lonsdale, Mark Davis, et al. 2018. “STRetch: Detecting and Discovering Pathogenic Short Tandem Repeat Expansions.” *Genome Biology* 19 (1): 121.

“Databank.Worldbank.Org.” 2021. Databank.Worldbank.Org. 2021. https://databank.worldbank.org/reports.aspx?source=2&series=SP.DYN.TFRT.IN.

Davey Smith, George, and Shah Ebrahim. 2003. “‘Mendelian Randomization’: Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease?” *International Journal of Epidemiology* 32 (1): 1–22.

De Baere, Elfride, Michael J. Dixon, Kent W. Small, Ethylin W. Jabs, Bart P. Leroy, Koenraad Devriendt, Yves Gillerot, et al. 2001. “Spectrum of FOXL2 Gene Mutations in Blepharophimosis-Ptosis-Epicanthus Inversus (BPES) Families Demonstrates a Genotype–Phenotype Correlation.” *Human Molecular Genetics* 10 (15): 1591–1600.

DeJesus-Hernandez, Mariely, Ian R. Mackenzie, Bradley F. Boeve, Adam L. Boxer, Matt Baker, Nicola J. Rutherford, Alexandra M. Nicholson, et al. 2011. “Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS.” *Neuron* 72 (2): 245–56.

Del Bo, R., C. Tiloca, V. Pensato, L. Corrado, A. Ratti, N. Ticozzi, S. Corti, et al. 2011. “Novel Optineurin Mutations in Patients with Familial and Sporadic Amyotrophic

Lateral Sclerosis." *Journal of Neurology, Neurosurgery, and Psychiatry* 82 (11): 1239–43.

Deng, Han-Xiang, Wenjie Chen, Seong-Tshool Hong, Kym M. Boycott, George H. Gorrie, Nailah Siddique, Yi Yang, et al. 2011. "Mutations in UBQLN2 Cause Dominant X-Linked Juvenile and Adult-Onset ALS and ALS/Dementia." *Nature* 477 (7363): 211–15.

Deng, Jianwen, Jiaxi Yu, Pidong Li, Xinghua Luan, Li Cao, Juan Zhao, Meng Yu, et al. 2020. "Expansion of GGC Repeat in GIPC1 Is Associated with Oculopharyngodistal Myopathy." *American Journal of Human Genetics* 106 (6): 793–804.

Depienne, Christel, and Jean-Louis Mandel. 2021. "30 Years of Repeat Expansion Disorders: What Have We Learned and What Are the Remaining Challenges?" *American Journal of Human Genetics* 108 (5): 764–85.

Dewan, Ramita, Ruth Chia, Jinhui Ding, Richard A. Hickman, Thor D. Stein, Yevgeniya Abramzon, Sarah Ahmed, et al. 2021. "Pathogenic Huntingtin Repeat Expansions in Patients with Frontotemporal Dementia and Amyotrophic Lateral Sclerosis." *Neuron* 109 (3): 448-460.e4.

Dick, Katherine J., Matthias Eckhardt, Coro Paisán-Ruiz, Aisha Alkhayat Alshehhi, Christos Proukakis, Naomi A. Sibtain, Helena Maier, et al. 2010. "Mutation of FA2H Underlies a Complicated Form of Hereditary Spastic Paraplegia (SPG35)." *Human Mutation* 31 (4): E1251-60.

Dols-Icardo, Oriol, Alberto García-Redondo, Ricardo Rojas-García, Daniel Borrego-Hernández, Ignacio Illán-Gala, José Luís Muñoz-Blanco, Alberto Rábano, et al. 2018. "Analysis of Known Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Genes Reveals a Substantial Genetic Burden in Patients Manifesting Both Diseases Not Carrying the C9orf72 Expansion Mutation." *Journal of Neurology, Neurosurgery, and Psychiatry* 89 (2): 162–68.

Dolzhenko, Egor, Mark F. Bennett, Phillip A. Richmond, Brett Trost, Sai Chen, Joke J. F. A. van Vugt, Charlotte Nguyen, et al. 2020. "ExpansionHunter Denovo: A Computational Method for Locating Known and Novel Repeat Expansions in Short-Read Sequencing Data." *Genome Biology* 21 (1): 102.

Dolzhenko, Egor, Viraj Deshpande, Felix Schlesinger, Peter Krusche, Roman Petrovski, Sai Chen, Dorothea Emig-Agius, et al. 2019. "ExpansionHunter: A Sequence-Graph-Based Tool to Analyze Variation in Short Tandem Repeat Regions." *Bioinformatics* 35 (22): 4754–56.

Dolzhenko, Egor, Joke J. F. A. van Vugt, Richard J. Shaw, Mitchell A. Bekritsky, Marka van Blitterswijk, Giuseppe Narzisi, Subramanian S. Ajay, et al. 2017. "Detection of Long Repeat Expansions from PCR-Free Whole-Genome Sequence Data." *Genome Research* 27 (11): 1895–1903.

Dolzhenko, Egor, Ben Weisburd, Kristina Ibanez Garikano, Indhu Shree Rajan Babu, Mark F. Bennett, Kimberley Billingsley, Ashley Carroll, et al. 2021. "REViewer: Haplotype-Resolved Visualization of Read Alignments in and around Tandem Repeats." *BioRxiv*. https://doi.org/10.1101/2021.10.20.465046.

Donnelly, Christopher J., Ping-Wu Zhang, Jacqueline T. Pham, Aaron R. Haeusler, Nipun A. Mistry, Svetlana Vidensky, Elizabeth L. Daley, et al. 2013. "RNA Toxicity from the ALS/FTD C9ORF72 Expansion Is Mitigated by Antisense Intervention." *Neuron* 80 (2): 415–28.

Dorai-Raj, Sundar. 2014. *Binom: Binomial Confidence Intervals For Several Parameterizations*. https://CRAN.R-project.org/package=binom.

Douville, Christopher, David L. Masica, Peter D. Stenson, David N. Cooper, Derek M. Gygax, Rick Kim, Michael Ryan, and Rachel Karchin. 2016. "Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel)." *Human Mutation* 37 (1): 28–35.

Dowle, Matt, and Arun Srinivasan. 2019. *Data.Table: Extension of `data.Frame`*. https://CRAN.R-project.org/package=data.table.

Dupré, Nicolas, Paul N. Valdmanis, Jean-Pierre Bouchard, and Guy A. Rouleau. 2007. "Autosomal Dominant Primary Lateral Sclerosis." *Neurology* 68 (14): 1156–57.

Eaglehouse, Yvonne L., Evelyn O. Talbott, Yuefang Chang, and Lewis H. Kuller. 2016. "Participation in Physical Activity and Risk for Amyotrophic Lateral Sclerosis Mortality Among Postmenopausal Women." *JAMA Neurology* 73 (3): 329–36.

Ebbert, Mark T. W., Stefan L. Farrugia, Jonathon P. Sens, Karen Jansen-West, Tania F. Gendron, Mercedes Prudencio, Ian J. McLaughlin, et al. 2018. "Long-Read Sequencing across the C9orf72 'GGGGCC' Repeat Expansion: Implications for Clinical Use and Genetic Discovery Efforts in Human Disease." *Molecular Neurodegeneration* 13 (1): 46.

Edgar, Ron, Michael Domrachev, and Alex E. Lash. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Research* 30 (1): 207–10.

Eklund, Aron, and James Trimble. 2021. "Beeswarm: The Bee Swarm Plot, an Alternative to Stripchart." https://CRAN.R-project.org/package=beeswarm.

Elden, Andrew C., Hyung-Jun Kim, Michael P. Hart, Alice S. Chen-Plotkin, Brian S. Johnson, Xiaodong Fang, Maria Armakola, et al. 2010. "Ataxin-2 Intermediate-Length Polyglutamine Expansions Are Associated with Increased Risk for ALS." *Nature* 466 (7310): 1069–75.

"EPACTS - Genome Analysis Wiki." n.d. Accessed February 10, 2019. https://genome.sph.umich.edu/wiki/EPACTS.

Erlich, Yaniv, Simon Edvardson, Emily Hodges, Shamir Zenvirt, Pramod Thekkat, Avraham Shaag, Talya Dor, Gregory J. Hannon, and Orly Elpeleg. 2011. "Exome Sequencing and Disease-Network Analysis of a Single Family Implicate a Mutation in KIF1A in Hereditary Spastic Paraparesis." *Genome Research* 21 (5): 658–64.

Es, Michael A. van, Jan H. Veldink, Christiaan G. J. Saris, Hylke M. Blauw, Paul W. J. van Vught, Anna Birve, Robin Lemmens, et al. 2009. "Genome-Wide Association Study Identifies 19p13.3 (UNC13A) and 9p21.2 as Susceptibility Loci for Sporadic Amyotrophic Lateral Sclerosis." *Nature Genetics* 41 (10): 1083–87.

Esteves, Typhaine, Alexandra Durr, Emeline Mundwiller, José L. Loureiro, Maxime Boutry, Michael A. Gonzalez, Julie Gauthier, et al. 2014. "Loss of Association of REEP2 with Membranes Leads to Hereditary Spastic Paraplegia." *The American Journal of Human Genetics* 94 (2): 268–77.

Fang, Ton, Ahmad Al Khleifat, Jacques-Henri Meurgey, Ashley Jones, P. Nigel Leigh, Gilbert Bensimon, and Ammar Al-Chalabi. 2018. "Stage at Which Riluzole Treatment Prolongs Survival in Patients with Amyotrophic Lateral Sclerosis: A Retrospective Analysis of Data from a Dose-Ranging Study." *Lancet Neurology* 17 (5): 416–22.

Fecto, Faisal, Jianhua Yan, S. Pavan Vemula, Erdong Liu, Yi Yang, Wenjie Chen, Jian Guo Zheng, et al. 2011. "SQSTM1 Mutations in Familial and Sporadic Amyotrophic Lateral Sclerosis." *Archives of Neurology* 68 (11): 1440–46.

Feliubadaló, Lidia, Alejandro Moles-Fernández, Marta Santamariña-Pena, Alysson T. Sánchez, Anael López-Novo, Luz-Marina Porras, Ana Blanco, et al. 2021. "A Collaborative Effort to Define Classification Criteria for ATM Variants in Hereditary Cancer Patients." *Clinical Chemistry* 67 (3): 518–33.

Ferrari, Raffaele, Dimitrios Kapogiannis, Edward D. Huey, Jordan Grafman, John Hardy, and Parastoo Momeni. 2010. "Novel Missense Mutation in Charged Multivesicular Body Protein 2B in a Patient with Frontotemporal Dementia." *Alzheimer Disease and Associated Disorders* 24 (4): 397–401.

Figlewicz, D. A., A. Krizus, M. G. Martinoli, V. Meininger, M. Dib, G. A. Rouleau, and J. P. Julien. 1994. “Variants of the Heavy Neurofilament Subunit Are Associated with the Development of Amyotrophic Lateral Sclerosis.” *Human Molecular Genetics* 3 (10): 1757–61.

Finkel, Richard S., Claudia A. Chiriboga, Jiri Vajsar, John W. Day, Jacqueline Montes, Darryl C. De Vivo, Kathie M. Bishop, et al. 2021. “Treatment of Infantile-Onset Spinal Muscular Atrophy with Nusinersen: Final Report of a Phase 2, Open-Label, Multicentre, Dose-Escalation Study.” *The Lancet. Child & Adolescent Health* 5 (7): 491–500.

Finkel, Richard S., Eugenio Mercuri, Basil T. Darras, Anne M. Connolly, Nancy L. Kuntz, Janbernd Kirschner, Claudia A. Chiriboga, et al. 2017. “Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy.” *The New England Journal of Medicine* 377 (18): 1723–32.

Finsterer, Josef, and Jean-Marc Burgunder. 2014. “Recent Progress in the Genetics of Motor Neuron Disease.” *European Journal of Medical Genetics* 57 (2–3): 103–12.

Florian, Rahel T., FAME consortium, Florian Kraft, Elsa Leitão, Sabine Kaya, Stephan Klebe, Eloi Magnin, et al. 2019. “Unstable TTTTA/TTTCA Expansions in MARCH6 Are Associated with Familial Adult Myoclonic Epilepsy Type 3.” *Nature Communications*. https://doi.org/10.1038/s41467-019-12763-9.

Fogh, Isabella, Antonia Ratti, Cinzia Gellera, Kuang Lin, Cinzia Tiloca, Valentina Moskvina, Lucia Corrado, et al. 2014. “A Genome-Wide Association Meta-Analysis Identifies a Novel Locus at 17q11.2 Associated with Sporadic Amyotrophic Lateral Sclerosis.” *Human Molecular Genetics* 23 (8): 2220–31.

Fortes-Lima, Cesar, Jonas Bybjerg-Grauholm, Lilia Caridad Marin-Padrón, Enrique Javier Gomez-Cabezas, Marie Bækvad-Hansen, Christine Søholm Hansen, Phuong Le, et al. 2018. “Exploring Cuba’s Population Structure and Demographic History Using Genome-Wide Data.” *Scientific Reports* 8 (1): 11422.

Fortuno, Cristina, Kristy Lee, Magali Olivier, Tina Pesaran, Phuong L. Mai, Kelvin C. de Andrade, Laura D. Attardi, et al. 2021. “Specifications of the ACMG/AMP Variant Interpretation Guidelines for Germline TP53 Variants.” *Human Mutation* 42 (3): 223–36.

Fratta, Pietro, Mark Poulter, Tammaryn Lashley, Jonathan D. Rohrer, James M. Polke, Jon Beck, Natalie Ryan, et al. 2013. “Homozygosity for the C9orf72 GGGGCC Repeat Expansion in Frontotemporal Dementia.” *Acta Neuropathologica* 126 (3): 401–9.

Freischmidt, Axel, Thomas Wieland, Benjamin Richter, Wolfgang Ruf, Veronique Schaeffer, Kathrin Müller, Nicolai Marroquin, et al. 2015. “Haploinsufficiency of TBK1 Causes Familial ALS and Fronto-Temporal Dementia.” *Nature Neuroscience* 18 (5): 631–36.

Gallo, Maura, Carmine Tomaino, Gianfranco Puccio, Francesca Frangipane, Sabrina A. M. Curcio, Livia Bernardi, Silvana Geracitano, et al. 2010. “Novel MAPT Val75Ala Mutation and PSEN2 Arg62Hys in Two Siblings with Frontotemporal Dementia.” *Neurological Sciences: Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology* 31 (1): 65–70.

Gallo, Valentina, Nicola Vanacore, H. Bas Bueno-de-Mesquita, Roel Vermeulen, Carol Brayne, Neil Pearce, Petra A. Wark, et al. 2016. “Physical Activity and Risk of Amyotrophic Lateral Sclerosis in a Prospective Cohort Study.” *European Journal of Epidemiology* 31 (3): 255–66.

García-Redondo, Alberto, Oriol Dols-Icardo, Ricard Rojas-García, Jesús Esteban-Pérez, Pilar Cordero-Vázquez, José Luis Muñoz-Blanco, Irene Catalina, et al. 2013. “Analysis of the C9orf72 Gene in Patients with Amyotrophic Lateral Sclerosis in Spain and Different Populations Worldwide.” *Human Mutation* 34 (1): 79–82.

GBD 2016 Motor Neuron Disease Collaborators. 2018. “Global, Regional, and National Burden of Motor Neuron Diseases 1990-2016: A Systematic Analysis for the Global Burden of Disease Study 2016.” *Lancet Neurology* 17 (12): 1083–97.

Gellera, Cinzia, Nicola Ticozzi, Viviana Pensato, Lorenzo Nanetti, Alessia Castucci, Barbara Castellotti, Giuseppe Lauria, Franco Taroni, Vincenzo Silani, and Caterina Mariotti. 2012. “ATAXIN2 CAG-Repeat Length in Italian Patients with Amyotrophic Lateral Sclerosis: Risk Factor or Variant Phenotype? Implication for Genetic Testing and Counseling.” *Neurobiology of Aging* 33 (8): 1847.e15-21.

Gendron, Tania F., Kevin F. Bieniek, Yong-Jie Zhang, Karen Jansen-West, Peter E. A. Ash, Thomas Caulfield, Lillian Daughrity, et al. 2013. “Antisense Transcripts of the Expanded C9ORF72 Hexanucleotide Repeat Form Nuclear RNA Foci and Undergo Repeat-Associated Non-ATG Translation in C9FTD/ALS.” *Acta Neuropathologica* 126 (6): 829–44.

Ghani, Mahdi, Anthony E. Lang, Lorne Zinman, Benedetta Nacmias, Sandro Sorbi, Valentina Bessi, Andrea Tedde, et al. 2015. “Mutation Analysis of Patients with Neurodegenerative Disorders Using NeuroX Array.” *Neurobiology of Aging* 36 (1): 545.e9-14.

Ghasemi, Mehdi, and Robert H. Brown Jr. 2018. “Genetics of Amyotrophic Lateral Sclerosis.” *Cold Spring Harbor Perspectives in Medicine* 8 (5). https://doi.org/10.1101/cshperspect.a024125.

Giannoccaro, Maria Pia, Anna Bartoletti-Stella, Silvia Piras, Annalisa Pession, Patrizia De Massis, Federico Oppi, Michelangelo Stanzani-Maserati, et al. 2017. “Multiple Variants in Families with Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Related to C9orf72 Repeat Expansion: Further Observations on Their Oligogenic Nature.” *Journal of Neurology* 264 (7): 1426–33.

Gijselinck, Ilse, Tim Van Langenhove, Julie van der Zee, Kristel Sleegers, Stéphanie Philtjens, Gernot Kleinberger, Jonathan Janssens, et al. 2012. “A C9orf72 Promoter Repeat Expansion in a Flanders-Belgian Cohort with Disorders of the Frontotemporal Lobar Degeneration-Amyotrophic Lateral Sclerosis Spectrum: A Gene Identification Study.” *Lancet Neurology* 11 (1): 54–65.

Goldman, J. S., J. M. Farmer, E. M. Wood, J. K. Johnson, A. Boxer, J. Neuhaus, C. Lomen-Hoerth, et al. 2005. “Comparison of Family Histories in FTLD Subtypes and Related Tauopathies.” *Neurology* 65 (11): 1817–19.

Goodman, F. R., C. Bacchelli, A. F. Brady, L. A. Brueton, J. P. Fryns, D. P. Mortlock, J. W. Innis, et al. 2000. “Novel HOXA13 Mutations and the Phenotypic Spectrum of Hand-Foot-Genital Syndrome.” *American Journal of Human Genetics* 67 (1): 197–202.

Gordon, P. H., B. Cheng, I. B. Katz, M. Pinto, A. P. Hays, H. Mitsumoto, and L. P. Rowland. 2006. “The Natural History of Primary Lateral Sclerosis.” *Neurology* 66 (5): 647–53.

Greenway, M. J., M. D. Alexander, S. Ennis, B. J. Traynor, B. Corr, E. Frost, A. Green, and O. Hardiman. 2004. “A Novel Candidate Region for ALS on Chromosome 14q11.2.” *Neurology* 63 (10): 1936–38.

Gros-Louis, François, Roxanne Larivière, Geneviève Gowing, Sandra Laurent, William Camu, Jean-Pierre Bouchard, Vincent Meininger, Guy A. Rouleau, and Jean-Pierre Julien. 2004. “A Frameshift Deletion in Peripherin Gene Associated with Amyotrophic Lateral Sclerosis.” *The Journal of Biological Chemistry* 279 (44): 45951–56.

Grosso, Valentina, Luca Marcolungo, Simone Maestri, Massimiliano Alfano, Denise Lavezzari, Barbara Iadarola, Alessandro Salviati, et al. 2021. “Characterization of FMR1 Repeat Expansion and Intragenic Variants by Indirect Sequence Capture.” *Frontiers in Genetics* 12 (September): 743230.

Gymrek, Melissa. 2017. “A Genomic View of Short Tandem Repeats.” *Current Opinion in Genetics & Development* 44 (June): 9–16.

Hadano, S., C. K. Hand, H. Osuga, Y. Yanagisawa, A. Otomo, R. S. Devon, N. Miyamoto, et al. 2001. “A Gene Encoding a Putative GTPase Regulator Is Mutated in Familial Amyotrophic Lateral Sclerosis 2.” *Nature Genetics* 29 (2): 166–73.

Hadley, Wickham, and D. Seidel. 2019. “Scales: Scale Functions for Visualization.” GitHub San Francisco: https://CRAN.R-project.org/package=scales.

Halman, Andreas, and Alicia Oshlack. 2020. “Accuracy of Short Tandem Repeats Genotyping Tools in Whole Exome Sequencing Data.” *F1000Research* 9 (March): 200.

Hanagasi, Hasmet A., Anamika Giri, Ece Kartal, Gamze Guven, Başar Bilgiç, Ann-Kathrin Hauser, Murat Emre, et al. 2016. “A Novel Homozygous DJ1 Mutation Causes Parkinsonism and ALS in a Turkish Family.” *Parkinsonism & Related Disorders* 29 (August): 117–20.

Hanein, Sylvain, Alexandra Dürr, Pascale Ribai, Sylvie Forlani, Anne-Louise Leutenegger, Isabelle Nelson, Marie-Claude Babron, et al. 2007. “A Novel Locus for Autosomal Dominant ‘Uncomplicated’ Hereditary Spastic Paraplegia Maps to Chromosome 8p21.1-Q13.3.” *Human Genetics* 122 (3–4): 261–73.

Harel, Tamar, Wan Hee Yoon, Caterina Garone, Shen Gu, Zeynep Coban-Akdemir, Mohammad K. Eldomery, Jennifer E. Posey, et al. 2016. “Recurrent DE Novo and Biallelic Variation of ATAD3A, Encoding a Mitochondrial Membrane Protein, Results in Distinct Neurological Syndromes.” *The American Journal of Human Genetics* 99 (4): 831–45.

Harrison, Steven M., Leslie G. Biesecker, and Heidi L. Rehm. 2019. “Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines.” *Et al [Current Protocols in Human Genetics]* 103 (1): e93.

Harwood, Ceryl A., Kate Westgate, Sue Gunstone, Soren Brage, Nicholas J. Wareham, Christopher J. McDermott, and Pamela J. Shaw. 2016. “Long-Term Physical Activity: An Exogenous Risk Factor for Sporadic Amyotrophic Lateral Sclerosis?” *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration* 17 (5–6): 377–84.

Hauge, X. Y., and M. Litt. 1993. “A Study of the Origin of ‘shadow Bands’ Seen When Typing Dinucleotide Repeat Polymorphisms by the PCR.” *Human Molecular Genetics* 2 (4): 411–15.

Hazan, J., N. Fonknechten, D. Mavel, C. Paternotte, D. Samson, F. Artiguenave, C. S. Davoine, et al. 1999. “Spastin, a New AAA Protein, Is Altered in the Most Frequent Form of Autosomal Dominant Spastic Paraplegia.” *Nature Genetics* 23 (3): 296–303.

Hentati, A., H. X. Deng, H. Zhai, W. Chen, Y. Yang, W. Y. Hung, A. C. Azim, et al. 2000. “Novel Mutations in Spastin Gene and Absence of Correlation with Age at Onset of Symptoms.” *Neurology* 55 (9): 1388–90.

Hewitt, Christopher, Janine Kirby, J. Robin Highley, Judith A. Hartley, Rachael Hibberd, Hannah C. Hollinger, Tim L. Williams, Paul G. Ince, Christopher J. McDermott, and Pamela J. Shaw. 2010. “Novel FUS/TLS Mutations and Pathology in Familial and Sporadic Amyotrophic Lateral Sclerosis.” *Archives of Neurology* 67 (4): 455–61.

Higgins, Julian P. T., and Simon G. Thompson. 2002. “Quantifying Heterogeneity in a Meta-Analysis.” *Statistics in Medicine* 21 (11): 1539–58.

Higgins, Julian P. T., Simon G. Thompson, Jonathan J. Deeks, and Douglas G. Altman. 2003. “Measuring Inconsistency in Meta-Analyses.” *BMJ* 327 (7414): 557–60.

Highnam, Gareth, Christopher Franck, Andy Martin, Calvin Stephens, Ashwin Puthige, and David Mittelman. 2013. “Accurate Human Microsatellite Genotypes from High-Throughput Resequencing Data Using Informed Error Profiles.” *Nucleic Acids Research* 41 (1): e32.

Hijmans, Robert J. 2021. “Raster: Geographic Data Analysis and Modeling.” https://CRAN.R-project.org/package=raster.

Hildebrand, Michael S., Hans-Henrik M. Dahl, John Anthony Damiano, Richard J. H. Smith, Ingrid E. Scheffer, and Samuel F. Berkovic. 2013. “Recent Advances in the Molecular Genetics of Epilepsy.” *Journal of Medical Genetics* 50 (5): 271–79.

Hirayanagi, Kimitoshi, Masayuki Sato, Natsumi Furuta, Kouki Makioka, and Yoshio Ikeda. 2016. “Juvenile-Onset Sporadic Amyotrophic Lateral Sclerosis with a Frameshift FUS Gene Mutation Presenting Unique Neuroradiological Findings and Cognitive Impairment.” *Internal Medicine* 55 (6): 689–93.

Holm, Ida Elisabeth, Elisabet Englund, Ian R. A. Mackenzie, Peter Johannsen, and Adrian M. Isaacs. 2007. “A Reassessment of the Neuropathology of Frontotemporal Dementia Linked to Chromosome 3.” *Journal of Neuropathology and Experimental Neurology* 66 (10): 884–91.

Holmes, Susan E., Elizabeth E. O’Hearn, Melvin G. McInnis, Daniel A. Gorelick-Feldman, John J. Kleiderlein, Colleen Callahan, Noeun G. Kwak, et al. 1999. “Expansion of a Novel CAG Trinucleotide Repeat in the 5′ Region of PPP2R2B Is Associated with SCA12.” *Nature Genetics* 23 (4): 391–92.

Hübers, Annemarie, Nicolai Marroquin, Birgit Schmoll, Stefan Vielhaber, Marlies Just, Benjamin Mayer, Josef Högel, et al. 2014. “Polymerase Chain Reaction and Southern Blot-Based Analysis of the C9orf72 Hexanucleotide Repeat in Different Motor Neuron Diseases.” *Neurobiology of Aging* 35 (5): 1214.e1-6.

Huisman, Mark H. B., Meinie Seelen, Sonja W. de Jong, Kirsten R. I. S. Dorresteijn, Perry T. C. van Doormaal, Anneke J. van der Kooi, Marianne de Visser, Helenius Jurgen Schelhaas, Leonard H. van den Berg, and Jan Herman Veldink. 2013. “Lifetime Physical Activity and the Risk of Amyotrophic Lateral Sclerosis.” *Journal of Neurology, Neurosurgery, and Psychiatry* 84 (9): 976–81.

Hutton, M., C. L. Lendon, P. Rizzu, M. Baker, S. Froelich, H. Houlden, S. Pickering-Brown, et al. 1998. “Association of Missense and 5’-Splice-Site Mutations in Tau with the Inherited Dementia FTDP-17.” *Nature* 393 (6686): 702–5.

Ingre, Caroline, Per M. Roos, Fredrik Piehl, Freya Kamel, and Fang Fang. 2015. “Risk Factors for Amyotrophic Lateral Sclerosis.” *Clinical Epidemiology* 7 (February): 181–93.

International HapMap Consortium. 2003. “The International HapMap Project.” *Nature* 426 (6968): 789–96.

Ishiura, Hiroyuki, Koichiro Doi, Jun Mitsui, Jun Yoshimura, Miho Kawabe Matsukawa, Asao Fujiyama, Yasuko Toyoshima, et al. 2018. “Expansions of Intronic TTTCA and TTTTA Repeats in Benign Adult Familial Myoclonic Epilepsy.” *Nature Genetics* 50 (4): 581–90.

Ishiura, Hiroyuki, Shota Shibata, Jun Yoshimura, Yuta Suzuki, Wei Qu, Koichiro Doi, M. Asem Almansour, et al. 2019. “Noncoding CGG Repeat Expansions in Neuronal Intranuclear Inclusion Disease, Oculopharyngodistal Myopathy and an Overlapping Disease.” *Nature Genetics* 51 (8): 1222–32.

Ismail, Azza, Johnathan Cooper-Knock, J. Robin Highley, Antonio Milano, Janine Kirby, Emily Goodall, James Lowe, et al. 2013. “Concurrence of Multiple Sclerosis and Amyotrophic Lateral Sclerosis in Patients with Hexanucleotide Repeat Expansions of C9ORF72.” *Journal of Neurology, Neurosurgery, and Psychiatry* 84 (1): 79–87.

Jarvik, Gail P., and Brian L. Browning. 2016. “Consideration of Cosegregation in the Pathogenicity Classification of Genomic Variants.” *American Journal of Human Genetics* 98 (6): 1077–81.

Jedrzejowska, Maria, Michal Milewski, Janusz Zimowski, Pawel Zagozdzon, Anna Kostera-Pruszczyk, Janina Borkowska, Danuta Sielska, Marta Jurek, and Irena Hausmanowa-

Petrusewicz. 2010. “Incidence of Spinal Muscular Atrophy in Poland--More Frequent than Predicted?” *Neuroepidemiology* 34 (3): 152–57.

Jian, Xueqiu, Eric Boerwinkle, and Xiaoming Liu. 2014. “In Silico Prediction of Splice-Altering Single Nucleotide Variants in the Human Genome.” *Nucleic Acids Research* 42 (22): 13534–44.

Johnson, Janel O., Jessica Mandrioli, Michael Benatar, Yevgeniya Abramzon, Vivianna M. Van Deerlin, John Q. Trojanowski, J. Raphael Gibbs, et al. 2010. “Exome Sequencing Reveals VCP Mutations as a Cause of Familial ALS.” *Neuron* 68 (5): 857–64.

Johnson, Janel O., Erik P. Pioro, Ashley Boehringer, Ruth Chia, Howard Feit, Alan E. Renton, Hannah A. Pliner, et al. 2014. “Mutations in the Matrin 3 Gene Cause Familial Amyotrophic Lateral Sclerosis.” *Nature Neuroscience* 17 (5): 664–66.

Johnston, Clare A., Biba R. Stanton, Martin R. Turner, Rebecca Gray, Ashley Hay-Ming Blunt, David Butt, Mary-Ann Ampong, Christopher E. Shaw, P. Nigel Leigh, and Ammar Al-Chalabi. 2006. “Amyotrophic Lateral Sclerosis in an Urban Setting: A Population Based Study of Inner City London.” *Journal of Neurology* 253 (12): 1642–43.

Jones, C., L. Penny, T. Mattina, S. Yu, E. Baker, L. Voullaire, W. Y. Langdon, G. R. Sutherland, R. I. Richards, and A. Tunnacliffe. 1995. “Association of a Chromosome Deletion Syndrome with a Fragile Site within the Proto-Oncogene CBL2.” *Nature* 376 (6536): 145–49.

Jouet, M., A. Rosenthal, G. Armstrong, J. MacFarlane, R. Stevenson, J. Paterson, A. Metzenberg, V. Ionasescu, K. Temple, and S. Kenwrick. 1994. “X-Linked Spastic Paraplegia (SPG1), MASA Syndrome and X-Linked Hydrocephalus Result from Mutations in the L1 Gene.” *Nature Genetics* 7 (3): 402–7.

Julian, Thomas H., Sarah Boddy, Mahjabin Islam, Julian Kurz, Katherine J. Whittaker, Tobias Moll, Calum Harvey, et al. 2021. “A Review of Mendelian Randomization in Amyotrophic Lateral Sclerosis.” *Brain: A Journal of Neurology*, November. https://doi.org/10.1093/brain/awab420.

Kanadia, Rahul N., Karen A. Johnstone, Ami Mankodi, Codrin Lungu, Charles A. Thornton, Douglas Esson, Adrian M. Timmers, William W. Hauswirth, and Maurice S. Swanson. 2003. “A Muscleblind Knockout Model for Myotonic Dystrophy.” *Science* 302 (5652): 1978–80.

Kancheva, Dahlia, Teodora Chamova, Velina Guergueltcheva, Vanio Mitev, Dimitar N. Azmanov, Luba Kalaydjieva, Ivailo Tournev, and Albena Jordanova. 2015. “Mosaic Dominant TUBB4A Mutation in an Inbred Family with Complicated Hereditary Spastic Paraplegia.” *Movement Disorders: Official Journal of the Movement Disorder Society* 30 (6): 854–58.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

Kawaguchi, Y., T. Okamoto, M. Taniwaki, M. Aizawa, M. Inoue, S. Katayama, H. Kawakami, S. Nakamura, M. Nishimura, and I. Akiguchi. 1994. “CAG Expansions in a Novel Gene for Machado-Joseph Disease at Chromosome 14q32.1.” *Nature Genetics* 8 (3): 221–28.

Kelly, Melissa A., Colleen Caleshu, Ana Morales, Jillian Buchan, Zena Wolf, Steven M. Harrison, Stuart Cook, et al. 2018. “Adaptation and Validation of the ACMG/AMP Variant Classification Framework for MYH7-Associated Inherited Cardiomyopathies: Recommendations by ClinGen’s Inherited Cardiomyopathy Expert Panel.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 20 (3): 351–59.

Kenna, Kevin P., Perry T. C. van Doormaal, Annelot M. Dekker, Nicola Ticozzi, Brendan J. Kenna, Frank P. Diekstra, Wouter van Rheenen, et al. 2016. "*NEK1* Variants Confer Susceptibility to Amyotrophic Lateral Sclerosis." *Nature Genetics* 48 (9): 1037–42.

Kenna, Kevin P., Russell L. McLaughlin, Susan Byrne, Marwa Elamin, Mark Heverin, Elaine M. Kenny, Paul Cormican, et al. 2013. "Delineating the Genetic Heterogeneity of ALS Using Targeted High-Throughput Sequencing." *Journal of Medical Genetics* 50 (11): 776–83.

Kenna, Kevin P., Russell L. McLaughlin, Orla Hardiman, and Daniel G. Bradley. 2013. "Using Reference Databases of Genetic Variation to Evaluate the Potential Pathogenicity of Candidate Disease Variants." *Human Mutation* 34 (6): 836–41.

Kent, Louisa, Thomas N. Vizard, Bradley N. Smith, Simon D. Topp, Caroline Vance, Athina Gkazi, Jack Miller, Christopher E. Shaw, and Kevin Talbot. 2014. "Autosomal Dominant Inheritance of Rapidly Progressive Amyotrophic Lateral Sclerosis Due to a Truncation Mutation in the Fused in Sarcoma (FUS) Gene." *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration* 15 (7–8): 557–62.

Kent, W. James, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. 2002. "The Human Genome Browser at UCSC." *Genome Research* 12 (6): 996–1006.

Kent-Braun, J. A., C. H. Walker, M. W. Weiner, and R. G. Miller. 1998. "Functional Significance of Upper and Lower Motor Neuron Impairment in Amyotrophic Lateral Sclerosis." *Muscle & Nerve* 21 (6): 762–68.

Khristich, Alexandra N., and Sergei M. Mirkin. 2020. "On the Wrong DNA Track: Molecular Mechanisms of Repeat-Mediated Genome Instability." *The Journal of Biological Chemistry* 295 (13): 4134–70.

Kim, Hong Joo, Nam Chul Kim, Yong-Dong Wang, Emily A. Scarborough, Jennifer Moore, Zamia Diaz, Kyle S. MacLea, et al. 2013. "Mutations in Prion-like Domains in HnRNPA2B1 and HnRNPA1 Cause Multisystem Proteinopathy and ALS." *Nature* 495 (7442): 467–73.

Kim, W-K, X. Liu, J. Sandner, M. Pasmantier, J. Andrews, L. P. Rowland, and H. Mitsumoto. 2009. "Study of 962 Patients Indicates Progressive Muscular Atrophy Is a Form of ALS." *Neurology* 73 (20): 1686–92.

Kim, Young-Eun, Ki-Wook Oh, Min-Jung Kwon, Won-Jun Choi, Seong-Il Oh, Chang-Seok Ki, and Seung Hyun Kim. 2015. "De Novo FUS Mutations in 2 Korean Patients with Sporadic Amyotrophic Lateral Sclerosis." *Neurobiology of Aging* 36 (3): 1604.e17-9.

Klebe, S., G. Stevanin, and C. Depienne. 2015. "Clinical and Genetic Heterogeneity in Hereditary Spastic Paraplegias: From SPG1 to SPG72 and Still Counting." *Revue Neurologique* 171 (6–7): 505–30.

Knight, S. J., A. V. Flannery, M. C. Hirst, L. Campbell, Z. Christodoulou, S. R. Phelps, J. Pointon, H. R. Middleton-Price, A. Barnicoat, and M. E. Pembrey. 1993. "Trinucleotide Repeat Amplification and Hypermethylation of a CpG Island in FRAXE Mental Retardation." *Cell* 74 (1): 127–34.

Kobayashi, Hatasu, Koji Abe, Tohru Matsuura, Yoshio Ikeda, Toshiaki Hitomi, Yuji Akechi, Toshiyuki Habu, Wanyang Liu, Hiroko Okuda, and Akio Koizumi. 2011. "Expansion of Intronic GGCCTG Hexanucleotide Repeat in NOP56 Causes SCA36, a Type of Spinocerebellar Ataxia Accompanied by Motor Neuron Involvement." *American Journal of Human Genetics* 89 (1): 121–30.

Koide, R., T. Ikeuchi, O. Onodera, H. Tanaka, S. Igarashi, K. Endo, H. Takahashi, R. Kondo, A. Ishikawa, and T. Hayashi. 1994. "Unstable Expansion of CAG Repeat in Hereditary Dentatorubral-Pallidoluysian Atrophy (DRPLA)." *Nature Genetics* 6 (1): 9–13.

Koide, R., S. Kobayashi, T. Shimohata, T. Ikeuchi, M. Maruyama, M. Saito, M. Yamada, H. Takahashi, and S. Tsuji. 1999. “A Neurological Disease Caused by an Expanded CAG Trinucleotide Repeat in the TATA-Binding Protein Gene: A New Polyglutamine Disease?” *Human Molecular Genetics* 8 (11): 2047–53.

Koob, M. D., M. L. Moseley, L. J. Schut, K. A. Benzow, T. D. Bird, J. W. Day, and L. P. Ranum. 1999. “An Untranslated CTG Expansion Causes a Novel Form of Spinocerebellar Ataxia (SCA8).” *Nature Genetics* 21 (4): 379–84.

Koppers, Max, Anna M. Blokhuis, Henk-Jan Westeneng, Margo L. Terpstra, Caroline A. C. Zundel, Renata Vieira de Sá, Raymond D. Schellevis, et al. 2015. “C9orf72 Ablation in Mice Does Not Cause Motor Neuron Degeneration or Motor Deficits.” *Annals of Neurology* 78 (3): 426–38.

Kovach, M. J., B. Waggoner, S. M. Leal, D. Gelber, R. Khardori, M. A. Levenstien, C. A. Shanks, et al. 2001. “Clinical Delineation and Localization to Chromosome 9p13.3-P12 of a Unique Dominant Disorder in Four Families: Hereditary Inclusion Body Myopathy, Paget Disease of Bone, and Frontotemporal Dementia.” *Molecular Genetics and Metabolism* 74 (4): 458–75.

Kozarewa, Iwanka, Zemin Ning, Michael A. Quail, Mandy J. Sanders, Matthew Berriman, and Daniel J. Turner. 2009. “Amplification-Free Illumina Sequencing-Library Preparation Facilitates Improved Mapping and Assembly of (G+C)-Biased Genomes.” *Nature Methods* 6 (4): 291–95.

Kuilenburg, André B. P. van, Maja Tarailo-Graovac, Phillip A. Richmond, Britt I. Drögemöller, Mahmoud A. Pouladi, René Leen, Koroboshka Brand-Arzamendi, et al. 2019. “Glutaminase Deficiency Caused by Short Tandem Repeat Expansion in GLS.” *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa1806627.

Kumar, Kishore R., Nicholas F. Blair, Himesha Vandebona, Christina Liang, Karl Ng, David M. Sharpe, Anne Grünewald, et al. 2013. “Targeted next Generation Sequencing in SPAST-Negative Hereditary Spastic Paraplegia.” *Journal of Neurology* 260 (10): 2516–22.

Kuuluvainen, Liina, Karri Kaivola, Saana Mönkäre, Hannu Laaksovirta, Manu Jokela, Bjarne Udd, Miko Valori, et al. 2019. “Oligogenic Basis of Sporadic ALS: The Example of SOD1 p.Ala90Val Mutation.” *Neurology. Genetics* 5 (3): e335.

Kwiatkowski, T. J., Jr, D. A. Bosco, A. L. Leclerc, E. Tamrazian, C. R. Vanderburg, C. Russ, A. Davis, et al. 2009. “Mutations in the FUS/TLS Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis.” *Science* 323 (5918): 1205–8.

Kwon, Min-Jung, Wonki Baek, Chang-Seok Ki, Hyun Young Kim, Seong-Ho Koh, Jong-Won Kim, and Seung Hyun Kim. 2012. “Screening of the SOD1, FUS, TARDBP, ANG, and OPTN Mutations in Korean Patients with Familial and Sporadic ALS.” *Neurobiology of Aging* 33 (5): 1017.e17-23.

La Spada, A. R., E. M. Wilson, D. B. Lubahn, A. E. Harding, and K. H. Fischbeck. 1991. “Androgen Receptor Gene Mutations in X-Linked Spinal and Bulbar Muscular Atrophy.” *Nature* 352 (6330): 77–79.

Laaksovirta, Hannu, Terhi Peuralinna, Jennifer C. Schymick, Sonja W. Scholz, Shaoi-Lin Lai, Liisa Myllykangas, Raimo Sulkava, et al. 2010. “Chromosome 9p21 in Amyotrophic Lateral Sclerosis in Finland: A Genome-Wide Association Study.” *Lancet Neurology* 9 (10): 978–85.

Lagier-Tourenne, Clotilde, Michael Baughn, Frank Rigo, Shuying Sun, Patrick Liu, Hai-Ri Li, Jie Jiang, et al. 2013. “Targeted Degradation of Sense and Antisense C9orf72 RNA Foci as Therapy for ALS and Frontotemporal Degeneration.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (47): E4530-9.

Lagorio, Ilaria, Federico Zara, Salvatore Striano, and Pasquale Striano. 2019. “Familial Adult Myoclonic Epilepsy: A New Expansion Repeats Disorder.” *Seizure: The Journal of the British Epilepsy Association* 67 (April): 73–77.

Lalioti, M. D., M. Mirotsou, C. Buresi, M. C. Peitsch, C. Rossier, R. Ouazzani, M. Baldy-Moulinier, A. Bottani, A. Malafosse, and S. E. Antonarakis. 1997. “Identification of Mutations in Cystatin B, the Gene Responsible for the Unverricht-Lundborg Type of Progressive Myoclonus Epilepsy (EPM1).” *American Journal of Human Genetics* 60 (2): 342–51.

Lamp, Merit, Paola Origone, Alessandro Geroldi, Simonetta Verdiani, Fabio Gotta, Claudia Caponnetto, Grazia Devigili, et al. 2018. “Twenty Years of Molecular Analyses in Amyotrophic Lateral Sclerosis: Genetic Landscape of Italian Patients.” *Neurobiology of Aging* 66 (June): 179.e5-179.e16.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921.

Landouré, Guida, Peng-Peng Zhu, Charles M. Lourenço, Janel O. Johnson, Camilo Toro, Katherine V. Bricceno, Carlo Rinaldi, et al. 2013. “Hereditary Spastic Paraplegia Type 43 (SPG43) Is Caused by Mutation in C19orf12.” *Human Mutation* 34 (10): 1357–60.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. “ClinVar: Improving Access to Variant Interpretations and Supporting Evidence.” *Nucleic Acids Research* 46 (D1): D1062–67.

Lashley, Tammaryn, Jonathan D. Rohrer, Colin Mahoney, Elizabeth Gordon, Jon Beck, Simon Mead, Jason Warren, Martin Rossor, and Tamas Revesz. 2014. “A Pathogenic Progranulin Mutation and C9orf72 Repeat Expansion in a Family with Frontotemporal Dementia.” *Neuropathology and Applied Neurobiology* 40 (4): 502–13.

Lattante, Serena, Maria Grazia Pomponi, Amelia Conte, Giuseppe Marangi, Giulia Bisogni, Agata Katia Patanella, Emiliana Meleo, et al. 2018. “ATXN1 Intermediate-Length Polyglutamine Expansions Are Associated with Amyotrophic Lateral Sclerosis.” *Neurobiology of Aging* 64 (April): 157.e1-157.e5.

Laumonnier, Frédéric, Nathalie Ronce, Ben C. J. Hamel, Paul Thomas, James Lespinasse, Martine Raynaud, Christine Paringaux, et al. 2002. “Transcription Factor SOX3 Is Involved in X-Linked Mental Retardation with Growth Hormone Deficiency.” *American Journal of Human Genetics* 71 (6): 1450–55.

Lawson, Raef. 2004. “Small Sample Confidence Intervals for the Odds Ratio.” *Communications in Statistics - Simulation and Computation* 33 (4): 1095–1113.

Lee, Youn-Bok, Han-Jou Chen, João N. Peres, Jorge Gomez-Deza, Jan Attig, Maja Stalekar, Claire Troakes, et al. 2013. “Hexanucleotide Repeats in ALS/FTD Form Length-Dependent RNA Foci, Sequester RNA Binding Proteins, and Are Neurotoxic.” *Cell Reports* 5 (5): 1178–86.

Lehman, Everett J., Misty J. Hein, Sherry L. Baron, and Christine M. Gersic. 2012. “Neurodegenerative Causes of Death among Retired National Football League Players.” *Neurology* 79 (19): 1970–74.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Li, Miaoxin, Philip Wing-Lok Ho, Shirley Yin-Yu Pang, Zero Ho-Man Tse, Michelle Hiu-Wai Kung, Pak-Chung Sham, and Shu-Leong Ho. 2014. “PMCA4 (ATP2B4) Mutation in Familial Spastic Paraplegia.” *PloS One* 9 (8): e104790.

Lin, Pengfei, Jianwei Li, Qiji Liu, Fei Mao, Jisheng Li, Rongfang Qiu, Huili Hu, et al. 2008. “A Missense Mutation in SLC33A1, Which Encodes the Acetyl-CoA Transporter, Causes Autosomal-Dominant Spastic Paraplegia (SPG42).” *The American Journal of Human Genetics* 83 (6): 752–59.

Lindblad, K., M. L. Savontaus, G. Stevanin, M. Holmberg, K. Digre, C. Zander, H. Ehrsson, et al. 1996. “An Expanded CAG Repeat Sequence in Spinocerebellar Ataxia Type 7.” *Genome Research* 6 (10): 965–71.

Liquori, C. L., K. Ricker, M. L. Moseley, J. F. Jacobsen, W. Kress, S. L. Naylor, J. W. Day, and L. P. Ranum. 2001. “Myotonic Dystrophy Type 2 Caused by a CCTG Expansion in Intron 1 of ZNF9.” *Science* 293 (5531): 864–67.

Liu, Fang, Qing Liu, Chao Xia Lu, Bo Cui, Xia Nan Guo, Rong Rong Wang, Ming Sheng Liu, Xiao Guang Li, Li-Ying Cui, and Xue Zhang. 2016. “Identification of a Novel Loss-of-Function C9orf72 Splice Site Mutation in a Patient with Amyotrophic Lateral Sclerosis.” *Neurobiology of Aging* 47 (November): 219.e1-219.e5.

Liu, Qing, Shi Shu, Rong Rong Wang, Fang Liu, Bo Cui, Xia Nan Guo, Chao Xia Lu, et al. 2016. “Whole-Exome Sequencing Identifies a Missense Mutation in HnRNPA1 in a Family with Flail Arm ALS.” *Neurology* 87 (17): 1763–69.

Liu, Xiaoming, Xueqiu Jian, and Eric Boerwinkle. 2011. “DbNSFP: A Lightweight Database of Human Nonsynonymous SNPs and Their Functional Predictions.” *Human Mutation* 32 (8): 894–99.

———. 2013. “DbNSFP v2.0: A Database of Human Non-Synonymous SNVs and Their Functional Predictions and Annotations.” *Human Mutation* 34 (9): E2393-402.

Liu, Xiaoming, Chunlei Wu, Chang Li, and Eric Boerwinkle. 2016. “DbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.” *Human Mutation* 37 (3): 235–41.

Lo Giudice, Temistocle, Federica Lombardi, Filippo Maria Santorelli, Toshitaka Kawarai, and Antonio Orlacchio. 2014. “Hereditary Spastic Paraplegia: Clinical-Genetic Characteristics and Evolving Molecular Mechanisms.” *Experimental Neurology* 261 (November): 518–39.

Logroscino, G., and M. Piccininni. 2019. “Amyotrophic Lateral Sclerosis Descriptive Epidemiology: The Origin of Geographic Difference.” *Neuroepidemiology* 52 (1–2): 93–103.

Logroscino, Giancarlo, Bryan J. Traynor, Orla Hardiman, Adriano Chiò, Douglas Mitchell, Robert J. Swingler, Andrea Millul, Emma Benn, Ettore Beghi, and EURALS. 2010. “Incidence of Amyotrophic Lateral Sclerosis in Europe.” *Journal of Neurology, Neurosurgery, and Psychiatry* 81 (4): 385–90.

Lomen-Hoerth, Catherine, Thomas Anderson, and Bruce Miller. 2002. “The Overlap of Amyotrophic Lateral Sclerosis and Frontotemporal Dementia.” *Neurology* 59 (7): 1077–79.

Longinetti, Elisa, and Fang Fang. 2019. “Epidemiology of Amyotrophic Lateral Sclerosis: An Update of Recent Literature.” *Current Opinion in Neurology* 32 (5): 771–76.

Lundh, Fredrik. 1999. “An Introduction to Tkinter.” *URL: Www. Pythonware. Com/Library/Tkinter/Introduction/Index. Htm*. http://www.tcltk.co.kr/files/TclTk_Introduction_To_Tkiner.pdf.

Luty, Agnes A., John B. J. Kwok, Carol Dobson-Stone, Clement T. Loy, Kirsten G. Coupland, Helena Karlström, Tomasz Sobow, et al. 2010. “Sigma Nonopioid Intracellular Receptor 1 Mutations Cause Frontotemporal Lobar Degeneration-Motor Neuron Disease.” *Annals of Neurology* 68 (5): 639–49.

Lynex, Clare N., Ian M. Carr, Jack P. Leek, Rajgopal Achuthan, Simon Mitchell, Eamonn R. Maher, C. Geoffrey Woods, David T. Bonthon, and Alex F. Markham. 2004. "Homozygosity for a Missense Mutation in the 67 KDa Isoform of Glutamate Decarboxylase in a Family with Autosomal Recessive Spastic Cerebral Palsy: Parallels with Stiff-Person Syndrome and Other Movement Disorders." *BMC Neurology* 4 (1): 20.

MacDonald, Marcy E., Scott A. Strobel, Karen M. Draths, Jennifer L. Wales, and Peter Dervan. 1993. "A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes. The Huntington's Disease Collaborative Research Group." *Cell* 72 (6): 971–83.

Mackenzie, Ian R., Thomas Arzberger, Elisabeth Kremmer, Dirk Troost, Stefan Lorenzl, Kohji Mori, Shih-Ming Weng, et al. 2013. "Dipeptide Repeat Protein Pathology in C9ORF72 Mutation Cases: Clinico-Pathological Correlations." *Acta Neuropathologica* 126 (6): 859–79.

Mahadevan, M., C. Tsilfidis, L. Sabourin, G. Shutler, C. Amemiya, G. Jansen, C. Neville, M. Narang, J. Barceló, and K. O'Hoy. 1992. "Myotonic Dystrophy Mutation: An Unstable CTG Repeat in the 3' Untranslated Region of the Gene." *Science* 255 (5049): 1253–55.

Malik, Indranil, Chase P. Kelley, Eric T. Wang, and Peter K. Todd. 2021. "Molecular Mechanisms Underlying Nucleotide Repeat Expansion Disorders." *Nature Reviews. Molecular Cell Biology* 22 (9): 589–607.

Mangiafico, Salvatore. 2021. "Rcompanion: Functions to Support Extension Education Program Evaluation." https://CRAN.R-project.org/package=rcompanion.

Mannan, Ashraf U., Philip Krawen, Simone M. Sauter, Johann Boehm, Agnieszka Chronowska, Walter Paulus, Juergen Neesen, and Wolfgang Engel. 2006. "ZFYVE27 (SPG33), a Novel Spastin-Binding Protein, Is Mutated in Hereditary Spastic Paraplegia." *The American Journal of Human Genetics* 79 (2): 351–57.

Margolis, R. L., E. O'Hearn, A. Rosenblatt, V. Willour, S. E. Holmes, M. L. Franz, C. Callahan, H. S. Hwang, J. C. Troncoso, and C. A. Ross. 2001. "A Disorder Similar to Huntington's Disease Is Associated with a Novel CAG Repeat Expansion." *Annals of Neurology* 50 (3): 373–80.

Margulies, Marcel, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, et al. 2005. "Genome Sequencing in Microfabricated High-Density Picolitre Reactors." *Nature* 437 (7057): 376–80.

Martin, Elodie, Rebecca Schüle, Katrien Smets, Agnès Rastetter, Amir Boukhris, José L. Loureiro, Michael A. Gonzalez, et al. 2013. "Loss of Function of Glucocerebrosidase GBA2 Is Responsible for Motor Neuron Defects in Hereditary Spastic Paraplegia." *The American Journal of Human Genetics* 92 (2): 238–44.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.Journal* 17 (1): 10–12.

Maruyama, Hirofumi, Hiroyuki Morino, Hidefumi Ito, Yuishin Izumi, Hidemasa Kato, Yasuhito Watanabe, Yoshimi Kinoshita, et al. 2010. "Mutations of Optineurin in Amyotrophic Lateral Sclerosis." *Nature* 465 (7295): 223–26.

Matsuura, T., T. Yamagata, D. L. Burgess, A. Rasmussen, R. P. Grewal, K. Watase, M. Khajavi, et al. 2000. "Large Expansion of the ATTCT Pentanucleotide Repeat in Spinocerebellar Ataxia Type 10." *Nature Genetics* 26 (2): 191–94.

Maxwell, Kara N., Steven N. Hart, Joseph Vijai, Kasmintan A. Schrader, Thomas P. Slavin, Tinu Thomas, Bradley Wubbenhorst, et al. 2016. "Evaluation of ACMG-Guideline-Based Variant Classification of Cancer Susceptibility and Non-Cancer-Associated Genes in Families Affected by Breast Cancer." *The American Journal of Human Genetics* 98 (5): 801–17.

McCann, Emily P., Lyndal Henden, Jennifer A. Fifita, Katharine Y. Zhang, Natalie Grima, Denis C. Bauer, Sandrine Chan Moi Fat, et al. 2020. “Evidence for Polygenic and Oligogenic Basis of Australian Sporadic Amyotrophic Lateral Sclerosis.” *Journal of Medical Genetics*, May. https://doi.org/10.1136/jmedgenet-2020-106866.

McCombe, Pamela A., and Robert D. Henderson. 2010. “Effects of Gender in Amyotrophic Lateral Sclerosis.” *Gender Medicine* 7 (6): 557–70.

McCorquodale, D. S., 3rd, U. Ozomaro, J. Huang, G. Montenegro, A. Kushman, L. Citrigno, J. Price, F. Speziani, M. A. Pericak-Vance, and S. Züchner. 2011. “Mutation Screening of Spastin, Atlastin, and REEP1 in Hereditary Spastic Paraplegia.” *Clinical Genetics* 79 (6): 523–30.

McDermott, Christopher J., Dewi Roberts, Janine Tomkins, Kate M. Bushby, and Pamela J. Shaw. 2003. “Spastin and Paraplegin Gene Analysis in Selected Cases of Motor Neurone Disease (MND).” *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders: Official Publication of the World Federation of Neurology, Research Group on Motor Neuron Diseases* 4 (2): 96–99.

McGuire, V., W. T. Longstreth Jr, T. D. Koepsell, and G. van Belle. 1996. “Incidence of Amyotrophic Lateral Sclerosis in Three Counties in Western Washington State.” *Neurology* 47 (2): 571–73.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303.

McLaughlin, Russell L., Kevin P. Kenna, Alice Vajda, Peter Bede, Marwa Elamin, Simon Cronin, Colette G. Donaghy, Daniel G. Bradley, and Orla Hardiman. 2015. “Second-Generation Irish Genome-Wide Association Study for Amyotrophic Lateral Sclerosis.” *Neurobiology of Aging* 36 (2): 1221.e7-13.

McLaughlin, Russell L., Kevin P. Kenna, Alice Vajda, Mark Heverin, Susan Byrne, Colette G. Donaghy, Simon Cronin, Daniel G. Bradley, and Orla Hardiman. 2015. “Homozygosity Mapping in an Irish ALS Case-Control Cohort Describes Local Demographic Phenomena and Points towards Potential Recessive Risk Loci.” *Genomics* 105 (4): 237–41.

McMurray, Cynthia T. 2010. “Mechanisms of Trinucleotide Repeat Instability during Human Development.” *Nature Reviews. Genetics* 11 (11): 786–99.

Mejzini, Rita, Loren L. Flynn, Ianthe L. Pitout, Sue Fletcher, Steve D. Wilton, and P. Anthony Akkari. 2019. “ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now?” *Frontiers in Neuroscience* 13 (December): 1310.

Meyer, T., A. Schwan, J. S. Dullinger, J. Brocke, K-T Hoffmann, C. H. Nolte, A. Hopt, et al. 2005. “Early-Onset ALS with Long-Term Survival Associated with Spastin Gene Mutation.” *Neurology* 65 (1): 141–43.

Miller, Robert G., J. D. Mitchell, and Dan H. Moore. 2012. “Riluzole for Amyotrophic Lateral Sclerosis (ALS)/Motor Neuron Disease (MND).” *Cochrane Database of Systematic Reviews*, no. 3 (March): CD001447.

Mills, Ryan E., W. Stephen Pittard, Julienne M. Mullaney, Umar Farooq, Todd H. Creasy, Anup A. Mahurkar, David M. Kemeza, et al. 2011. “Natural Genetic Variation Caused by Small Insertions and Deletions in the Human Genome.” *Genome Research* 21 (6): 830–39.

Minikel, Eric Vallabh, Sonia M. Vallabh, Monkol Lek, Karol Estrada, Kaitlin E. Samocha, J. Fah Sathirapongsasuti, Cory Y. McLean, et al. 2016. “Quantifying Prion Disease Penetrance Using Large Population Control Cohorts.” *Science Translational Medicine* 8 (322): 322ra9.

Mitchell, John, Praveen Paul, Han-Jou Chen, Alex Morris, Miles Payling, Mario Falchi, James Habgood, et al. 2010. “Familial Amyotrophic Lateral Sclerosis Is Associated with a Mutation in D-Amino Acid Oxidase." *Proceedings of the National Academy of Sciences of the United States of America* 107 (16): 7556–61.

Mitra, Ileena, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E. Lohmueller, and Melissa Gymrek. 2020. "Genome-Wide Patterns of de Novo Tandem Repeat Mutations and Their Contribution to Autism Spectrum Disorders." *BioRxiv*. bioRxiv. https://doi.org/10.1101/2020.03.04.974170.

Mitsumoto, Hiroshi, Peter L. Nagy, Chris Gennings, Jennifer Murphy, Howard Andrews, Raymond Goetz, Mary Kay Floeter, et al. 2015. "Phenotypic and Molecular Analyses of Primary Lateral Sclerosis." *Neurology. Genetics* 1 (1): e3.

Mizielinska, Sarah, Sebastian Grönke, Teresa Niccoli, Charlotte E. Ridler, Emma L. Clayton, Anny Devoy, Thomas Moens, et al. 2014. "C9orf72 Repeat Expansions Cause Neurodegeneration in Drosophila through Arginine-Rich Proteins." *Science* 345 (6201): 1192–94.

Mizielinska, Sarah, Tammaryn Lashley, Frances E. Norona, Emma L. Clayton, Charlotte E. Ridler, Pietro Fratta, and Adrian M. Isaacs. 2013. "C9orf72 Frontotemporal Lobar Degeneration Is Characterised by Frequent Neuronal Sense and Antisense RNA Foci." *Acta Neuropathologica* 126 (6): 845–57.

Mohammadi, Leila, Maaike P. Vreeswijk, Rogier Oldenburg, Ans van den Ouweland, Jan C. Oosterwijk, Annemarie H. van der Hout, Nicoline Hoogerbrugge, et al. 2009. "A Simple Method for Co-Segregation Analysis to Evaluate the Pathogenicity of Unclassified Variants; BRCA1 and BRCA2 as an Example." *BMC Cancer* 9 (June): 211.

Mok, Kin, Bryan J. Traynor, Jennifer Schymick, Pentti J. Tienari, Hannu Laaksovirta, Terhi Peuralinna, Liisa Myllykangas, et al. 2012. "Chromosome 9 ALS and FTD Locus Is Probably Derived from a Single Founder." *Neurobiology of Aging* 33 (1): 209.e3-8.

Montenegro, Gladys, Adriana P. Rebelo, James Connell, Rachel Allison, Carla Babalini, Michela D'Aloia, Pasqua Montieri, et al. 2012. "Mutations in the ER-Shaping Protein Reticulon 2 Cause the Axon-Degenerative Disorder Hereditary Spastic Paraplegia Type 12." *The Journal of Clinical Investigation* 122 (2): 538–44.

Mootha, V. Vinod, Imran Hussain, Khrishen Cunnusamy, Eric Graham, Xin Gong, Sudha Neelam, Chao Xing, Ralf Kittler, and W. Matthew Petroll. 2015. "TCF4 Triplet Repeat Expansion and Nuclear RNA Foci in Fuchs' Endothelial Corneal Dystrophy." *Investigative Ophthalmology & Visual Science* 56 (3): 2003–11.

Morais, Sara, Laure Raymond, Mathilde Mairey, Paula Coutinho, Eva Brandão, Paula Ribeiro, José Leal Loureiro, et al. 2017. "Massive Sequencing of 70 Genes Reveals a Myriad of Missing Genes or Mechanisms to Be Uncovered in Hereditary Spastic Paraplegias." *European Journal of Human Genetics: EJHG* 25 (11): 1217–28.

Morales Ana, Kinnamon Daniel D., Jordan Elizabeth, Platt Julia, Vatta Matteo, Dorschner Michael O., Starkey Carl A., et al. 2020. "Variant Interpretation for Dilated Cardiomyopathy." *Circulation: Genomic and Precision Medicine* 13 (2): e002480.

Morgan, Sarah, Aleksey Shatunov, William Sproviero, Ashley R. Jones, Maryam Shoai, Deborah Hughes, Ahmad Al Khleifat, et al. 2017. "A Comprehensive Analysis of Rare Genetic Variation in Amyotrophic Lateral Sclerosis in the UK." *Brain: A Journal of Neurology* 140 (6): 1611–18.

Mori, Kohji, Thomas Arzberger, Friedrich A. Grässer, Ilse Gijselinck, Stephanie May, Kristin Rentzsch, Shih-Ming Weng, et al. 2013. "Bidirectional Transcripts of the Expanded C9orf72 Hexanucleotide Repeat Are Translated into Aggregating Dipeptide Repeat Proteins." *Acta Neuropathologica* 126 (6): 881–93.

Mori, Kohji, Shih-Ming Weng, Thomas Arzberger, Stephanie May, Kristin Rentzsch, Elisabeth Kremmer, Bettina Schmid, et al. 2013. "The C9orf72 GGGGCC Repeat Is

Translated into Aggregating Dipeptide-Repeat Proteins in FTLD/ALS." *Science* 339 (6125): 1335–38.

Mousavi, Nima, Sharona Shleizer-Burko, Richard Yanicky, and Melissa Gymrek. 2019. "Profiling the Genome-Wide Landscape of Tandem Repeat Expansions." *Nucleic Acids Research* 47 (15): e90.

Münch, C., R. Sedlmeier, T. Meyer, V. Homberg, A. D. Sperfeld, A. Kurt, J. Prudlo, et al. 2004. "Point Mutations of the P150 Subunit of Dynactin (DCTN1) Gene in ALS." *Neurology* 63 (4): 724–26.

Mundlos, S., F. Otto, C. Mundlos, J. B. Mulliken, A. S. Aylsworth, S. Albright, D. Lindhout, et al. 1997. "Mutations Involving the Transcription Factor CBFA1 Cause Cleidocranial Dysplasia." *Cell* 89 (5): 773–79.

Muratet, François, Elisa Teyssou, Guillaume Banneau, Véronique Danel-Brunaud, Etienne Allart, Jean-Christophe Antoine, Jean-Philippe Camdessanché, et al. 2019. "Spastic Paraplegia Due to Recessive or Dominant Mutations in ERLIN2 Can Convert to ALS." *Neurology Genetics* 5 (6). https://ng.neurology.org/content/5/6/e374.abstract.

Murch, Susan J., Paul Alan Cox, and Sandra Anne Banack. 2004. "A Mechanism for Slow Release of Biomagnified Cyanobacterial Neurotoxins and Neurodegenerative Disease in Guam." *Proceedings of the National Academy of Sciences of the United States of America* 101 (33): 12228–31.

Murphy, Natalie A., Karissa C. Arthur, Pentti J. Tienari, Henry Houlden, Adriano Chiò, and Bryan J. Traynor. 2017. "Age-Related Penetrance of the C9orf72 Repeat Expansion." *Scientific Reports* 7 (1): 2116.

Nanetti, L., S. Baratta, M. Panzeri, C. Tomasello, C. Lovati, J. Azzollini, C. Gellera, D. Di Bella, F. Taroni, and C. Mariotti. 2012. "Novel and Recurrent Spastin Mutations in a Large Series of SPG4 Italian Families." *Neuroscience Letters* 528 (1): 42–45.

Neumann, Manuela, Deepak M. Sampathu, Linda K. Kwong, Adam C. Truax, Matthew C. Micsenyi, Thomas T. Chou, Jennifer Bruce, et al. 2006. "Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis." *Science* 314 (5796): 130–33.

Nguyen, Hung Phuoc, Christine Van Broeckhoven, and Julie van der Zee. 2018. "ALS Genes in the Genomic Era and Their Implications for FTD." *Trends in Genetics: TIG* 34 (6): 404–23.

Nguyen, Hung Phuoc, Sara Van Mossevelde, Lubina Dillen, Jan L. De Bleecker, Matthieu Moisse, Philip Van Damme, Christine Van Broeckhoven, Julie van der Zee, and BELNEU Consortium. 2018. "*NEK1* Genetic Variability in a Belgian Cohort of ALS and ALS-FTD Patients." *Neurobiology of Aging* 61 (January): 255.e1-255.e7.

Nicolas, Aude, Kevin P. Kenna, Alan E. Renton, Nicola Ticozzi, Faraz Faghri, Ruth Chia, Janice A. Dominov, et al. 2018. "Genome-Wide Analyses Identify KIF5A as a Novel ALS Gene." *Neuron* 97 (6): 1268-1283.e6.

Nicolau, Stefan, Megan A. Waldrop, Anne M. Connolly, and Jerry R. Mendell. 2021. "Spinal Muscular Atrophy." *Seminars in Pediatric Neurology* 37 (April): 100878.

Nishimura, Agnes L., Miguel Mitne-Neto, Helga C. A. Silva, Antônio Richieri-Costa, Susan Middleton, Duilio Cascio, Fernando Kok, et al. 2004. "A Mutation in the Vesicle-Trafficking Protein VAPB Causes Late-Onset Spinal Muscular Atrophy and Amyotrophic Lateral Sclerosis." *American Journal of Human Genetics* 75 (5): 822–31.

Novak, R. L., and A. C. Phillips. 2008. "Adenoviral-Mediated Rybp Expression Promotes Tumor Cell-Specific Apoptosis." *Cancer Gene Therapy* 15 (11): 713–22.

Novarino, Gaia, Ali G. Fenstermaker, Maha S. Zaki, Matan Hofree, Jennifer L. Silhavy, Andrew D. Heiberg, Mostafa Abdellateef, et al. 2014. "Exome Sequencing Links Corticospinal Motor Neuron Disease to Common Neurodegenerative Disorders." *Science (New York, N.Y.)* 343 (6170): 506–11.

Oberlé, I., F. Rousseau, D. Heitz, C. Kretz, D. Devys, A. Hanauer, J. Boué, M. F. Bertheas, and J. L. Mandel. 1991. "Instability of a 550-Base Pair DNA Segment and Abnormal Methylation in Fragile X Syndrome." *Science* 252 (5009): 1097–1102.

Okamoto, Kazushi, Tameko Kihira, Tomoyoshi Kondo, Gen Kobashi, Masakazu Washio, Satoshi Sasaki, Tetsuji Yokoyama, et al. 2009. "Lifestyle Factors and Risk of Amyotrophic Lateral Sclerosis: A Case-Control Study in Japan." *Annals of Epidemiology* 19 (6): 359–64.

ONEI. 2021. "Cuban 2012 Census." Www.Onei.Gob.Cu. July 20, 2021. http://www.onei.gob.cu/sites/default/files/informe_nacional_censo_0.pdf.

Onyike, Chiadi U., and Janine Diehl-Schmid. 2013. "The Epidemiology of Frontotemporal Dementia." *International Review of Psychiatry (Abingdon, England)* 25 (2): 130–37.

Opie-Martin, Sarah, Robyn E. Wootton, Ashley Budu-Aggrey, Aleksey Shatunov, Ashley R. Jones, Alfredo Iacoangeli, Ahmad Al Khleifat, George Davey-Smith, and Ammar Al-Chalabi. 2020. "Relationship between Smoking and ALS: Mendelian Randomisation Interrogation of Causality." *Journal of Neurology, Neurosurgery, and Psychiatry* 91 (12): 1312–15.

Orlacchio, Antonio, Carla Babalini, Antonella Borreca, Clarice Patrono, Roberto Massa, Sarenur Basaran, Renato P. Munhoz, et al. 2010. "SPATACSIN Mutations Cause Autosomal Recessive Juvenile Amyotrophic Lateral Sclerosis." *Brain: A Journal of Neurology* 133 (Pt 2): 591–98.

Orr, H. T., M. Y. Chung, S. Banfi, T. J. Kwiatkowski Jr, A. Servadio, A. L. Beaudet, A. E. McCall, L. A. Duvick, L. P. Ranum, and H. Y. Zoghbi. 1993. "Expansion of an Unstable Trinucleotide CAG Repeat in Spinocerebellar Ataxia Type 1." *Nature Genetics* 4 (3): 221–26.

Orthmann-Murphy, Jennifer L., Ettore Salsano, Charles K. Abrams, Alberto Bizzi, Graziella Uziel, Mona M. Freidin, Eleonora Lamantea, Massimo Zeviani, Steven S. Scherer, and Davide Pareyson. 2009. "Hereditary Spastic Paraplegia Is a Novel Phenotype for GJA12/GJC2 Mutations." *Brain: A Journal of Neurology* 132 (Pt 2): 426–38.

O'Toole, O., B. J. Traynor, P. Brennan, C. Sheehan, E. Frost, B. Corr, and O. Hardiman. 2008. "Epidemiology and Clinical Features of Amyotrophic Lateral Sclerosis in Ireland between 1995 and 2004." *Journal of Neurology, Neurosurgery, and Psychiatry* 79 (1): 30–32.

Oza, Andrea M., Marina T. DiStefano, Sarah E. Hemphill, Brandon J. Cushman, Andrew R. Grant, Rebecca K. Siegert, Jun Shen, et al. 2018. "Expert Specification of the ACMG/AMP Variant Interpretation Guidelines for Genetic Hearing Loss." *Human Mutation* 39 (11): 1593–1613.

Oz-Levi, Danit, Bruria Ben-Zeev, Elizabeth K. Ruzzo, Yuki Hitomi, Amir Gelman, Kimberly Pelak, Yair Anikster, et al. 2012. "Mutation in TECPR2 Reveals a Role for Autophagy in Hereditary Spastic Paraparesis." *The American Journal of Human Genetics* 91 (6): 1065–72.

Özoğuz, Aslıhan, Özgün Uyan, Güneş Birdal, Ceren Iskender, Ece Kartal, Suna Lahut, Özgür Ömür, et al. 2015. "The Distinct Genetic Pattern of ALS in Turkey and Novel Mutations." *Neurobiology of Aging* 36 (4): 1764.e9-1764.e18.

Paila, Umadevi, Brad A. Chapman, Rory Kirchner, and Aaron R. Quinlan. 2013. "GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations." *PLoS Computational Biology* 9 (7): e1003153.

Pang, Shirley Yin-Yu, Jacob Shujui Hsu, Kay-Cheong Teo, Yan Li, Michelle H. W. Kung, Kathryn S. E. Cheah, Danny Chan, et al. 2017. "Burden of Rare Variants in ALS Genes Influences Survival in Familial and Sporadic ALS." *Neurobiology of Aging* 58 (October): 238.e9-238.e15.

Panza, Emanuele, Juan M. Escamilla-Honrubia, Clara Marco-Marín, Nadine Gougeard, Giuseppe De Michele, Vincenzo Brescia Morra, Rocco Liguori, et al. 2016.

"ALDH18A1 Gene Mutations Cause Dominant Spastic Paraplegia SPG9: Loss of Function Effect and Plausibility of a Dominant Negative Mechanism." *Brain: A Journal of Neurology*. Oxford University Press (OUP).

Parkinson, N., P. G. Ince, M. O. Smith, R. Highley, G. Skibinski, P. M. Andersen, K. E. Morrison, et al. 2006. "ALS Phenotypes with Mutations in CHMP2B (Charged Multivesicular Body Protein 2B)." *Neurology* 67 (6): 1074–77.

Parodi, L., S. Fenu, G. Stevanin, and A. Durr. 2017. "Hereditary Spastic Paraplegia: More than an Upper Motor Neuron Disease." *Revue Neurologique* 173 (5): 352–60.

Parra, H. J., D. Arveiler, A. E. Evans, J. P. Cambou, P. Amouyel, A. Bingham, D. McMaster, P. Schaffer, P. Douste-Blazy, and G. Luc. 1992. "A Case-Control Study of Lipoprotein Particles in Two Populations at Contrasting Risk for Coronary Heart Disease. The ECTIM Study." *Arteriosclerosis and Thrombosis: A Journal of Vascular Biology / American Heart Association* 12 (6): 701–7.

Patel, Heema, Harold Cross, Christos Proukakis, Ruth Hershberger, Peer Bork, Francesca D. Ciccarelli, Michael A. Patton, Victor A. McKusick, and Andrew H. Crosby. 2002. "SPG20 Is Mutated in Troyer Syndrome, an Hereditary Spastic Paraplegia." *Nature Genetics* 31 (4): 347–48.

Paulson, Henry. 2018. "Chapter 9 - Repeat Expansion Diseases." In *Handbook of Clinical Neurology*, edited by Daniel H. Geschwind, Henry L. Paulson, and Christine Klein, 147:105–23. Elsevier.

Payne, Alex, Nadine Holmes, Vardhman Rakyan, and Matthew Loose. 2018. "Whale Watching with BulkVis: A Graphical Viewer for Oxford Nanopore Bulk Fast5 Files." *BioRxiv*. https://doi.org/10.1101/312256.

Pedersen, Brent S., and Aaron R. Quinlan. 2018. "Mosdepth: Quick Coverage Calculation for Genomes and Exomes." *Bioinformatics* 34 (5): 867–68.

Perucca, Piero, Melanie Bahlo, and Samuel F. Berkovic. 2020. "The Genetics of Epilepsy." *Annual Review of Genomics and Human Genetics* 21 (August): 205–30.

Peters, Owen M., Mehdi Ghasemi, and Robert H. Brown Jr. 2015. "Emerging Mechanisms of Molecular Pathology in ALS." *The Journal of Clinical Investigation* 125 (5): 1767–79.

Petrov, Dmitry, Colin Mansfield, Alain Moussy, and Olivier Hermine. 2017. "ALS Clinical Trials Review: 20 Years of Failure. Are We Any Closer to Registering a New Treatment?" *Frontiers in Aging Neuroscience* 9 (March): 68.

Philips, A. V., L. T. Timchenko, and T. A. Cooper. 1998. "Disruption of Splicing Regulated by a CUG-Binding Protein in Myotonic Dystrophy." *Science* 280 (5364): 737–41.

Phukan, Julie, Marwa Elamin, Peter Bede, Norah Jordan, Laura Gallagher, Susan Byrne, Catherine Lynch, Niall Pender, and Orla Hardiman. 2012. "The Syndrome of Cognitive Impairment in Amyotrophic Lateral Sclerosis: A Population-Based Study." *Journal of Neurology, Neurosurgery, and Psychiatry* 83 (1): 102–8.

Pirity, Melinda K., Joseph Locker, and Nicole Schreiber-Agus. 2005. "Rybp/DEDAF Is Required for Early Postimplantation and for Central Nervous System Development." *Molecular and Cellular Biology* 25 (16): 7193–7202.

Polushina, Tatiana, Niladri Banerjee, Sudheer Giddaluru, Francesco Bettella, Thomas Espeseth, Astri J. Lundervold, Srdjan Djurovic, et al. 2021. "Identification of Pleiotropy at the Gene Level between Psychiatric Disorders and Related Traits." *Translational Psychiatry* 11 (1): 410.

Praline, Julien, Anne-Marie Guennoc, Patrick Vourc'h, Bertrand De Toffol, and Philippe Corcia. 2010. "Primary Lateral Sclerosis May Occur within Familial Amyotrophic Lateral Sclerosis Pedigrees." *Amyotrophic Lateral Sclerosis: Official Publication of the World Federation of Neurology Research Group on Motor Neuron Diseases* 11 (1–2): 154–56.

Prior, Thomas W., Pamela J. Snyder, Britton D. Rink, Dennis K. Pearl, Robert E. Pyatt, David C. Mihal, Todd Conlan, et al. 2010. "Newborn and Carrier Screening for Spinal Muscular Atrophy." *American Journal of Medical Genetics. Part A* 152A (7): 1608–16.

Project MinE ALS Sequencing Consortium. 2018. "Project MinE: Study Design and Pilot Analyses of a Large-Scale Whole-Genome Sequencing Study in Amyotrophic Lateral Sclerosis." *European Journal of Human Genetics: EJHG* 26 (10): 1537–46.

Puls, Imke, Catherine Jonnakuty, Bernadette H. LaMonte, Erika L. F. Holzbaur, Mariko Tokito, Eric Mann, Mary Kay Floeter, et al. 2003. "Mutant Dynactin in Motor Neuron Disease." *Nature Genetics* 33 (4): 455–56.

Pulst, S. M., A. Nechiporuk, T. Nechiporuk, S. Gispert, X. N. Chen, I. Lopes-Cendes, S. Pearlman, et al. 1996. "Moderate Expansion of a Normally Biallelic Trinucleotide Repeat in Spinocerebellar Ataxia Type 2." *Nature Genetics* 14 (3): 269–76.

Pupillo, Elisabetta, Paolo Messina, Giorgia Giussani, Giancarlo Logroscino, Stefano Zoccolella, Adriano Chiò, Andrea Calvo, et al. 2014. "Physical Activity and Amyotrophic Lateral Sclerosis: A European Population-Based Case-Control Study." *Annals of Neurology* 75 (5): 708–16.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81 (3): 559–75.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. https://www.R-project.org/.

Rafehi, Haloom, David J. Szmulewicz, Mark F. Bennett, Nara L. M. Sobreira, Kate Pope, Katherine R. Smith, Greta Gillies, et al. 2019. "Bioinformatics-Based Identification of Expanded Repeats: A Non-Reference Intronic Pentamer Expansion in RFC1 Causes CANVAS." *The American Journal of Human Genetics* 105 (1): 151–65.

Rainier, Shirley, Carron Sher, Orit Reish, Donald Thomas, and John K. Fink. 2006. "De Novo Occurrence of Novel SPG3A/Atlastin Mutation Presenting as Cerebral Palsy." *Archives of Neurology* 63 (3): 445–47.

Rajan-Babu, Indhu-Shree, Junran Peng, Readman Chiu, IMAGINE Study, CAUSES Study, Arezoo Mohajeri, Egor Dolzhenko, Michael A. Eberle, Inanc Birol, and Jan M. Friedman. 2020. "Genome-Wide Sequencing as a First-Tier Screening Test for Short Tandem Repeat Expansions." *BioRxiv*. https://doi.org/10.1101/2020.06.06.137356.

Rañola, John Michael O., Quanhui Liu, Elisabeth A. Rosenthal, and Brian H. Shirts. 2018. "A Comparison of Cosegregation Analysis Methods for the Clinical Setting." *Familial Cancer* 17 (2): 295–302.

Rascovsky, Katya, John R. Hodges, David Knopman, Mario F. Mendez, Joel H. Kramer, John Neuhaus, John C. van Swieten, et al. 2011. "Sensitivity of Revised Diagnostic Criteria for the Behavioural Variant of Frontotemporal Dementia." *Brain: A Journal of Neurology* 134 (Pt 9): 2456–77.

Raux, G., R. Gantier, C. Thomas-Anterion, J. Boulliat, P. Verpillat, D. Hannequin, A. Brice, T. Frebourg, and D. Campion. 2000. "Dementia with Prominent Frontotemporal Features Associated with L113P Presenilin 1 Mutation." *Neurology* 55 (10): 1577–78.

Raza, M. Hashim, Rafael Mattera, Robert Morell, Eduardo Sainz, Rachel Rahn, Joanne Gutierrez, Emily Paris, et al. 2015. "Association between Rare Variants in AP4E1, a Component of Intracellular Trafficking, and Persistent Stuttering." *American Journal of Human Genetics* 97 (5): 715–25.

Reid, Evan, Mark Kloos, Allison Ashley-Koch, Lori Hughes, Simon Bevan, Ingrid K. Svenson, Felicia Lennon Graham, et al. 2002. “A Kinesin Heavy Chain (KIF5A) Mutation in Hereditary Spastic Paraplegia (SPG10).” *The American Journal of Human Genetics* 71 (5): 1189–94.

Renton, Alan E., Elisa Majounie, Adrian Waite, Javier Simón-Sánchez, Sara Rollinson, J. Raphael Gibbs, Jennifer C. Schymick, et al. 2011. “A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD.” *Neuron* 72 (2): 257–68.

Rheenen, Wouter van, Marka van Blitterswijk, Mark H. B. Huisman, Lotte Vlam, Perry T. C. van Doormaal, Meinie Seelen, Jelena Medic, et al. 2012. “Hexanucleotide Repeat Expansions in C9ORF72 in the Spectrum of Motor Neuron Diseases.” *Neurology* 79 (9): 878–82.

Rheenen, Wouter van, Aleksey Shatunov, Annelot M. Dekker, Russell L. McLaughlin, Frank P. Diekstra, Sara L. Pulit, Rick A. A. van der Spek, et al. 2016. “Genome-Wide Association Analyses Identify New Risk Variants and the Genetic Architecture of Amyotrophic Lateral Sclerosis.” *Nature Genetics* 48 (9): 1043–48.

Rheenen, Wouter van, Rick A. A. van der Spek, Mark K. Bakker, Joke J. F. A. van Vugt, Paul J. Hop, Ramona A. J. Zwamborn, Niek de Klein, et al. 2021. “Common and Rare Variant Association Analyses in Amyotrophic Lateral Sclerosis Identify 15 Risk Loci with Distinct Genetic Architectures and Neuron-Specific Biology.” *Nature Genetics* 53 (12): 1636–48.

Richard, Pascale, Capucine Trollet, Tanya Stojkovic, Alix de Becdelievre, Sophie Perie, Jean Pouget, Bruno Eymard, and Neurologists of French Neuromuscular Reference Centers CORNEMUS and FILNEMUS. 2017. “Correlation between PABPN1 Genotype and Disease Severity in Oculopharyngeal Muscular Dystrophy.” *Neurology* 88 (4): 359–65.

Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, et al. 2015. “Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 17 (5): 405–24.

Riku, Yuichi, Naoki Atsuta, Mari Yoshida, Shinsui Tatsumi, Yasushi Iwasaki, Maya Mimuro, Hirohisa Watanabe, et al. 2014. “Differential Motor Neuron Involvement in Progressive Muscular Atrophy: A Comparative Study with Amyotrophic Lateral Sclerosis.” *BMJ Open* 4 (5): e005213.

Rohrer, J. D., R. Guerreiro, J. Vandrovcova, J. Uphill, D. Reiman, J. Beck, A. M. Isaacs, et al. 2009. “The Heritability and Genetics of Frontotemporal Lobar Degeneration.” *Neurology* 73 (18): 1451–56.

Romanet, Pauline, Marie-Françoise Odou, Marie-Odile North, Alexandru Saveanu, Lucie Coppin, Eric Pasmant, Amira Mohamed, et al. 2019. “Proposition of Adjustments to the ACMG-AMP Framework for the Interpretation of MEN1 Missense Variants.” *Human Mutation* 40 (6): 661–74.

Rosen, D. R., T. Siddique, D. Patterson, D. A. Figlewicz, P. Sapp, A. Hentati, D. Donaldson, J. Goto, J. P. O’Regan, and H. X. Deng. 1993. “Mutations in Cu/Zn Superoxide Dismutase Gene Are Associated with Familial Amyotrophic Lateral Sclerosis.” *Nature* 362 (6415): 59–62.

Ruano, Luis, Claudia Melo, M. Carolina Silva, and Paula Coutinho. 2014. “The Global Epidemiology of Hereditary Ataxia and Spastic Paraplegia: A Systematic Review of Prevalence Studies.” *Neuroepidemiology* 42 (3): 174–83.

Rutherford, Nicola J., Mariely DeJesus-Hernandez, Matt C. Baker, Thomas B. Kryston, Patricia E. Brown, Catherine Lomen-Hoerth, Kevin Boylan, Zbigniew K. Wszolek, and Rosa Rademakers. 2012. "C9ORF72 Hexanucleotide Repeat Expansions in Patients with ALS from the Coriell Cell Repository." *Neurology* 79 (5): 482–83.

Rutherford, Nicola J., Nicole A. Finch, Mariely DeJesus-Hernandez, Richard J. P. Crook, Catherine Lomen-Hoerth, Zbigniew K. Wszolek, Ryan J. Uitti, Neill R. Graff-Radford, and Rosa Rademakers. 2012. "Pathogenicity of Exonic Indels in Fused in Sarcoma in Amyotrophic Lateral Sclerosis." *Neurobiology of Aging* 33 (2): 424.e23-4.

Ryan, M., M. Doherty, M. Heverin, N. Pender, R. McLaughlin, and O. Hardiman. 2018. "Oligogenic and Discordant Inheritance: A Population Based Genomic Study of Irish Kindreds Carrying the C9orf72 Repeat Expansion." In *EUROPEAN JOURNAL OF NEUROLOGY*, 25:509–509. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.

Ryan, Marie, Mark Heverin, Mark A. Doherty, Nicola Davis, Emma M. Corr, Alice Vajda, Niall Pender, Russell McLaughlin, and Orla Hardiman. 2018. "Determining the Incidence of Familiality in ALS: A Study of Temporal Trends in Ireland from 1994 to 2016." *Neurology. Genetics* 4 (3): e239.

Ryan, Marie, Mark Heverin, Russell L. McLaughlin, and Orla Hardiman. 2019. "Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis." *JAMA Neurology*, July. https://doi.org/10.1001/jamaneurol.2019.2044.

Ryan, Marie, Tatiana Zaldívar Vaillant, Russell L. McLaughlin, Mark A. Doherty, James Rooney, Mark Heverin, Joel Gutierrez, et al. 2019. "Comparison of the Clinical and Genetic Features of Amyotrophic Lateral Sclerosis across Cuban, Uruguayan and Irish Clinic-Based Populations." *Journal of Neurology, Neurosurgery, and Psychiatry*, March. https://doi.org/10.1136/jnnp-2018-319838.

Sabatelli, M., A. Conte, and M. Zollino. 2013. "Clinical and Genetic Heterogeneity of Amyotrophic Lateral Sclerosis." *Clinical Genetics* 83 (5): 408–16.

Saez-Atienzar, Sara, Clifton L. Dalgard, Jinhui Ding, Adriano Chiò, Camile Alba, Dan N. Hupalo, Matthew D. Wilkerson, et al. 2020. "Identification of a Pathogenic Intronic KIF5A Mutation in an ALS-FTD Kindred." *Neurology* 95 (22): 1015–18.

Sala Frigerio, Carlo, Pierre Lau, Claire Troakes, Vincent Deramecourt, Patrick Gele, Peter Van Loo, Thierry Voet, and Bart De Strooper. 2015. "On the Identification of Low Allele Frequency Mosaic Mutations in the Brains of Alzheimer's Disease Patients." *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 11 (11): 1265–76.

Sambuughin, Nyamkhishig, Lev G. Goldfarb, Tatiana M. Sivtseva, Tatiana K. Davydova, Vsevolod A. Vladimirtsev, Vladimir L. Osakovskiy, Al'bina P. Danilova, et al. 2015. "Adult-Onset Autosomal Dominant Spastic Paraplegia Linked to a GTPase-Effector Domain Mutation of Dynamin 2." *BMC Neurology* 15 (1): 223.

Saugier-Veber, P., A. Munnich, D. Bonneau, J. M. Rozet, M. Le Merrer, R. Gil, and O. Boespflug-Tanguy. 1994. "X-Linked Spastic Paraplegia and Pelizaeus-Merzbacher Disease Are Allelic Disorders at the Proteolipid Protein Locus." *Nature Genetics* 6 (3): 257–62.

Schmitz, Boris, Peter Vischer, Eva Brand, Klaus Schmidt-Petersen, Adelheid Korb-Pap, Katrin Guske, Johanna Nedele, et al. 2013. "Increased Monocyte Adhesion by Endothelial Expression of VCAM-1 Missense Variation in Vitro." *Atherosclerosis* 230 (2): 185–90.

Schulert, Grant S., Mingce Zhang, Ndate Fall, Ammar Husami, Diane Kissell, Andrew Hanosh, Kejian Zhang, et al. 2016. "Whole-Exome Sequencing Reveals Mutations in Genes Linked to Hemophagocytic Lymphohistiocytosis and Macrophage Activation Syndrome in Fatal Cases of H1N1 Influenza." *The Journal of Infectious Diseases* 213 (7): 1180–88.

Schuurs-Hoeijmakers, Janneke H. M., Michael T. Geraghty, Erik-Jan Kamsteeg, Salma Ben-Salem, Susanne T. de Bot, Bonnie Nijhof, Ilse I. G. M. van de Vondervoort, et al. 2012. "Mutations in DDHD2, Encoding an Intracellular Phospholipase A(1), Cause a Recessive Form of Complex Hereditary Spastic Paraplegia." *The American Journal of Human Genetics* 91 (6): 1073–81.

Schwartz, Charles E., Melanie M. May, Nancy J. Carpenter, R. Curtis Rogers, Judith Martin, Martin G. Bialer, Jewell Ward, et al. 2005. "Allan-Herndon-Dudley Syndrome and the Monocarboxylate Transporter 8 (MCT8) Gene." *The American Journal of Human Genetics* 77 (1): 41–53.

Scoles, Daniel R., Mi H. T. Ho, Warunee Dansithong, Lance T. Pflieger, Lance W. Petersen, Khanh K. Thai, and Stefan M. Pulst. 2015. "Repeat Associated Non-AUG Translation (RAN Translation) Dependent on Sequence Downstream of the ATXN2 CAG Repeat." *PloS One* 10 (6): e0128769.

Sharma, Rajesh, Saeeda Bhatti, Mariluz Gomez, Rhonda M. Clark, Cynthia Murray, Tetsuo Ashizawa, and Sanjay I. Bidichandani. 2002. "The GAA Triplet-Repeat Sequence in Friedreich Ataxia Shows a High Level of Somatic Instability in Vivo, with a Significant Predilection for Large Contractions." *Human Molecular Genetics* 11 (18): 2175–87.

Sherry, S. T., M. Ward, and K. Sirotkin. 1999. "DbSNP-Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation." *Genome Research* 9 (8): 677–79.

Shimazaki, Haruo, Junko Honda, Tametou Naoi, Michito Namekawa, Imaharu Nakano, Masahide Yazaki, Katsuya Nakamura, et al. 2014. "Autosomal-Recessive Complicated Spastic Paraplegia with a Novel Lysosomal Trafficking Regulator Gene Mutation." *Journal of Neurology, Neurosurgery, and Psychiatry* 85 (9): 1024–28.

Shimazaki, Haruo, Yoshihisa Takiyama, Hiroyuki Ishiura, Chika Sakai, Yuichi Matsushima, Hideyuki Hatakeyama, Junko Honda, et al. 2012. "A Homozygous Mutation of C12orf65 Causes Spastic Paraplegia with Optic Atrophy and Neuropathy (SPG55)." *Journal of Medical Genetics* 49 (12): 777–84.

Sim, Ngak-Leng, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C. Ng. 2012. "SIFT Web Server: Predicting Effects of Amino Acid Substitutions on Proteins." *Nucleic Acids Research* 40 (Web Server issue): W452-7.

Simpson, Claire L., Robin Lemmens, Katarzyna Miskiewicz, Wendy J. Broom, Valerie K. Hansen, Paul W. J. van Vught, John E. Landers, et al. 2009. "Variants of the Elongator Protein 3 (ELP3) Gene Are Associated with Motor Neuron Degeneration." *Human Molecular Genetics* 18 (3): 472–81.

Simpson, Michael A., Harold Cross, Christos Proukakis, Anna Pryde, Ruth Hershberger, Arnaud Chatonnet, Michael A. Patton, and Andrew H. Crosby. 2003. "Maspardin Is Mutated in Mast Syndrome, a Complicated Form of Hereditary Spastic Paraplegia Associated with Dementia." *The American Journal of Human Genetics* 73 (5): 1147–56.

Sinnwell, Jason P., Terry M. Therneau, and Daniel J. Schaid. 2014. "The Kinship2 R Package for Pedigree Data." *Human Heredity* 78 (2): 91–93.

Sivadasan, Rajeeve, Daniel Hornburg, Carsten Drepper, Nicolas Frank, Sibylle Jablonka, Anna Hansel, Xenia Lojewski, et al. 2016. "C9ORF72 Interaction with Cofilin Modulates Actin Dynamics in Motor Neurons." *Nature Neuroscience* 19 (12): 1610–18.

Skibinski, Gaia, Nicholas J. Parkinson, Jeremy M. Brown, Lisa Chakrabarti, Sarah L. Lloyd, Holger Hummerich, Jørgen E. Nielsen, et al. 2005. "Mutations in the Endosomal ESCRTIII-Complex Subunit CHMP2B in Frontotemporal Dementia." *Nature Genetics* 37 (8): 806–8.

Słabicki, Mikołaj, Mirko Theis, Dragomir B. Krastev, Sergey Samsonov, Emeline Mundwiller, Magno Junqueira, Maciej Paszkowski-Rogacz, et al. 2010. "A Genome-Scale DNA Repair RNAi Screen Identifies SPG48 as a Novel Gene Associated with Hereditary Spastic Paraplegia." *PLoS Biology* 8 (6): e1000408.

Smith, Bradley N., Stephen Newhouse, Aleksey Shatunov, Caroline Vance, Simon Topp, Lauren Johnson, Jack Miller, et al. 2013. "The C9ORF72 Expansion Mutation Is a Common Cause of ALS+/-FTD in Europe and Has a Single Founder." *European Journal of Human Genetics: EJHG* 21 (1): 102–8.

Sone, Jun, Satomi Mitsuhashi, Atsushi Fujita, Takeshi Mizuguchi, Kohei Hamanaka, Keiko Mori, Haruki Koike, et al. 2019. "Long-Read Sequencing Identifies GGC Repeat Expansions in NOTCH2NLC Associated with Neuronal Intranuclear Inclusion Disease." *Nature Genetics* 51 (8): 1215–21.

Souza, Paulo Victor Sgobbi de, Wladimir Bocca Vieira de Rezende Pinto, Gabriel Novaes de Rezende Batistella, Thiago Bortholin, and Acary Souza Bulle Oliveira. 2017. "Hereditary Spastic Paraplegia: Clinical and Genetic Hallmarks." *Cerebellum* 16 (2): 525–51.

Spargo, Thomas P., Sarah Opie-Martin, Cathryn M. Lewis, Alfredo Iacoangeli, and Ammar Al-Chalabi. 2021. "Calculating Variant Penetrance Using Family History of Disease and Population Data Authorship." *BioRxiv*. medRxiv. https://doi.org/10.1101/2021.03.16.21253691.

Spek, Rick A. A. van der, Wouter van Rheenen, Sara L. Pulit, Kevin P. Kenna, Leonard H. van den Berg, Jan H. Veldink, and Project MinE ALS Sequencing Consortium¶. 2019. "The Project MinE Databrowser: Bringing Large-Scale Whole-Genome Sequencing in ALS to Researchers and the Public." *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration* 20 (5–6): 432–40.

Spek, Rick A. A. van der, Wouter van Rheenen, Sara L. Pulit, Kevin P. Kenna, Russell L. McLaughlin, Matthieu Moisse, Annelot M. Dekker, et al. 2019. "The Project MinE Databrowser: Bringing Large-Scale Whole-Genome Sequencing in ALS to Researchers and the Public." *BioRxiv*. https://doi.org/10.1101/377911.

Sproviero, William, Aleksey Shatunov, Daniel Stahl, Maryam Shoai, Wouter van Rheenen, Ashley R. Jones, Safa Al-Sarraj, et al. 2017. "ATXN2 Trinucleotide Repeat Length Correlates with Risk of ALS." *Neurobiology of Aging* 51 (March): 178.e1-178.e9.

Stenson, Peter D., Matthew Mort, Edward V. Ball, Katy Evans, Matthew Hayden, Sally Heywood, Michelle Hussain, Andrew D. Phillips, and David N. Cooper. 2017. "The Human Gene Mutation Database: Towards a Comprehensive Repository of Inherited Mutation Data for Medical Research, Genetic Diagnosis and next-Generation Sequencing Studies." *Human Genetics* 136 (6): 665–77.

Stevanin, Giovanni, Filippo M. Santorelli, Hamid Azzedine, Paula Coutinho, Jacques Chomilier, Paola S. Denora, Elodie Martin, et al. 2007. "Mutations in SPG11, Encoding Spatacsin, Are a Major Cause of Spastic Paraplegia with Thin Corpus Callosum." *Nature Genetics* 39 (3): 366–72.

Stewart, Heather, Nicola J. Rutherford, Hannah Briemberg, Charles Krieger, Neil Cashman, Marife Fabros, Matt Baker, et al. 2012. "Clinical and Pathological Features of Amyotrophic Lateral Sclerosis Caused by Mutation in the C9ORF72 Gene on Chromosome 9p." *Acta Neuropathologica* 123 (3): 409–17.

Stoddart, David, Andrew J. Heron, Ellina Mikhailova, Giovanni Maglia, and Hagan Bayley. 2009. "Single-Nucleotide Discrimination in Immobilized DNA Oligonucleotides with a Biological Nanopore." *Proceedings of the National Academy of Sciences of the United States of America* 106 (19): 7702–7.

Stokholm, Jette, Thomas W. Teasdale, Peter Johannsen, Jorgen E. Nielsen, Troels Tolstrup Nielsen, Adrian Isaacs, Jerry M. Brown, Anders Gade, and Frontotemporal dementia Research in Jutland Association (FReJA) consortium. 2013. "Cognitive Impairment in the Preclinical Stage of Dementia in FTD-3 CHMP2B Mutation Carriers: A Longitudinal Prospective Study." *Journal of Neurology, Neurosurgery, and Psychiatry* 84 (2): 170–76.

Strømme, Petter, Marie E. Mangelsdorf, Marie A. Shaw, Karen M. Lower, Suzanne M. E. Lewis, Helene Bruyere, Viggo Lütcherath, et al. 2002. "Mutations in the Human Ortholog of Aristaless Cause X-Linked Mental Retardation and Epilepsy." *Nature Genetics* 30 (4): 441–45.

Taioli, Emanuela. 2007. "All Causes of Mortality in Male Professional Soccer Players." *European Journal of Public Health* 17 (6): 600–604.

Takahashi, Yuji, Yoko Fukuda, Jun Yoshimura, Atsushi Toyoda, Kari Kurppa, Hiroyoko Moritoyo, Veronique V. Belzil, et al. 2013. "ERBB4 Mutations That Disrupt the Neuregulin-ErbB4 Pathway Cause Amyotrophic Lateral Sclerosis Type 19." *American Journal of Human Genetics* 93 (5): 900–905.

Tan, Adrian, Gonçalo R. Abecasis, and Hyun Min Kang. 2015. "Unified Representation of Genetic Variants." *Bioinformatics* 31 (13): 2202–4.

Tang, Haibao, Ewen F. Kirkness, Christoph Lippert, William H. Biggs, Martin Fabani, Ernesto Guzman, Smriti Ramakrishnan, et al. 2017. "Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes." *American Journal of Human Genetics* 101 (5): 700–715.

Tankard, Rick M., Mark F. Bennett, Peter Degorski, Martin B. Delatycki, Paul J. Lockhart, and Melanie Bahlo. 2018. "Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data." *American Journal of Human Genetics* 103 (6): 858–73.

Tazelaar, Gijs H. P., Steven Boeynaems, Mathias De Decker, Joke J. F. A. van Vugt, Lindy Kool, H. Stephan Goedee, Russell L. McLaughlin, et al. 2020. "ATXN1 Repeat Expansions Confer Risk for Amyotrophic Lateral Sclerosis and Contribute to TDP-43 Mislocalization." *Brain Communications* 2 (2): fcaa064.

Tazelaar, Gijs H. P., Annelot M. Dekker, Joke J. F. A. van Vugt, Rick A. van der Spek, Henk-Jan Westeneng, Lindy J. B. G. Kool, Kevin P. Kenna, et al. 2019. "Association of NIPA1 Repeat Expansions with Amyotrophic Lateral Sclerosis in a Large International Cohort." *Neurobiology of Aging* 74 (February): 234.e9-234.e15.

Team, R. Core. 2014. "R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing [Internet]. Vienna, Austria; 2014."

Tesson, Christelle, Magdalena Nawara, Mustafa A. M. Salih, Rodrigue Rossignol, Maha S. Zaki, Mohammed Al Balwi, Rebecca Schule, et al. 2012. "Alteration of Fatty-Acid-Metabolizing Enzymes Affects Mitochondrial Form and Function in Hereditary Spastic Paraplegia." *The American Journal of Human Genetics* 91 (6): 1051–64.

Therrien, Martine, Guy A. Rouleau, Patrick A. Dion, and J. Alex Parker. 2013. "Deletion of C9ORF72 Results in Motor Neuron Degeneration and Stress Sensitivity in C. Elegans." *PloS One* 8 (12): e83450.

Thomas, Quentin, Giulia Coarelli, Anna Heinzmann, Isabelle Le Ber, Maria Del Mar Amador, and Alexandra Durr. 2021. "Questioning the Causality of HTT CAG-Repeat Expansions in FTD/ALS." *Neuron*.

Thompson, Deborah, Douglas F. Easton, and David E. Goldgar. 2003. "A Full-Likelihood Method for the Evaluation of Causality of Sequence Variants from Family Data." *American Journal of Human Genetics* 73 (3): 652–55.

Tian, Yun, Jun-Ling Wang, Wen Huang, Sheng Zeng, Bin Jiao, Zhen Liu, Zhao Chen, et al. 2019. "Expansion of Human-Specific GGC Repeat in Neuronal Intranuclear Inclusion Disease-Related Disorders." *American Journal of Human Genetics* 105 (1): 166–76.

Ticozzi, N., C. Vance, A. L. Leclerc, P. Keagle, J. D. Glass, D. McKenna-Yasek, P. C. Sapp, et al. 2011. "Mutational Analysis Reveals the FUS Homolog TAF15 as a Candidate

Gene for Familial Amyotrophic Lateral Sclerosis." *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics* 156B (3): 285–90.

Ticozzi, Nicola, Cinzia Tiloca, Daniela Calini, Stella Gagliardi, Alessandra Altieri, Claudia Colombrita, Cristina Cereda, et al. 2014. "C9orf72 Repeat Expansions Are Restricted to the ALS-FTD Spectrum." *Neurobiology of Aging* 35 (4): 936.e13-7.

Todd, Peter K., Seok Yoon Oh, Amy Krans, Fang He, Chantal Sellier, Michelle Frazer, Abigail J. Renoux, et al. 2013. "CGG Repeat-Associated Translation Mediates Neurodegeneration in Fragile X Tremor Ataxia Syndrome." *Neuron* 78 (3): 440–55.

Tran, Helene, Sandra Almeida, Jill Moore, Tania F. Gendron, Umadevi Chalasani, Yubing Lu, Xing Du, et al. 2015. "Differential Toxicity of Nuclear RNA Foci versus Dipeptide Repeat Proteins in a Drosophila Model of C9ORF72 FTD/ALS." *Neuron* 87 (6): 1207–14.

Traynor, B. J., M. B. Codd, B. Corr, C. Forde, E. Frost, and O. Hardiman. 1999. "Incidence and Prevalence of ALS in Ireland, 1995-1997: A Population-Based Study." *Neurology* 52 (3): 504–9.

Tsaousidou, Maria K., Karim Ouahchi, Tom T. Warner, Yi Yang, Michael A. Simpson, Nigel G. Laing, Philip A. Wilkinson, et al. 2008. "Sequence Alterations within CYP7B1 Implicate Defective Cholesterol Homeostasis in Motor-Neuron Degeneration." *The American Journal of Human Genetics* 82 (2): 510–15.

Tunca, Ceren, Fulya Akçimen, Cemre Coşkun, Aslı Gündoğdu-Eken, Cemile Kocoglu, Betül Çevik, Can Ebru Bekircan-Kurt, Ersin Tan, and A. Nazlı Başak. 2018. "ERLIN1 Mutations Cause Teenage-Onset Slowly Progressive ALS in a Large Turkish Pedigree." *European Journal of Human Genetics: EJHG* 26 (5): 745–48.

Turner, Martin R., Richard J. Barohn, Philippe Corcia, John K. Fink, Matthew B. Harms, Matthew C. Kiernan, John Ravits, et al. 2020. "Primary Lateral Sclerosis: Consensus Diagnostic Criteria." *Journal of Neurology, Neurosurgery, and Psychiatry* 91 (4): 373–77.

Turner, Stephen D. n.d. "Qqman: An R Package for Visualizing GWAS Results Using Q-Q and Manhattan Plots." https://doi.org/10.1101/005165.

Urwin, Hazel, Astrid Authier, Jorgen E. Nielsen, Daniel Metcalf, Caroline Powell, Kristina Froud, Denise S. Malcolm, et al. 2010. "Disruption of Endocytic Trafficking in Frontotemporal Dementia with CHMP2B Mutations." *Human Molecular Genetics* 19 (11): 2228–38.

Utsch, Boris, Karl Becker, Detlef Brock, Michael J. Lentze, Frank Bidlingmaier, and Michael Ludwig. 2002. "A Novel Stable Polyalanine [Poly(A)] Expansion in the HOXA13 Gene Associated with Hand-Foot-Genital Syndrome: Proper Function of Poly(A)-Harbouring Transcription Factors Depends on a Critical Repeat Length?" *Human Genetics* 110 (5): 488–94.

Vajda, Alice, Russell L. McLaughlin, Mark Heverin, Owen Thorpe, Sharon Abrahams, Ammar Al-Chalabi, and Orla Hardiman. 2017. "Genetic Testing in ALS: A Survey of Current Practices." *Neurology* 88 (10): 991–99.

Valdmanis, Paul N., Nicolas Dupré, and Guy A. Rouleau. 2008. "A Locus for Primary Lateral Sclerosis on Chromosome 4ptel-4p16.1." *Archives of Neurology* 65 (3): 383–86.

Valdmanis, Paul N., Inge A. Meijer, Annie Reynolds, Adrienne Lei, Patrick MacLeod, David Schlesinger, Mayana Zatz, et al. 2007. "Mutations in the KIAA0196 Gene at the SPG8 Locus Cause Hereditary Spastic Paraplegia." *The American Journal of Human Genetics* 80 (1): 152–61.

Van Damme, P., J. H. Veldink, M. van Blitterswijk, A. Corveleyn, P. W. J. van Vught, V. Thijs, B. Dubois, G. Matthijs, L. H. van den Berg, and W. Robberecht. 2011.

"Expanded ATXN2 CAG Repeat Size in ALS Identifies Genetic Overlap between ALS and SCA2." *Neurology* 76 (24): 2066–72.

Van der Auwera, Geraldine A., Mauricio O. Carneiro, Chris Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. "From FastQ Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 43: 11.10.1-33.

Van Rossum, Guido, and Fred L. Drake Jr. 1995. *Python Reference Manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Vance, Caroline, Boris Rogelj, Tibor Hortobágyi, Kurt J. De Vos, Agnes Lumi Nishimura, Jemeen Sreedharan, Xun Hu, et al. 2009. "Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6." *Science* 323 (5918): 1208–11.

Vázquez, M. C., C. Ketzoián, C. Legnani, I. Rega, N. Sánchez, A. Perna, M. Penela, X. Aguirrezábal, M. Druet-Cabanac, and M. Medici. 2008. "Incidence and Prevalence of Amyotrophic Lateral Sclerosis in Uruguay: A Population-Based Study." *Neuroepidemiology* 30 (2): 105–11.

Veltman, Joris A., and Han G. Brunner. 2012. "De Novo Mutations in Human Genetic Disease." *Nature Reviews. Genetics* 13 (8): 565–75.

Verkerk, A. J., M. Pieretti, J. S. Sutcliffe, Y. H. Fu, D. P. Kuhl, A. Pizzuti, O. Reiner, S. Richards, M. F. Victoria, and F. P. Zhang. 1991. "Identification of a Gene (FMR-1) Containing a CGG Repeat Coincident with a Breakpoint Cluster Region Exhibiting Length Variation in Fragile X Syndrome." *Cell* 65 (5): 905–14.

Visser, Anne E., James P. K. Rooney, Fabrizio D'Ovidio, Henk-Jan Westeneng, Roel C. H. Vermeulen, Ettore Beghi, Adriano Chiò, et al. 2018. "Multicentre, Cross-Cultural, Population-Based, Case-Control Study of Physical Activity as Risk Factor for Amyotrophic Lateral Sclerosis." *Journal of Neurology, Neurosurgery, and Psychiatry* 89 (8): 797–803.

Visser, Jeldican, Renske M. van den Berg-Vos, Hessel Franssen, Leonard H. van den Berg, John H. Wokke, J. M. Vianney de Jong, Rebecca Holman, Rob J. de Haan, and Marianne de Visser. 2007. "Disease Course and Prognostic Factors of Progressive Muscular Atrophy." *Archives of Neurology* 64 (4): 522–28.

Vries, Bálint S. de, Laura M. M. Rustemeijer, Anneke J. van der Kooi, Joost Raaphorst, Carin D. Schröder, Tanja C. W. Nijboer, Jeroen Hendrikse, Jan H. Veldink, Leonard H. van den Berg, and Michael A. van Es. 2017. "A Case Series of PLS Patients with Frontotemporal Dementia and Overview of the Literature." *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration* 18 (7–8): 534–48.

Waite, Adrian J., Dirk Bäumer, Simon East, James Neal, Huw R. Morris, Olaf Ansorge, and Derek J. Blake. 2014. "Reduced C9orf72 Protein Levels in Frontal Cortex of Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration Brain with the C9ORF72 Hexanucleotide Repeat Expansion." *Neurobiology of Aging* 35 (7): 1779.e5-1779.e13.

Wali, Gautam, Kishore Raj Kumar, Erandhi Liyanage, Ryan L. Davis, Alan Mackay-Sim, and Carolyn M. Sue. 2020. "Mitochondrial Function in Hereditary Spastic Paraplegia: Deficits in SPG7 but Not SPAST Patient-Derived Stem Cells." *Frontiers in Neuroscience* 14 (August): 820.

Walton, Clare, Rachel King, Lindsay Rechtman, Wendy Kaye, Emmanuelle Leray, Ruth Ann Marrie, Neil Robertson, et al. 2020. "Rising Prevalence of Multiple Sclerosis Worldwide: Insights from the Atlas of MS, Third Edition." *Multiple Sclerosis (Houndmills, Basingstoke, England)* 26 (14): 1816–21.

Wang, Hao, Éilis J. O'Reilly, Marc G. Weisskopf, Giancarlo Logroscino, Marji L. McCullough, Michael J. Thun, Arthur Schatzkin, Laurence N. Kolonel, and Alberto

Ascherio. 2011. “Smoking and Risk of Amyotrophic Lateral Sclerosis: A Pooled Analysis of 5 Prospective Cohorts.” *Archives of Neurology* 68 (2): 207–13.

Wang, Ming-Dong, James Gomes, Neil R. Cashman, Julian Little, and Daniel Krewski. 2014. “Intermediate CAG Repeat Expansion in the ATXN2 Gene Is a Unique Genetic Risk Factor for ALS--a Systematic Review and Meta-Analysis of Observational Studies.” *PloS One* 9 (8): e105534.

Wang, Xue-Bin, Ning-Hua Cui, Jia-Jia Gao, Xue-Ping Qiu, and Fang Zheng. 2014. “SMN1 Duplications Contribute to Sporadic Amyotrophic Lateral Sclerosis Susceptibility: Evidence from a Meta-Analysis.” *Journal of the Neurological Sciences* 340 (1–2): 63–68.

Warren, J. D., J. D. Rohrer, and M. N. Rossor. 2013. “Frontotemporal Dementia.” *BMJ (Clinical Research Ed.)* 347 (aug12 3): f4827–f4827.

Wen, Xinmei, Wenjia Zhu, Nan L. Xia, Qianwen Li, Li Di, Shu Zhang, Hai Chen, et al. 2021. “Missense Mutations of Codon 116 in the SOD1 Gene Cause Rapid Progressive Familial ALS and Predict Short Viability with PMA Phenotype.” *Frontiers in Genetics* 12 (November): 776831.

Wickham, H. 2011. “The Split-Apply-Combine Strategy for Data Analysis.” *Journal of Statistical Software*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.182.5667&rep=rep1&type=pdf.

Wickham, Hadley. 2016. “Ggplot2: Elegant Graphics for Data Analysis.” Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Lionel Henry. 2020. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Wilcox, Emma H., Mahdi Sarmady, Bryan Wulf, Matt W. Wright, Heidi L. Rehm, Leslie G. Biesecker, and Ahmad N. Abou Tayoun. 2021. “Evaluating the Impact of in Silico Predictors on Clinical Variant Classification.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, December. https://doi.org/10.1016/j.gim.2021.11.018.

Willems, Thomas, Dina Zielinski, Jie Yuan, Assaf Gordon, Melissa Gymrek, and Yaniv Erlich. 2017. “Genome-Wide Profiling of Heritable and de Novo STR Variations.” *Nature Methods* 14 (6): 590–92.

Windpassinger, Christian, Michaela Auer-Grumbach, Joy Irobi, Heema Patel, Erwin Petek, Gerd Hörl, Roland Malli, et al. 2004. “Heterozygous Missense Mutations in BSCL2 Are Associated with Distal Hereditary Motor Neuropathy and Silver Syndrome.” *Nature Genetics* 36 (3): 271–76.

Winnepenninckx, Birgitta, Kim Debacker, Jacqueline Ramsay, Dominique Smeets, Arie Smits, David R. FitzPatrick, and R. Frank Kooy. 2007. “CGG-Repeat Expansion in the DIP2B Gene Is Associated with the Fragile Site FRA12A on Chromosome 12q13.1.” *American Journal of Human Genetics* 80 (2): 221–31.

Woollacott, Ione O. C., and Simon Mead. 2014. “The C9ORF72 Expansion Mutation: Gene Structure, Phenotypic and Diagnostic Issues.” *Acta Neuropathologica* 127 (3): 319–32.

Wu, Chi-Hong, Claudia Fallini, Nicola Ticozzi, Pamela J. Keagle, Peter C. Sapp, Katarzyna Piotrowska, Patrick Lowe, et al. 2012. “Mutations in the Profilin 1 Gene Cause Familial Amyotrophic Lateral Sclerosis.” *Nature* 488 (7412): 499–503.

Wu, Michael C., Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. 2011. “Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test.” *American Journal of Human Genetics* 89 (1): 82–93.

Xi, Zhengrui, Lorne Zinman, Danielle Moreno, Jennifer Schymick, Yan Liang, Christine Sato, Yonglan Zheng, et al. 2013. “Hypermethylation of the CpG Island near the G4C2 Repeat in ALS with a C9orf72 Expansion.” *American Journal of Human Genetics* 92 (6): 981–89.

Xiao, Shangxi, Laura MacNair, Philip McGoldrick, Paul M. McKeever, Jesse R. McLean, Ming Zhang, Julia Keith, Lorne Zinman, Ekaterina Rogaeva, and Janice Robertson. 2015. “Isoform-Specific Antibodies Reveal Distinct Subcellular Localizations of C9orf72 in Amyotrophic Lateral Sclerosis.” *Annals of Neurology* 78 (4): 568–83.

Xie, Yihui, Joe Cheng, and Xianying Tan. 2020. *DT: A Wrapper of the JavaScript Library “DataTables.”* https://CRAN.R-project.org/package=DT.

Xiong, Hui Y., Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, et al. 2015. “RNA Splicing. The Human Splicing Code Reveals New Insights into the Genetic Determinants of Disease.” *Science* 347 (6218): 1254806.

Yan, J., H-X Deng, N. Siddique, F. Fecto, W. Chen, Y. Yang, E. Liu, et al. 2010. “Frameshift and Novel Mutations in FUS in Familial Amyotrophic Lateral Sclerosis and ALS/Dementia.” *Neurology* 75 (9): 807–14.

Yang, Y., A. Hentati, H. X. Deng, O. Dabbagh, T. Sasaki, M. Hirano, W. Y. Hung, et al. 2001. “The Gene Encoding Alsin, a Protein with Three Guanine-Nucleotide Exchange Factor Domains, Is Mutated in a Form of Recessive Amyotrophic Lateral Sclerosis.” *Nature Genetics* 29 (2): 160–65.

Yang, Yi, Lei Zhang, David R. Lynch, Thomas Lukas, Kreshnik Ahmeti, Patrick M. A. Sleiman, Eanna Ryan, et al. 2016. “Compound Heterozygote Mutations in SPG7 in a Family with Adult-Onset Primary Lateral Sclerosis.” *Neurology. Genetics* 2 (2): e60.

Yeetong, Patra, Monnat Pongpanich, Chalurmpon Srichomthong, Adjima Assawapitaksakul, Varote Shotelersuk, Nithiphut Tantirukdham, Chaipat Chunharas, Kanya Suphapeetiporn, and Vorasuk Shotelersuk. 2019. “TTTCA Repeat Insertions in an Intron of YEATS2 in Benign Adult Familial Myoclonic Epilepsy Type 4.” *Brain: A Journal of Neurology* 142 (11): 3360–66.

Zaldivar, T., J. Gutierrez, G. Lara, M. Carbonara, G. Logroscino, and O. Hardiman. 2009. “Reduced Frequency of ALS in an Ethnically Mixed Population: A Population-Based Mortality Study.” *Neurology* 72 (19): 1640–45.

Zee, Julie van der, Ilse Gijselinck, Sara Van Mossevelde, Federica Perrone, Lubina Dillen, Bavo Heeman, Veerle Bäumer, et al. 2017. “TBK1 Mutation Spectrum in an Extended European Patient Cohort with Frontotemporal Dementia and Amyotrophic Lateral Sclerosis.” *Human Mutation* 38 (3): 297–309.

Zeng, Sheng, Mei-Yun Zhang, Xue-Jing Wang, Zheng-Mao Hu, Jin-Chen Li, Nan Li, Jun-Ling Wang, et al. 2019. “Long-Read Sequencing Identified Intronic Repeat Expansions in SAMD12 from Chinese Pedigrees Affected with Familial Cortical Myoclonic Tremor with Epilepsy.” *Journal of Medical Genetics* 56 (4): 265–70.

Zhang, Hang, Wanshi Cai, Siyu Chen, Jialong Liang, Zhanjun Wang, Yuting Ren, Wenxiu Liu, Xiaolan Zhang, Zhongsheng Sun, and Xusheng Huang. 2018. “Screening for Possible Oligogenic Pathogenesis in Chinese Sporadic ALS Patients.” *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration* 19 (5–6): 419–25.

Zhao, X., D. Alvarado, S. Rainier, R. Lemons, P. Hedera, C. H. Weber, T. Tukel, et al. 2001. “Mutations in a Newly Identified GTPase Gene Cause Autosomal Dominant Hereditary Spastic Paraplegia.” *Nature Genetics* 29 (3): 326–31.

Zhu, Xiaonian, Meng Yan, Wei Luo, Wei Liu, Yuan Ren, Chunhua Bei, Guifang Tang, Ruiling Chen, and Shengkui Tan. 2017. “Expression and Clinical Significance of PcG-Associated Protein RYBP in Hepatocellular Carcinoma.” *Oncology Letters* 13 (1): 141–50.

Zhuchenko, O., J. Bailey, P. Bonnen, T. Ashizawa, D. W. Stockton, C. Amos, W. B. Dobyns, S. H. Subramony, H. Y. Zoghbi, and C. C. Lee. 1997. “Autosomal Dominant Cerebellar Ataxia (SCA6) Associated with Small Polyglutamine Expansions in the Alpha 1A-Voltage-Dependent Calcium Channel.” *Nature Genetics* 15 (1): 62–69.

Zivony-Elboum, Yifat, Wendy Westbroek, Nehama Kfir, David Savitzki, Yishay Shoval, Assnat Bloom, Raya Rod, et al. 2012. “A Founder Mutation in Vps37A Causes Autosomal Recessive Complex Hereditary Spastic Paraparesis.” *Journal of Medical Genetics* 49 (7): 462–72.

Zou, Zhang-Yu, Li-Ying Cui, Qing Sun, Xiao-Guang Li, Ming-Sheng Liu, Yan Xu, Yan Zhou, and Xun-Zhe Yang. 2013. “De Novo FUS Gene Mutations Are Associated with Juvenile-Onset Sporadic Amyotrophic Lateral Sclerosis in China.” *Neurobiology of Aging* 34 (4): 1312.e1-8.

Zou, Zhang-Yu, Ming-Sheng Liu, Xiao-Guang Li, and Li-Ying Cui. 2016. “The Distinctive Genetic Architecture of ALS in Mainland China.” *Journal of Neurology, Neurosurgery, and Psychiatry* 87 (8): 906–7.

Zou, Zhang-Yu, Zhi-Rui Zhou, Chun-Hui Che, Chang-Yun Liu, Rao-Li He, and Hua-Pin Huang. 2017. “Genetic Epidemiology of Amyotrophic Lateral Sclerosis: A Systematic Review and Meta-Analysis.” *Journal of Neurology, Neurosurgery, and Psychiatry* 88 (7): 540–49.

Zu, Tao, Brian Gibbens, Noelle S. Doty, Mário Gomes-Pereira, Aline Huguet, Matthew D. Stone, Jamie Margolis, et al. 2011. “Non-ATG-Initiated Translation Directed by Microsatellite Expansions.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (1): 260–65.

Zu, Tao, Yuanjing Liu, Monica Bañez-Coronel, Tammy Reid, Olga Pletnikova, Jada Lewis, Timothy M. Miller, et al. 2013. “RAN Proteins and RNA Foci from Antisense Transcripts in C9ORF72 ALS and Frontotemporal Dementia.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (51): E4968-77.

Züchner, Stephan, Gaofeng Wang, Khanh-Nhat Tran-Viet, Martha A. Nance, Perry C. Gaskell, Jeffery M. Vance, Allison E. Ashley-Koch, and Margaret A. Pericak-Vance. 2006. “Mutations in the Novel Mitochondrial Protein REEP1 Cause Hereditary Spastic Paraplegia Type 31.” *The American Journal of Human Genetics* 79 (2): 365–69.

Appendix

Supplementary Tables

Chapter 2

Supplementary Table S2.1: HGMD phenotypes screened
Amyotrophic lateral sclerosis
Amyotrophic lateral sclerosis 4, juvenile
Amyotrophic lateral sclerosis and frontotemporal dementia
Amyotrophic lateral sclerosis and parkinson disease
Amyotrophic lateral sclerosis, association with
Amyotrophic lateral sclerosis, autosomal recessive
Amyotrophic lateral sclerosis & cognitive decline
Amyotrophic lateral sclerosis & dementia
Amyotrophic lateral sclerosis, familial
Amyotrophic lateral sclerosis, flail arm variant
Amyotrophic lateral sclerosis / frontotemporal dementia
Amyotrophic lateral sclerosis/Frontotemporal dementia
Amyotrophic lateral sclerosis, increased risk
Amyotrophic lateral sclerosis, increased risk, association with
Amyotrophic lateral sclerosis, increased survival, association with
Amyotrophic lateral sclerosis, juvenile
Amyotrophic lateral sclerosis, juvenile with basophilic inclusion
Amyotrophic lateral sclerosis, late onset, association with
Amyotrophic lateral sclerosis, modifier of
Amyotrophic lateral sclerosis, phenotype modifier
Amyotrophic lateral sclerosis, PMA variant
Amyotrophic lateral sclerosis, predisposition to
Amyotrophic lateral sclerosis, progression
Amyotrophic lateral sclerosis, reduced disease severity
Amyotrophic lateral sclerosis, sporadic
Amyotrophic lateral sclerosis, susceptibility to, association with
Amyotrophic lateral sclerosis type 19
Amyotrophic lateral sclerosis with aphasia
Frontotemporal dementia
Frontotemporal dementia / amyotrophic lateral sclerosis
Frontotemporal dementia - amyotrophic lateral sclerosis, association with
Frontotemporal dementia, association with
Frontotemporal dementia, behavioural variant
Frontotemporal dementia/corticobasal degeneration
Frontotemporal dementia, increased risk
Frontotemporal dementia, in GRN mutation carriers, association with
Frontotemporal dementia-like syndrome
Frontotemporal dementia, right temporal lobe variant
Frontotemporal dementia, supranuclear gaze palsy & chorea
Frontotemporal dementia, with parkinsonism
Frontotemporal dementia with parkinsonism and pick body-like inclusions
Frontotemporal lobar degeneration
Frontotemporal lobar degeneration / amyotrophic lateral sclerosis
Frontotemporal lobar degeneration, behavioural variant
Frontotemporal lobar degeneration - motor neuron disease
IBMPFD / Amyotrophic lateral sclerosis
Motor neuron disease
Motor neuron disease, association with
Motor neuron disease, juvenile
Motor neuron disease, lower
Motor neuron disease, lower-predominant
Motor neuron disease, paraparesis
Motor neuron disease, progressive
Motor neuron disease, scoliosis, chest deformity

Supplementary Table S2.2: Review articles included in screening

PMID	Note	Reference
24630593	NA	Wang *et al.*, 2014
28017481	NA	Sproviero *et al.*, 2017
28270533	Table 1	Ghasemi and Brown, 2018
27982040	Supplemental table	Al-Chalabi, van den Berg and Veldink, 2017
21989245	NA	Andersen and Al-Chalabi, 2011
23379621	NA	Sabatelli, Conte and Zollino, 2013
24503148	NA	Finsterer and Burgunder, 2014
28522837	NA	Murphy *et al.*, 2017
28057713	NA	Zou *et al.*, 2017

Supplementary Table S2.3: *C9orf72* Control Cohorts (1/2)

PMID	Number of Controls	Number of Positive Controls	Country
21944778	909	0	USA
22154785	856	0	Belgium
22228244	0	0	Canada
22300873	0	0	England
22366791	0	0	England
22366793	0	0	USA
22406228	2585	5	Global
22418734	619	0	Italy
22445326	228	0	Greece
22499346	580	0	France
22637429	0	0	Kii_Peninsula
22645277	748	0	Netherlands
22722621	0	0	Italy
22773853	0	0	Italy
22815561	0	0	USA
22818528	182	0	Japan
22936364	248	0	Spain
22941224	4	0	Canada
23012445	180	0	Japan
23088937	0	0	South_Korea
23100398	245	0	Italy
23254636	270	0	France
23284068	216	0	Spain
23338682	0	0	Belgium
23435409	0	0	Italy
23869403	100	0	China
23870417	384	0	Belgium
23881933	311	0	Ireland
23962495	10	0	Iran
24064469	201	0	Italy
24269022	150	0	China
24325798	0	0	Italy
24445580	0	0	Australia
25108559	0	0	Australia
25123918	700	0	Sardinia
25179228	0	0	UK
25382069	0	0	USA
25585530	0	0	Slovenia
25681989	200	0	Turkey
26142124	1062	0	China
26176978	0	0	Italy
26254955	223	0	Russia
26362943	0	0	Germany
26519472	355	0	China
26725464	632	0	China
26742954	191	0	Japan
26823199	0	0	Japan
27311648	300	0	China
27439681	146	0	China
27480424	4	0	New Zealand
27557666	0	0	Sweden
27632209	0	0	Turkey
27790088	0	0	Germany
27978769	0	0	Brazil
28089114	0	0	Scotland
28105640	0	0	Australia
28160950	0	0	Japan
28222900	0	0	Hungary
28264768	0	0	Italy
28429524	500	0	China
28444446	0	0	Serbia
28749476	0	0	Germany
29033165	0	0	Japan

Supplementary Table S2.3: *C9orf72* Control Cohorts (2/2)

PMID	Number of Controls	Number of Positive Controls	Country
29476165	0	0	Finland
29525178	0	0	Italy
29650794	0	0	Germany
29748150	82	0	South_East_Asia
29861044	150	0	Portugal
29930232	0	0	Croatia
30054183	0	0	China
30054184	0	0	South_Korea
30528349	51	0	Greece
30599136	0	0	USA
30846540	0	0	Cuba
31537715	0	0	USA
31914217	0	0	USA
32166880	0	0	China
32409511	0	0	Australia

Supplementary Table S2.4: Excluded ACMG categories

Category	Description	Justification
PS3	Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product	Although TDP-43-positive aggregates are a common postmortem feature in ALS, and tau/TDP-43 deposits are frequently observed in FTD, there is little consensus on whether these inclusions are causative or emergent features of the disease. Furthermore, there are no universally established functional assays to assess the pathogenicity of potential ALS of FTD variants.
PM3	For recessive disorders, detected in trans with a pathogenic variant	ALS is an oligogenic disease (McCann et al. 2020,Cooper-Knock et al. 2017, van Blitterswijk et al. 2012), wherein even highly pathogenic variants may require additional variants to lead to disease.
PP5	Reputable source recently reports variant as pathogenic but the evidence is not available to the laboratory to perform an independent evaluation	This is an agnostic analysis of variants wherein previous variant classifications are purposefully disregarded.
BS2	Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder with full penetrance expected at an early age	ALS is a late onset disease with variants that exhibit reduced penetrance. A healthy adult carrying a variant may indicate that it is benign but may also indicate a presymptomatic individual or reduced variant penetrance.
BS3	Well-established in vitro or in vivo functional studies shows no damaging effect on protein function or splicing	There are no universally established pathogenicity assays for ALS which can be uniformly applied across different genes. As such this category could not be objectively assessed when performing an agnostic analysis of all variants.
BP2	Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder; or observed in cis with a pathogenic variant in any inheritance pattern	ALS is an oligogenic disease (McCann et al. 2020,Cooper-Knock et al. 2017, van Blitterswijk et al. 2012), wherein even highly pathogenic variants may require additional variants to lead to disease.
BP5	Variant found in a case with an alternate molecular basis for disease	ALS is an oligogenic disease (McCann et al. 2020,Cooper-Knock et al. 2017, van Blitterswijk et al. 2012), wherein even highly pathogenic variants may require additional variants to lead to disease.
BP6	Reputable source recently reports variant as benign but the evidence is not available to the laboratory to perform an independent evaluation.	This is an agnostic analysis of variants wherein previous variant classifications are purposefully disregarded.

Supplementary Table S2.5: Independent ACMG categories (1/3)

Category	Description	ACMG Categorisation	Treatment	Justification	Methods
PS2	De novo (both maternity and paternity confirmed) in a patient with the disease and no family history	StrP	StrP	NA	Variants marked as PM6 ModP are screened for confirmed parentage. If parentage is confirmed these variants are marked as PS2 StrP and PM6 reverts to null.
PS4	The prevalence of the variant in affected individuals is significantly increased compared to the prevalence in controls	StrP	ModP	PS4 relies on the frequency of variants in the Project MinE control cohort, which has previously undergone genome wide testing (van der Spek et al. 2019). To avoid retesting the same dataset, no significance testing is performed here. The category is downgraded to reflect this.	Variants present in the literature and the three ALS patient databases ALSdb, ALSVS and the ProjectMinE case cohort and which were absent in the ProjectMinE control cohort were designated PS4 ModP.
PM2	Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes or ExAC	ModP	StrP SupP	As variants are frequently absent in gnomAD, the ClinGen consortium propose reducing PM2 to SupP. Rather than compensating for this by altering the criteria required for a likely pathogenic classification as proposed by ClinGen, variants which are merely absent are distinguished from variants with robust statistical support.	Variants which are present in the literature and absent in gnomAD are classed as PM2 StrP. Fisher exact tests comparing the AF of variants in the literature to the AF of variants in gnomAD were performed. A p-value threshold of $1.9x10^{-5}$ was identified by dividing 0.05 by the number of testable variants (variants present in both the population subset of our data and in the gnomAD controls subset). Variants that are significantly more common in the literature are classed as PM2 StrP.
PM4	Protein length changes due to in-frame deletions/insertions in a non-repeat region or stop-loss variants	ModP	ModP	NA	PM4 ModP is assigned for in-frame INDELs falling in a non-repeat region as defined by the UCSC RepeatMasker tract
PM6	Assumed de novo, but without confirmation of paternity and maternity	ModP	ModP	NA	Variants are preliminarily classed as PM6 ModP if two independent cases are linked to the same de novo variant. Genes carrying a preliminary PM6 ModP de novo variant are classed as de novo susceptible, these genes are rescreened and variants with only one de novo case are marked as PM6 ModP. If variants marked as PM6 ModP have confirmed parentage, PS2 is upgraded to StrP and PM6 reverts to null.
PP1	Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease	SupP ModP StrP	SupP ModP StrP	NA	As outlined in methods, the counting meioses method of (Jarvik and Browning 2016) is the most appropriate method of quantifying segregation evidence from the available data. The cutoffs suggested by Jarvik et. al. are used (Supplementary Table S2.7).

Supplementary Table S2.5: Independent ACMG categories (2/3)

Category	Description	ACMG Categorisation	Treatment	Justification	Methods
PP3 BP4	Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc)	SupP SupB	SupP SupB	NA	The suggested pathogenic and benign cutoffs are used for in silico pathogenicity prediction tools where available. Otherwise suggested thresholds from dbNSFPv4.0a or from a review of the in silico literature were used (Li et al. 2018). Thresholds used are available in supplementary table s8. For in silico predictions coding SNVs are classified as per (Ghosh et al. 2017). MutationTaster (Schwarz et al. 2014), Mcap (Jagadeesh et al. 2016), and CADD scores are checked for pathogenic agreement. VEST4, REVEL (Ioannidis et al. 2016), and MetaSVM (Kim et al. 2017) scores are checked for benign agreement. If a variant has both pathogenic and benign agreement, PP3 and BP4 are marked as null. INDELs are checked for pathogenic or benign agreement with CADD, SIFT INDEL and VEST4. Splicing variants are screened for pathogenic or benign agreement with AdaBoost, randomForest and CADD. For variants that either do not fit one of the above categories (e.g. intronic variants) or variants which do not have a prediction for one of the three tools against which it is screened, all calls from all tools are checked. A categorisation is made if predictions are available for three or more tools and they are in pathogenic or benign agreement.
PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic etiology	SupP	StrP	This category is only loosely described in the ACMG guidelines; however here robust quantitative statistical evidence is replied upon to test whether carriers of a variant share a common phenotype indicating a common molecular mechanism.	As described in methods, a Kruskal-Wallis test is used to test whether carriers of the variant of interest display significantly early or late disease-onset relative to the rest of the collected cohort.
BA1	Allele frequency too high in reference databases	SAB	SAB StrB	Account for varying strengths of evidence	A variant is assigned as BA1 SAB if the gnomAD AF is greater than or equal to 0.01. A variant is assigned as StrB if the gnomAD AF is below 0.01, the penetrance is less than 1% and the variant is not homozygous in any reported individual in the literature.
BS1	Allele frequency is greater than expected for disorder	StrB	StrB	NA	A variant is assigned BS1 StrB if the Project MinE control AF is greater than the Project MinE case AF.

Supplementary Table S2.5: Independent ACMG categories (3/3)					
Category	**Description**	**ACMG Categorisation**	**Treatment**	**Justification**	**Methods**
BS4	Lack of segregation in affected members of a family	StrB	SupB	The oligogenic nature of ALS (Cooper-Knock et al. 2017; McCann et al. 2020; van Blitterswijk et al. 2012) implies that a variant may not segregate entirely in a pedigree but may still be influencing disease where present.	BS4 SupB was assigned if any affected individual was homozygous for the reference allele.
BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease	SupB	SupB	NA	BP1 SupB is assigned if a missense variant is present in a gene with a gnomAD constraint missense z score below -2, which strongly indicates that the gene is tolerant of missense variants.
BP3	In-frame deletions/insertions in a repetitive region without a known function	SupB	SupB	NA	BP3 SupB was assigned for in-frame INDELs falling in a repetitive region as predicted by the UCSC RepeatMasker tract
BP7	A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved	SupB	SupB	NA	BP7 SupB is assigned if the variant is predicted to be synonymous on the most severely affected transcript.

StrP: Strong pathogenicModP: Modetate pathogenicSupP: Supporting pathogenic SAB: Stand-along benignStrB: Strong BenignSupB: Supporting benign

Supplementary Table S2.6: Dependent ACMG categories

Category	Description	ACMG Categorisation	Treatment	Justification	Methods
PVS1	Null variant (nonsense, frameshift, canonical +/−1 or 2 splice sites, initiation codon, single or multi-exon deletion) in a gene where loss of function (LOF) is a known mechanism of disease	VStrP	VStrP StrP ModP SupP	The original ACMG guidelines did not take into account the varying strengths of evidence that can contribute to this categorisation (Abou Tayoun et al. 2018)	Null variants are those with assigned impacts: splice_acceptor_variant, stop_gained, frameshift_variant, initiator_codon_variant, splice_donor_variant, start_lost or stop_lost. The process of assigning PVS1 is outlined in Supplementary Figure S2.2
PS1	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change	StrP	StrP	NA	Variant impact was assigned using gemini and SnpEff as described in methods. Following the first round of independent ACMG assessment, missense variants with the same amino acid change as variants deemed 'P' or 'LP' were assigned PS1 StrP
PM1	Located in a mutational hot spot and/or critical and well-established functional domain (e.g. active site of an enzyme) without benign variation	ModP	ModP	NA	Variants are assigned PM1 ModP if they are a missense variant falling in an InterPro domain which contains more than one pathogenic or likely pathogenic variant from the initial independent ACMG screen and no benign or likely benign variants
PM5	Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before	ModP	ModP	NA	Variants are assigned as PM5 ModP if they are a novel missense change at an amino acid residue found to be pathogenic or likely pathogenic following the first round independent screen
PP2	Missense variant in a gene that has a low rate of benign missense variation and where missense variants are a common mechanism of disease	SupP	SupP	NA	Genes with a low rate of benign variation were defined as those with a gnomAD constraint missense z score >2. Genes where missense variants are a known mechanism of disease are defined as those with a pathogenic or likely pathogenic variant from the initial independent ACMG screen or genes with more than one missense variant with strong or moderate segregation

VStrP: Very strong pathogenicStrP: Strong pathogenicModP: Modetate pathogenicSupP: Supporting pathogenic

Supplementary Table S2.7: Meioses Count Thresholds for PP1

	Single Family	>1 Families
Strong evidence	>1/16	>1/8
Moderate evidence	≤1/16	≤1/8
Supporting evidence	≤1/8	≤1/4

Supplementary Table S2.8: Cutoffs used for *in silico* prediction software

Software	Tool
CADD	15
Provean	-2.5
VEST	0.5
REVEL	0.4
MetaSVM	0
MutationTaster	0.31733
MCap	0.025
AdaBoost	0.6
RandomForest	0.6

Supplementary Table S2.9: Oligogenic carriers

Variant1	Variant2	PMID	Phenotype
C9orf72:c.-45+163GGGGCC[>24]	*FUS*:c.1474C>T(p.[R492C])	26176978	ALS-FTD
C9orf72:c.-45+163GGGGCC[>24]	*GRN*:c.87_90dupCTGC(p.[C31fs])	24286341	FTD-MND
C9orf72:c.-45+163GGGGCC[>24]	*OPTN*:c.1403T>G(p.[M468R])	29080331	ALS-FTD
C9orf72:c.-45+163GGGGCC[>24]	*OPTN*:c.1403T>G(p.[M468R])	29080331	ALS-FTD

Note: previous research has indicated that carriers of certain variant combinations either develop ALS or FTD (Nguyen, Van Broeckhoven and van der Zee, 2018) . This table outlines individuals in the journALS database who contradict this finding.

Supplementary Table S.2.10: Discordant pedigrees

Pedigree	PMID	Discordant Variant	Note
20460594_1	20460594	*SOD1*:c.301G>A(p.[E101K])	For confirmation DNA was recollected and checked independently by three separate labs using three separate sets of primers
20460594_2	20460594	*SOD1*:c.301G>A(p.[E101K])	For confirmation DNA was recollected and checked independently by three separate labs using three separate sets of primers
20460594_3	20460594	*SOD1*:c.272A>C(p.[D91A])	Where D91A was present in cases it was homozygous. For confirmation DNA was recollected and checked independently by three separate labs using three separate sets of primers
20460594_4	20460594	*SOD1*:c.272A>C(p.[D91A])	Where D91A was present in cases it was homozygous. For confirmation DNA was recollected and checked independently by three separate labs using three separate sets of primers
22550220_1	22550220	*C9orf72*:c.-45+163GGGGCC[>24] *TARDBP*:c.1144G>A(p.[A382T])	Pedigree has two segregating pathogenic variants
22645277_1	22645277	*TARDBP*:c.1055A>G(p.[N352S])	Pedigree also has a partially segregating *ANG*:c.122A>T(p.[K41I]) VUS
22645277_4	22645277	*C9orf72*:c.-45+163GGGGCC[>24]	Pedigree also has a segregating *TARDBP*:c.1055A>G(p.[N352S]), which is present in all affected individuals who were screened for the variant.
26839080_1	26839080	*C9orf72*:c.-45+163GGGGCC[>24]	This discordance of this pedigree is ambigous. The pedigree also has a segregating *SQSTM1*:c.1175C>T(p.[P392L]) variant. The pedigree exhibits Paget's Disease of Bone, Cognitive impairment from childhood encephalopathy, FTD, and Parkinson's disease. There is a single individual who does not have the *C9orf72* repeat expansion however they only exhibit PDB and cognitivie impairment but not FTD.
32223976_1	32223976	*SOD1*:c.14C>T(p.[A5V])	The pedigree also has a discordantly segregating *OPTN*:c.138G>C(p.[E46D]) VUS

Supplementary Table S.2.11: Minimal reporting guidelines for future integration

Category	Explanation
Population Matched Controls	This study has demonstrated the significant geographic heterogeneity that variants can exhibit. Population databases such as gnomAD may be depleted for the population of interest. It is important to know if an identified variant is enriched in your ALS/FTD cohort or in your population in general
Pedigrees	Clearly identify all relevant members of a pedigree
	Distinguish cases from controls
	Distinguish sequenced individuals from unsequenced
	Distinguish variant carriers from non variant carriers
	List AOO/ age at death / current age/ disease duration where applicable
	Outline if pedigree has been reported before
Cohort selection	For screening studies the preference should be for an unbiased cohort representative of the overall study population
	If the cohort is biased please state any biases e.g. Were they previously negatively screened for any genes / variants, a particular family history, a specific AOO, a specific sub-phenotype?
Phenotype Reporting	Details of individual phenotypes as well as a summary of the overall cohort e.g. ALS-FRS, family history
Previous Reports	Clearly state whether a pedigree/ individual / cohort has been previously reported
De novo	If a variant has been found to be *de novo* is the parentage confirmed
Cohort size	Clearly state size of study cohort

Chapter 3

Supplementary Table S3.1: Repeat Availability Across Software

Gene	Motif	EH2	EH3	exSTRa	GangSTR	HipSTR	RepeatSeq	STRetch	TREDPARSE
AFF2	CCG	No	Yes	Yes	Yes	No	Yes	Yes	Yes
AR	CAG	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
ARX	GCG	No	No	No	Yes	Yes	Yes	Yes	Yes
ATN1	CAG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ATXN1	CAG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ATXN10	ATTCT	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
ATXN2	CAG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ATXN3	CAG	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
ATXN7	CAG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ATXN8OS	CTG.CAG	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
C9orf72	GGGGCC	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CACNA1A	CAG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CBL	CCG	Yes	Yes	No	Yes	No	Yes	No	Yes
CNBP	CCTG/CAGG	No	Yes	Yes	Yes	No	Yes	Yes	Yes
CSTB	C4GC4GCG	Yes	Yes	Yes	Yes	No	No	No	Yes
DIP2B	GGC	No	Yes	No	Yes	No	Yes	Yes	Yes
DMPK	CAG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FMR1	CGG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FOXL2	GCN	No	No	No	No	No	No	No	Yes
FXN	GAA	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
GIPC1	CCG	No	Yes	No	Yes	No	No	Yes	Yes
GLS	GCA	No	Yes	Yes	Yes	No	Yes	Yes	Yes
HOXA13	GCN	No	No	No	No	No	No	No	Yes
HOXD13	GCN	No	No	No	No	Yes	No	No	Yes
HTT	CAG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
JPH3	CTG/CAG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LRP12	CGG	No	No	Yes	No	No	No	Yes	Yes
MARCHF6	TTTTA(TTTCA)$_N$TTTTA	No	No	Yes	No	No	No	No	No
NIPA1	CGC	No	Yes	No	Yes	No	Yes	No	Yes
NOP56	GGCCTG	No	Yes	Yes	Yes	Yes	No	Yes	Yes
NOTCH2NLA	CGG	No	No	Yes	No	No	No	No	Yes
NUTM2B	CGG/CCG	No	No	Yes	Yes	No	No	No	Yes
PAPBN1	GCN	No	Yes	No	Yes	No	Yes	Yes	Yes
PHOX2B	GCN	No	Yes	No	No	No	Yes	Yes	Yes
PPP2R2B	CAG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RAPGEF2	TTTTA(TTTCA)$_N$TTTTA	No	No	Yes	No	No	No	No	No
RFC1	AAAAG	No	Yes	Yes	Yes	done	Yes	Yes	Yes
RUNX2	GCN	No	No	No	No	No	No	No	Yes
SAMD12	TTTTA(TTTCA)$_N$TTTTA	No	No	Yes	No	No	No	No	No
SOX3	GCN	No	No	No	No	Yes	No	No	Yes
STARD7	TTTTA(TTTCA)$_N$TTTTA	No	No	Yes	No	No	No	No	No
TBP	CAN	No	Yes	Yes	Yes	No	Yes	Yes	Yes
TCF4	CTG	No	Yes	Yes	No	No	Yes	Yes	Yes
TNRC6A	TTTTA(TTTCA)$_N$TTTTA	No	No	Yes	No	No	No	No	No
YEATS2	TTTTA(TTTCA)$_N$TTTTA	No	No	Yes	No	No	No	No	No
ZIC2	GCN	No	No	No	No	Yes	No	No	Yes

Supplementary Table S.3.2 : Samples with Significant Expansions as Predicted by exSTRa (1/2)

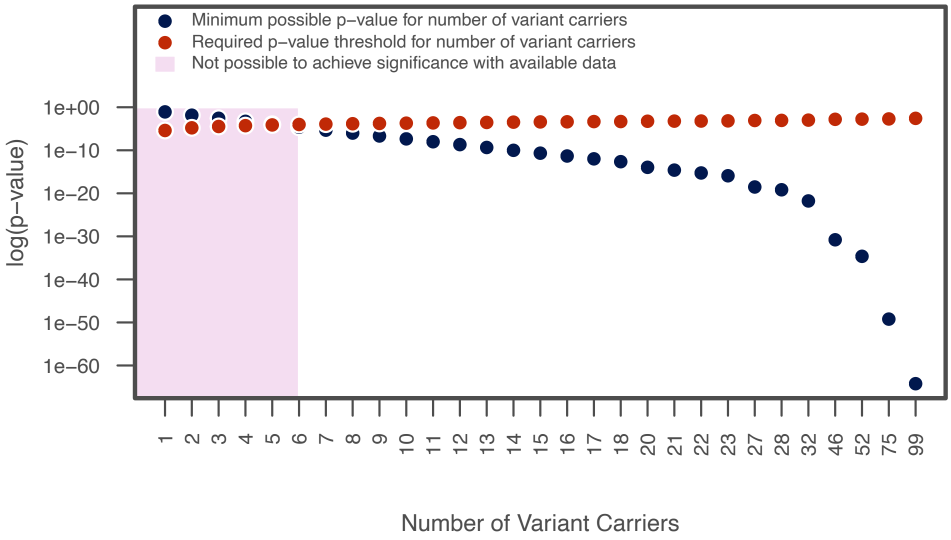
		exSTRa p-value	STRetch Significant	ExpansionHunter version 3			GangSTR (Target Mode)			ExpansionHunter version 2			TREDPARSE		
Sample	**Gene**			**Patient**	**Paternal**	**Maternal**	**Patient**	**Paternal**	**Maternal**	**Patient**	**Paternal**	**Maternal**	**Patient**	**Paternal**	**Maternal**
EP5A	*AR*	4.81248E-06	No	21/23	21	23/24	20/22	20	2/22	21/23	21	23/24	20/22	20	22/24
EP6A	*AR*	4.81248E-06	No	21/25	9	21/27	20/24	N/A	20/25	21/25	24	21/27	20/24	32	20/26
EP7A	*AR*	4.81248E-06	No	19/21	N/A	N/A	N/A	N/A	N/A	21/21	N/A	N/A	20/20	N/A	N/A
EP8A	*ATN1*	0.001301295	No	21/29	18/27	14/21	17/25	15/22	10/16	21/29	N/A	14/21	N/A	N/A	N/A
EP5A	*ATN1*	4.76665E-06	No	19/20	19/20	19/21	15/16	15/16	15/17	19/20	19/20	19/21	15/16	15/16	15/17
EP6A	*ATN1*	9.5333E-06	No	14/19	19/19	14/19	10/15	15/17	10/15	14/19	19/21	14/19	10/15	15/17	10/15
EP7A	*ATN1*	4.76665E-06	No	19/19	N/A	N/A	15/15	N/A	N/A	19/19	N/A	N/A	15/15	N/A	N/A
EP5A	*ATXN1*	4.78946E-06	No	31/31	30/31	31/31	N/A	N/A	N/A	N/A	N/A	N/A	30/30	29/30	30/30
EP6A	*ATXN1*	4.78946E-06	No	29/31	27/30	28/29	N/A	N/A	N/A	N/A	N/A	N/A	28/30	21/36	27/28
EP7A	*ATXN1*	4.78946E-06	No	29/31	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	28/30	N/A	N/A
EP6A	*ATXN2*	0.000541221	No	19/22	N/A	22/22	N/A	N/A	N/A	N/A	N/A	N/A	19/22	N/A	N/A
EP5A	*ATXN3*	4.76665E-06	No	25/28	18/28	11/25	N/A	N/A	N/A	N/A	N/A	N/A	22/25	15/25	8/22
EP6A	*ATXN3*	9.5333E-06	No	11/35	N/A	11/35	N/A	N/A	N/A	N/A	N/A	N/A	8/30	8/8	8/32
EP7A	*ATXN3*	0.000352732	No	11/35	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	8/32	N/A	N/A
EP5A	*ATXN7*	0.000404548	No	10/10	10/10	10/10	N/A	N/A	N/A	N/A	N/A	N/A	10/10	10/10	10/10
EP6A	*ATXN7*	0.000114103	No	10/10	N/A	10/19	N/A	N/A	N/A	N/A	N/A	N/A	10/10	10/10	10/10
EP7A	*ATXN7*	0.000269698	No	3/10	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	10/10	N/A	N/A
EP9A	*ATXN8OS*	0.001167829	No	26/32	N/A	N/A	15/21	15/22	15/15	N/A	N/A	N/A	15/22	15/21	15/15
EP10A	*ATXN8OS*	0.000185899	No	N/A	N/A	N/A	14/22	15/20	15/16	N/A	N/A	N/A	14/19	15/21	15/16
EP5A	*ATXN8OS*	1.43E-05	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	15/16	9/15	15/16
EP7A	*ATXN8OS*	4.76665E-06	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	15/17	N/A	N/A
EP11A	*DMPK*	3.35262E-05	No	20/23	N/A	N/A	N/A	N/A	N/A	20/23	N/A	N/A	20/23	N/A	N/A
EP5A	*DMPK*	4.78946E-06	No	35/42	40/42	8/35	N/A	N/A	N/A	35/35	40/42	8/35	N/A	N/A	N/A
EP12A	*DMPK*	0.00085796	No	5/22	5/22	N/A	N/A	N/A	N/A	5/22	5/22	N/A	5/22	5/22	N/A
EP13A	*DMPK*	0.001143947	No	12/25	5/25	11/12	N/A	N/A	N/A	12/25	5/25	11/12	12/25	5/25	11/12
EP5A	*FMR1*	0.000118095	No	N/A	N/A	N/A	N/A	N/A	N/A	17/17	3	6/6	17/17	7	6/6
EP6A	*FMR1*	6.74828E-05	No	N/A	N/A	N/A	N/A	N/A	N/A	13/13	N/A	12/12	14/14	N/A	15/21
EP7A	*FMR1*	0.000579228	No	N/A	N/A	N/A	N/A	N/A	N/A	7/7	N/A	N/A	16/16	N/A	N/A
EP9A	*FXN*	0.0004078	No	9/18	N/A	N/A	9/16	N/A	N/A	9/18	N/A	N/A	9/18	N/A	N/A
EP14A	*GLS*	0.000801934	No	N/A	N/A	N/A	14/14	14/17	14/14	N/A	N/A	N/A	N/A	8/9	8/8
EP5A	*GLS*	7.39372E-05	No	14/15	N/A	14/15	11/14	14/14	11/14	N/A	N/A	N/A	14/15	8/15	14/15
EP6A	*GLS*	2.27499E-05	No	13/18	N/A	15/18	14/14	14/15	14/14	N/A	N/A	N/A	14/18	14/14	15/18
EP7A	*GLS*	4.54998E-05	No	14/18	N/A	N/A	14/14	N/A	N/A	N/A	N/A	N/A	14/18	N/A	N/A
EP15A	*HTT*	0.000115057	No	N/A	N/A	N/A	N/A	N/A	N/A	17/26	17/17	17/20	22/22	17/18	20/20
EP5A	*HTT*	1.15057E-05	No	19/22	16/22	19/28	19/21	N/A	N/A	19/22	16/22	19/28	19/22	16/22	19/28
EP5A	*JPH3*	4.76665E-06	No	11/14	14/14	11/14	11/14	14/14	14/14	11/14	14/14	11/14	11/14	14/14	11/14
EP6A	*JPH3*	4.76665E-06	No	14/14	14/15	14/14	14/14	14/15	14/15	14/14	14/15	14/14	14/14	14/15	14/14
EP7A	*JPH3*	4.76665E-06	No	14/14	N/A	N/A	14/14	N/A	N/A	14/14	N/A	N/A	14/14	N/A	N/A
EP5A	*LRP12*	4.33123E-05	No	N/A	N/A	N/A	4/4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EP6A	*LRP12*	0.000611185	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EP5A	*NOTCH2*	9.76582E-05	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EP6A	*NOTCH2*	0.000126956	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EP7A	*NOTCH2*	0.000102541	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EP6A	*NUTM2B*	0.000667314	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EP7A	*NUTM2B*	0.001031303	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Supplementary Table S.3.2 : Samples with Significant Expansions as Predicted by exSTRa (2/2)

		exSTRa p-value	STRetch Significant	ExpansionHunter version 3			GangSTR (Target Mode)			ExpansionHunter version 2			TREDPARSE		
Sample	**Gene**			**Patient**	**Paternal**	**Maternal**	**Patient**	**Paternal**	**Maternal**	**Patient**	**Paternal**	**Maternal**	**Patient**	**Paternal**	**Maternal**
EP16A	*RFC1*	5.26815E-05	No	9/38	9/36	9/33	9/9	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EP13A	*SAMD12*	0.000476644	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EP5A	*TBP*	4.76665E-06	No	36/37	37/37	36/37	N/A	N/A	N/A	N/A	N/A	N/A	37/38	38/38	37/38
EP6A	*TBP*	4.76665E-06	No	37/37	49/77	37/37	N/A	N/A	N/A	N/A	N/A	N/A	37/38	24/136	38/66
EP7A	*TBP*	4.76665E-06	No	34/37	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	35/38	N/A	N/A
EP17A	*TCF4*	0.000410341	No	11/20	11/17	12/20	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EP18A	*TCF4*	0.000572568	No	20/20	9/19	15/17	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EP19A	*YEATS2*	5.00476E-05	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

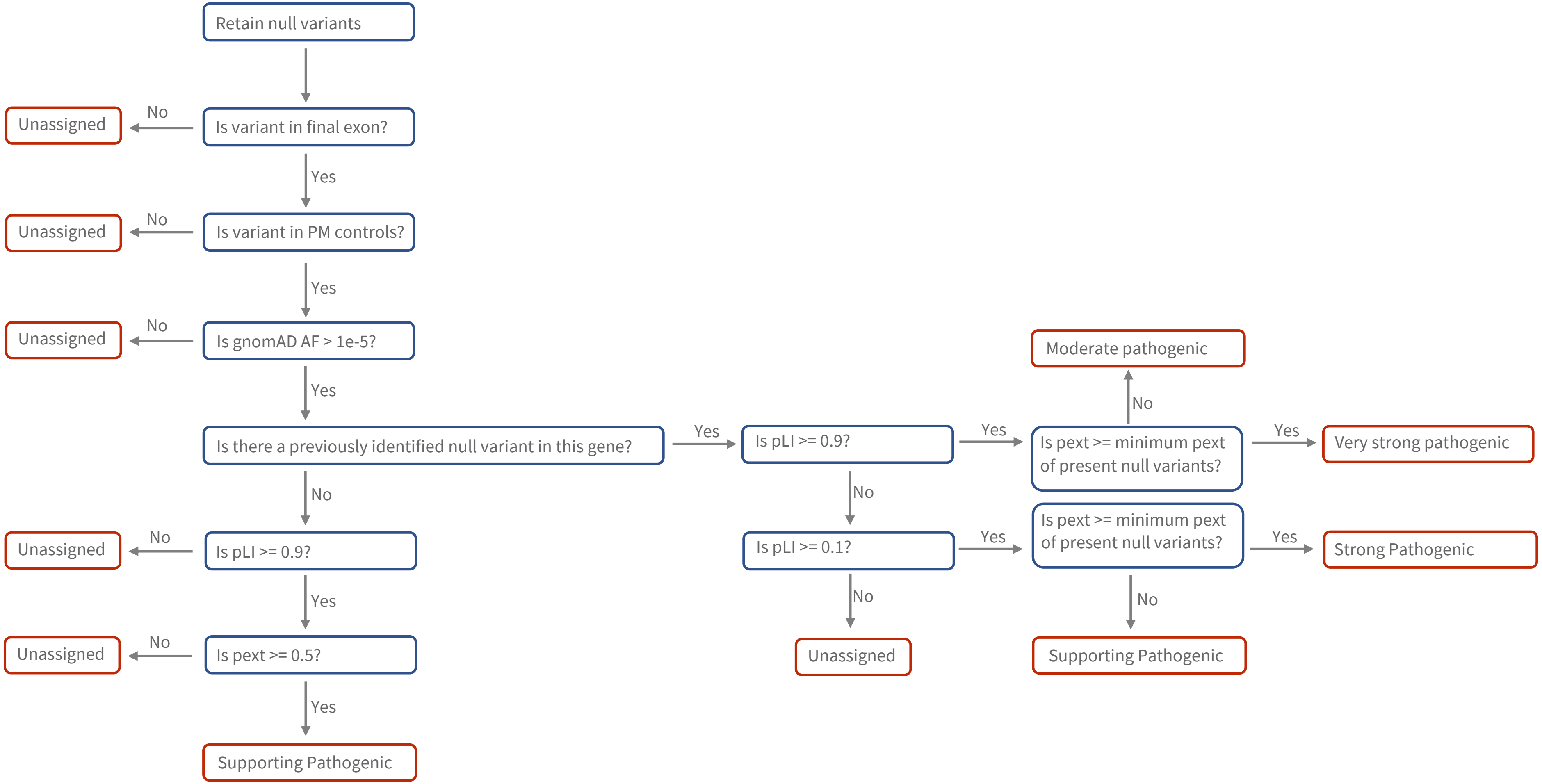
Supplementary Figures

Chapter 2



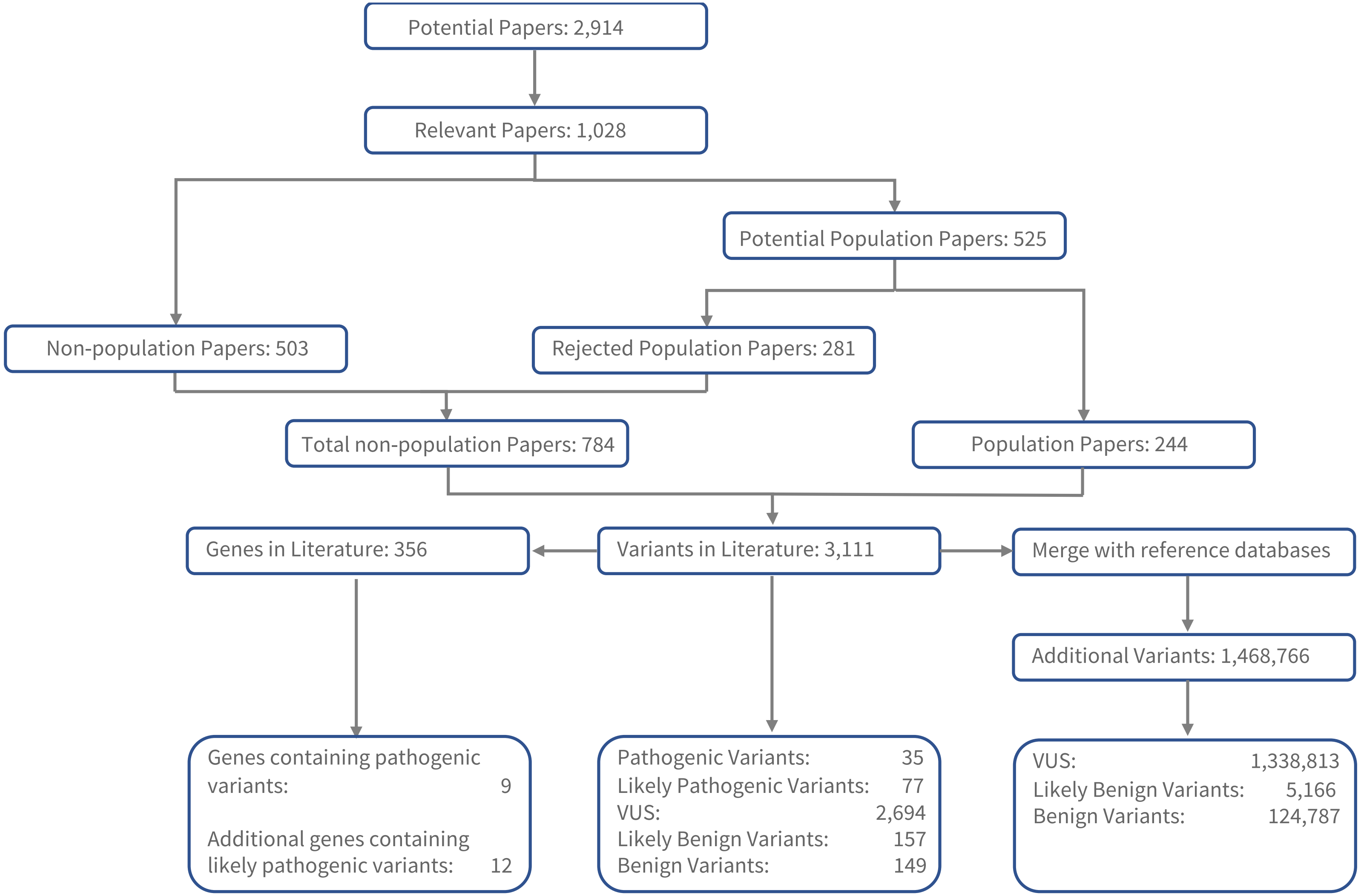
Supplementary Figure 2.1: Identifying p-value threshold for age of onset comparisons

Variant carriers can be categorised based on phenotype (all, ALS, FTD), sex (all, male, female) and family history (all, sporadic, familial); thus 27 tests comparing the age of onset of carriers of a particular variant to the remainder of the cohort can be conducted per variant. It is demonstrated that if a category has below six variant carriers it is impossible to achieve a significant p-value after correcting for the number of tests performed. Only categories with six or more variant carriers are tested and a p-value of 9.75x10^{-5} is required.



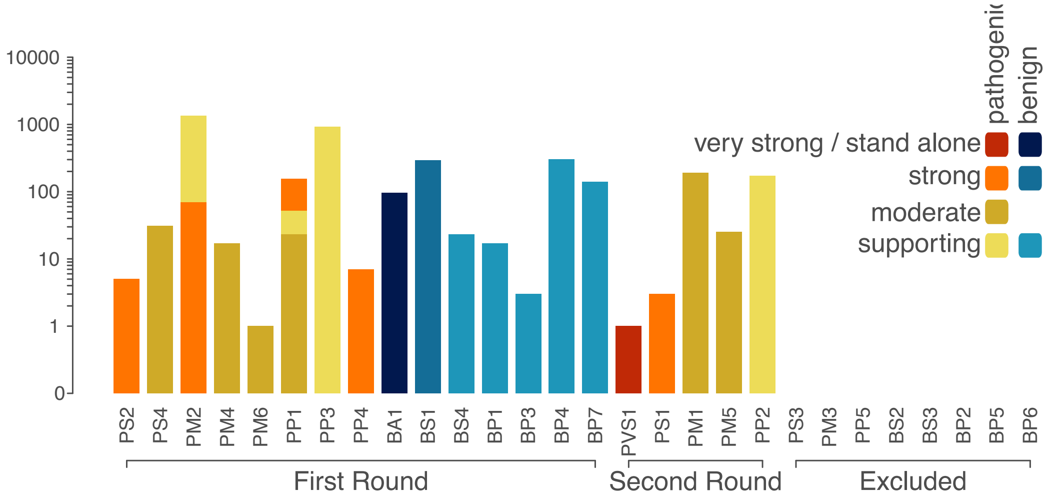
Supplementary Figure 2.2: Workflow for assigning ACMG category PVS1

Various gene properties are taken into account when assigning ACMG category PVS1. Variants are excluded if they fall in the final exon as per (Abou Tayoun et al. 2018). Variants common in gnomAD are excluded. The gnomAD probability of loss of function intolerance (pLI) score for a gene indicates how resilient a gene is to null variants. The gnomAD proportion expressed across transcripts (pext) score is a useful predictor of pathogenicity for null variants (Cummings et al. 2020).


Potential Papers: 2,914
Relevant Papers: 1,028
Potential Population Papers: 525
Non-population Papers: 503
Rejected Population Papers: 281
Total non-population Papers: 784
Population Papers: 244
Genes in Literature: 356
Variants in Literature: 3,111
Merge with reference databases
Additional Variants: 1,468,766
Genes containing pathogenic variants: 9
Additional genes containing likely pathogenic variants: 12
Pathogenic Variants: 35
Likely Pathogenic Variants: 77
VUS: 2,694
Likely Benign Variants: 157
Benign Variants: 149
VUS: 1,338,813
Likely Benign Variants: 5,166
Benign Variants: 124,787

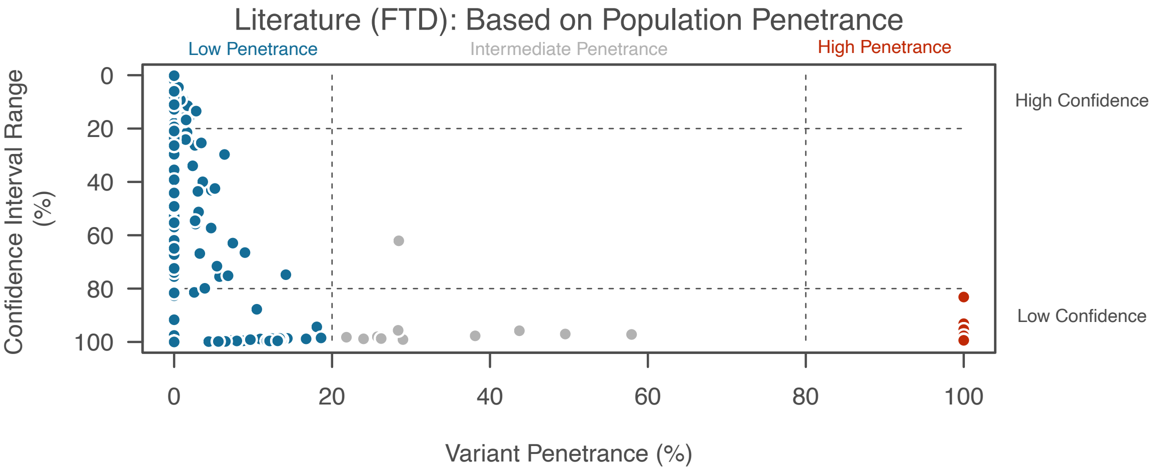
Supplementary Figure 2.3: Study workflow

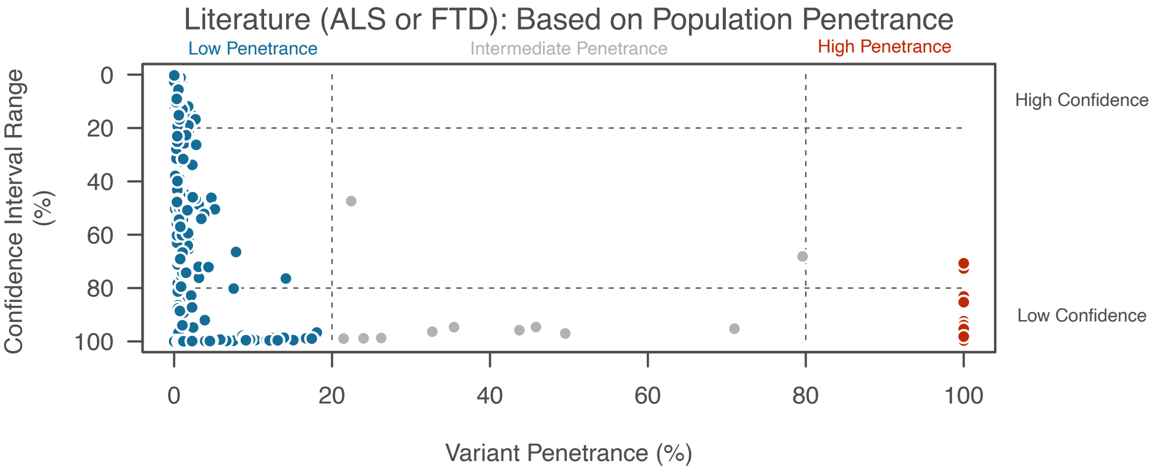
Figure outlining the filtering and processing of data in this study



Supplementary Figure 2.4: ACMG categories

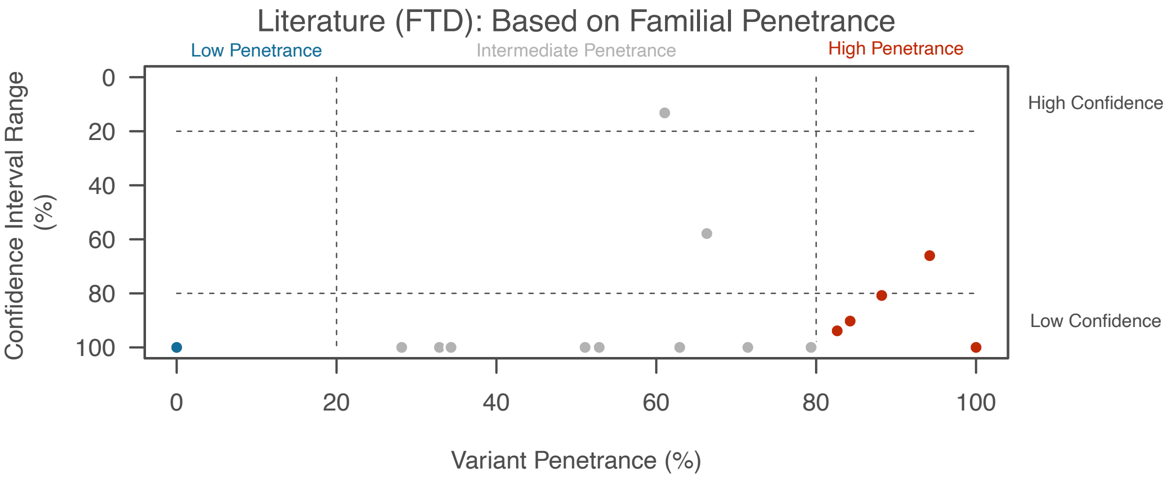
Plot displays the number of times each category was fulfilled when applying ACMG categorisation to our dataset

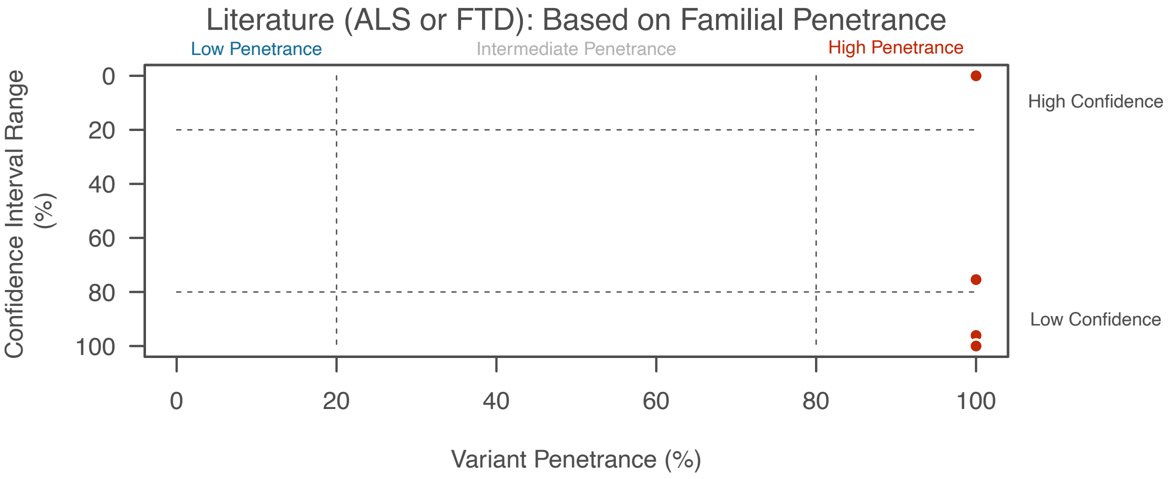




Supplementary Figure 2.5: FTD and ALS-FTD population penetrance estimates

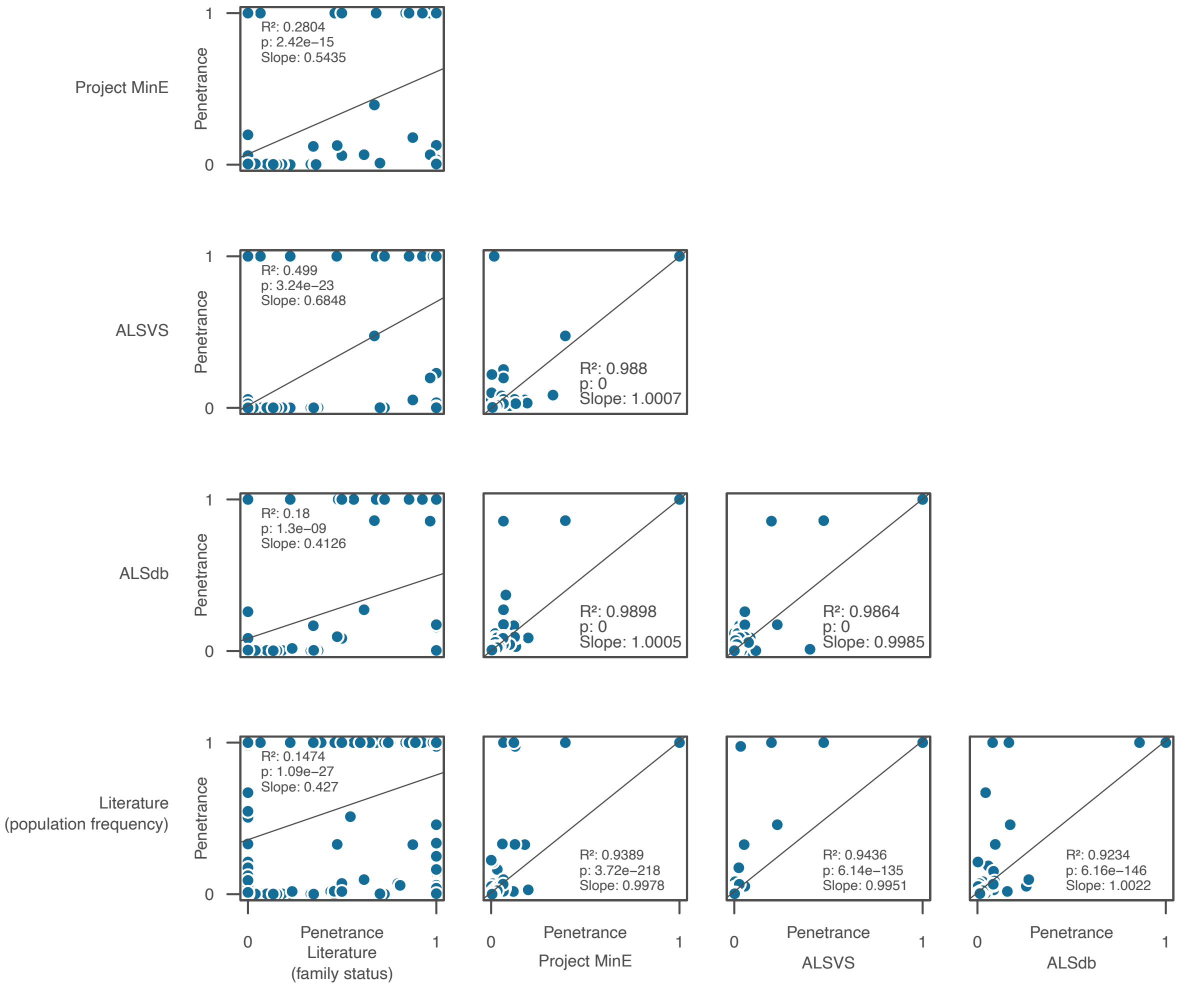
A) The FTD population penetrance estimates are shown here for 791 variants that had an FTD AF calculated from the literature and an available gnomAD AF. The majority of these variants have low penetrance with high confidence. Due to the high lifetime risk of FTD and the low AF of each variant, this method struggles to confidently identify intermediate and high penetrance variants. B) The lifetime risk of developing ALS or FTD is calculated via the population penetrance method for 649 variants which had both and ALS and an FTD AF calculated from the literature.





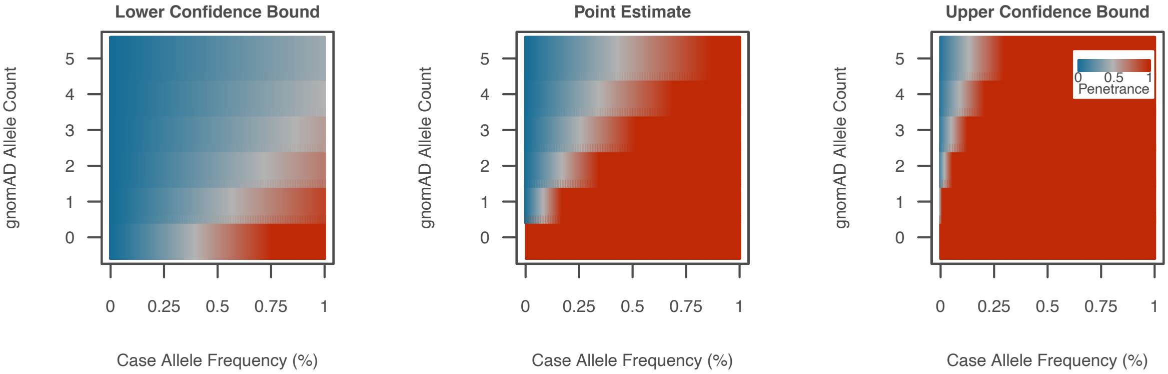
Supplementary Figure 2.6: FTD and ALS-FTD familial penetrance estimates

A) The FTD familial penetrance estimates are shown here for 104 variants have a calculated AF in fFTD and sFTD cases. B) The lifetime risk of developing ALS or FTD is calculated via the familial penetrance method for 10 variants which have an AF calculated in fALS, sALS, fFTD and sFTD cases.



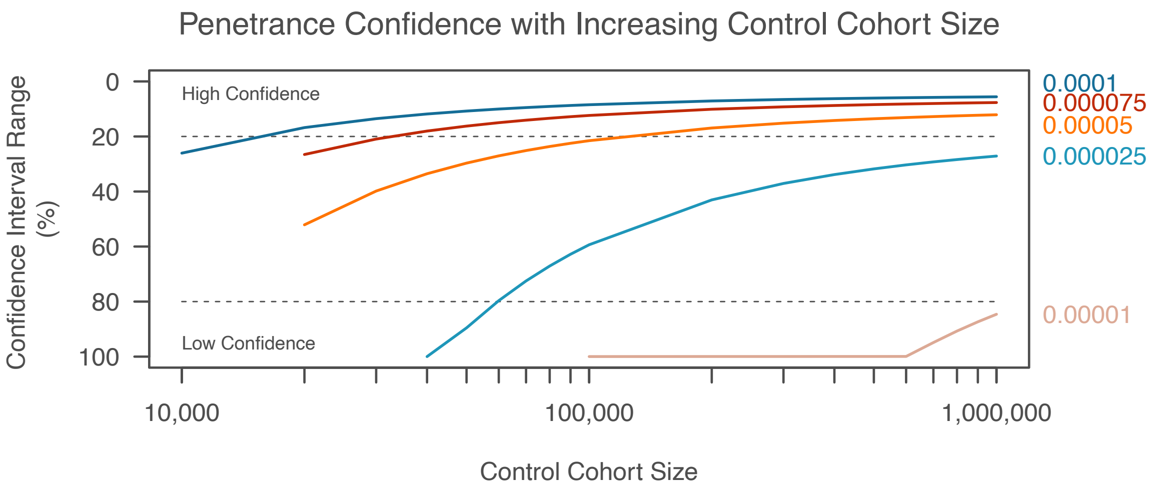
Supplementary Figure 2.7: Penetrance estimate comparisons

ALS penetrance estimates are calculated via the population penetrance method for AFs observed in the literature, the Project MinE case series, ALSdb and ALSVS. These are compared to each other and to the familial penetrance estimates calculated based on the AF in fALS and sALS cases. Population penetrance estimates from different datasets correlate well, highlighting the reliability of the literature collection. There is less correlation when comparing to the familial penetrance method, this reflects the inherent large confidence intervals of these two methods.



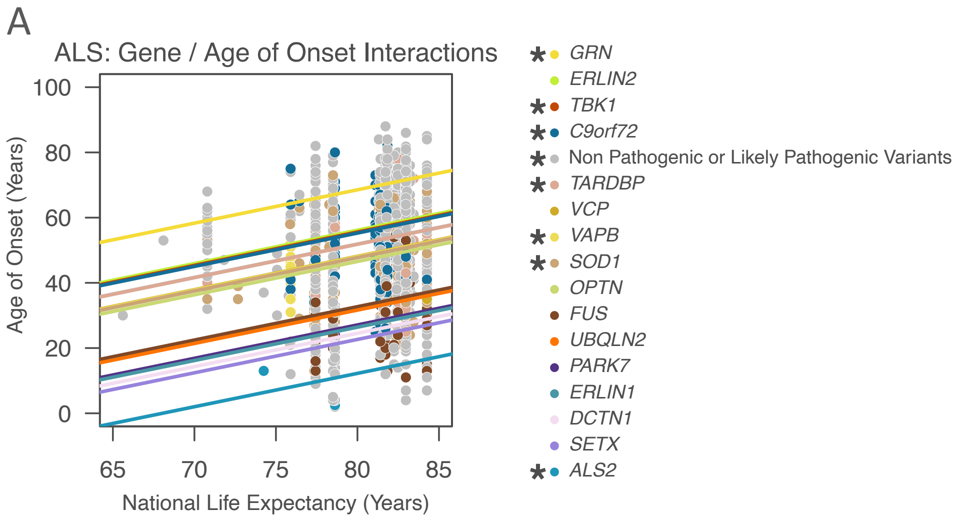
Supplementary Figure 2.8: ALS population penetrance modelling

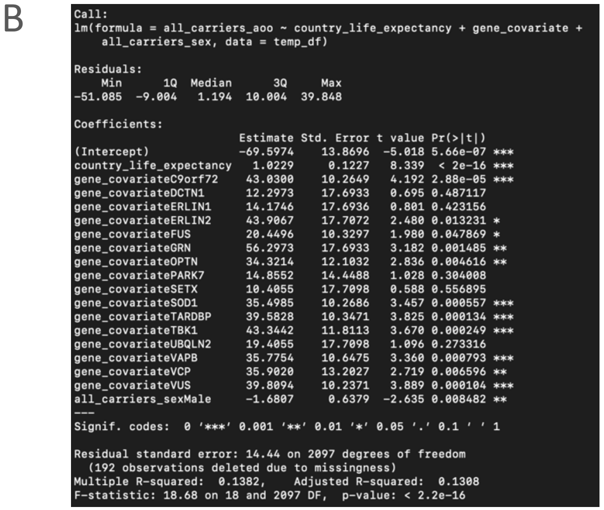
We calculate that even a dataset of 15,000 cases (the target size of Project MinE) will struggle to confidently identify high and intermediate penetrance variants due to the high lifetime risk of ALS.



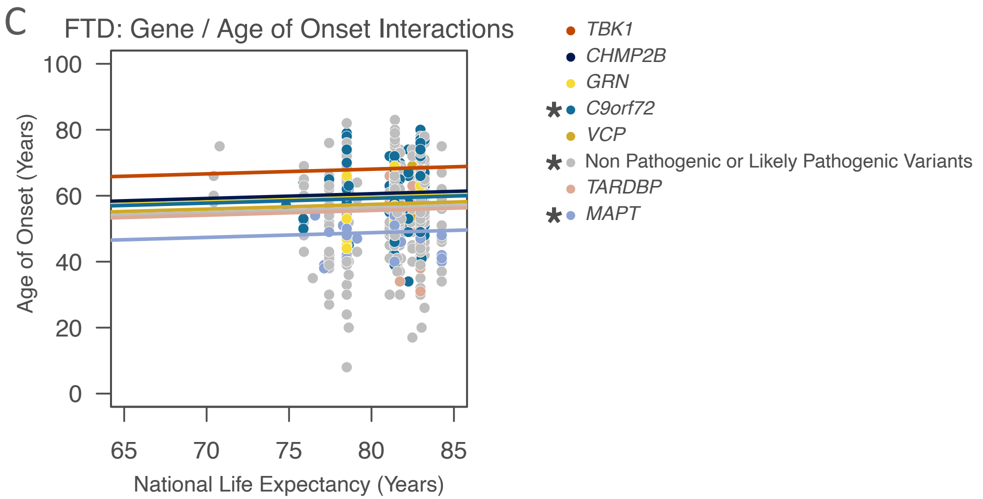
Supplementary Figure 2.9: Relationship between control cohort size and penetrance estimates confidence

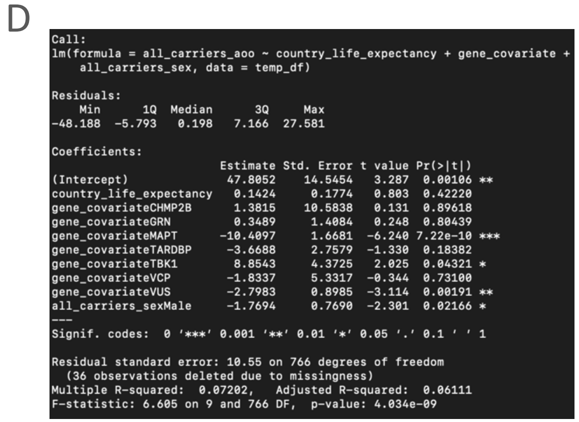
The figure demonstrates that increasing the size of the available control cohort can increase the confidence with which penetrance estimates can be calculated without increasing the size of case cohorts. The range of penetrance confidence estimates are plotted for hypothetical variants with a fixed case AF of 1x10^{-3} and with a lifetime risk of 1/400 and control AFs ranging from 1x10^{-4} to 1x10^{-5}.





```
Call:
lm(formula = all_carriers_aoo ~ country_life_expectancy + gene_covariate +
    all_carriers_sex, data = temp_df)

Residuals:
    Min      1Q  Median      3Q     Max
-51.085  -9.004   1.194  10.004  39.848

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -69.5974    13.8696  -5.018 5.66e-07 ***
country_life_expectancy   1.0229     0.1227   8.339  < 2e-16 ***
gene_covariateC9orf72    43.0300    10.2649   4.192 2.88e-05 ***
gene_covariateDCTN1      12.2973    17.6933   0.695 0.487117
gene_covariateERLIN1     14.1746    17.6936   0.801 0.423156
gene_covariateERLIN2     43.9067    17.7072   2.480 0.013231 *
gene_covariateFUS        20.4496    10.3297   1.980 0.047869 *
gene_covariateGRN        56.2973    17.6933   3.182 0.001485 **
gene_covariateOPTN       34.3214    12.1032   2.836 0.004616 **
gene_covariatePARK7      14.8552    14.4488   1.028 0.304008
gene_covariateSETX       10.4055    17.7098   0.588 0.556895
gene_covariateSOD1       35.4985    10.2686   3.457 0.000557 ***
gene_covariateTARDBP     39.5828    10.3471   3.825 0.000134 ***
gene_covariateTBK1       43.3442    11.8113   3.670 0.000249 ***
gene_covariateUBQLN2     19.4055    17.7098   1.096 0.273316
gene_covariateVAPB       35.7754    10.6475   3.360 0.000793 ***
gene_covariateVCP        35.9020    13.2027   2.719 0.006596 **
gene_covariateVUS        39.8094    10.2371   3.889 0.000104 ***
all_carriers_sexMale     -1.6807     0.6379  -2.635 0.008482 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.44 on 2097 degrees of freedom
  (192 observations deleted due to missingness)
Multiple R-squared:  0.1382,    Adjusted R-squared:  0.1308
F-statistic: 18.68 on 18 and 2097 DF,  p-value: < 2.2e-16
```

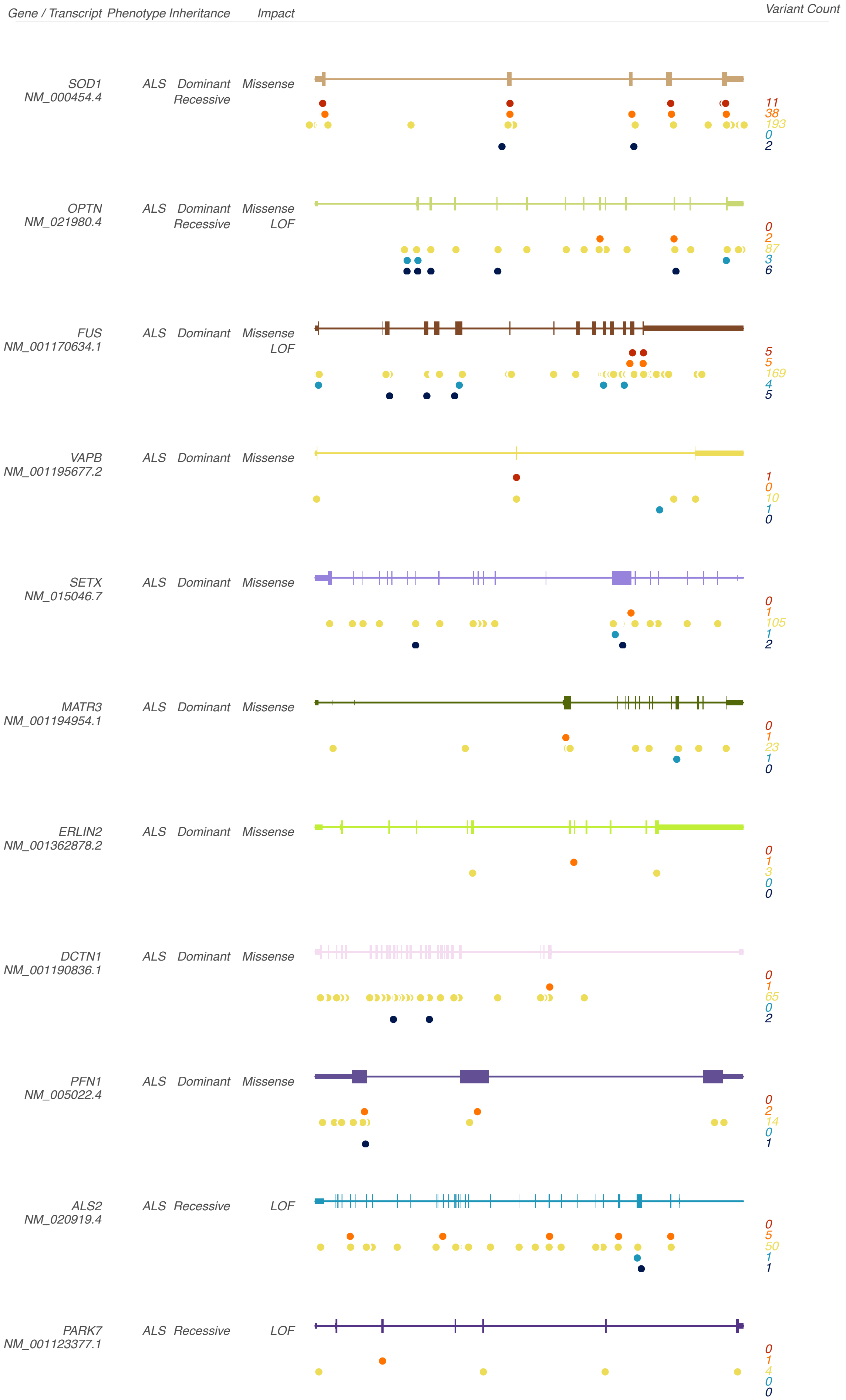


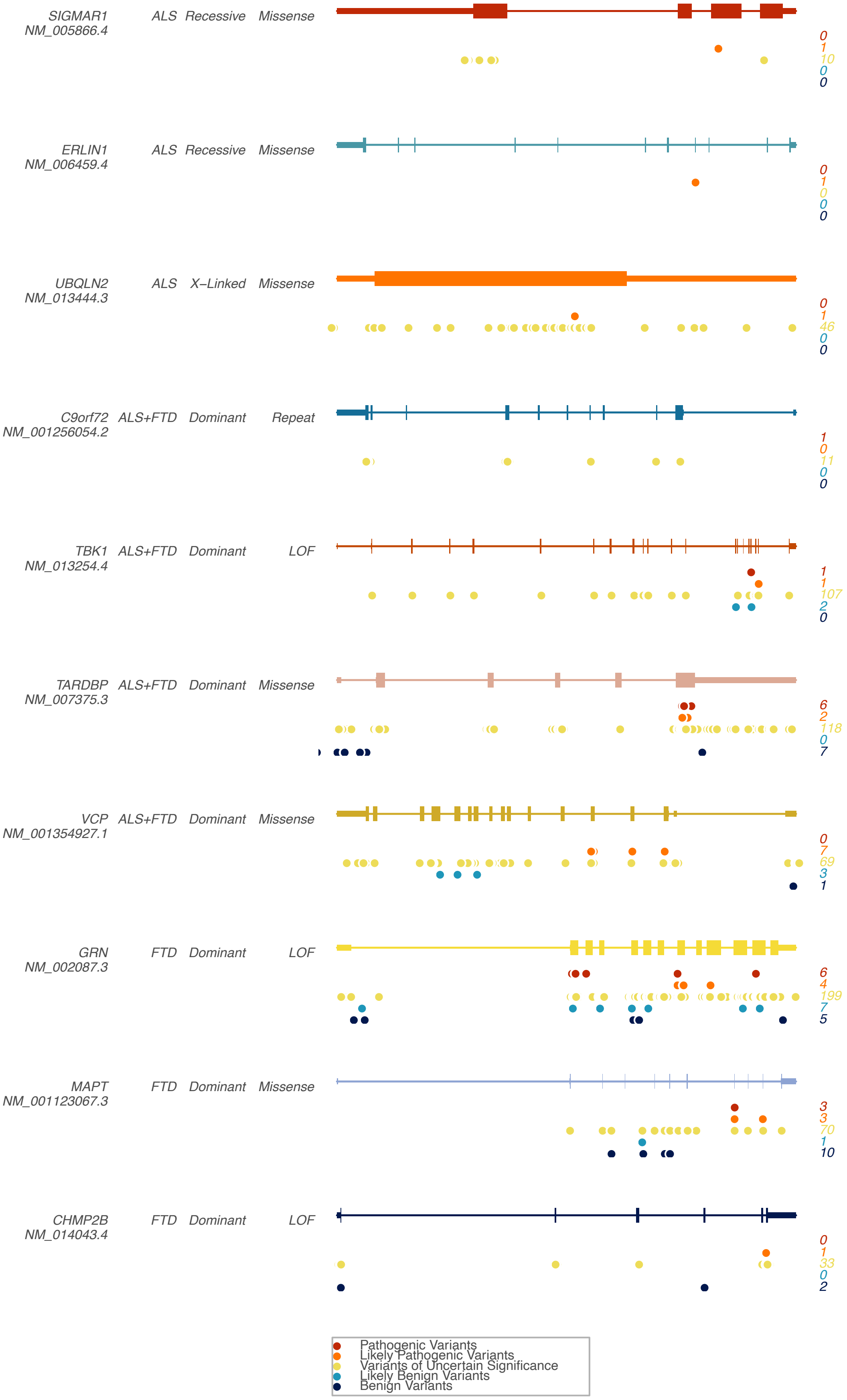


```
Call:
lm(formula = all_carriers_aoo ~ country_life_expectancy + gene_covariate +
    all_carriers_sex, data = temp_df)

Residuals:
    Min      1Q  Median      3Q     Max
-48.188  -5.793   0.198   7.166  27.581

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              47.8052    14.5454   3.287  0.00106 **
country_life_expectancy   0.1424     0.1774   0.803  0.42220
gene_covariateCHMP2B      1.3815    10.5838   0.131  0.89618
gene_covariateGRN         0.3489     1.4084   0.248  0.80439
gene_covariateMAPT      -10.4097     1.6681  -6.240 7.22e-10 ***
gene_covariateTARDBP     -3.6688     2.7579  -1.330  0.18382
gene_covariateTBK1        8.8543     4.3725   2.025  0.04321 *
gene_covariateVCP        -1.8337     5.3317  -0.344  0.73100
gene_covariateVUS        -2.7983     0.8985  -3.114  0.00191 **
all_carriers_sexMale     -1.7694     0.7690  -2.301  0.02166 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.55 on 766 degrees of freedom
  (36 observations deleted due to missingness)
Multiple R-squared:  0.07202,   Adjusted R-squared:  0.06111
F-statistic: 6.605 on 9 and 766 DF,  p-value: 4.034e-09
```

Supplementary Figure 2.10: Age of onset life expectancy regression with covariates

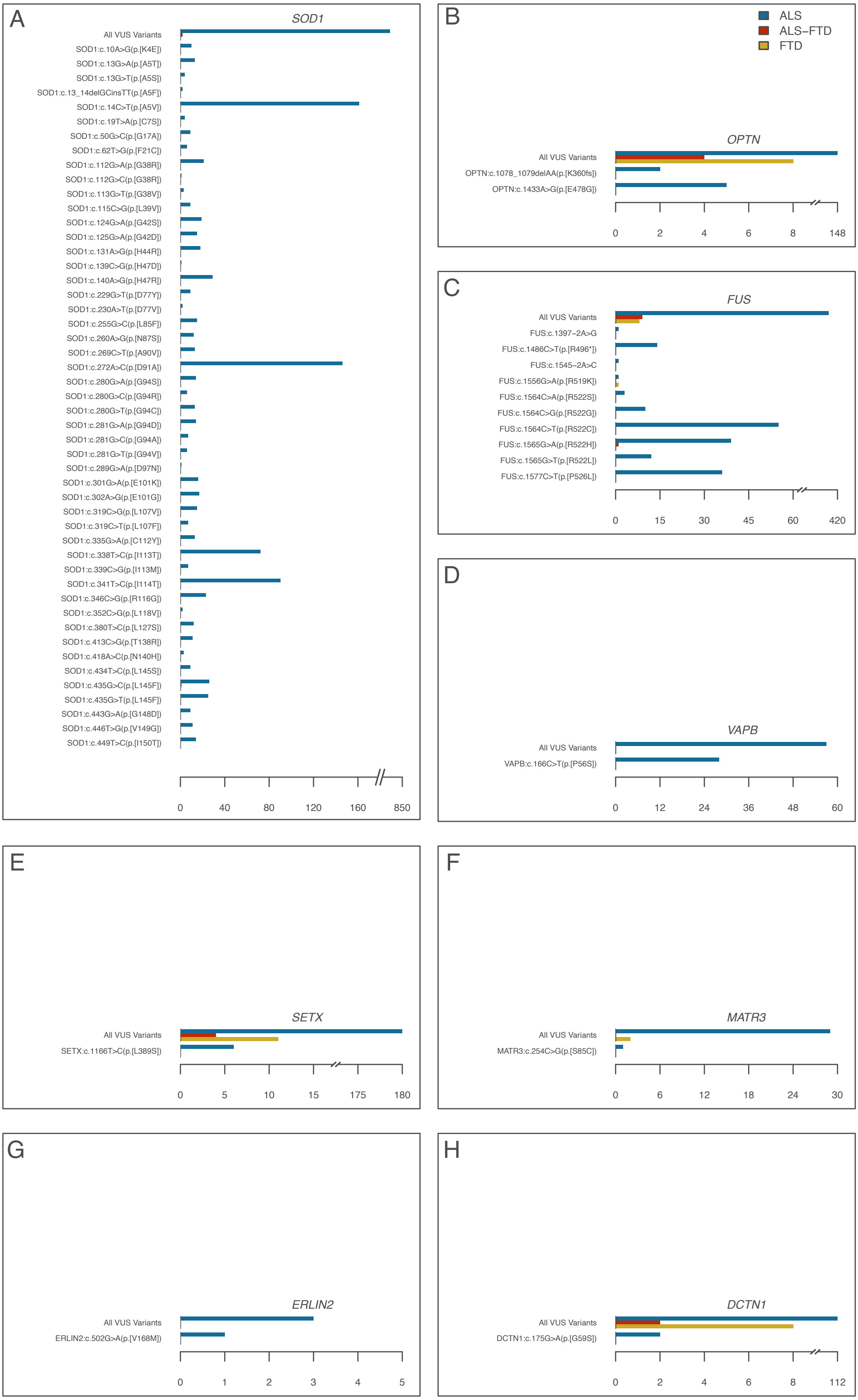
The AOO for ALS patients (A and B) and FTD patients (C and D) is regressed against the life expectancy for each country including sex and gene as covariates. B and C display the R output for each regression.

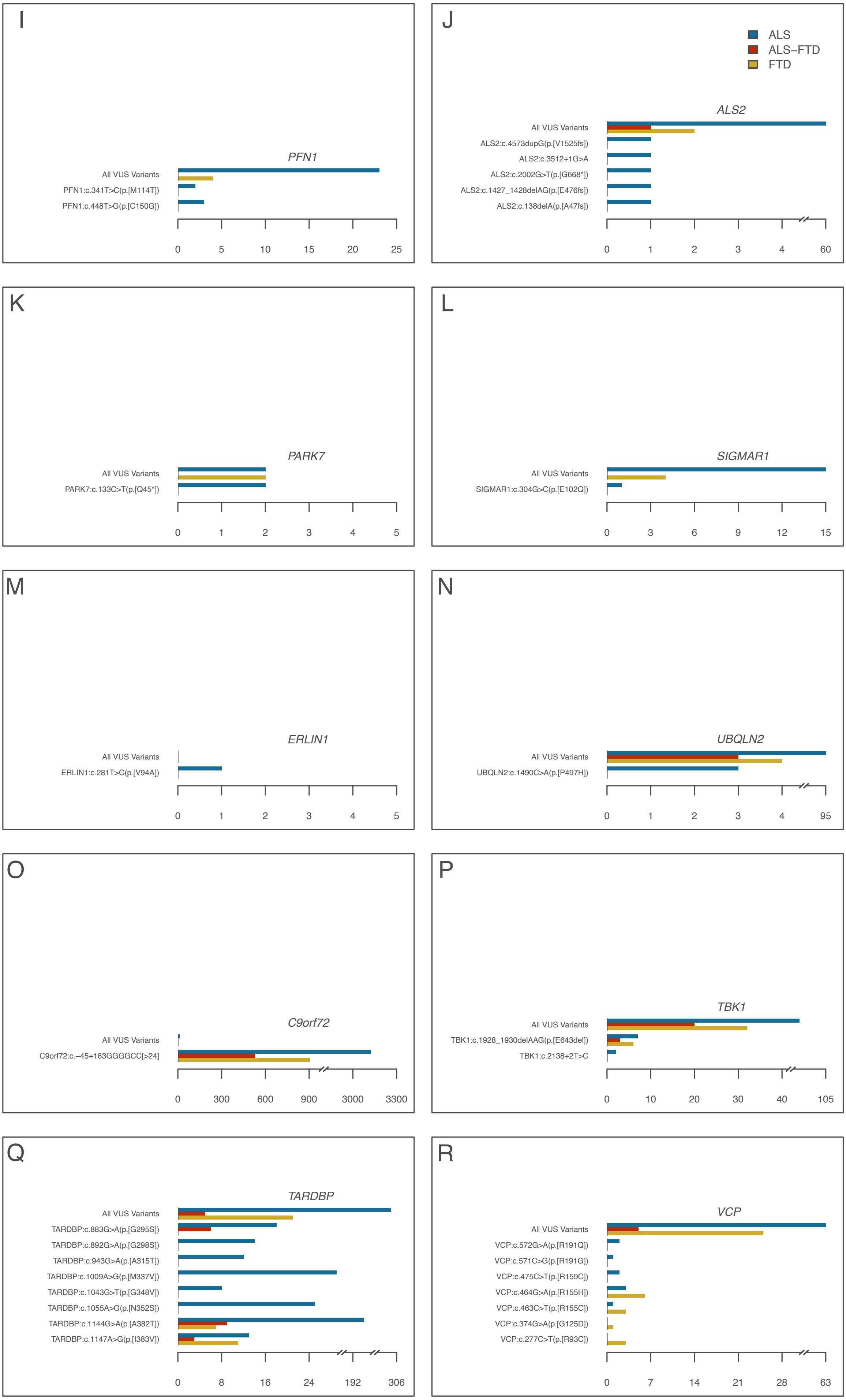

Gene / Transcript
Phenotype
Inheritance
Impact
Variant Count
SOD1
NM_000454.4
ALS
Dominant
Recessive
Missense
11
38
193
0
2
OPTN
NM_021980.4
ALS
Dominant
Recessive
Missense
LOF
0
2
87
3
6
FUS
NM_001170634.1
ALS
Dominant
Missense
LOF
5
5
169
4
5
VAPB
NM_001195677.2
ALS
Dominant
Missense
1
0
10
1
0
SETX
NM_015046.7
ALS
Dominant
Missense
0
1
105
1
2
MATR3
NM_001194954.1
ALS
Dominant
Missense
0
1
23
1
0
ERLIN2
NM_001362878.2
ALS
Dominant
Missense
0
1
3
0
0
DCTN1
NM_001190836.1
ALS
Dominant
Missense
0
1
65
0
2
PFN1
NM_005022.4
ALS
Dominant
Missense
0
2
14
0
1
ALS2
NM_020919.4
ALS
Recessive
LOF
0
5
50
1
1
PARK7
NM_001123377.1
ALS
Recessive
LOF
0
1
4
0
0

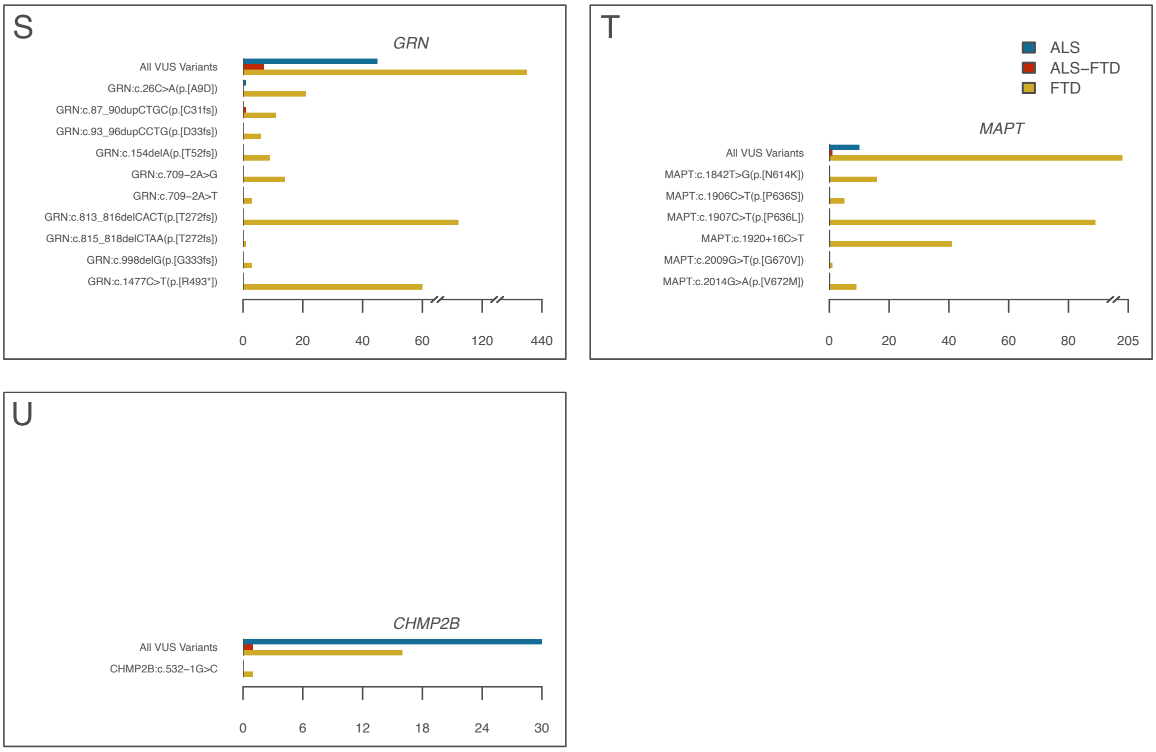


Supplementary Figure 2.11: Properties of genes carrying pathogenic and likely pathogenic variants

Displayed are the genes which are observed to carry pathogenic of likely pathogenic variants and the location and classification of variants in these genes.

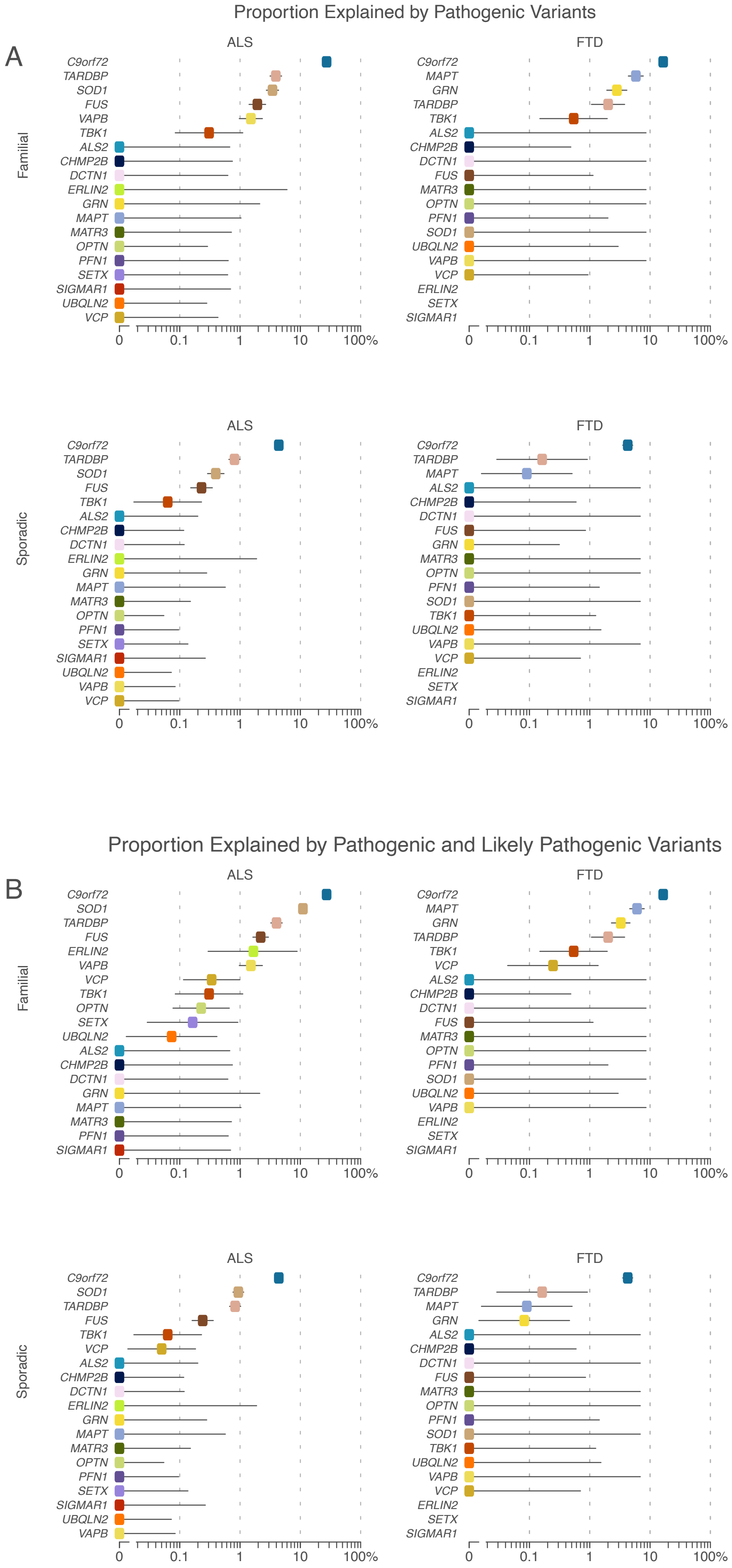

A
SOD1
All VUS Variants
SOD1:c.10A>G(p.[K4E])
SOD1:c.13G>A(p.[A5T])
SOD1:c.13G>T(p.[A5S])
SOD1:c.13_14delGCinsTT(p.[A5F])
SOD1:c.14C>T(p.[A5V])
SOD1:c.19T>A(p.[C7S])
SOD1:c.50G>C(p.[G17A])
SOD1:c.62T>G(p.[F21C])
SOD1:c.112G>A(p.[G38R])
SOD1:c.112G>C(p.[G38R])
SOD1:c.113G>T(p.[G38V])
SOD1:c.115C>G(p.[L39V])
SOD1:c.124G>A(p.[G42S])
SOD1:c.125G>A(p.[G42D])
SOD1:c.131A>G(p.[H44R])
SOD1:c.139C>G(p.[H47D])
SOD1:c.140A>G(p.[H47R])
SOD1:c.229G>T(p.[D77Y])
SOD1:c.230A>T(p.[D77V])
SOD1:c.255G>C(p.[L85F])
SOD1:c.260A>G(p.[N87S])
SOD1:c.269C>T(p.[A90V])
SOD1:c.272A>C(p.[D91A])
SOD1:c.280G>A(p.[G94S])
SOD1:c.280G>C(p.[G94R])
SOD1:c.280G>T(p.[G94C])
SOD1:c.281G>A(p.[G94D])
SOD1:c.281G>C(p.[G94A])
SOD1:c.281G>T(p.[G94V])
SOD1:c.289G>A(p.[D97N])
SOD1:c.301G>A(p.[E101K])
SOD1:c.302A>G(p.[E101G])
SOD1:c.319C>G(p.[L107V])
SOD1:c.319C>T(p.[L107F])
SOD1:c.335G>A(p.[C112Y])
SOD1:c.338T>C(p.[I113T])
SOD1:c.339C>G(p.[I113M])
SOD1:c.341T>C(p.[I114T])
SOD1:c.346C>G(p.[R116G])
SOD1:c.352C>G(p.[L118V])
SOD1:c.380T>C(p.[L127S])
SOD1:c.413C>G(p.[T138R])
SOD1:c.418A>C(p.[N140H])
SOD1:c.434T>C(p.[L145S])
SOD1:c.435G>C(p.[L145F])
SOD1:c.435G>T(p.[L145F])
SOD1:c.443G>A(p.[G148D])
SOD1:c.446T>G(p.[V149G])
SOD1:c.449T>C(p.[I150T])
0
40
80
120
160
850
B
🟦 ALS
🟥 ALS−FTD
🟨 FTD
OPTN
All VUS Variants
OPTN:c.1078_1079delAA(p.[K360fs])
OPTN:c.1433A>G(p.[E478G])
0
2
4
6
8
148
C
FUS
All VUS Variants
FUS:c.1397−2A>G
FUS:c.1486C>T(p.[R496*])
FUS:c.1545−2A>C
FUS:c.1556G>A(p.[R519K])
FUS:c.1564C>A(p.[R522S])
FUS:c.1564C>G(p.[R522G])
FUS:c.1564C>T(p.[R522C])
FUS:c.1565G>A(p.[R522H])
FUS:c.1565G>T(p.[R522L])
FUS:c.1577C>T(p.[P526L])
0
15
30
45
60
420
D
VAPB
All VUS Variants
VAPB:c.166C>T(p.[P56S])
0
12
24
36
48
60
E
SETX
All VUS Variants
SETX:c.1166T>C(p.[L389S])
0
5
10
15
175
180
F
MATR3
All VUS Variants
MATR3:c.254C>G(p.[S85C])
0
6
12
18
24
30
G
ERLIN2
All VUS Variants
ERLIN2:c.502G>A(p.[V168M])
0
1
2
3
4
5
H
DCTN1
All VUS Variants
DCTN1:c.175G>A(p.[G59S])
0
2
4
6
8
112

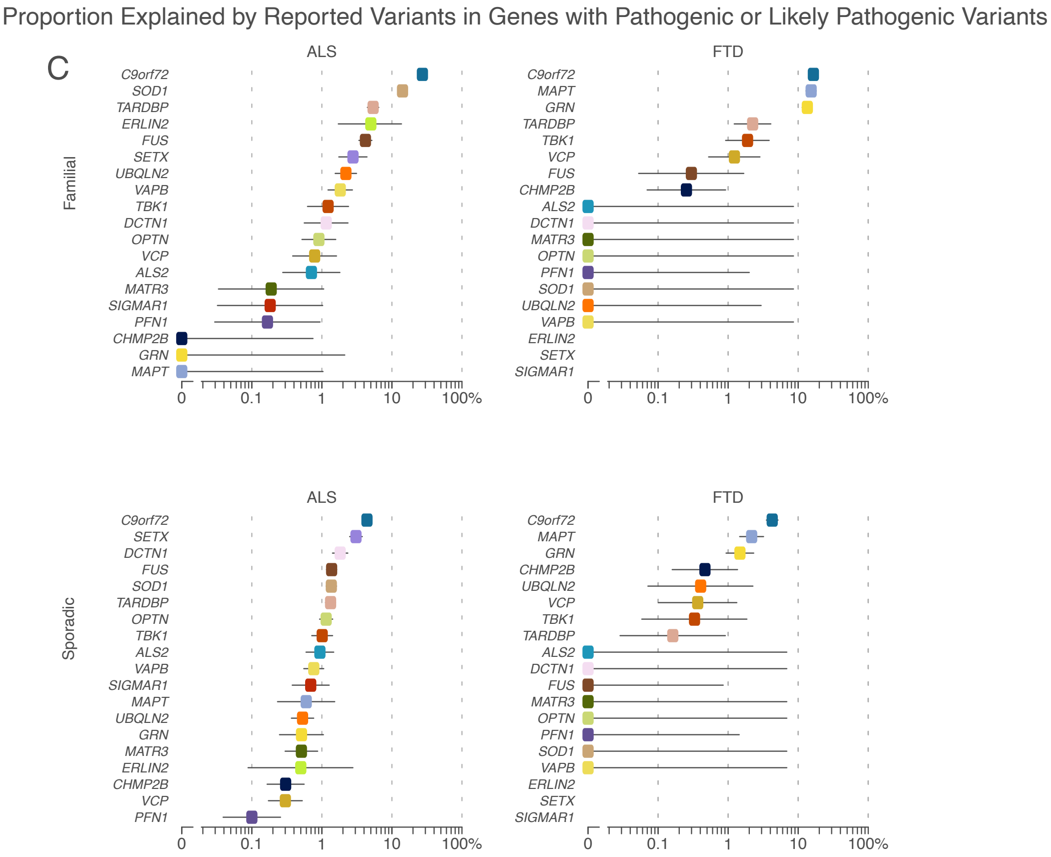

I
PFN1
All VUS Variants
PFN1:c.341T>C(p.[M114T])
PFN1:c.448T>G(p.[C150G])
0
5
10
15
20
25
J
ALS
ALS–FTD
FTD
ALS2
All VUS Variants
ALS2:c.4573dupG(p.[V1525fs])
ALS2:c.3512+1G>A
ALS2:c.2002G>T(p.[G668*])
ALS2:c.1427_1428delAG(p.[E476fs])
ALS2:c.138delA(p.[A47fs])
0
1
2
3
4
60
K
PARK7
All VUS Variants
PARK7:c.133C>T(p.[Q45*])
0
1
2
3
4
5
L
SIGMAR1
All VUS Variants
SIGMAR1:c.304G>C(p.[E102Q])
0
3
6
9
12
15
M
ERLIN1
All VUS Variants
ERLIN1:c.281T>C(p.[V94A])
0
1
2
3
4
5
N
UBQLN2
All VUS Variants
UBQLN2:c.1490C>A(p.[P497H])
0
1
2
3
4
95
O
C9orf72
All VUS Variants
C9orf72:c.–45+163GGGGCC[>24]
0
300
600
900
3000
3300
P
TBK1
All VUS Variants
TBK1:c.1928_1930delAAG(p.[E643del])
TBK1:c.2138+2T>C
0
10
20
30
40
105
Q
TARDBP
All VUS Variants
TARDBP:c.883G>A(p.[G295S])
TARDBP:c.892G>A(p.[G298S])
TARDBP:c.943G>A(p.[A315T])
TARDBP:c.1009A>G(p.[M337V])
TARDBP:c.1043G>T(p.[G348V])
TARDBP:c.1055A>G(p.[N352S])
TARDBP:c.1144G>A(p.[A382T])
TARDBP:c.1147A>G(p.[I383V])
0
8
16
24
192
306
R
VCP
All VUS Variants
VCP:c.572G>A(p.[R191Q])
VCP:c.571C>G(p.[R191G])
VCP:c.475C>T(p.[R159C])
VCP:c.464G>A(p.[R155H])
VCP:c.463C>T(p.[R155C])
VCP:c.374G>A(p.[G125D])
VCP:c.277C>T(p.[R93C])
0
7
14
21
28
63



Supplementary Figure 2.12: Phenotypes of carriers of pathogenic and likely pathogenic variants

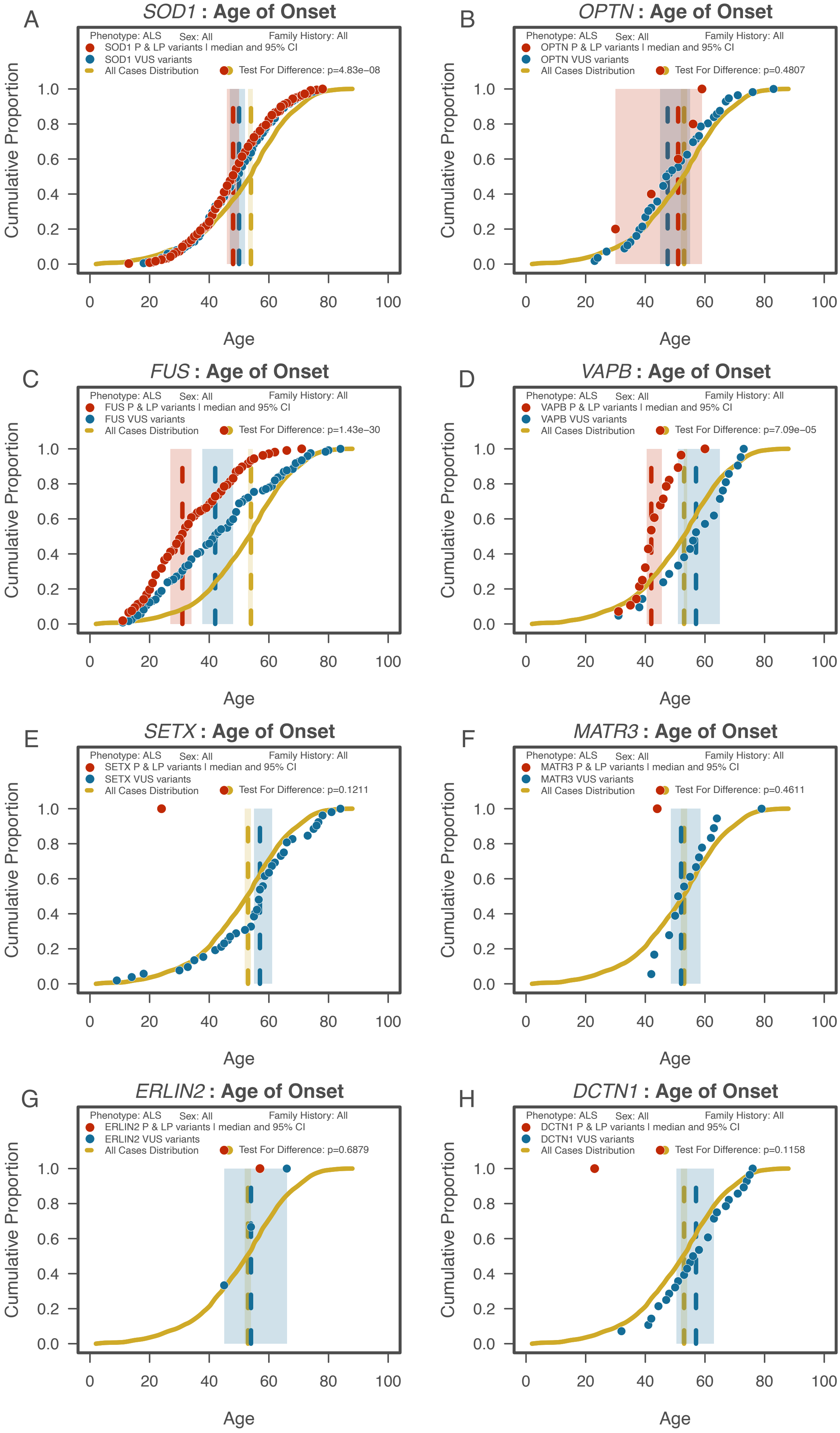
For each gene with an observed pathogenic or likely pathogenic variant the primary phenotype of variant carriers in that gene are displayed. Variants classified as either pathogenic or likely pathogenic are listed individually and other VUS variants are amalgamated.

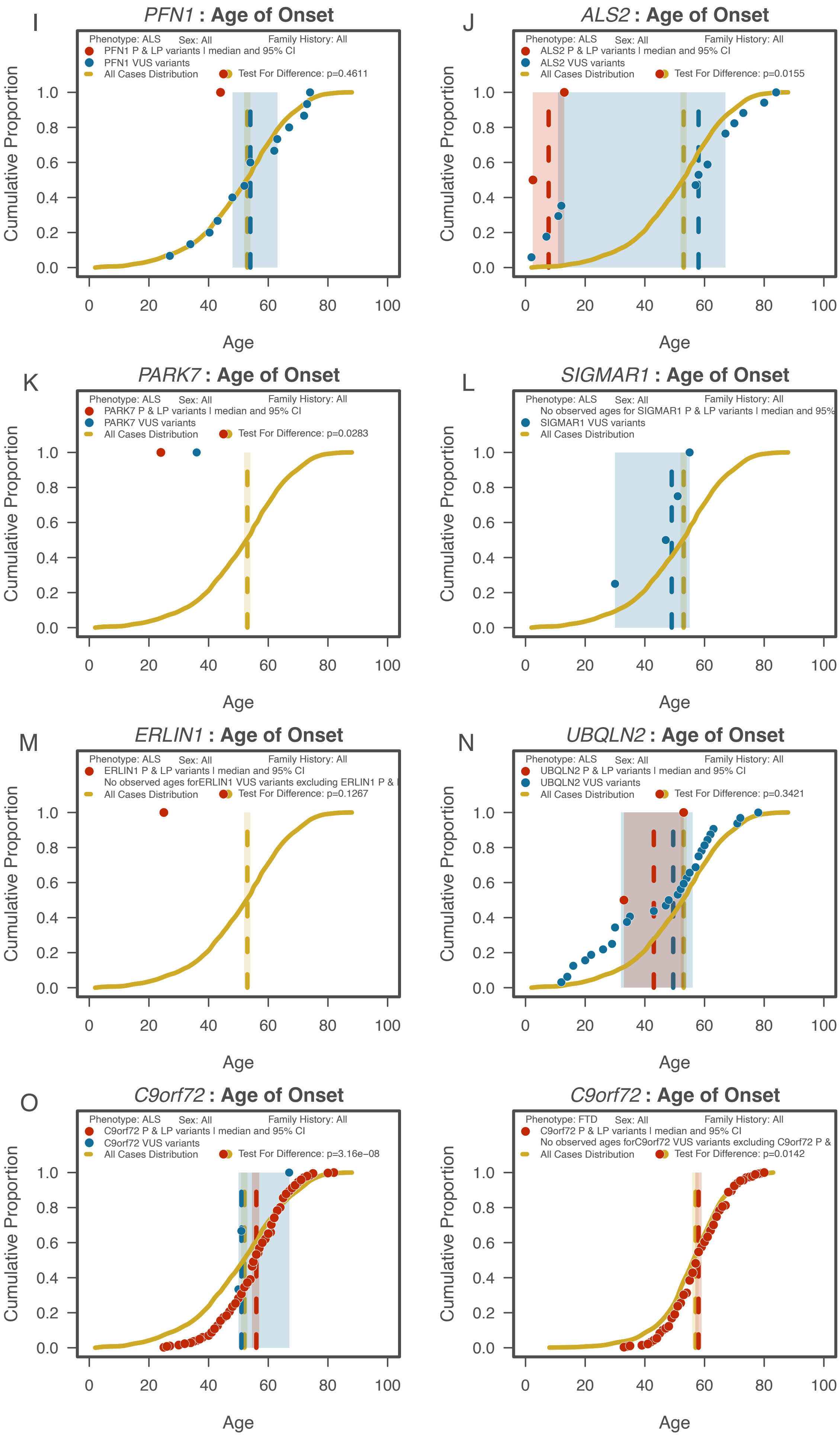

Proportion Explained by Pathogenic Variants
A
Familial
ALS
C9orf72
TARDBP
SOD1
FUS
VAPB
TBK1
ALS2
CHMP2B
DCTN1
ERLIN2
GRN
MAPT
MATR3
OPTN
PFN1
SETX
SIGMAR1
UBQLN2
VCP
0
0.1
1
10
100%
FTD
C9orf72
MAPT
GRN
TARDBP
TBK1
ALS2
CHMP2B
DCTN1
FUS
MATR3
OPTN
PFN1
SOD1
UBQLN2
VAPB
VCP
ERLIN2
SETX
SIGMAR1
0
0.1
1
10
100%
Sporadic
ALS
C9orf72
TARDBP
SOD1
FUS
TBK1
ALS2
CHMP2B
DCTN1
ERLIN2
GRN
MAPT
MATR3
OPTN
PFN1
SETX
SIGMAR1
UBQLN2
VAPB
VCP
0
0.1
1
10
100%
FTD
C9orf72
TARDBP
MAPT
ALS2
CHMP2B
DCTN1
FUS
GRN
MATR3
OPTN
PFN1
SOD1
TBK1
UBQLN2
VAPB
VCP
ERLIN2
SETX
SIGMAR1
0
0.1
1
10
100%
Proportion Explained by Pathogenic and Likely Pathogenic Variants
B
Familial
ALS
C9orf72
SOD1
TARDBP
FUS
ERLIN2
VAPB
VCP
TBK1
OPTN
SETX
UBQLN2
ALS2
CHMP2B
DCTN1
GRN
MAPT
MATR3
PFN1
SIGMAR1
0
0.1
1
10
100%
FTD
C9orf72
MAPT
GRN
TARDBP
TBK1
VCP
ALS2
CHMP2B
DCTN1
FUS
MATR3
OPTN
PFN1
SOD1
UBQLN2
VAPB
ERLIN2
SETX
SIGMAR1
0
0.1
1
10
100%
Sporadic
ALS
C9orf72
SOD1
TARDBP
FUS
TBK1
VCP
ALS2
CHMP2B
DCTN1
ERLIN2
GRN
MAPT
MATR3
OPTN
PFN1
SETX
SIGMAR1
UBQLN2
VAPB
0
0.1
1
10
100%
FTD
C9orf72
TARDBP
MAPT
GRN
ALS2
CHMP2B
DCTN1
FUS
MATR3
OPTN
PFN1
SOD1
TBK1
UBQLN2
VAPB
VCP
ERLIN2
SETX
SIGMAR1
0
0.1
1
10
100%

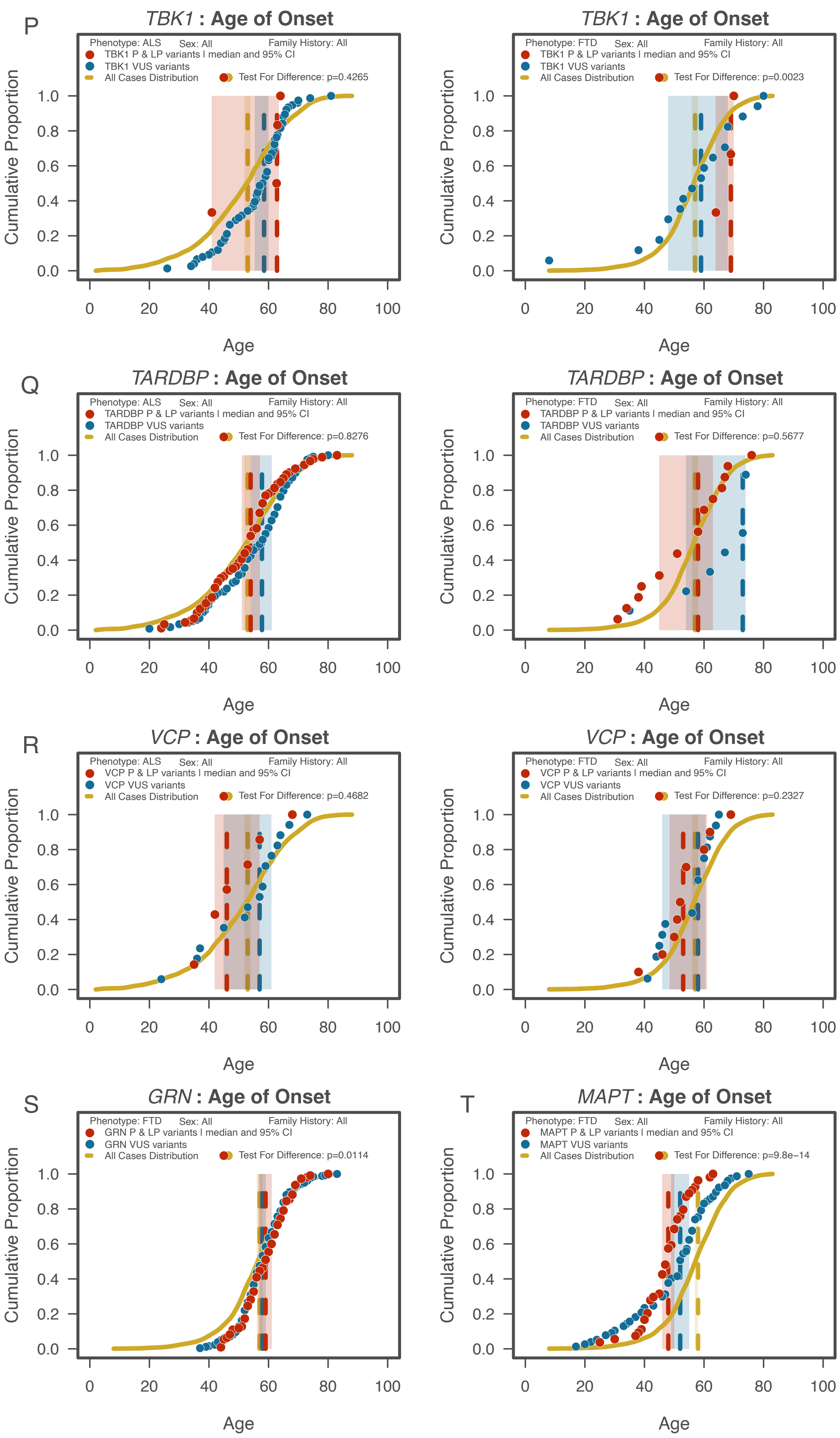


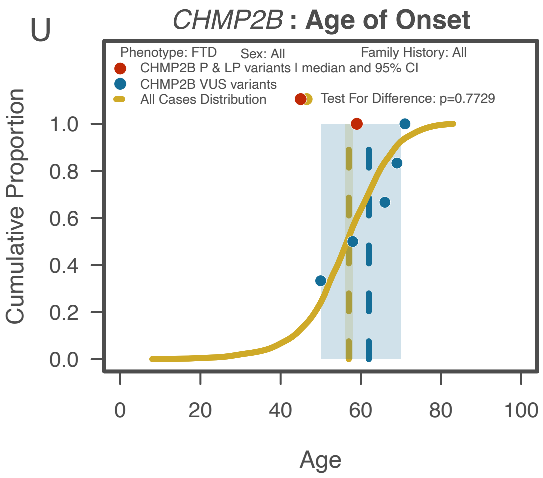
Supplementary Figure 2.13: Detailed proportion of explained ALS and FTD cases

A detailed breakdown of the overall proportion of global ALS and FTD cases with an explained genetic cause varies if considering A) pathogenic variants, B) pathogenic and likely pathogenic variants, or C) all reported variants in genes with observed pathogenic or likely pathogenic variants.


A
SOD1 : Age of Onset
Phenotype: ALS Sex: All Family History: All
SOD1 P & LP variants | median and 95% CI
SOD1 VUS variants
All Cases Distribution
Test For Difference: p=4.83e−08
B
OPTN : Age of Onset
Phenotype: ALS Sex: All Family History: All
OPTN P & LP variants | median and 95% CI
OPTN VUS variants
All Cases Distribution
Test For Difference: p=0.4807
C
FUS : Age of Onset
Phenotype: ALS Sex: All Family History: All
FUS P & LP variants | median and 95% CI
FUS VUS variants
All Cases Distribution
Test For Difference: p=1.43e−30
D
VAPB : Age of Onset
Phenotype: ALS Sex: All Family History: All
VAPB P & LP variants | median and 95% CI
VAPB VUS variants
All Cases Distribution
Test For Difference: p=7.09e−05
E
SETX : Age of Onset
Phenotype: ALS Sex: All Family History: All
SETX P & LP variants | median and 95% CI
SETX VUS variants
All Cases Distribution
Test For Difference: p=0.1211
F
MATR3 : Age of Onset
Phenotype: ALS Sex: All Family History: All
MATR3 P & LP variants | median and 95% CI
MATR3 VUS variants
All Cases Distribution
Test For Difference: p=0.4611
G
ERLIN2 : Age of Onset
Phenotype: ALS Sex: All Family History: All
ERLIN2 P & LP variants | median and 95% CI
ERLIN2 VUS variants
All Cases Distribution
Test For Difference: p=0.6879
H
DCTN1 : Age of Onset
Phenotype: ALS Sex: All Family History: All
DCTN1 P & LP variants | median and 95% CI
DCTN1 VUS variants
All Cases Distribution
Test For Difference: p=0.1158
Cumulative Proportion
0.0
0.2
0.4
0.6
0.8
1.0
Age
0
20
40
60
80
100


I
PFN1 : Age of Onset
Phenotype: ALS Sex: All Family History: All
PFN1 P & LP variants I median and 95% CI
PFN1 VUS variants
All Cases Distribution
Test For Difference: p=0.4611
Cumulative Proportion
Age
J
ALS2 : Age of Onset
Phenotype: ALS Sex: All Family History: All
ALS2 P & LP variants I median and 95% CI
ALS2 VUS variants
All Cases Distribution
Test For Difference: p=0.0155
K
PARK7 : Age of Onset
Phenotype: ALS Sex: All Family History: All
PARK7 P & LP variants I median and 95% CI
PARK7 VUS variants
All Cases Distribution
Test For Difference: p=0.0283
L
SIGMAR1 : Age of Onset
Phenotype: ALS Sex: All Family History: All
No observed ages for SIGMAR1 P & LP variants I median and 95%
SIGMAR1 VUS variants
All Cases Distribution
M
ERLIN1 : Age of Onset
Phenotype: ALS Sex: All Family History: All
ERLIN1 P & LP variants I median and 95% CI
No observed ages forERLIN1 VUS variants excluding ERLIN1 P &
All Cases Distribution
Test For Difference: p=0.1267
N
UBQLN2 : Age of Onset
Phenotype: ALS Sex: All Family History: All
UBQLN2 P & LP variants I median and 95% CI
UBQLN2 VUS variants
All Cases Distribution
Test For Difference: p=0.3421
O
C9orf72 : Age of Onset
Phenotype: ALS Sex: All Family History: All
C9orf72 P & LP variants I median and 95% CI
C9orf72 VUS variants
All Cases Distribution
Test For Difference: p=3.16e−08
C9orf72 : Age of Onset
Phenotype: FTD Sex: All Family History: All
C9orf72 P & LP variants I median and 95% CI
No observed ages forC9orf72 VUS variants excluding C9orf72 P &
All Cases Distribution
Test For Difference: p=0.0142
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100

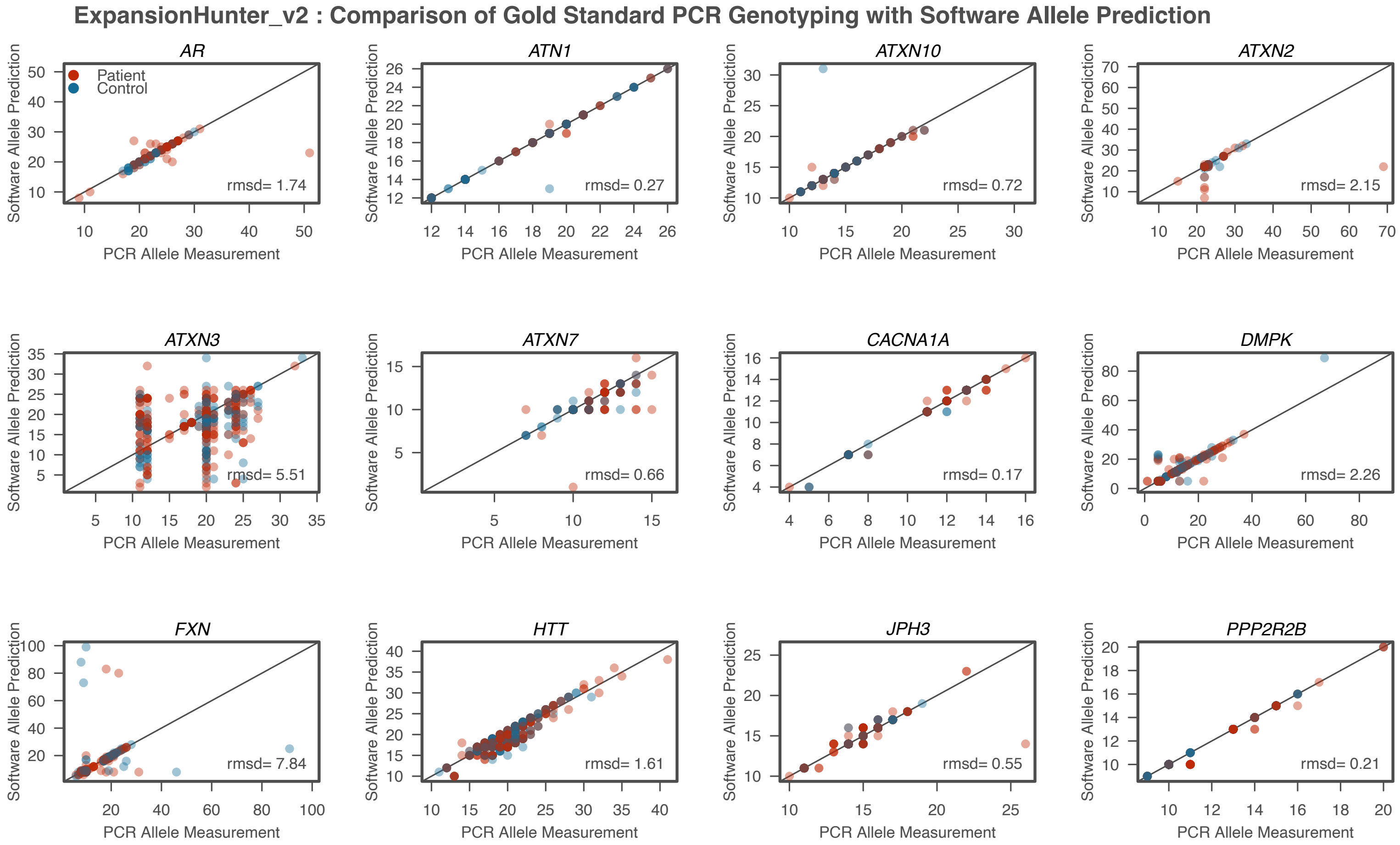

P
TBK1 : Age of Onset
Phenotype: ALS Sex: All Family History: All
TBK1 P & LP variants I median and 95% CI
TBK1 VUS variants
All Cases Distribution
Test For Difference: p=0.4265
TBK1 : Age of Onset
Phenotype: FTD Sex: All Family History: All
TBK1 P & LP variants I median and 95% CI
TBK1 VUS variants
All Cases Distribution
Test For Difference: p=0.0023
Q
TARDBP : Age of Onset
Phenotype: ALS Sex: All Family History: All
TARDBP P & LP variants I median and 95% CI
TARDBP VUS variants
All Cases Distribution
Test For Difference: p=0.8276
TARDBP : Age of Onset
Phenotype: FTD Sex: All Family History: All
TARDBP P & LP variants I median and 95% CI
TARDBP VUS variants
All Cases Distribution
Test For Difference: p=0.5677
R
VCP : Age of Onset
Phenotype: ALS Sex: All Family History: All
VCP P & LP variants I median and 95% CI
VCP VUS variants
All Cases Distribution
Test For Difference: p=0.4682
VCP : Age of Onset
Phenotype: FTD Sex: All Family History: All
VCP P & LP variants I median and 95% CI
VCP VUS variants
All Cases Distribution
Test For Difference: p=0.2327
S
GRN : Age of Onset
Phenotype: FTD Sex: All Family History: All
GRN P & LP variants I median and 95% CI
GRN VUS variants
All Cases Distribution
Test For Difference: p=0.0114
T
MAPT : Age of Onset
Phenotype: FTD Sex: All Family History: All
MAPT P & LP variants I median and 95% CI
MAPT VUS variants
All Cases Distribution
Test For Difference: p=9.8e−14
Cumulative Proportion
0.0 0.2 0.4 0.6 0.8 1.0
Age
0 20 40 60 80 100



Supplementary Figure 2.14: Age of onset for variant carriers

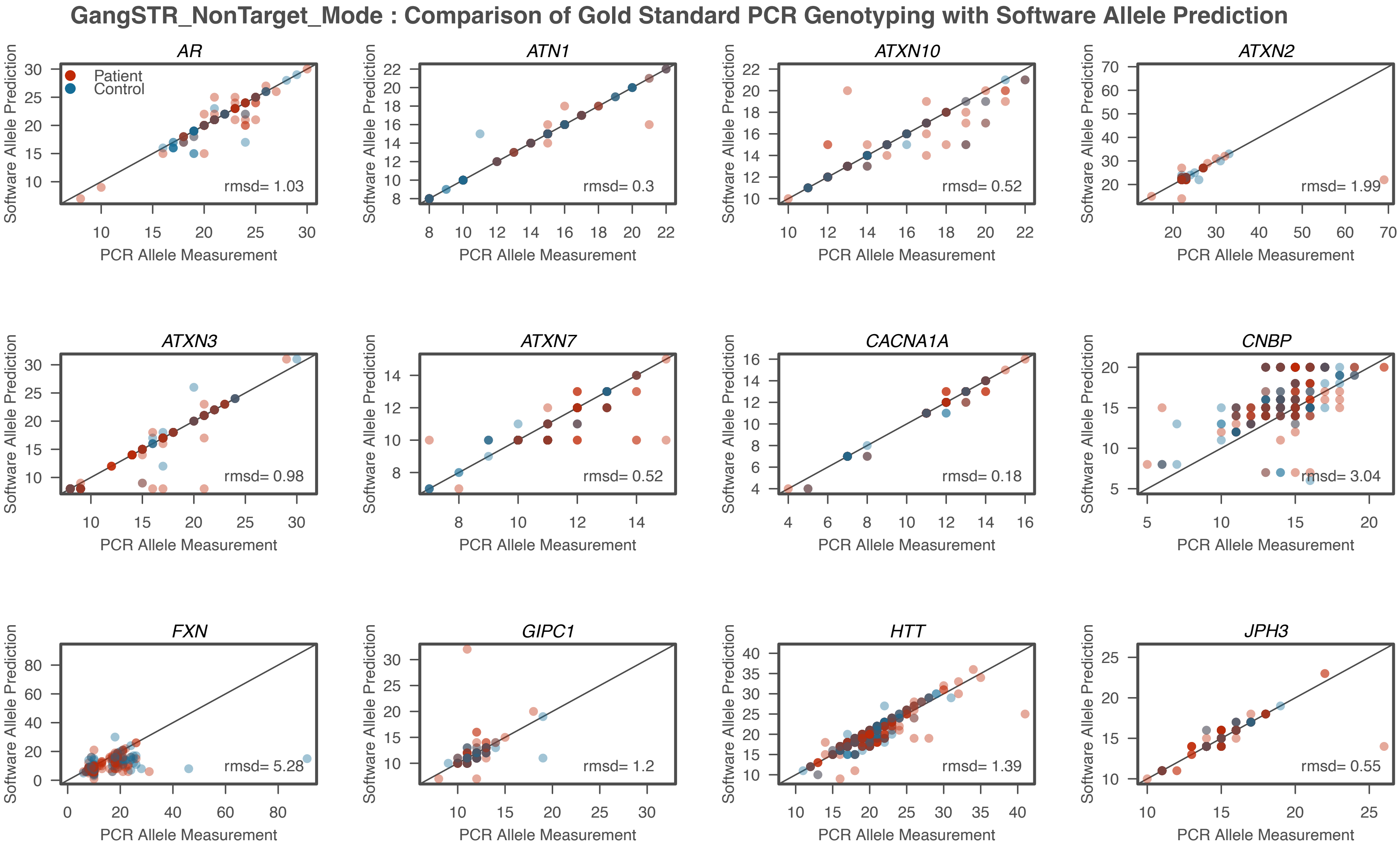
Plots display the age of onset for carriers of P and LP variants (red), relative to carriers of other variants in the same gene (blue), and the rest of cohort (yellow). P-values are displayed for Kruskal-Wallis tests comparing P and LP variant carries to the rest of the cohort. Where P and LP variants are observed in both ALS and FTD cases these are shown on separate plots (O) *C9orf72*, P) TBK1, Q) TARDBP, R) VCP)

Chapter 3



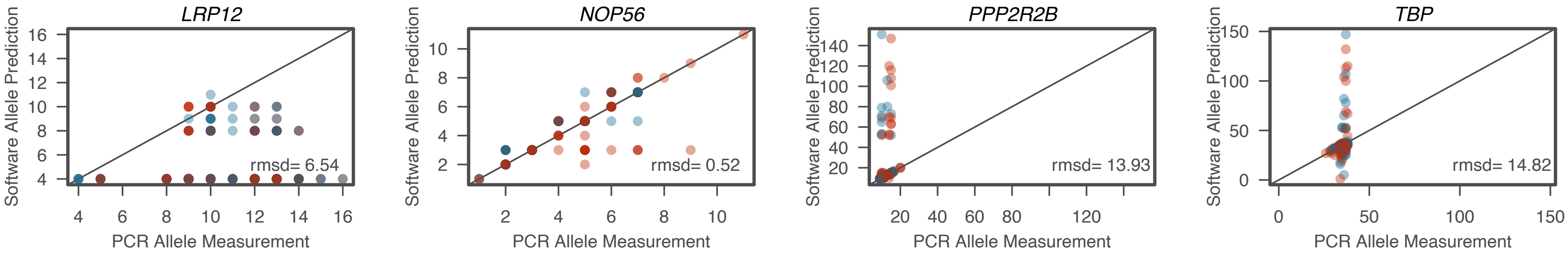
Supplementary Figure 3.1: ExpansionHunter v2 comparison of gold standard PCR genotyping with in silico predictions

Gold standard PCR genotypes are compared to predicted alleles using the software ExpansionHunter 2.



GangSTR_NonTarget_Mode : Comparison of Gold Standard PCR Genotyping with Software Allele Prediction

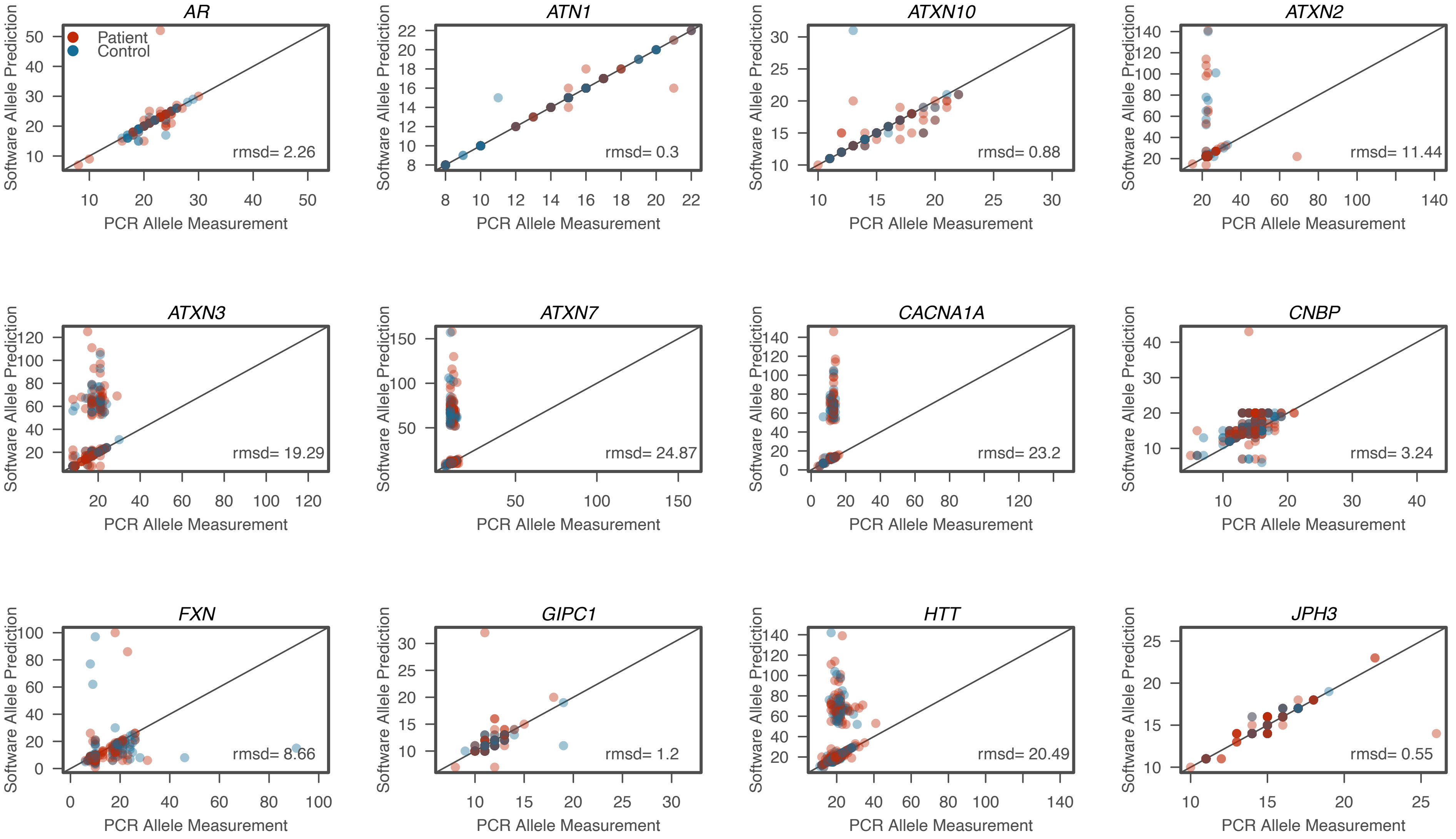
AR *ATN1* *ATXN10* *ATXN2* *ATXN3* *ATXN7* *CACNA1A* *CNBP* *FXN* *GIPC1* *HTT* *JPH3*

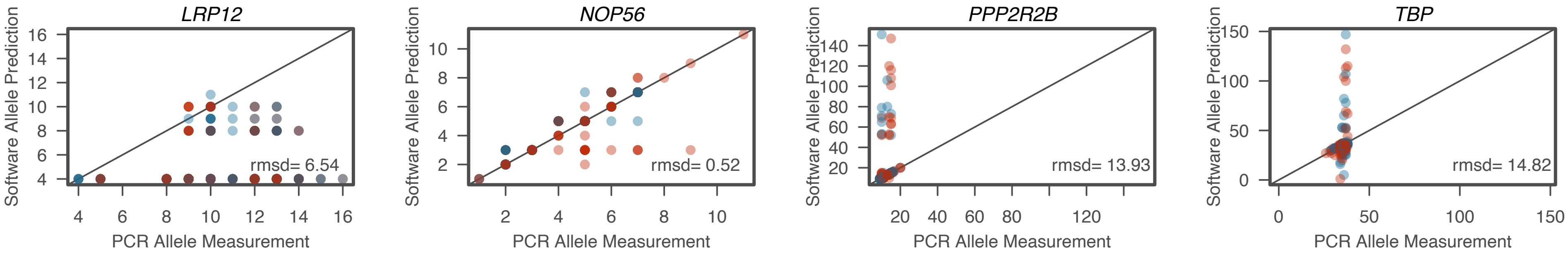


Supplementary Figure 3.2: GangSTR (target) comparison of gold standard PCR genotyping with in silico predictions

Gold standard PCR genotypes are compared to predicted alleles using the software GangSTR (targeted).

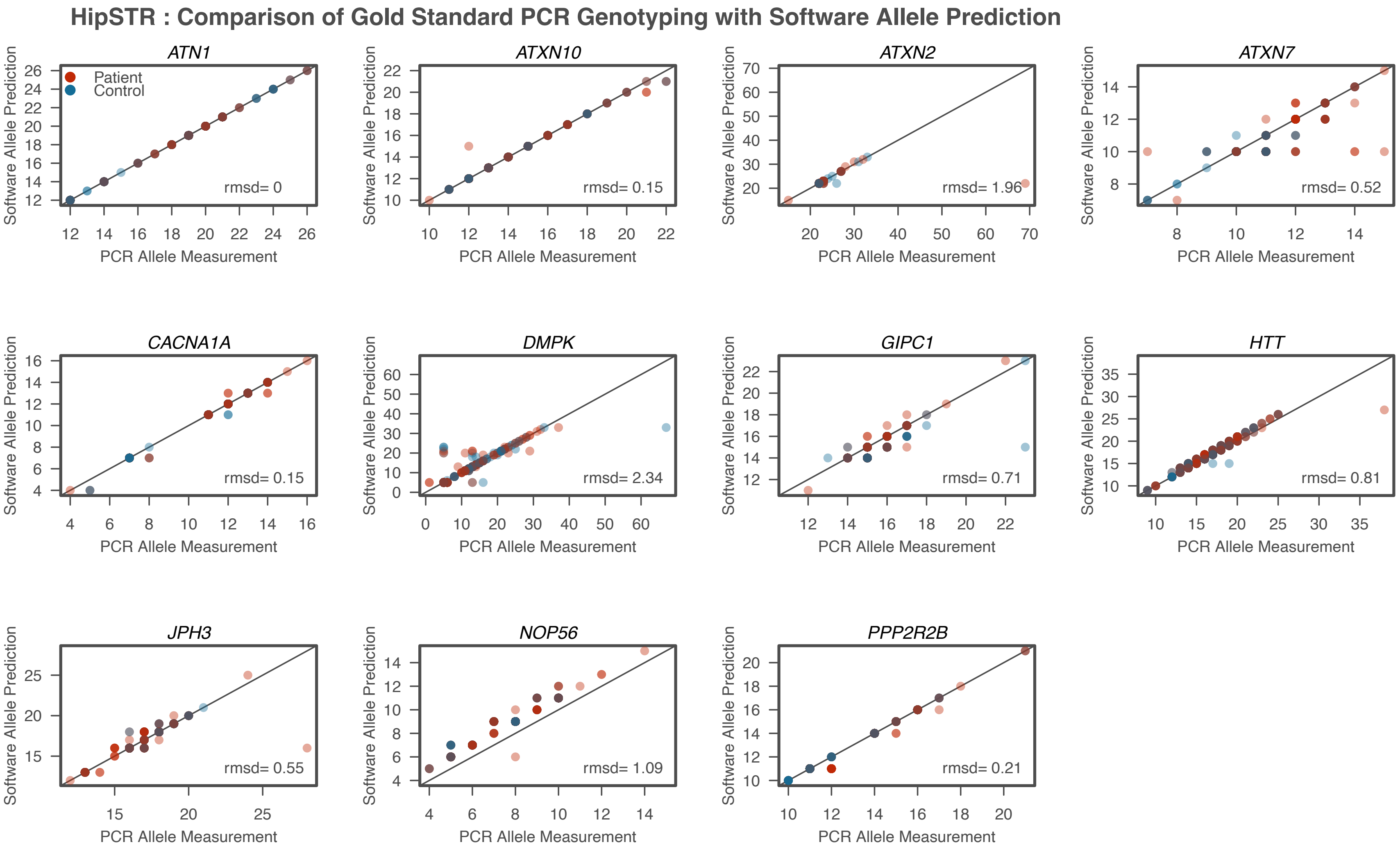
GangSTR_Target_Mode : Comparison of Gold Standard PCR Genotyping with Software Allele Prediction





Supplementary Figure 3.3: GangSTR (target) comparison of gold standard PCR genotyping with in silico predictions

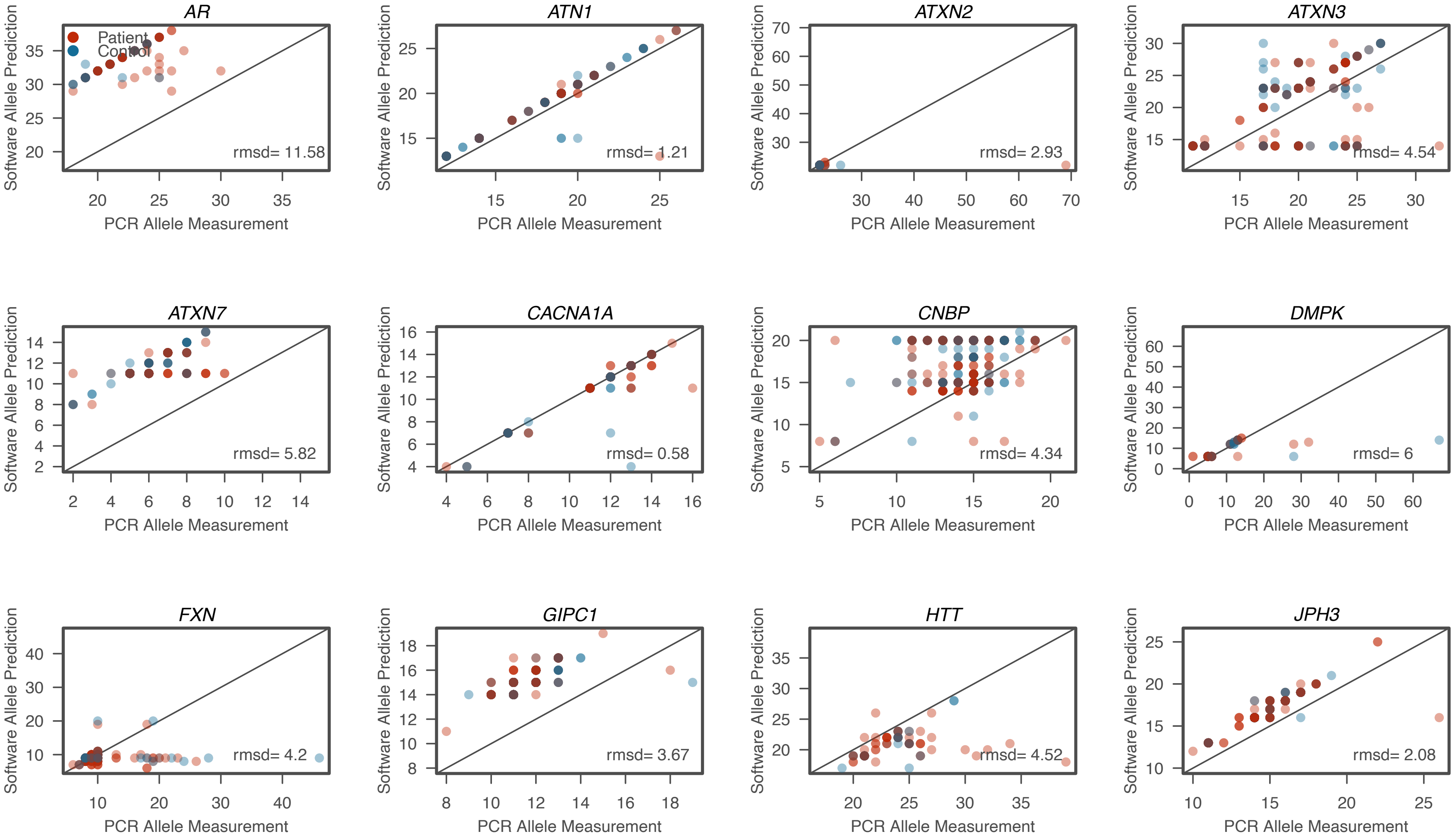
Gold standard PCR genotypes are compared to predicted alleles using the software GangSTR (targeted).



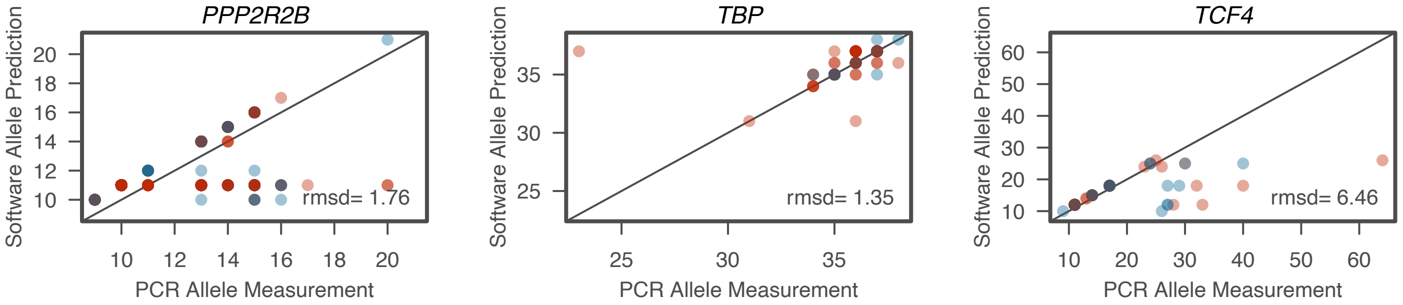
Supplementary Figure 3.4: HipSTR comparison of gold standard PCR genotyping with in silico predictions

Gold standard PCR genotypes are compared to predicted alleles using the software HipSTR

RepeatSeq : Comparison of Gold Standard PCR Genotyping with Software Allele Prediction

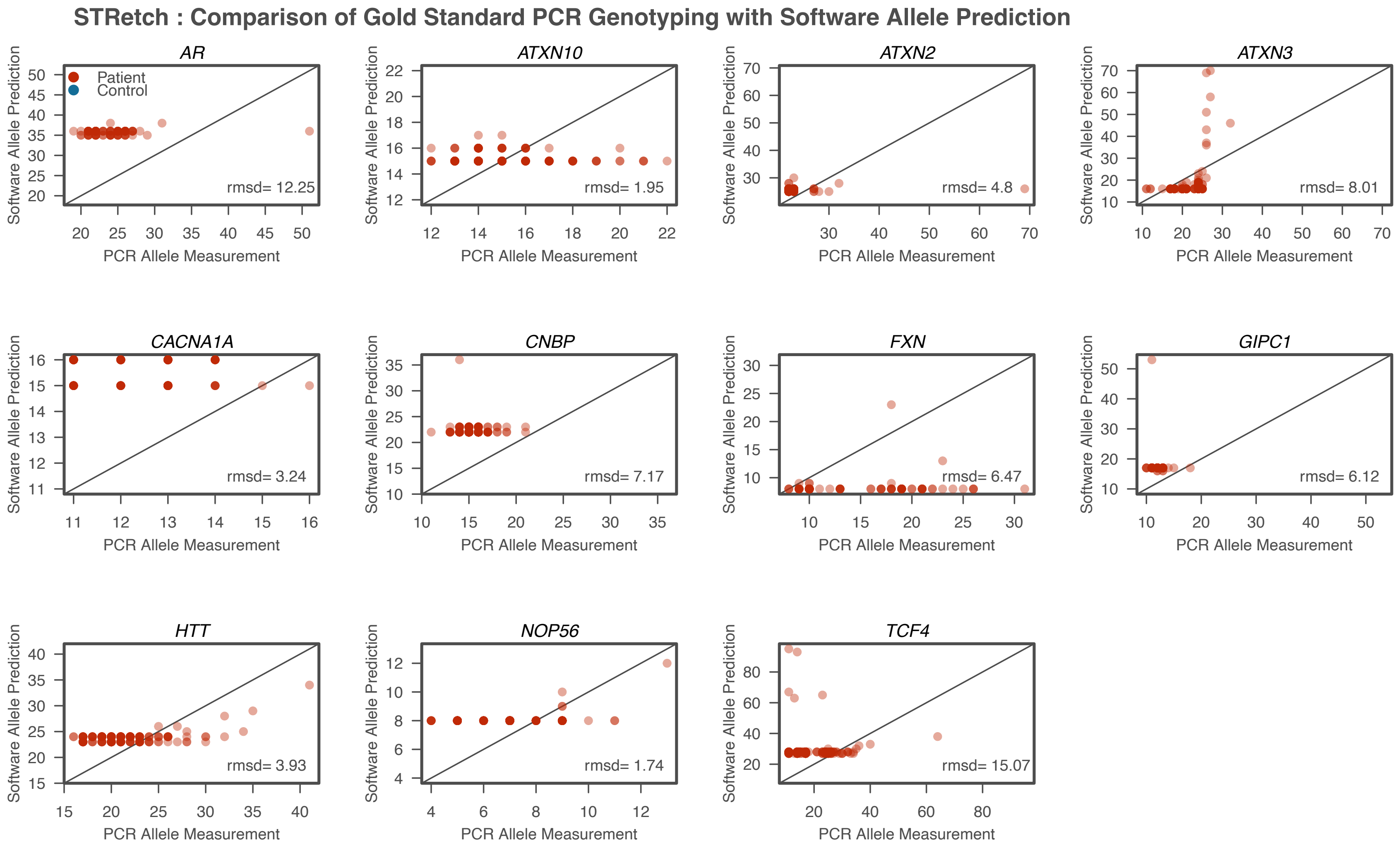


AR, *ATN1*, *ATXN2*, *ATXN3*, *ATXN7*, *CACNA1A*, *CNBP*, *DMPK*, *FXN*, *GIPC1*, *HTT*, *JPH3*



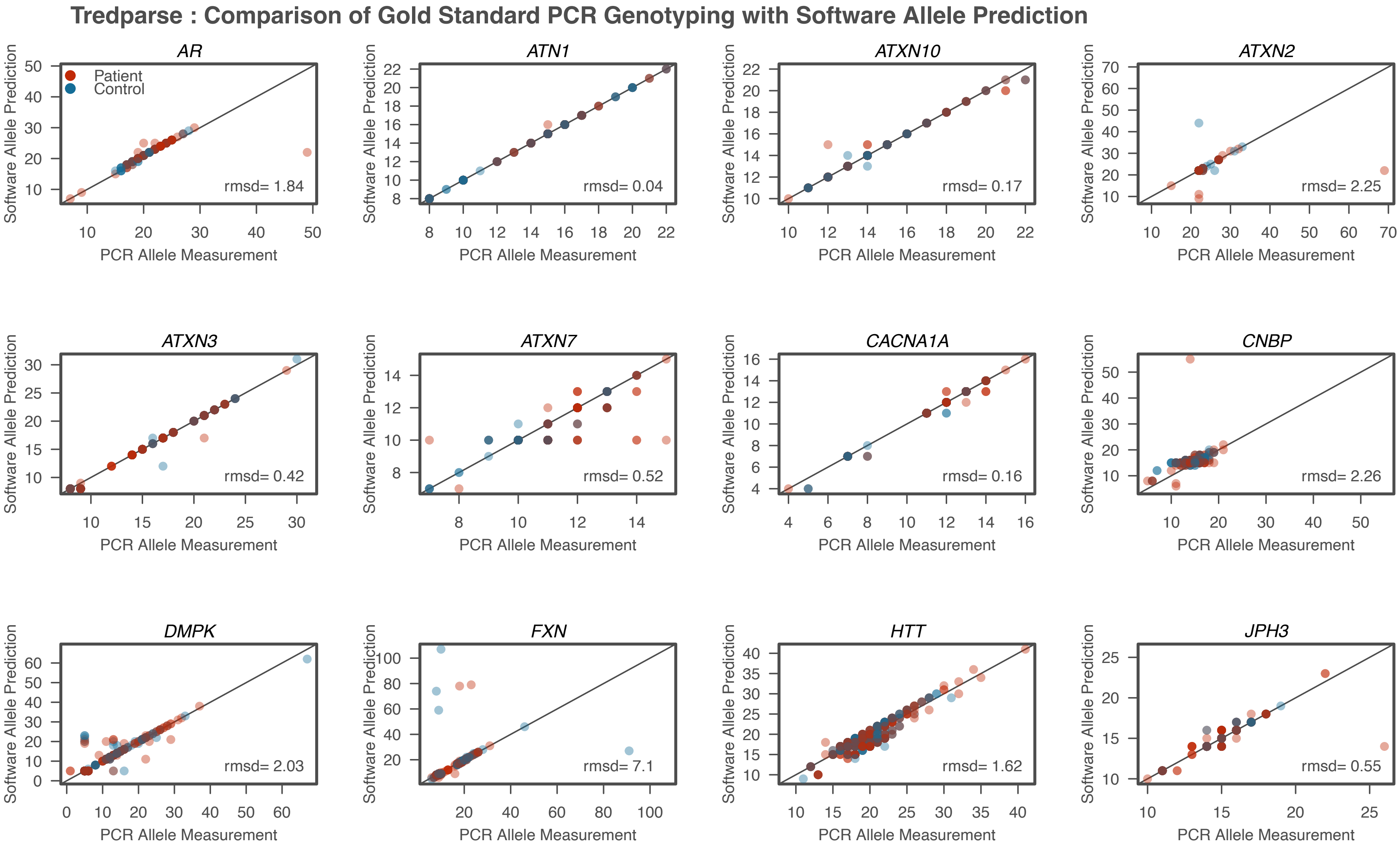
Supplementary Figure 3.5: RepeatSeq: comparison of gold standard PCR genotyping with in silico predictions

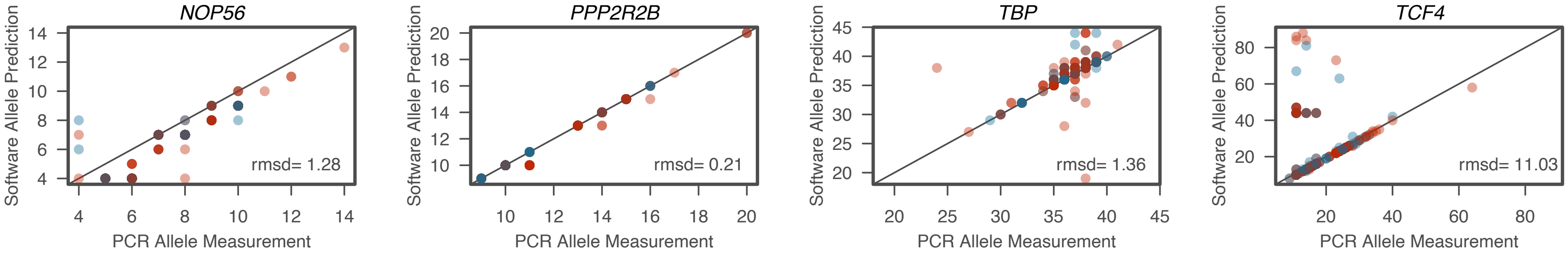
Gold standard PCR genotypes are compared to predicted alleles using the software HipSTR



Supplementary Figure 3.6: STRetch comparison of gold standard PCR genotyping with in silico predictions

Gold standard PCR genotypes are compared to predicted alleles using the software STRetch

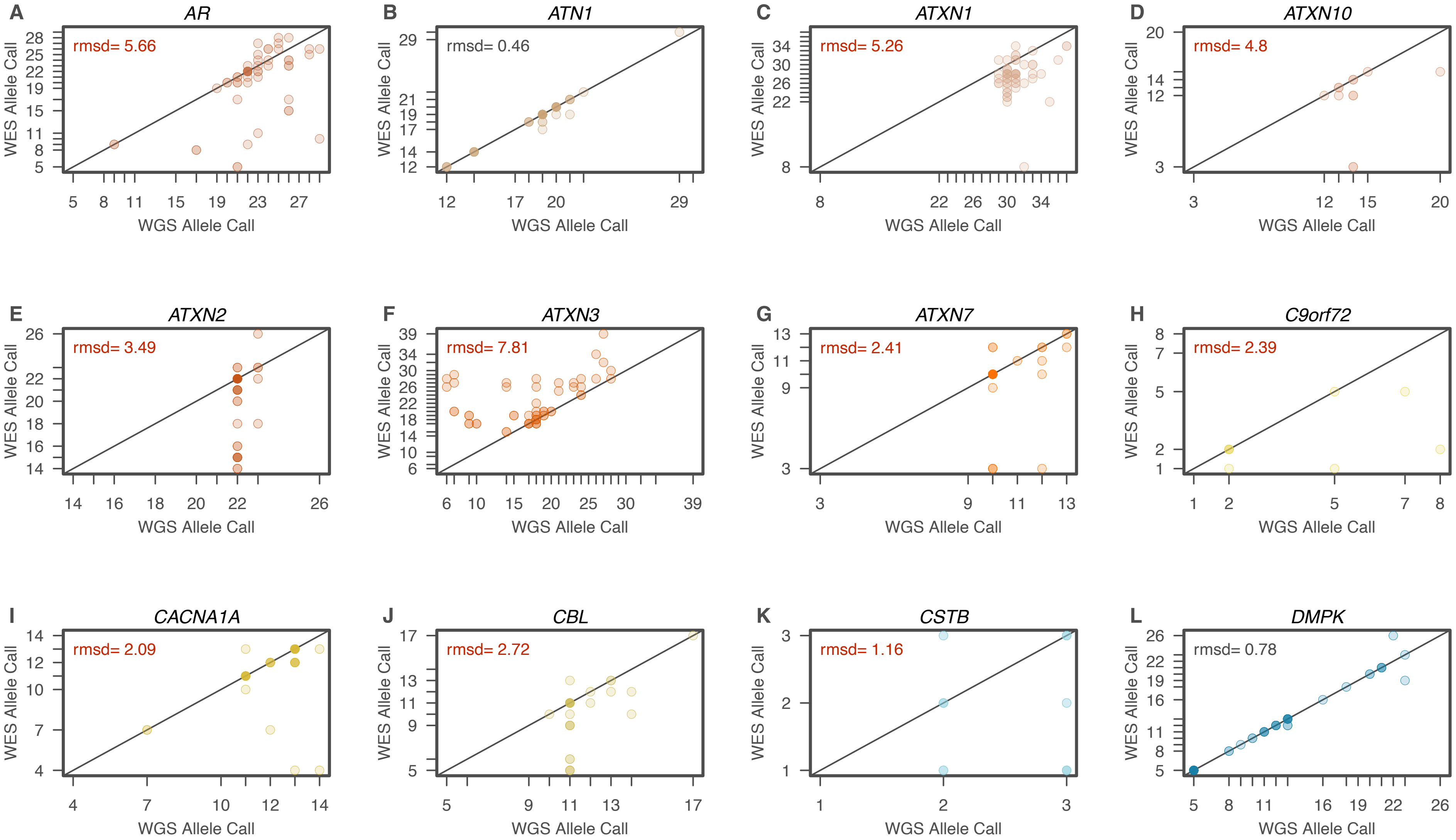

Tredparse : Comparison of Gold Standard PCR Genotyping with Software Allele Prediction
Patient
Control
AR
rmsd= 1.84
ATN1
rmsd= 0.04
ATXN10
rmsd= 0.17
ATXN2
rmsd= 2.25
ATXN3
rmsd= 0.42
ATXN7
rmsd= 0.52
CACNA1A
rmsd= 0.16
CNBP
rmsd= 2.26
DMPK
rmsd= 2.03
FXN
rmsd= 7.1
HTT
rmsd= 1.62
JPH3
rmsd= 0.55
Software Allele Prediction
PCR Allele Measurement

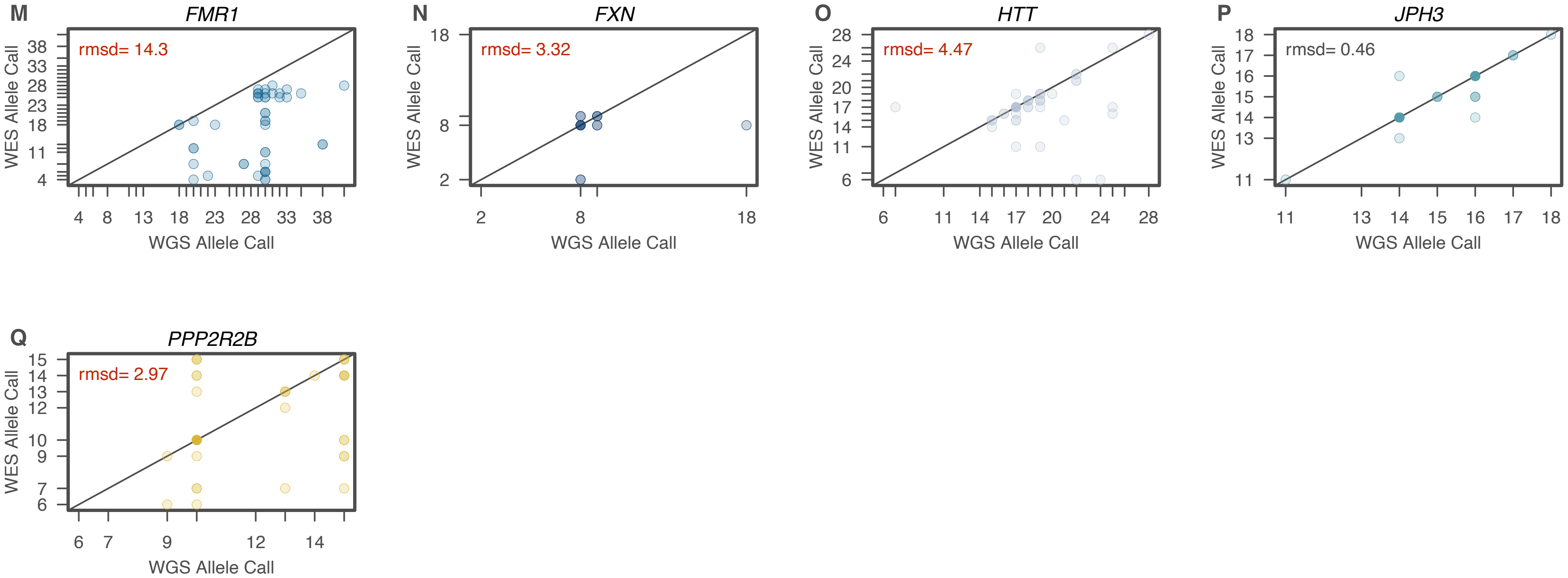


Supplementary Figure 3.7: TREDPARSE comparison of gold standard PCR genotyping with in silico predictions

Gold standard PCR genotypes are compared to predicted alleles using the software TREDPARSE

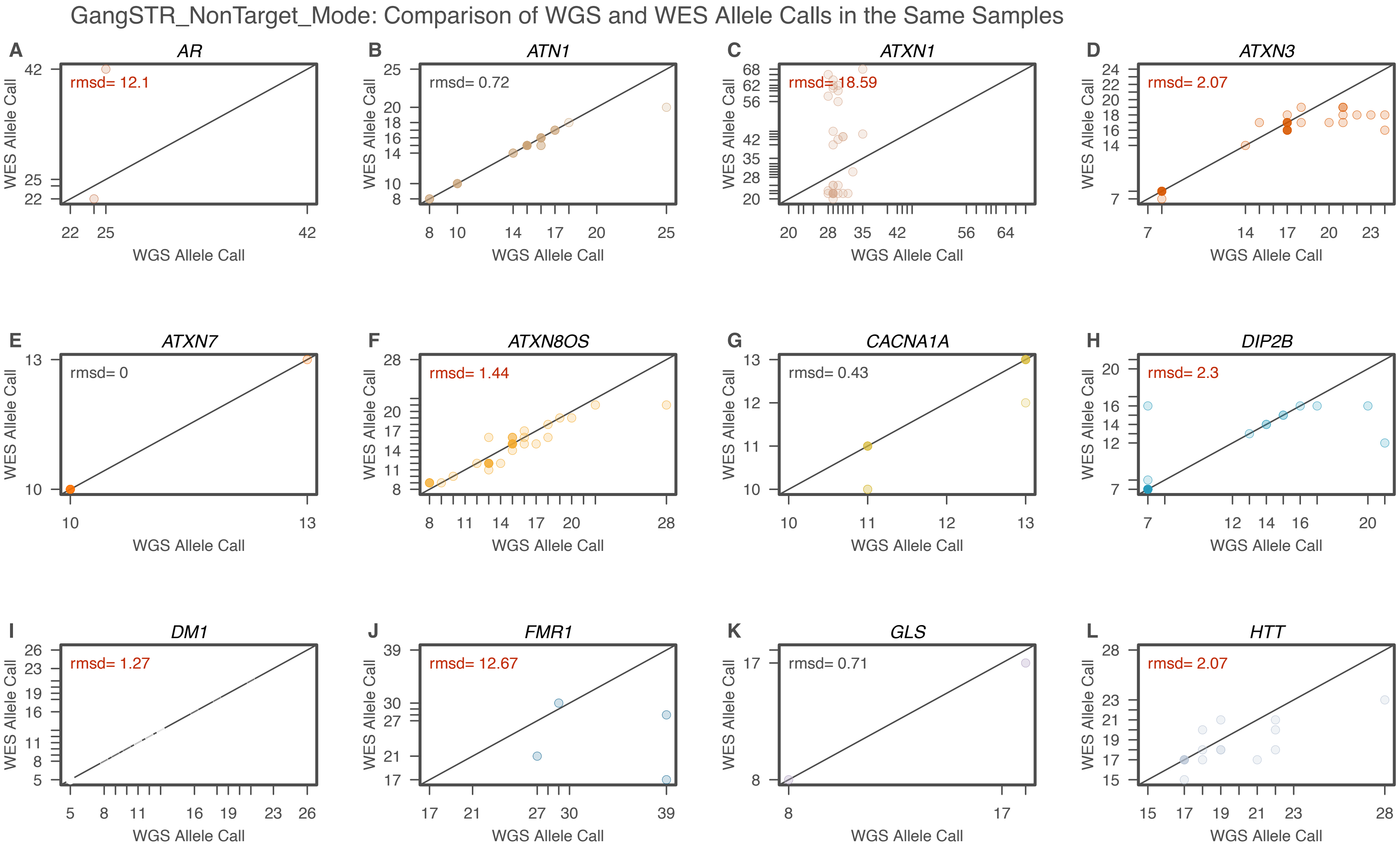
ExpansionHunter_v2: Comparison of WGS and WES Allele Calls in the Same Samples

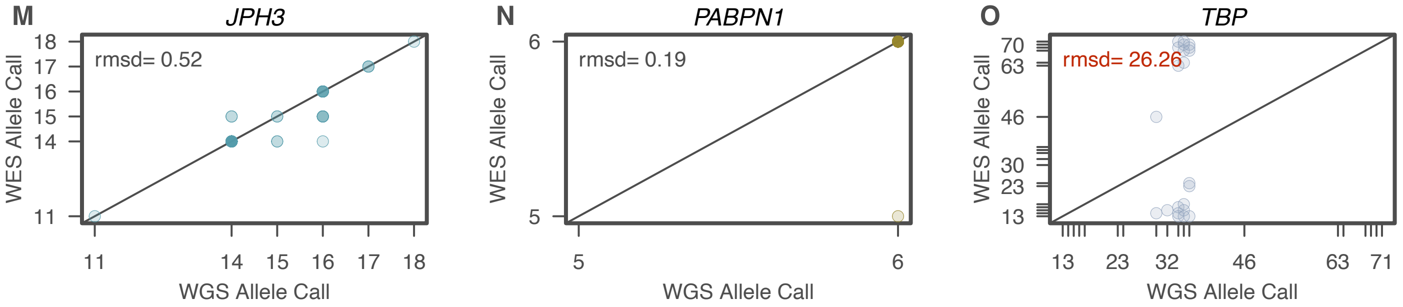




Supplementary Figure 3.8: ExpansionHunter v2: comparison of genotype calls from samples sequenced with WES and WGS

Gold standard PCR genotypes are compared to predicted alleles using the software ExpansionHunter 2

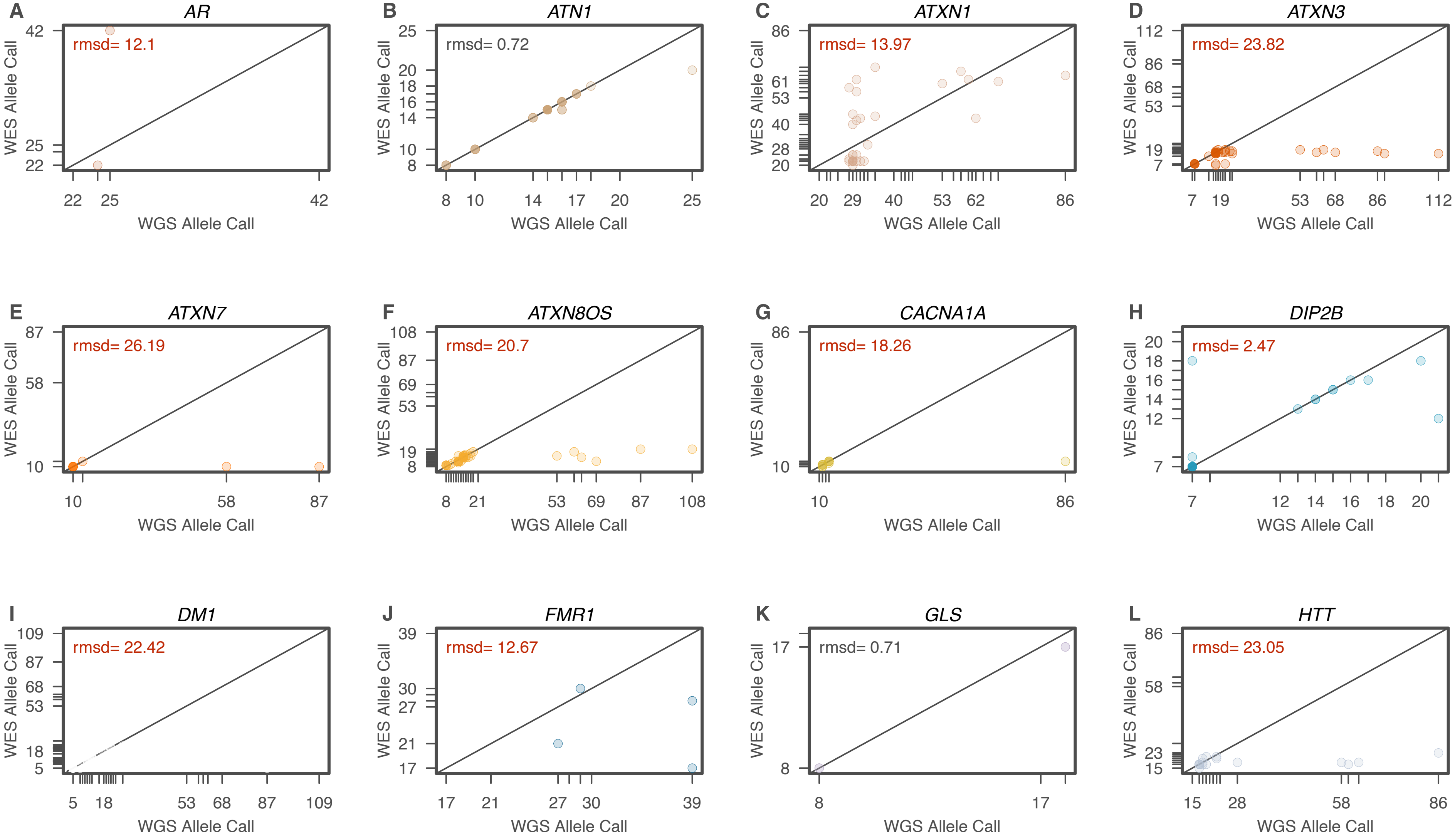

GangSTR_NonTarget_Mode: Comparison of WGS and WES Allele Calls in the Same Samples
A
AR
rmsd= 12.1
42
25
22
22
25
42
B
ATN1
rmsd= 0.72
25
20
18
16
14
10
8
8
10
14
17
20
25
C
ATXN1
rmsd= 18.59
68
62
56
42
35
28
20
20
28
35
42
56
64
D
ATXN3
rmsd= 2.07
24
22
20
18
16
14
7
7
14
17
20
23
E
ATXN7
rmsd= 0
13
10
10
13
F
ATXN8OS
rmsd= 1.44
28
20
17
14
11
8
8
11
14
17
20
28
G
CACNA1A
rmsd= 0.43
13
12
11
10
10
11
12
13
H
DIP2B
rmsd= 2.3
20
16
14
12
7
7
12
14
16
20
I
DM1
rmsd= 1.27
26
23
19
16
11
8
5
5
8
11
16
19
23
26
J
FMR1
rmsd= 12.67
39
30
27
21
17
17
21
27
30
39
K
GLS
rmsd= 0.71
17
8
8
17
L
HTT
rmsd= 2.07
28
23
21
19
17
15
15
17
19
21
23
28
WES Allele Call
WGS Allele Call

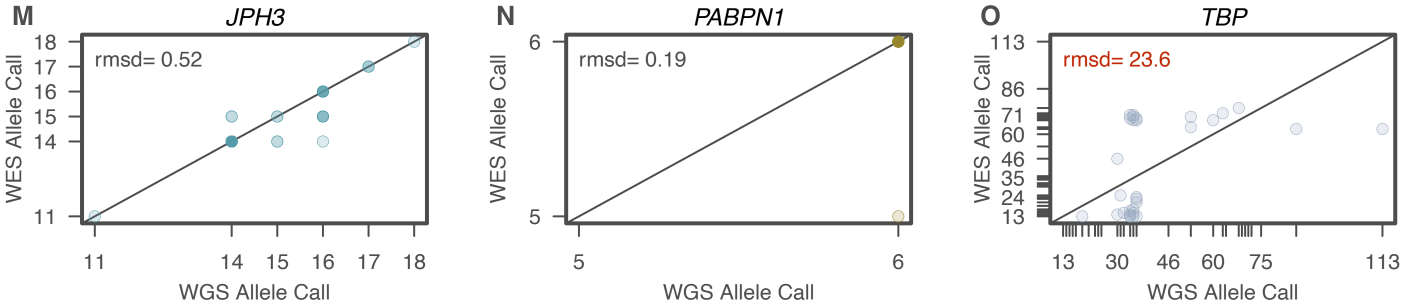


Supplementary Figure 3.9: GangSTR (Genome-Wide): : comparison of genotype calls from Samples sequenced with WES and WGS

Gold standard PCR genotypes are compared to predicted alleles using the software GangSTR (Genome-Wide)

GangSTR_Target_Mode: Comparison of WGS and WES Allele Calls in the Same Samples

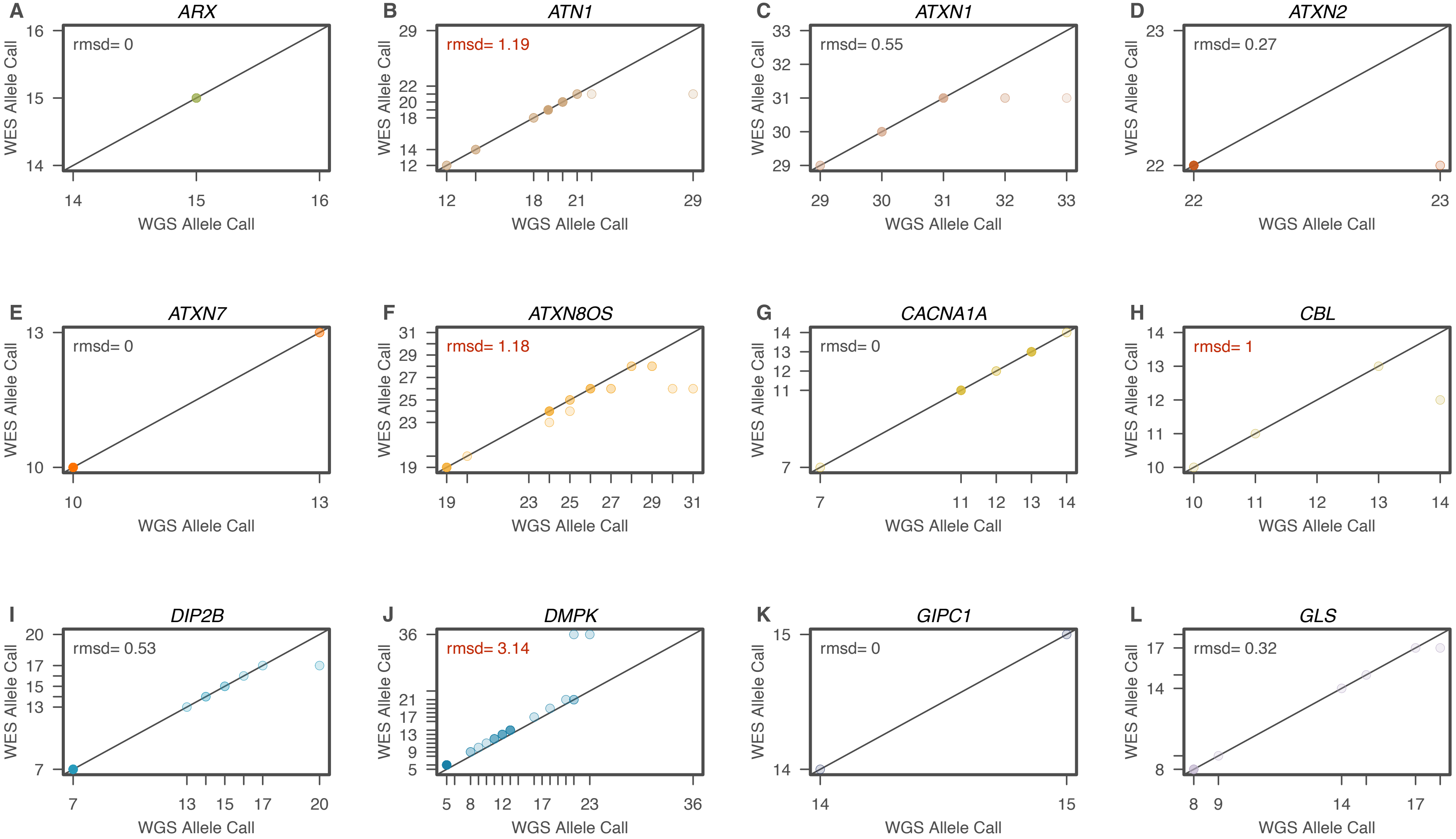


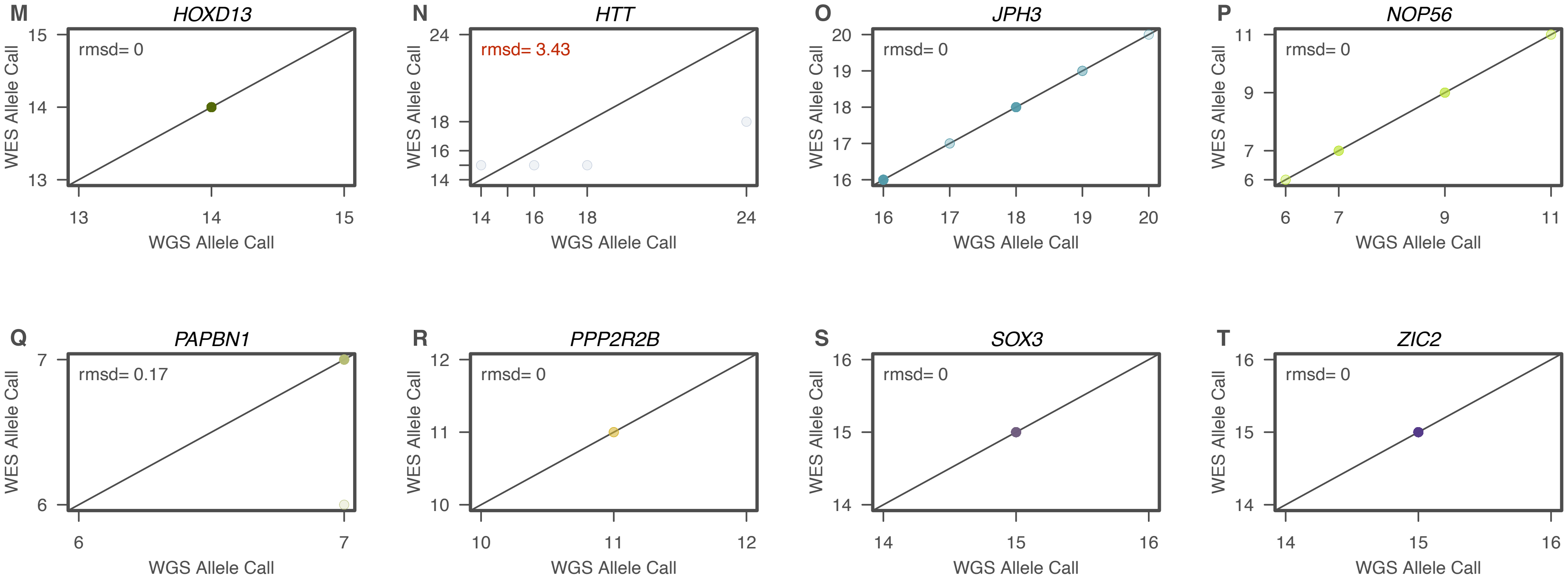


Supplementary Figure 3.10: GangSTR (Target): : comparison of genotype calls from samples sequenced with WES and WGS

Gold standard PCR genotypes are compared to predicted alleles using the software GangSTR (Target)

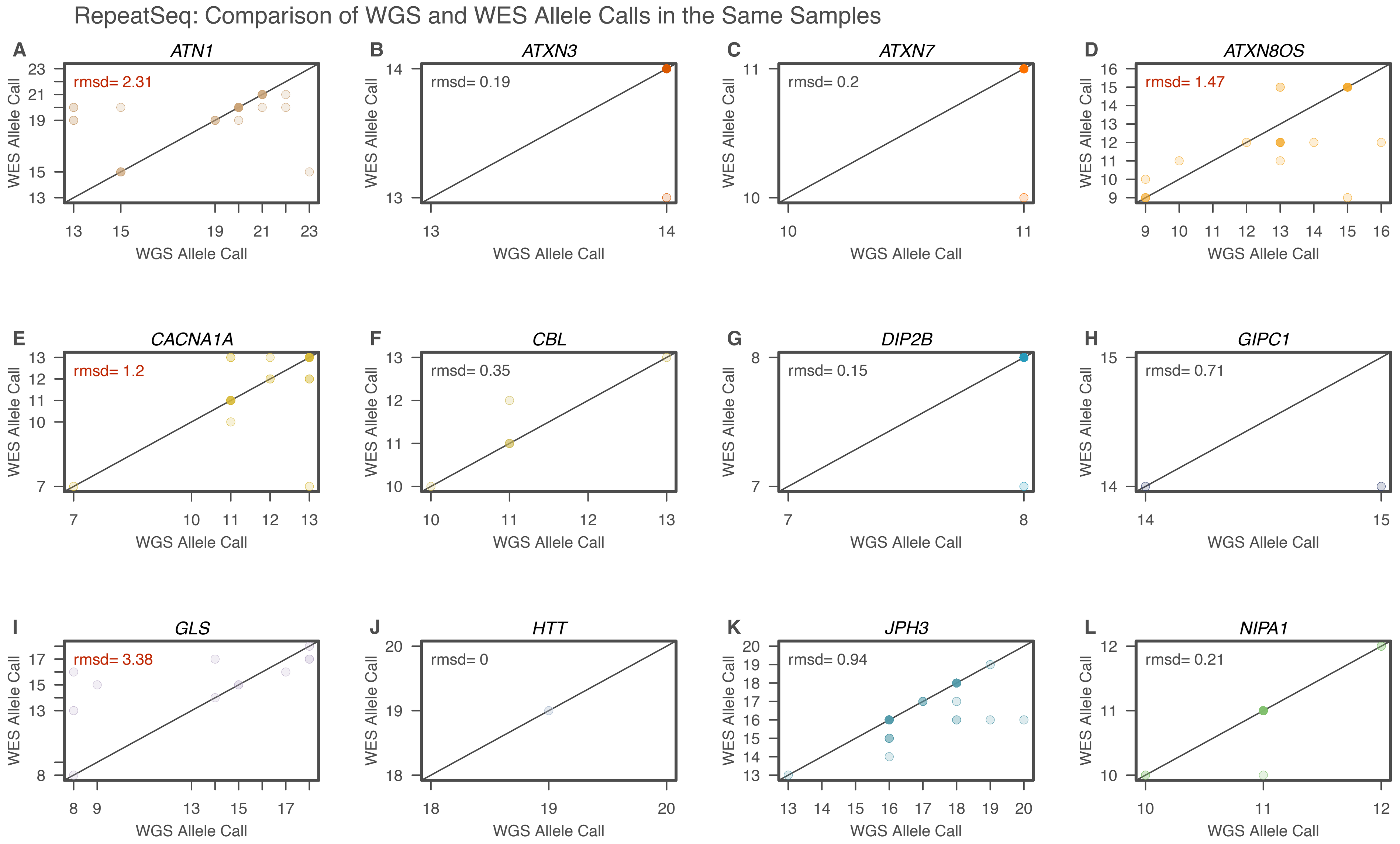
HipSTR: Comparison of WGS and WES Allele Calls in the Same Samples

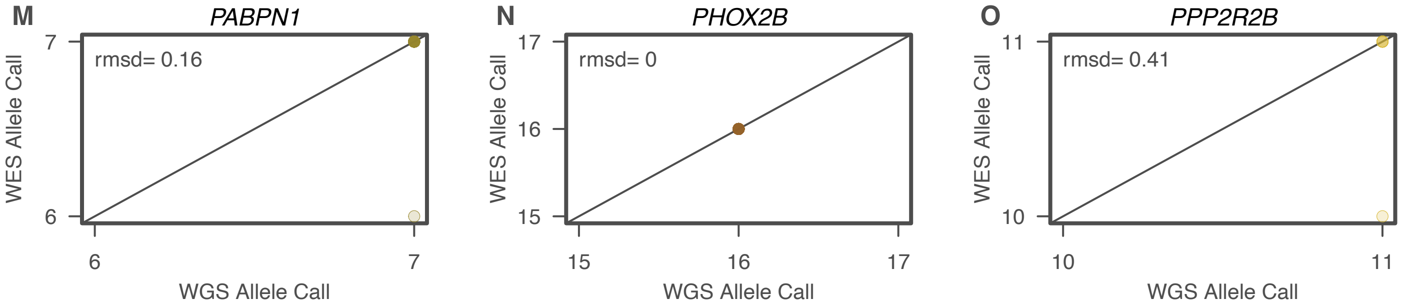




Supplementary Figure 3.11: HipSTR : comparison of genotype calls from samples sequenced with WES and WGS

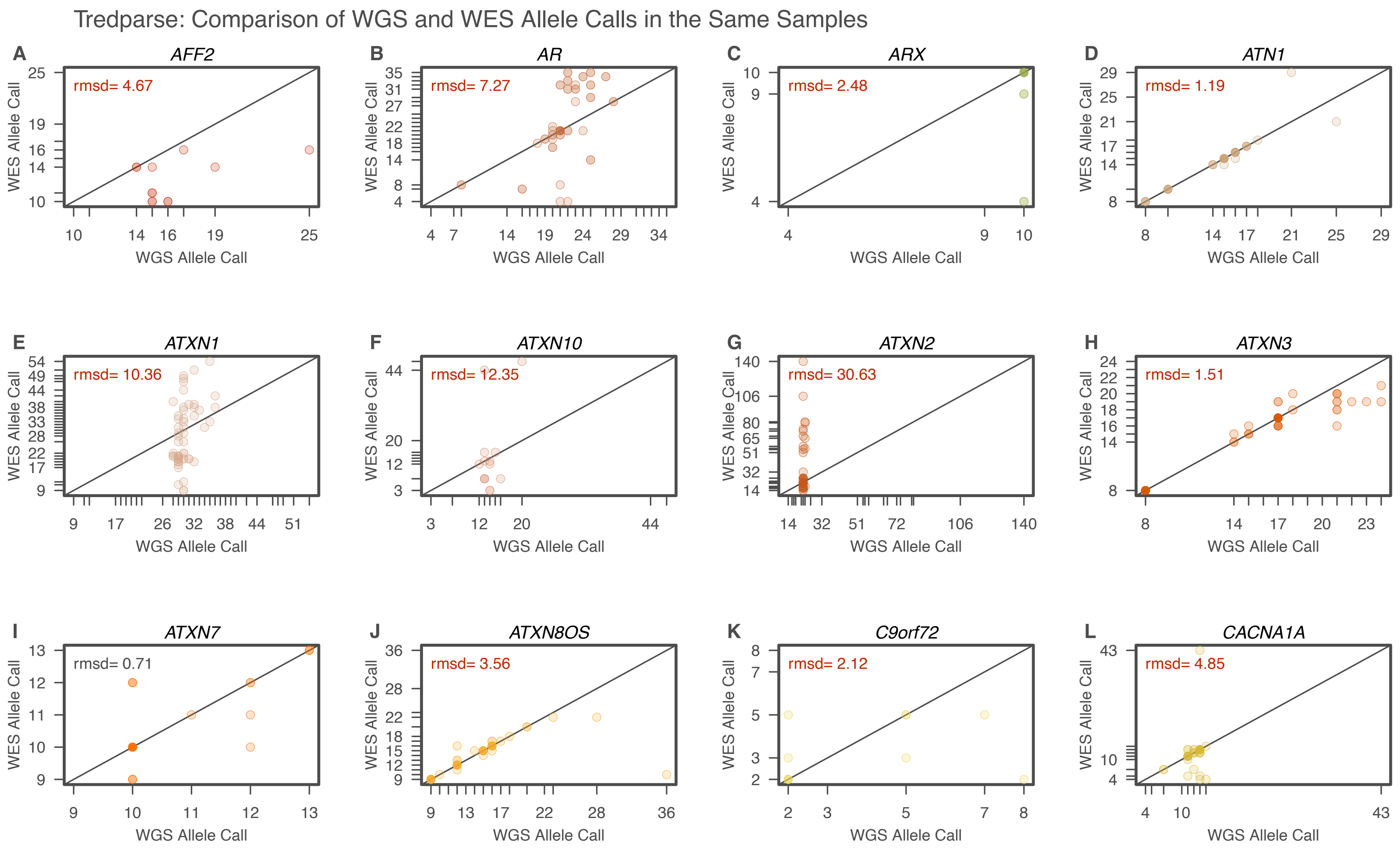
Gold standard PCR genotypes are compared to predicted alleles using the software HipSTR

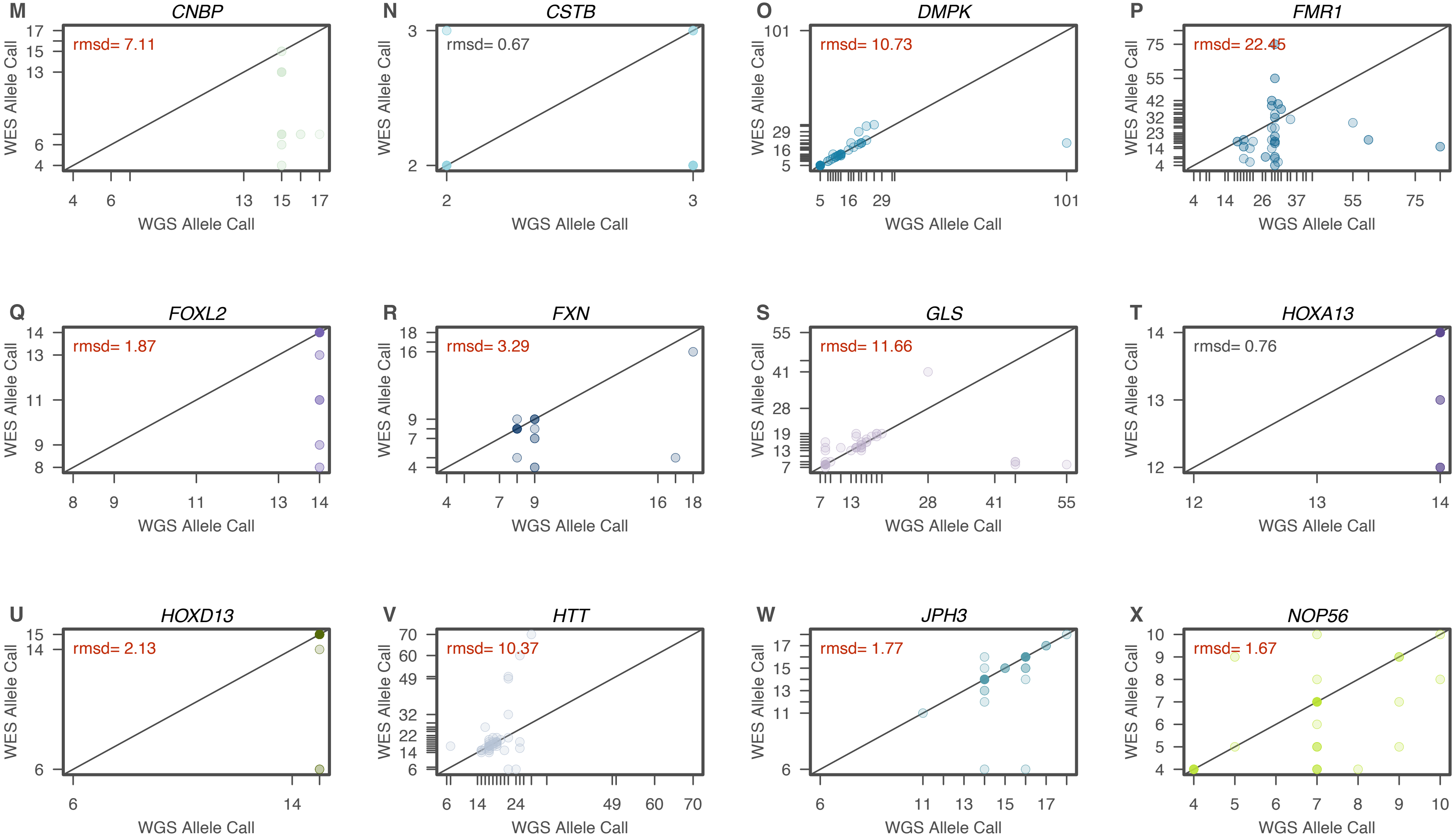

RepeatSeq: Comparison of WGS and WES Allele Calls in the Same Samples
A
ATN1
rmsd= 2.31
23
21
19
15
13
13
15
19
21
23
B
ATXN3
rmsd= 0.19
14
13
13
14
C
ATXN7
rmsd= 0.2
11
10
10
11
D
ATXN8OS
rmsd= 1.47
16
15
14
13
12
11
10
9
9
10
11
12
13
14
15
16
E
CACNA1A
rmsd= 1.2
13
12
11
10
7
7
10
11
12
13
F
CBL
rmsd= 0.35
13
12
11
10
10
11
12
13
G
DIP2B
rmsd= 0.15
8
7
7
8
H
GIPC1
rmsd= 0.71
15
14
14
15
I
GLS
rmsd= 3.38
17
15
13
8
8
9
13
15
17
J
HTT
rmsd= 0
20
19
18
18
19
20
K
JPH3
rmsd= 0.94
20
19
18
17
16
15
14
13
13
14
15
16
17
18
19
20
L
NIPA1
rmsd= 0.21
12
11
10
10
11
12
WES Allele Call
WGS Allele Call

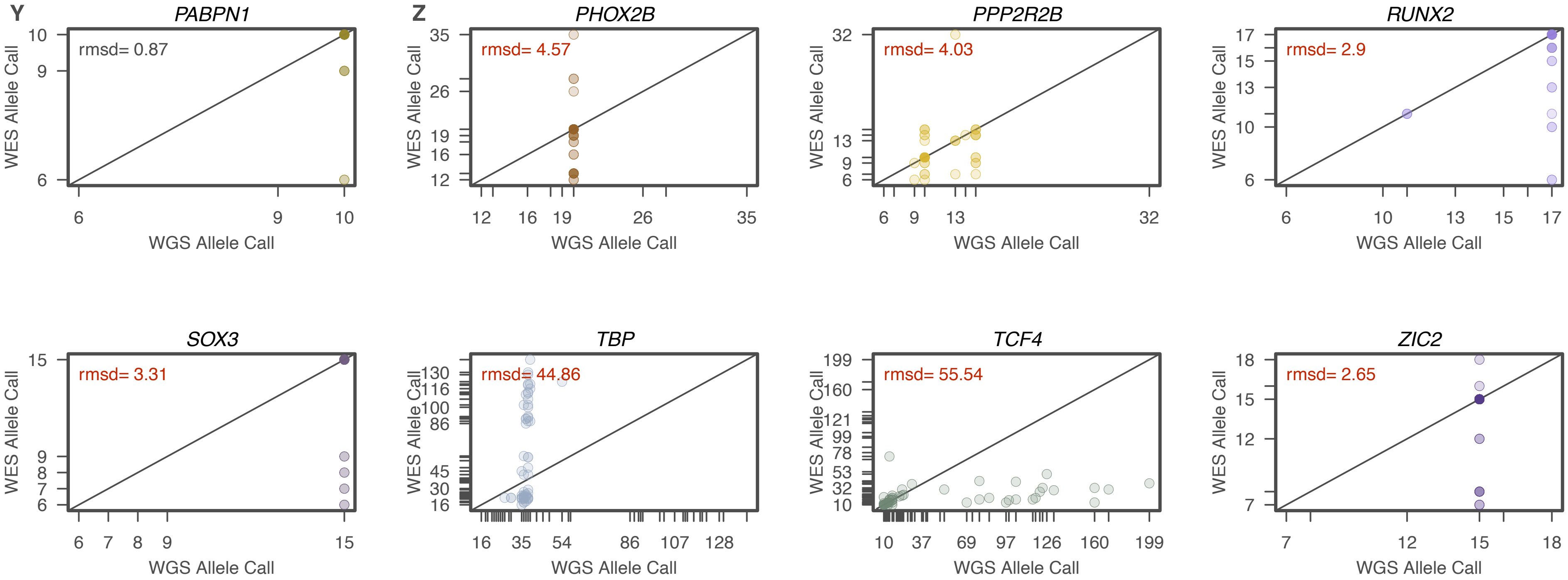


Supplementary Figure 3.12: RepeatSeq: comparison of genotype calls from samples sequenced with WES and WGS

Gold standard PCR genotypes are compared to predicted alleles using the software RepeatSeq

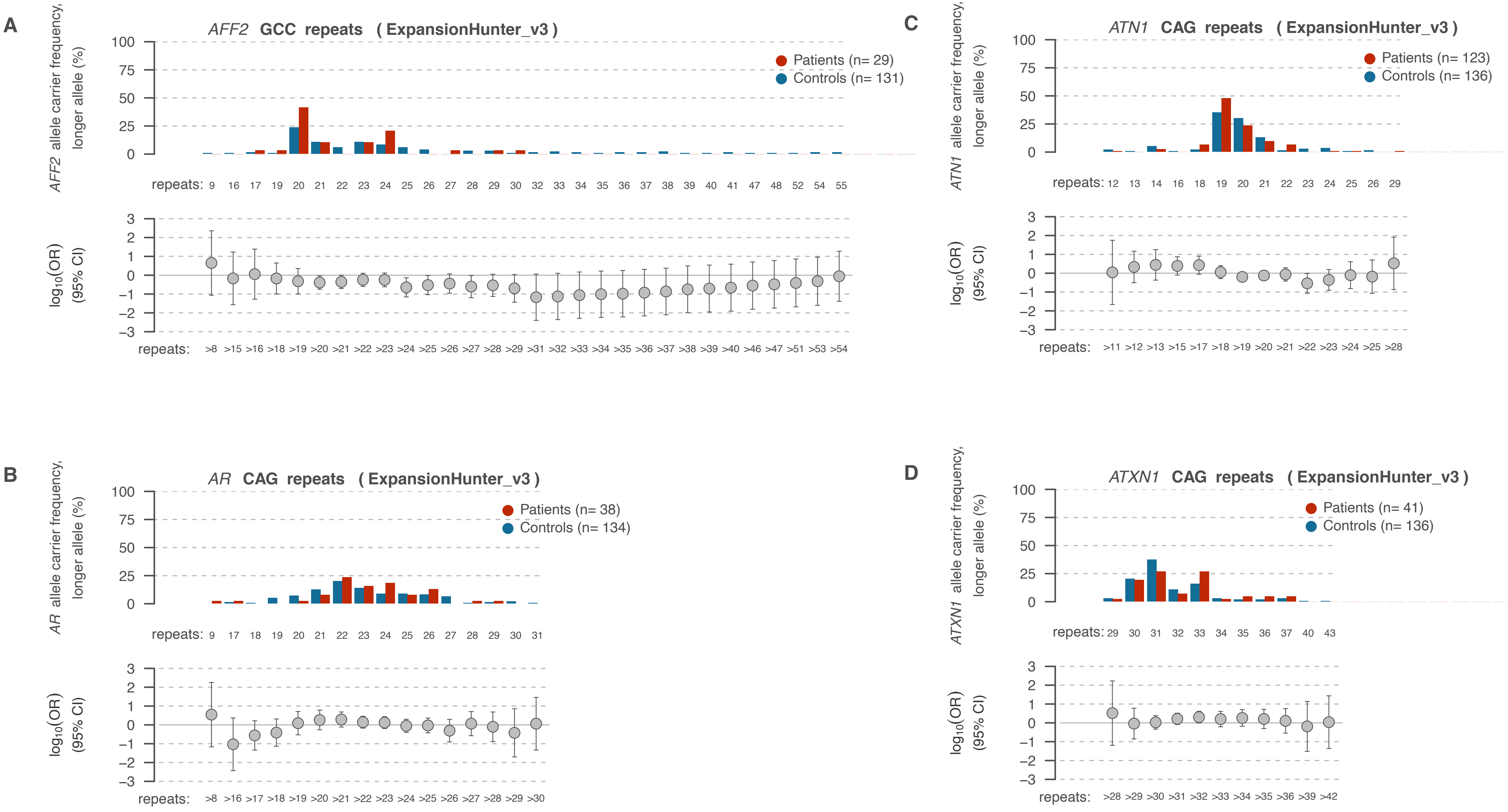

Tredparse: Comparison of WGS and WES Allele Calls in the Same Samples
A
AFF2
rmsd= 4.67
WES Allele Call
25
19
16
14
10
WGS Allele Call
10
14
16
19
25
B
AR
rmsd= 7.27
WES Allele Call
35
31
27
22
18
14
8
4
WGS Allele Call
4
7
14
19
24
29
34
C
ARX
rmsd= 2.48
WES Allele Call
10
9
4
WGS Allele Call
4
9
10
D
ATN1
rmsd= 1.19
WES Allele Call
29
25
21
17
14
8
WGS Allele Call
8
14
17
21
25
29
E
ATXN1
rmsd= 10.36
WES Allele Call
54
49
44
38
33
28
22
17
9
WGS Allele Call
9
17
26
32
38
44
51
F
ATXN10
rmsd= 12.35
WES Allele Call
44
20
12
3
WGS Allele Call
3
12
20
44
G
ATXN2
rmsd= 30.63
WES Allele Call
140
106
80
65
51
32
14
WGS Allele Call
14
32
51
72
106
140
H
ATXN3
rmsd= 1.51
WES Allele Call
24
22
20
18
16
14
8
WGS Allele Call
8
14
17
20
23
I
ATXN7
rmsd= 0.71
WES Allele Call
13
12
11
10
9
WGS Allele Call
9
10
11
12
13
J
ATXN8OS
rmsd= 3.56
WES Allele Call
36
28
22
18
15
12
9
WGS Allele Call
9
13
17
22
28
36
K
C9orf72
rmsd= 2.12
WES Allele Call
8
7
5
3
2
WGS Allele Call
2
3
5
7
8
L
CACNA1A
rmsd= 4.85
WES Allele Call
43
10
4
WGS Allele Call
4
10
43

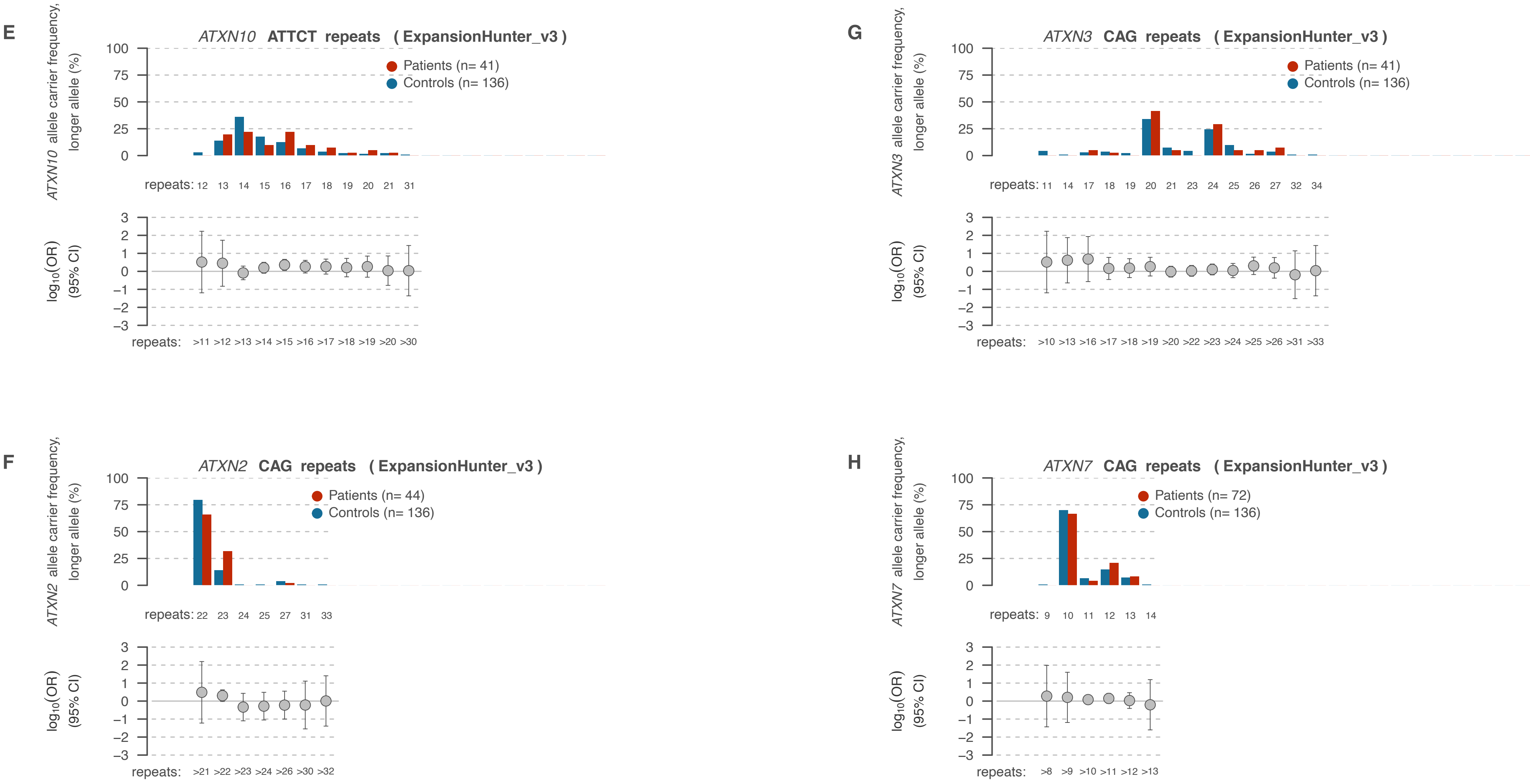

M
CNBP
rmsd= 7.11
WES Allele Call
17
15
13
6
4
4
6
13
15
17
WGS Allele Call
N
CSTB
rmsd= 0.67
3
2
2
3
O
DMPK
rmsd= 10.73
101
29
16
5
5
16
29
101
P
FMR1
rmsd= 22.45
75
55
42
32
23
14
4
4
14
26
37
55
75
Q
FOXL2
rmsd= 1.87
14
13
11
9
8
8
9
11
13
14
R
FXN
rmsd= 3.29
18
16
9
7
4
4
7
9
16
18
S
GLS
rmsd= 11.66
55
41
28
19
13
7
7
13
28
41
55
T
HOXA13
rmsd= 0.76
14
13
12
12
13
14
U
HOXD13
rmsd= 2.13
15
14
6
6
14
V
HTT
rmsd= 10.37
70
60
49
32
22
14
6
6
14
24
49
60
70
W
JPH3
rmsd= 1.77
17
15
13
11
6
6
11
13
15
17
X
NOP56
rmsd= 1.67
10
9
8
7
6
5
4
4
5
6
7
8
9
10

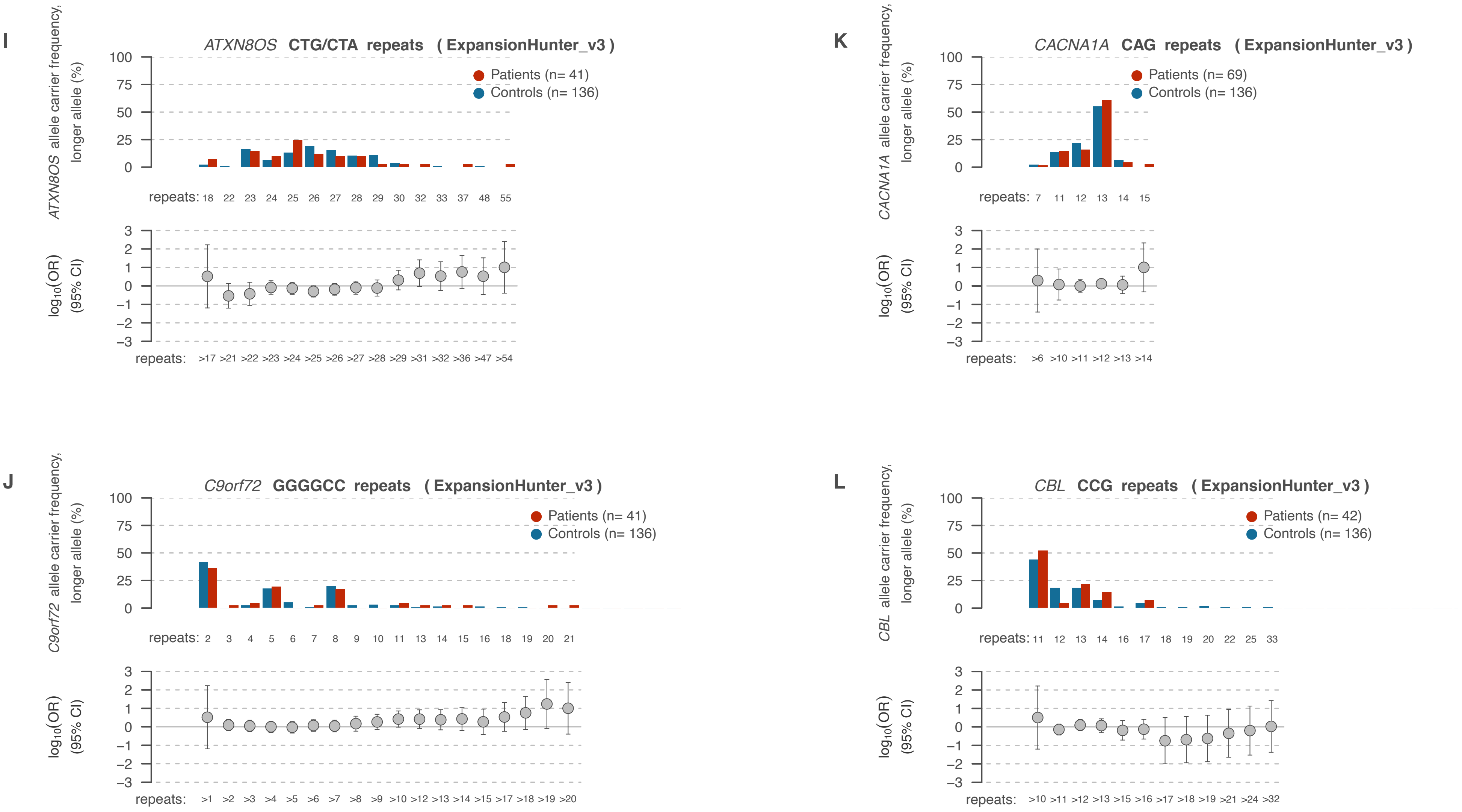


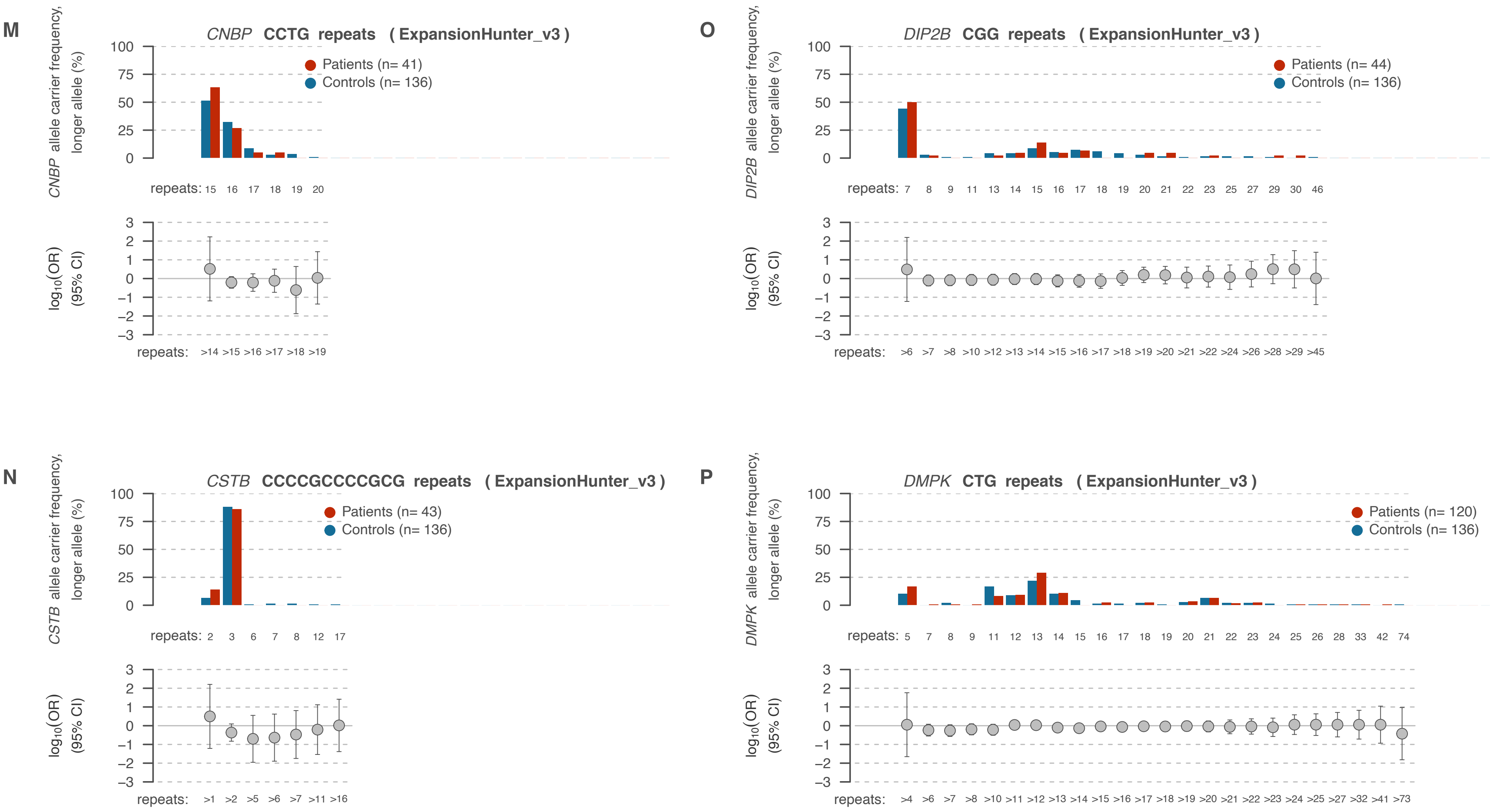
Supplementary Figure 3.13: TREDPARSE: comparison of genotype calls from samples sequenced with WES and WGS

Gold standard PCR genotypes are compared to predicted alleles using the software TREDPARSE

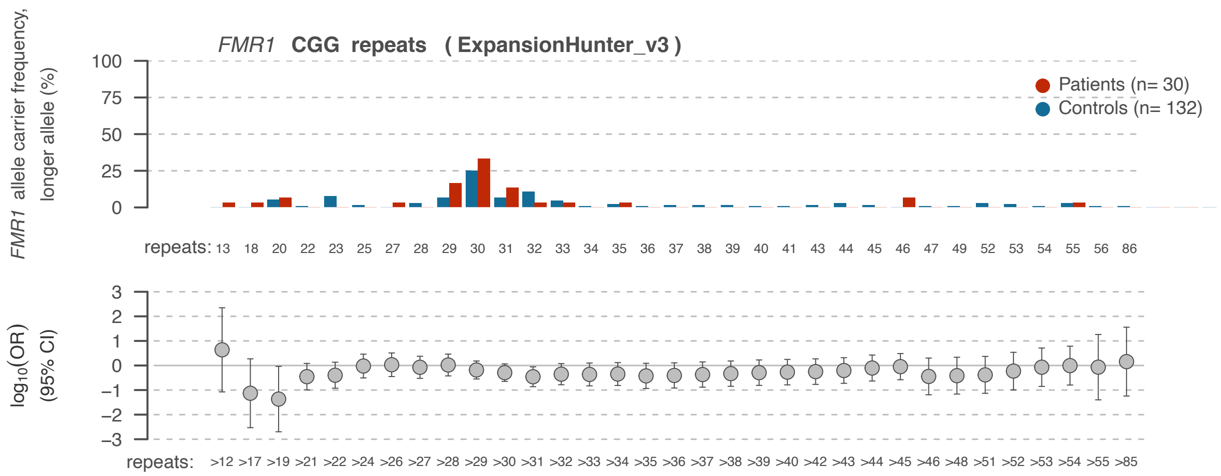

A
AFF2 GCC repeats (ExpansionHunter_v3)
AFF2 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 29)
Controls (n= 131)
repeats: 9 16 17 19 20 21 22 23 24 25 26 27 28 29 30 32 33 34 35 36 37 38 39 40 41 47 48 52 54 55
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >15 >16 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >29 >31 >32 >33 >34 >35 >36 >37 >38 >39 >40 >46 >47 >51 >53 >54
B
AR CAG repeats (ExpansionHunter_v3)
AR allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 38)
Controls (n= 134)
repeats: 9 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >29 >30
C
ATN1 CAG repeats (ExpansionHunter_v3)
ATN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 123)
Controls (n= 136)
repeats: 12 13 14 16 18 19 20 21 22 23 24 25 26 29
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >11 >12 >13 >15 >17 >18 >19 >20 >21 >22 >23 >24 >25 >28
D
ATXN1 CAG repeats (ExpansionHunter_v3)
ATXN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 29 30 31 32 33 34 35 36 37 40 43
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >28 >29 >30 >31 >32 >33 >34 >35 >36 >39 >42


E
ATXN10 ATTCT repeats (ExpansionHunter_v3)
ATXN10 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 136)
repeats: 12 13 14 15 16 17 18 19 20 21 31
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >30
G
ATXN3 CAG repeats (ExpansionHunter_v3)
ATXN3 allele carrier frequency, longer allele (%)
● Patients (n= 41)
● Controls (n= 136)
repeats: 11 14 17 18 19 20 21 23 24 25 26 27 32 34
log10(OR) (95% CI)
repeats: >10 >13 >16 >17 >18 >19 >20 >22 >23 >24 >25 >26 >31 >33
F
ATXN2 CAG repeats (ExpansionHunter_v3)
ATXN2 allele carrier frequency, longer allele (%)
● Patients (n= 44)
● Controls (n= 136)
repeats: 22 23 24 25 27 31 33
log10(OR) (95% CI)
repeats: >21 >22 >23 >24 >26 >30 >32
H
ATXN7 CAG repeats (ExpansionHunter_v3)
ATXN7 allele carrier frequency, longer allele (%)
● Patients (n= 72)
● Controls (n= 136)
repeats: 9 10 11 12 13 14
log10(OR) (95% CI)
repeats: >8 >9 >10 >11 >12 >13

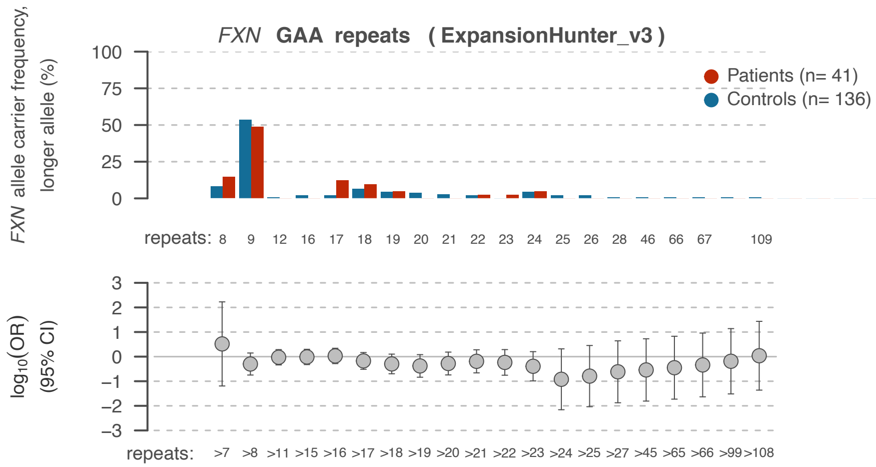

I
ATXN8OS CTG/CTA repeats (ExpansionHunter_v3)
ATXN8OS allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 18 22 23 24 25 26 27 28 29 30 32 33 37 48 55
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >17 >21 >22 >23 >24 >25 >26 >27 >28 >29 >31 >32 >36 >47 >54
K
CACNA1A CAG repeats (ExpansionHunter_v3)
CACNA1A allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 69)
Controls (n= 136)
repeats: 7 11 12 13 14 15
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >10 >11 >12 >13 >14
J
C9orf72 GGGGCC repeats (ExpansionHunter_v3)
C9orf72 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 2 3 4 5 6 7 8 9 10 11 13 14 15 16 18 19 20 21
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >1 >2 >3 >4 >5 >6 >7 >8 >9 >10 >12 >13 >14 >15 >17 >18 >19 >20
L
CBL CCG repeats (ExpansionHunter_v3)
CBL allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 42)
Controls (n= 136)
repeats: 11 12 13 14 16 17 18 19 20 22 25 33
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >10 >11 >12 >13 >15 >16 >17 >18 >19 >21 >24 >32

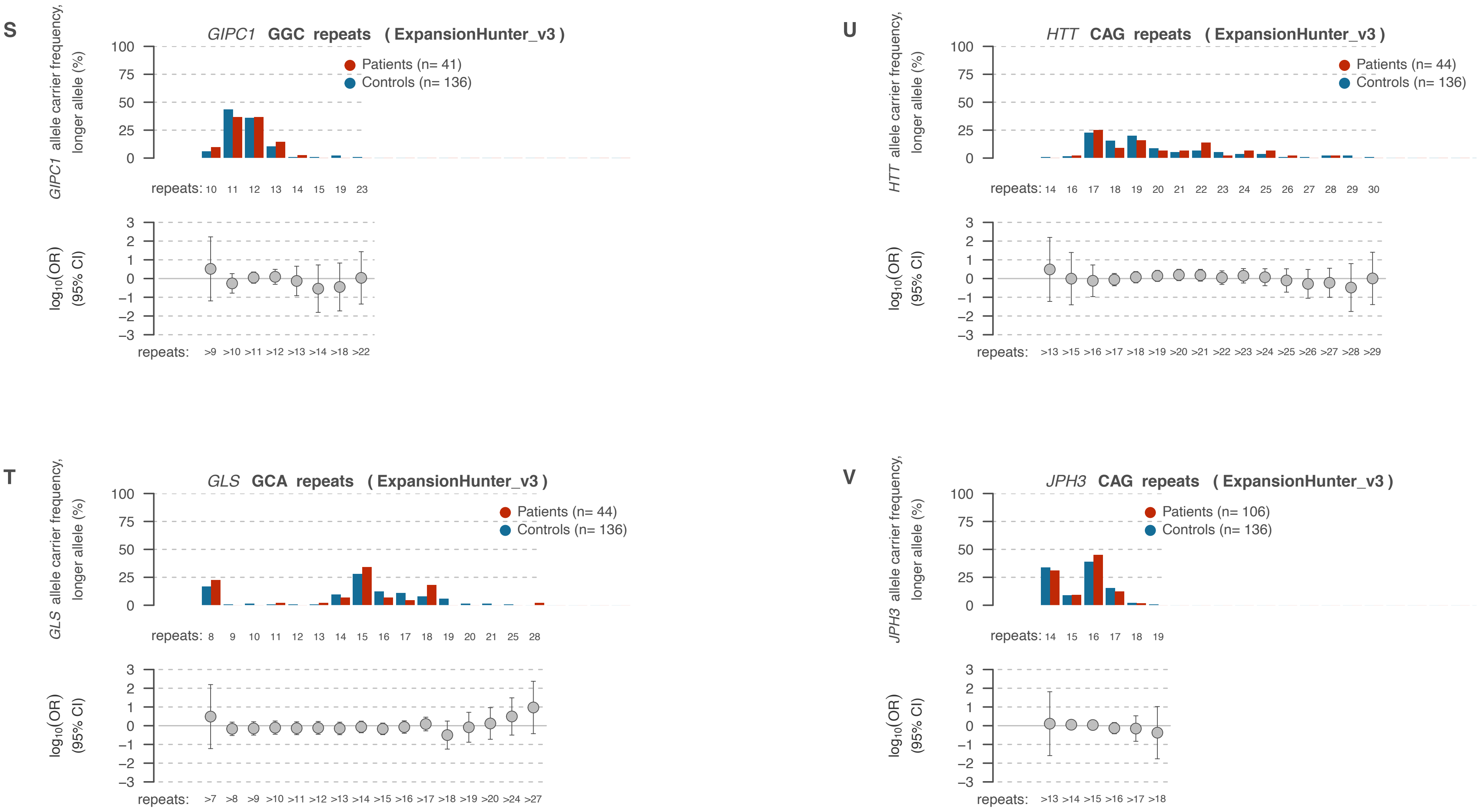

M
CNBP CCTG repeats (ExpansionHunter_v3)
CNBP allele carrier frequency, longer allele (%)
Patients (n= 41)
Controls (n= 136)
100
75
50
25
0
repeats: 15 16 17 18 19 20
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >14 >15 >16 >17 >18 >19
O
DIP2B CGG repeats (ExpansionHunter_v3)
DIP2B allele carrier frequency, longer allele (%)
Patients (n= 44)
Controls (n= 136)
repeats: 7 8 9 11 13 14 15 16 17 18 19 20 21 22 23 25 27 29 30 46
repeats: >6 >7 >8 >10 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >24 >26 >28 >29 >45
N
CSTB CCCCGCCCCGCG repeats (ExpansionHunter_v3)
CSTB allele carrier frequency, longer allele (%)
Patients (n= 43)
Controls (n= 136)
repeats: 2 3 6 7 8 12 17
repeats: >1 >2 >5 >6 >7 >11 >16
P
DMPK CTG repeats (ExpansionHunter_v3)
DMPK allele carrier frequency, longer allele (%)
Patients (n= 120)
Controls (n= 136)
repeats: 5 7 8 9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 33 42 74
repeats: >4 >6 >7 >8 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >27 >32 >41 >73

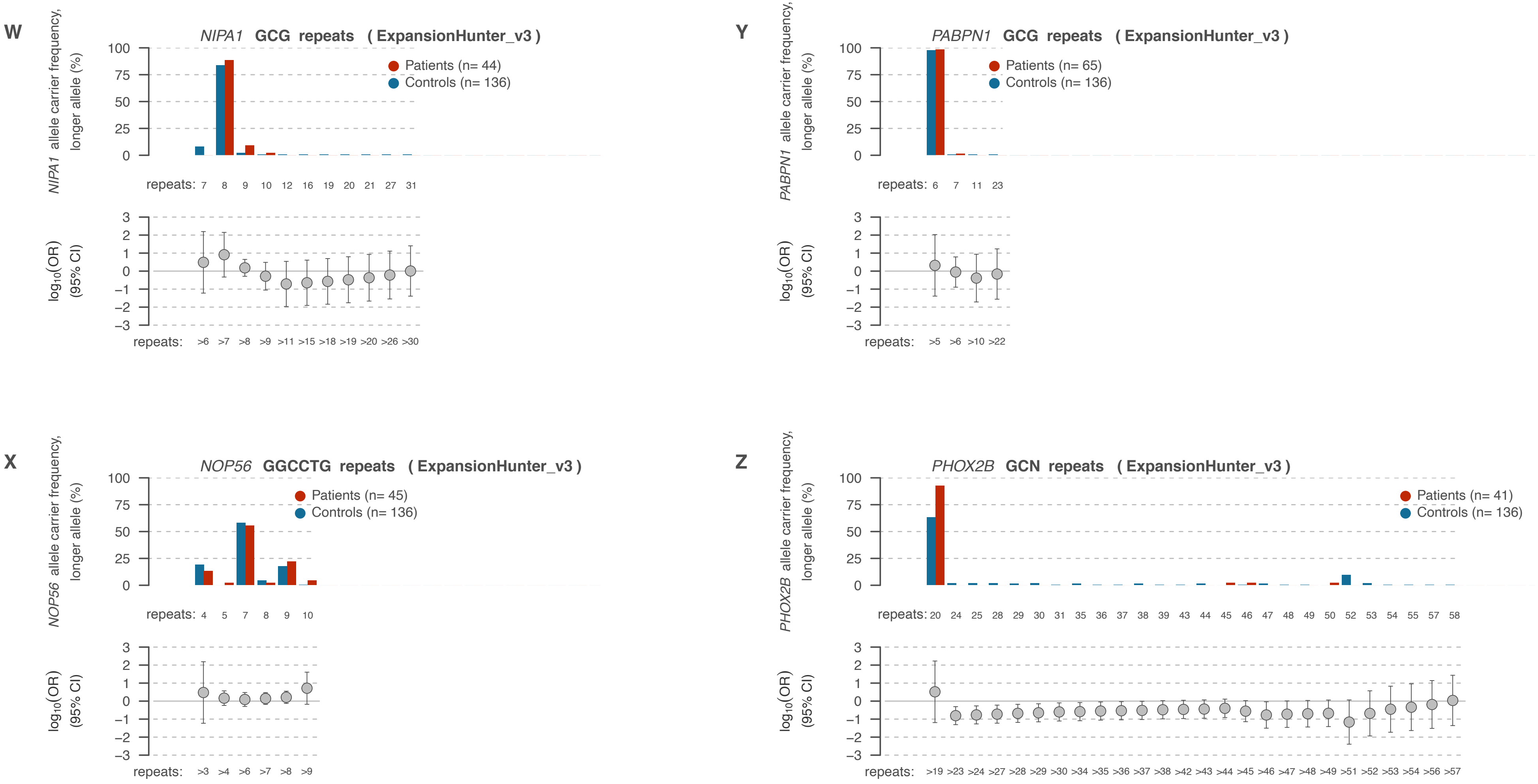
Q


FMR1 CGG repeats (ExpansionHunter_v3)
FMR1 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 30)
● Controls (n= 132)
repeats: 13 18 20 22 23 25 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 43 44 45 46 47 49 52 53 54 55 56 86
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >12 >17 >19 >21 >22 >24 >26 >27 >28 >29 >30 >31 >32 >33 >34 >35 >36 >37 >38 >39 >40 >42 >43 >44 >45 >46 >48 >51 >52 >53 >54 >55 >85

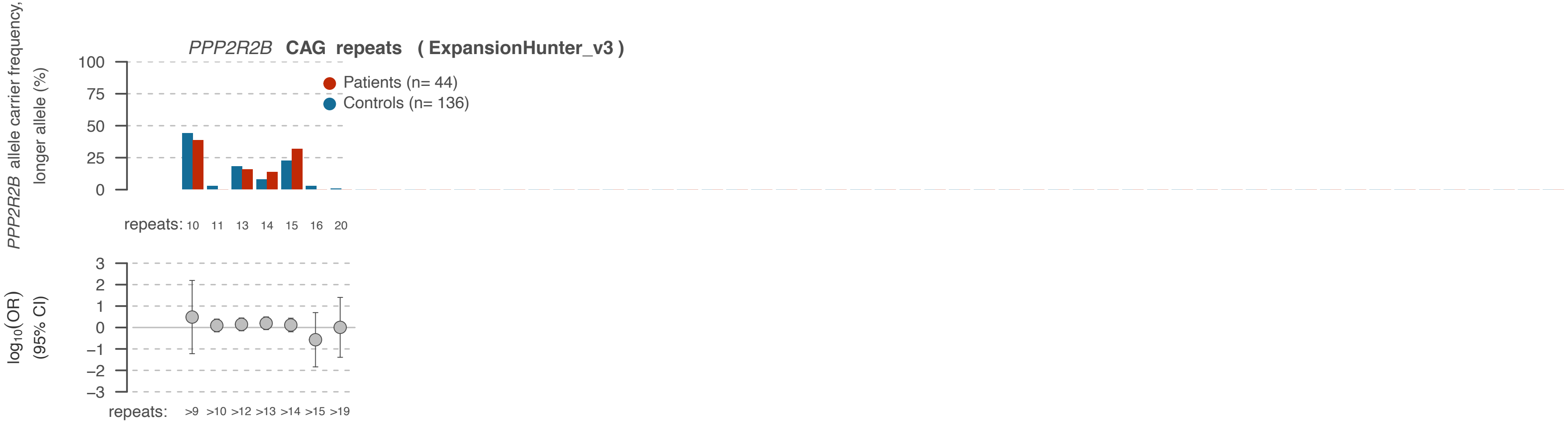
R


FXN GAA repeats (ExpansionHunter_v3)
FXN allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 136)
repeats: 8 9 12 16 17 18 19 20 21 22 23 24 25 26 28 46 66 67 109
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >11 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >27 >45 >65 >66 >99 >108


S
GIPC1 GGC repeats (ExpansionHunter_v3)
GIPC1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 10 11 12 13 14 15 19 23
log10(OR) (95% CI)
3 2 1 0 −1 −2 −3
repeats: >9 >10 >11 >12 >13 >14 >18 >22
U
HTT CAG repeats (ExpansionHunter_v3)
HTT allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 44)
Controls (n= 136)
repeats: 14 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
log10(OR) (95% CI)
3 2 1 0 −1 −2 −3
repeats: >13 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >29
T
GLS GCA repeats (ExpansionHunter_v3)
GLS allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 44)
Controls (n= 136)
repeats: 8 9 10 11 12 13 14 15 16 17 18 19 20 21 25 28
log10(OR) (95% CI)
3 2 1 0 −1 −2 −3
repeats: >7 >8 >9 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >24 >27
V
JPH3 CAG repeats (ExpansionHunter_v3)
JPH3 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 106)
Controls (n= 136)
repeats: 14 15 16 17 18 19
log10(OR) (95% CI)
3 2 1 0 −1 −2 −3
repeats: >13 >14 >15 >16 >17 >18

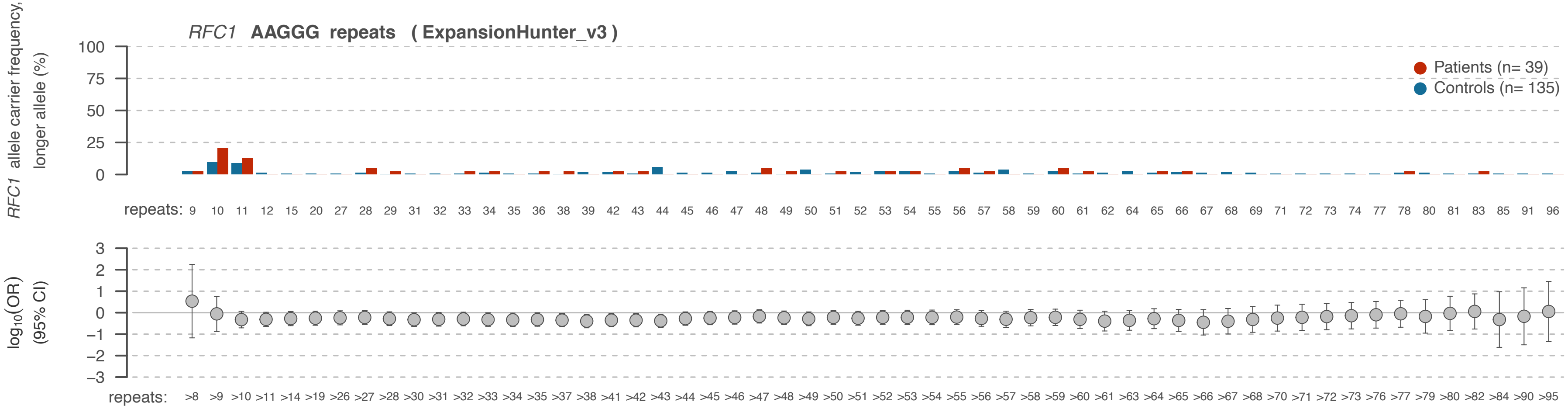

W
NIPA1 GCG repeats (ExpansionHunter_v3)
NIPA1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 44)
Controls (n= 136)
repeats: 7 8 9 10 12 16 19 20 21 27 31
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >7 >8 >9 >11 >15 >18 >19 >20 >26 >30
Y
PABPN1 GCG repeats (ExpansionHunter_v3)
PABPN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 65)
Controls (n= 136)
repeats: 6 7 11 23
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >5 >6 >10 >22
X
NOP56 GGCCTG repeats (ExpansionHunter_v3)
NOP56 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 45)
Controls (n= 136)
repeats: 4 5 7 8 9 10
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >3 >4 >6 >7 >8 >9
Z
PHOX2B GCN repeats (ExpansionHunter_v3)
PHOX2B allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 20 24 25 28 29 30 31 35 36 37 38 39 43 44 45 46 47 48 49 50 52 53 54 55 57 58
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >19 >23 >24 >27 >28 >29 >30 >34 >35 >36 >37 >38 >42 >43 >44 >45 >46 >47 >48 >49 >51 >52 >53 >54 >56 >57

A.2

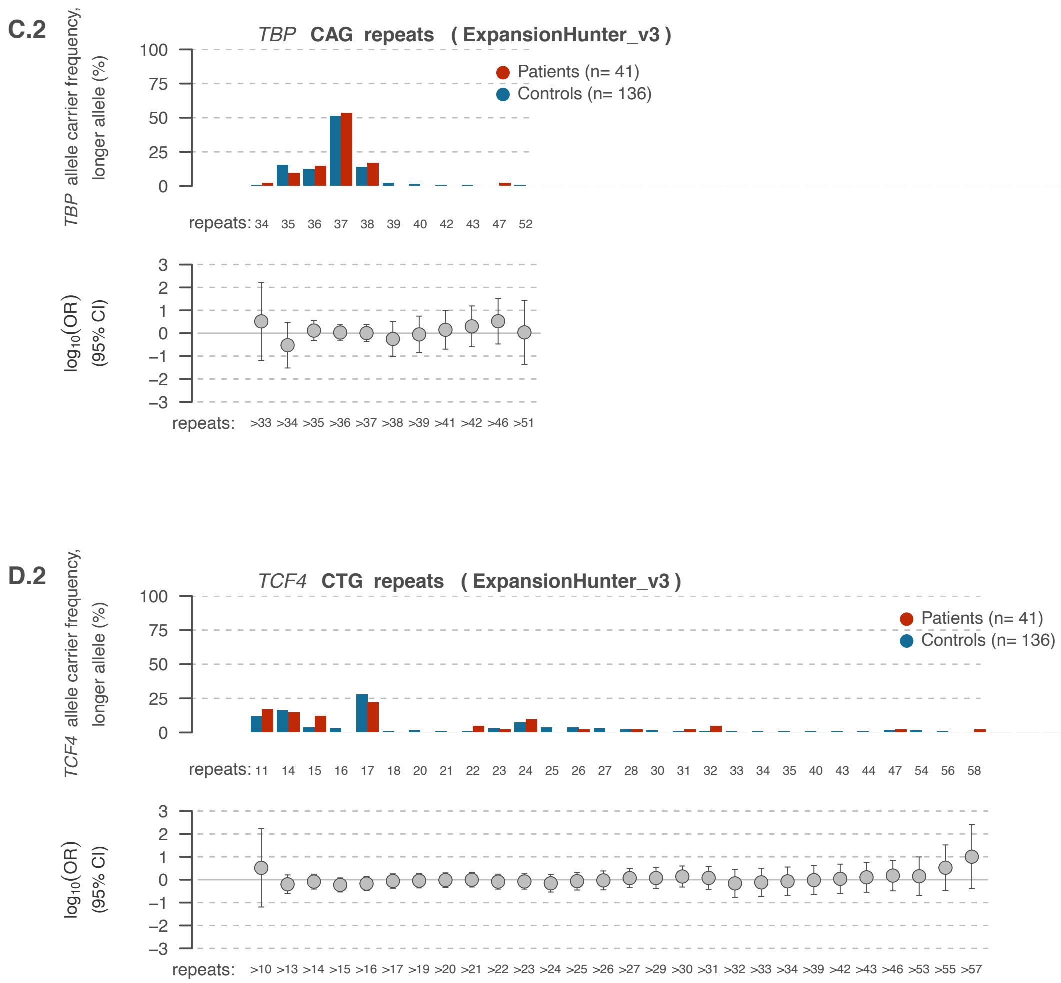


PPP2R2B CAG repeats (ExpansionHunter_v3)
PPP2R2B allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 44)
Controls (n= 136)
repeats: 10 11 13 14 15 16 20
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >9 >10 >12 >13 >14 >15 >19

B.2

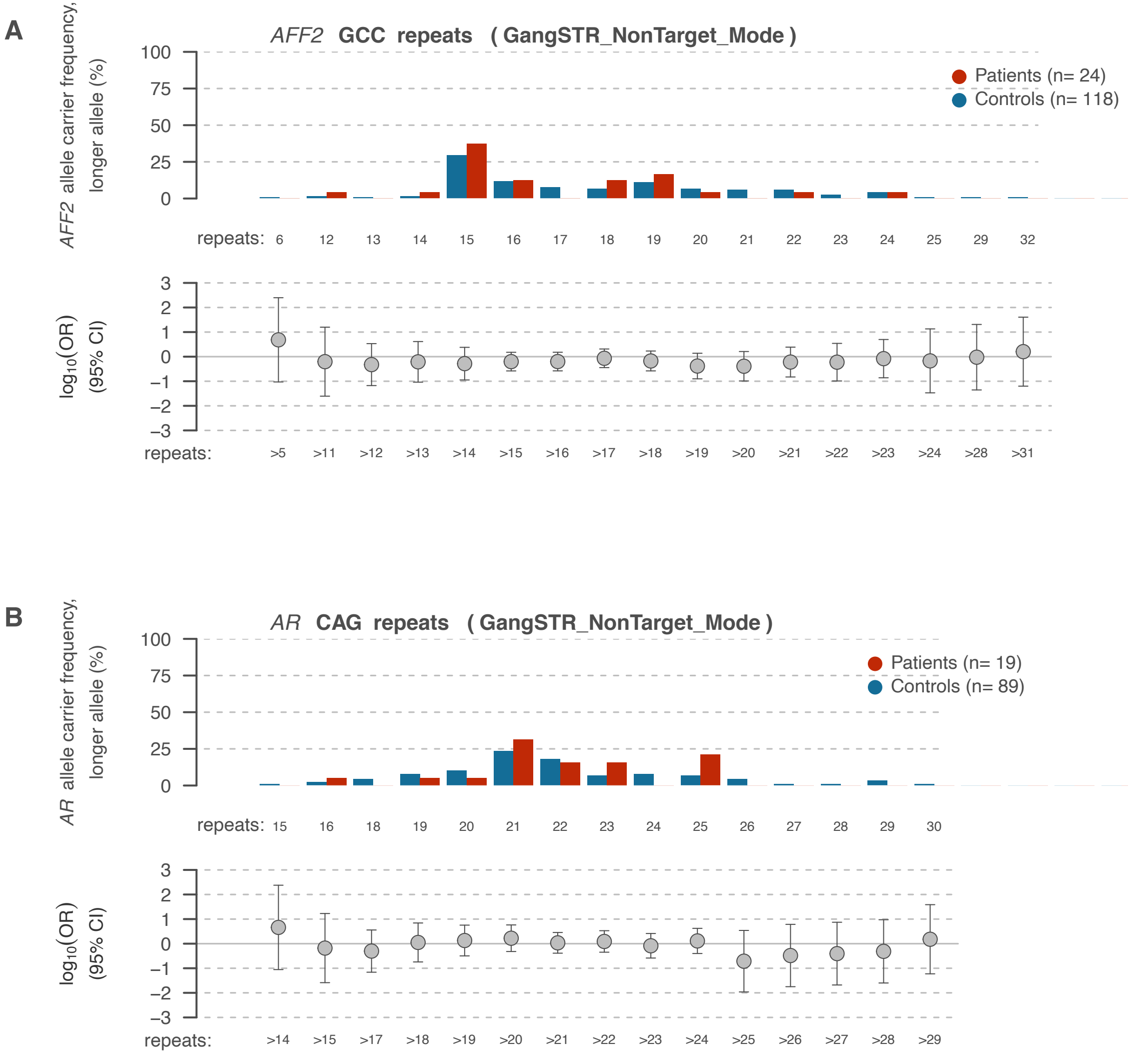


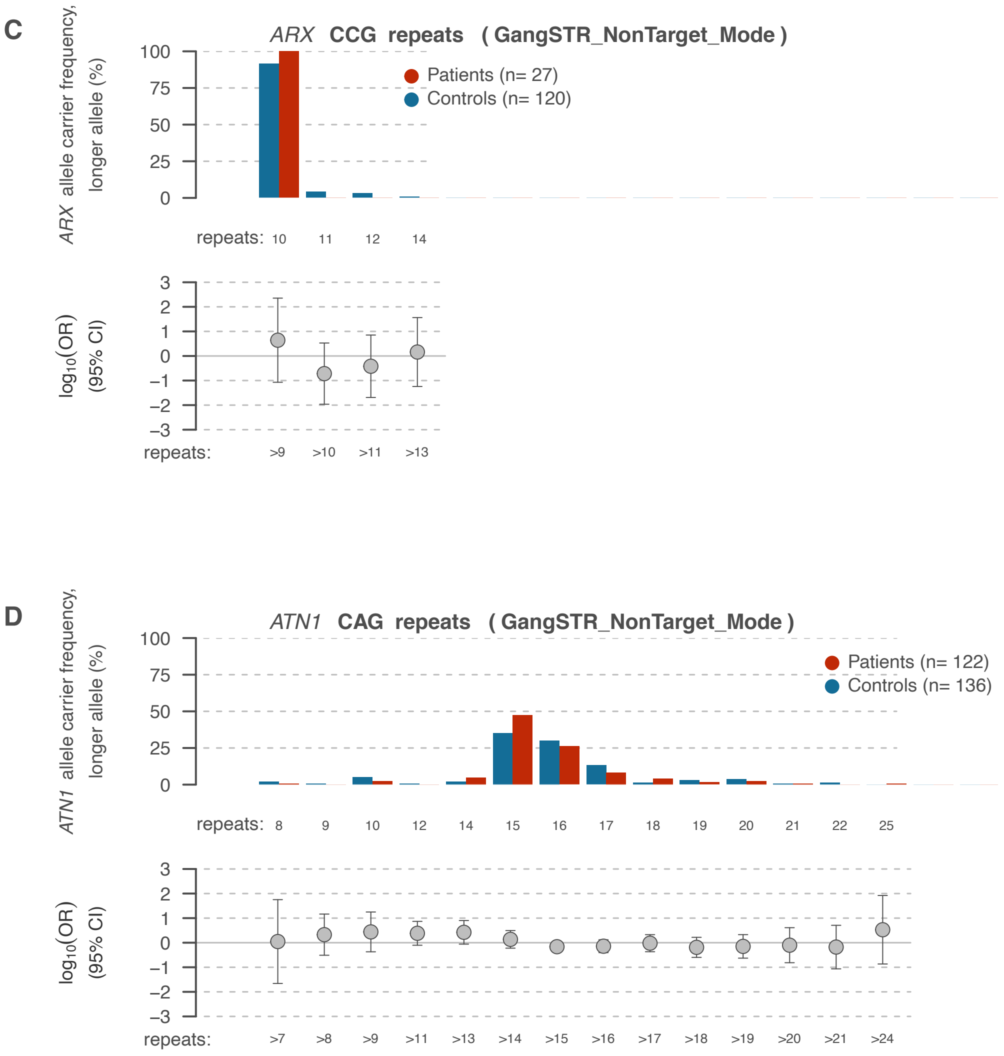
RFC1 AAGGG repeats (ExpansionHunter_v3)
RFC1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 39)
Controls (n= 135)
repeats: 9 10 11 12 15 20 27 28 29 31 32 33 34 35 36 38 39 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 64 65 66 67 68 69 71 72 73 74 77 78 80 81 83 85 91 96
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >9 >10 >11 >14 >19 >26 >27 >28 >30 >31 >32 >33 >34 >35 >37 >38 >41 >42 >43 >44 >45 >46 >47 >48 >49 >50 >51 >52 >53 >54 >55 >56 >57 >58 >59 >60 >61 >63 >64 >65 >66 >67 >68 >70 >71 >72 >73 >76 >77 >79 >80 >82 >84 >90 >95

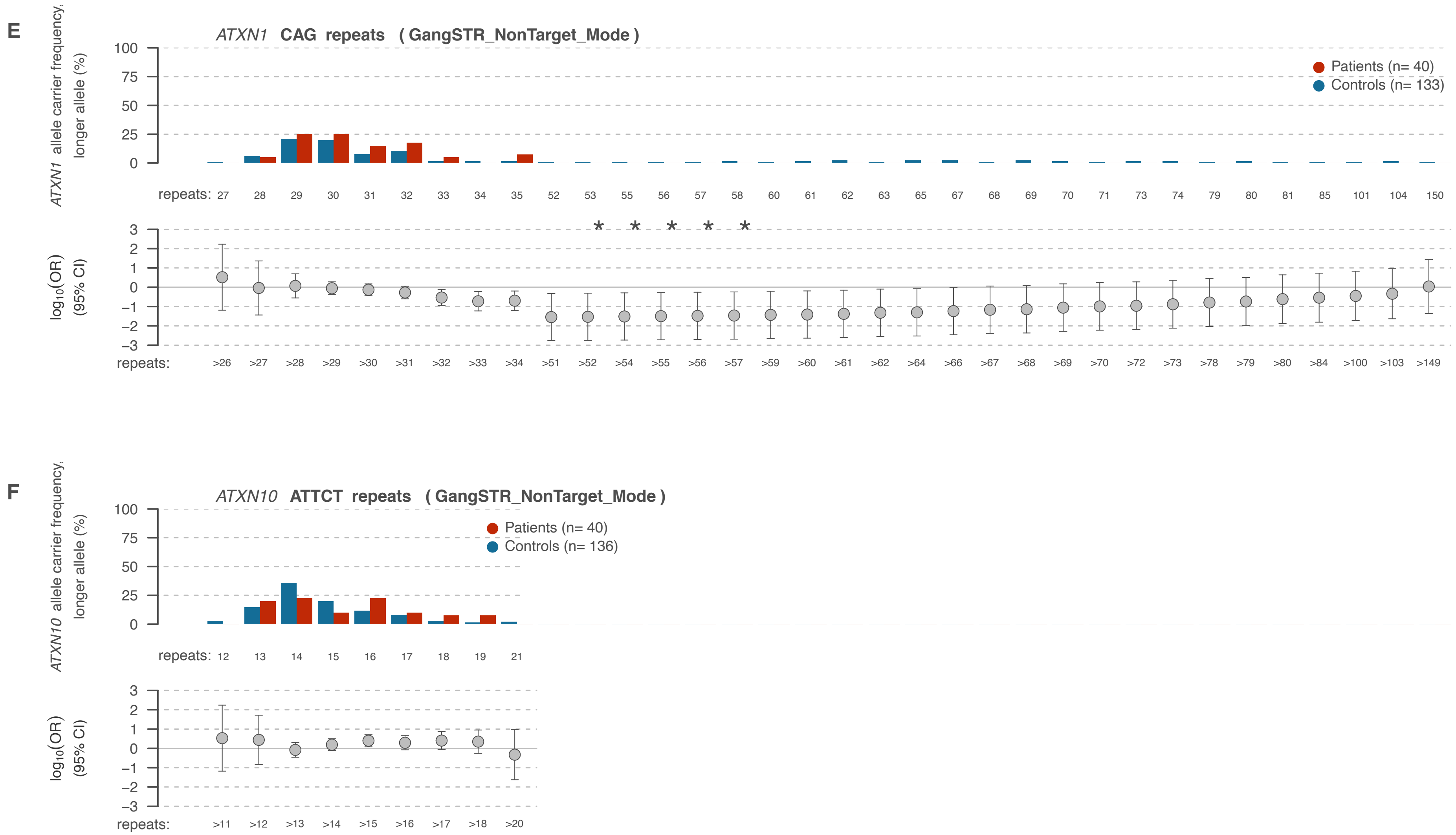


Supplementary Figure 3.14: ExpansionHunter3 prediction of STR lengths in epilepsy patients

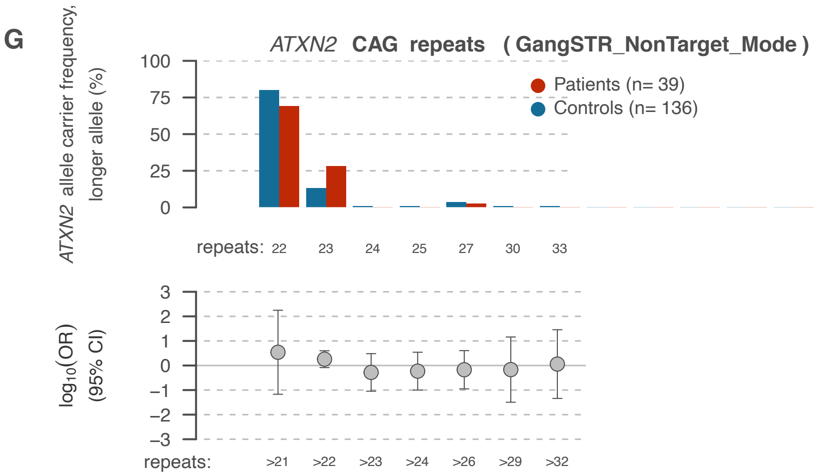
For each gene genotyped with ExpansionHunter3 the allele lengths in epilepsy patients are compared to 136 Irish controls. The upper plot shows the predicted allele lengths and the lower plot shows the OR. An asterisks indicate a significant OR. The epilepsy results include PCR-free WGS samples, PCR WGS samples and WES sample if an RMSD below one was observed when comparing WES results to WGS results for a given gene.

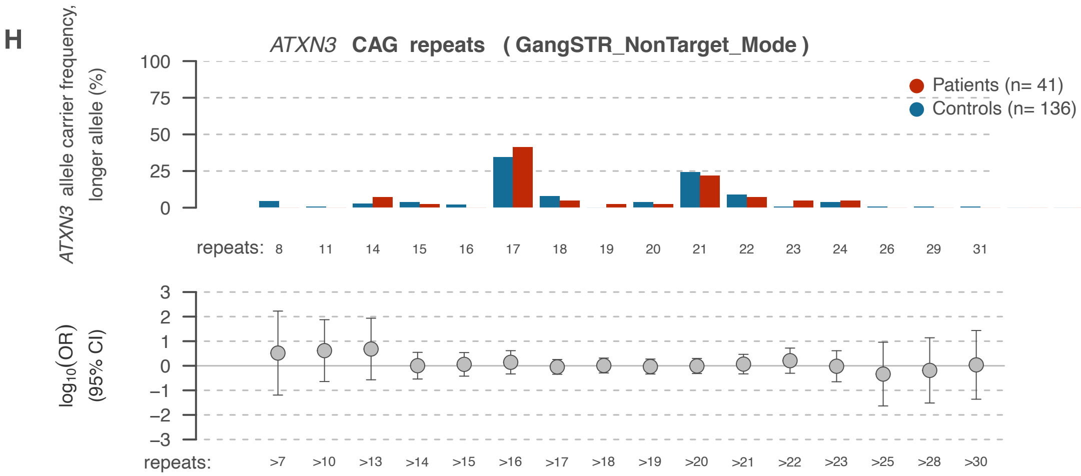

A
AFF2 GCC repeats (GangSTR_NonTarget_Mode)
AFF2 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 24)
● Controls (n= 118)
repeats: 6 12 13 14 15 16 17 18 19 20 21 22 23 24 25 29 32
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >5 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >28 >31
B
AR CAG repeats (GangSTR_NonTarget_Mode)
AR allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 19)
● Controls (n= 89)
repeats: 15 16 18 19 20 21 22 23 24 25 26 27 28 29 30
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >14 >15 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >29

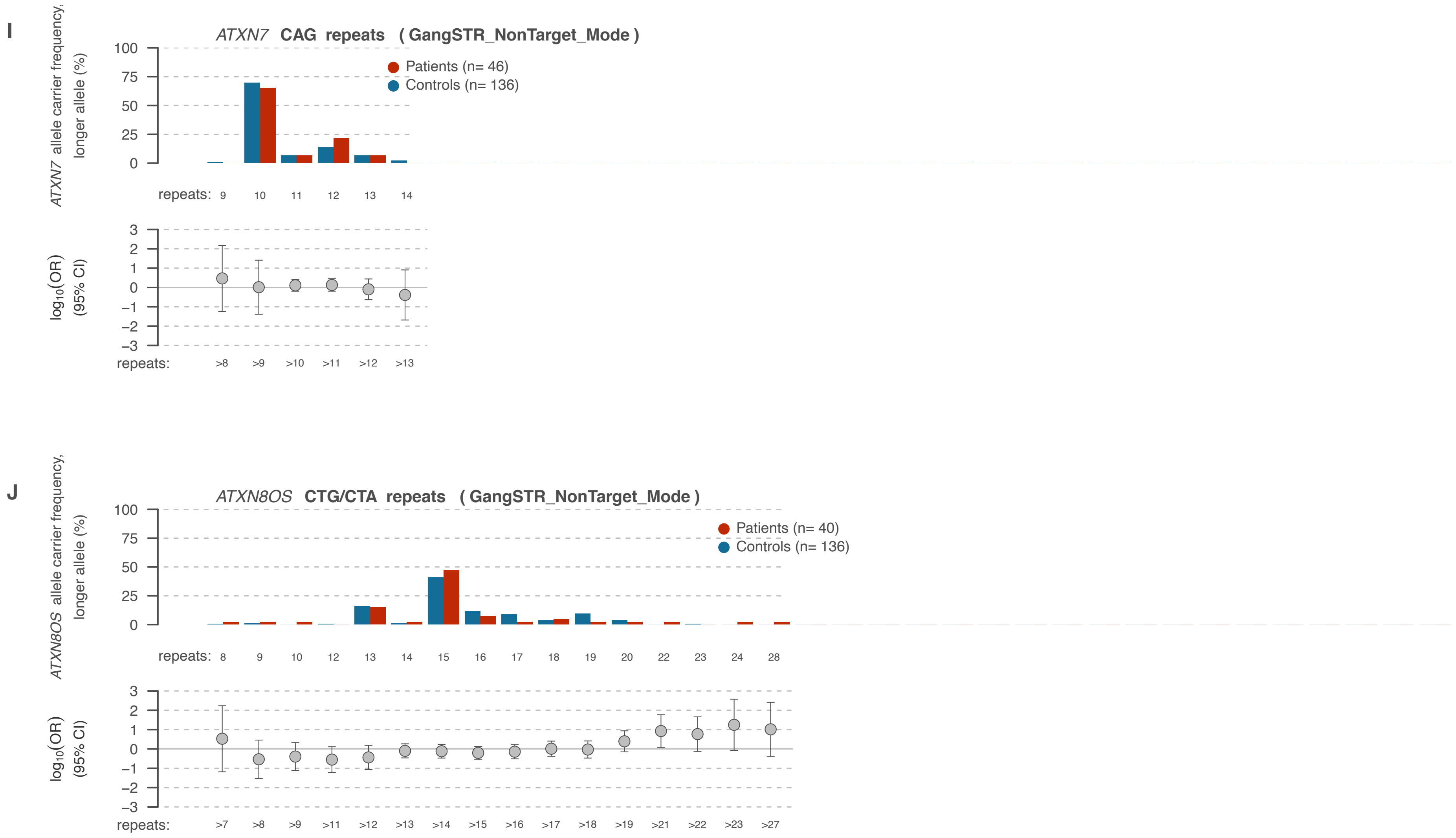

C
ARX CCG repeats (GangSTR_NonTarget_Mode)
Patients (n= 27)
Controls (n= 120)
ARX allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 10 11 12 14
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >9 >10 >11 >13
D
ATN1 CAG repeats (GangSTR_NonTarget_Mode)
Patients (n= 122)
Controls (n= 136)
ATN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 8 9 10 12 14 15 16 17 18 19 20 21 22 25
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >9 >11 >13 >14 >15 >16 >17 >18 >19 >20 >21 >24

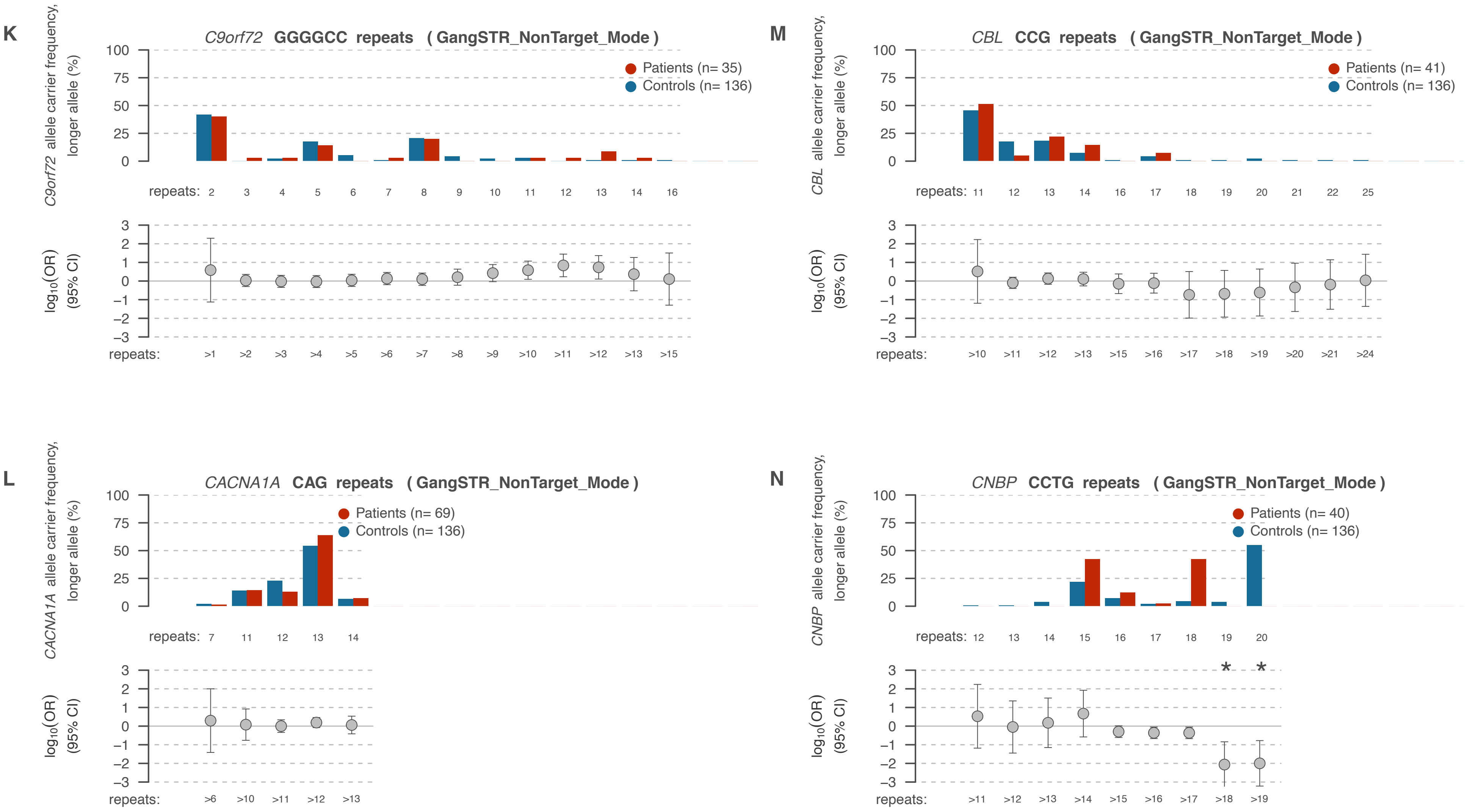


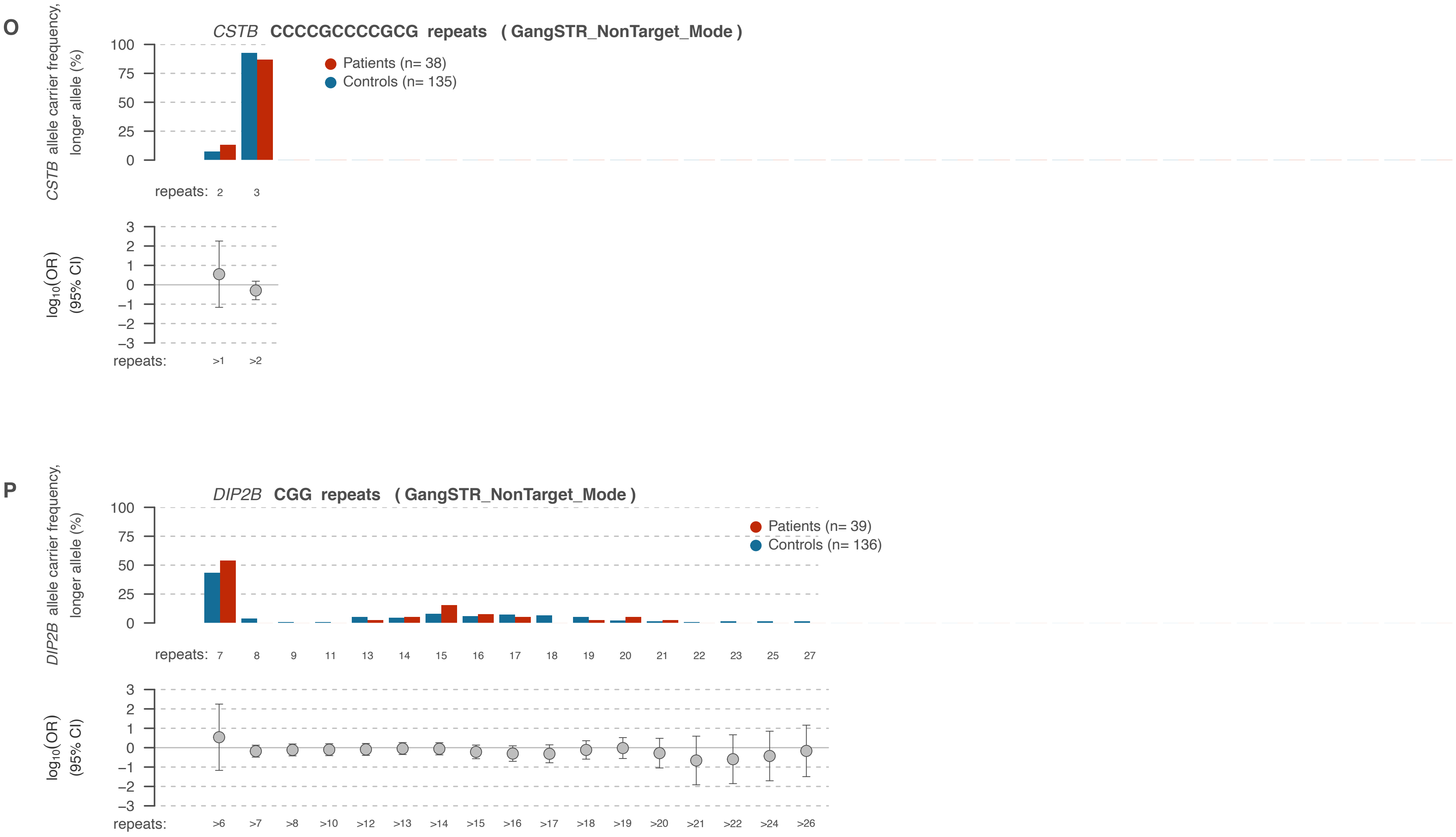
E
ATXN1 CAG repeats (GangSTR_NonTarget_Mode)
ATXN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 40)
Controls (n= 133)
repeats: 27 28 29 30 31 32 33 34 35 52 53 55 56 57 58 60 61 62 63 65 67 68 69 70 71 73 74 79 80 81 85 101 104 150
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
* * * * *
repeats: >26 >27 >28 >29 >30 >31 >32 >33 >34 >51 >52 >54 >55 >56 >57 >59 >60 >61 >62 >64 >66 >67 >68 >69 >70 >72 >73 >78 >79 >80 >84 >100 >103 >149
F
ATXN10 ATTCT repeats (GangSTR_NonTarget_Mode)
ATXN10 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 40)
Controls (n= 136)
repeats: 12 13 14 15 16 17 18 19 21
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >11 >12 >13 >14 >15 >16 >17 >18 >20


G
ATXN2 CAG repeats (GangSTR_NonTarget_Mode)
ATXN2 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 39)
● Controls (n= 136)
repeats: 22 23 24 25 27 30 33
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >21 >22 >23 >24 >26 >29 >32

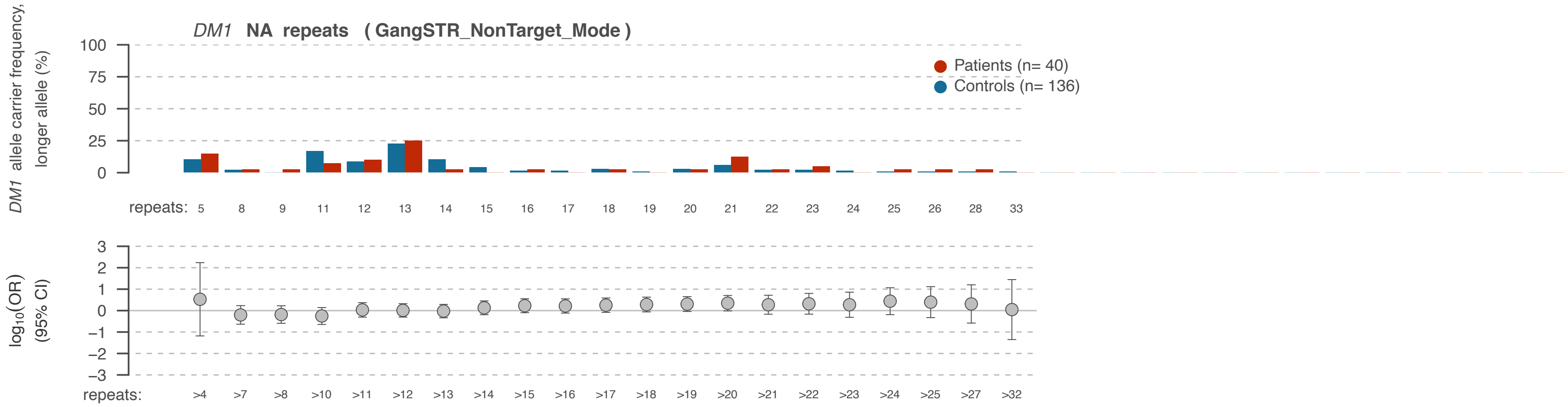

H
ATXN3 CAG repeats (GangSTR_NonTarget_Mode)
ATXN3 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 136)
repeats: 8 11 14 15 16 17 18 19 20 21 22 23 24 26 29 31
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >10 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >25 >28 >30


I
ATXN7 CAG repeats (GangSTR_NonTarget_Mode)
Patients (n= 46)
Controls (n= 136)
ATXN7 allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 9 10 11 12 13 14
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >9 >10 >11 >12 >13
J
ATXN8OS CTG/CTA repeats (GangSTR_NonTarget_Mode)
Patients (n= 40)
Controls (n= 136)
ATXN8OS allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 8 9 10 12 13 14 15 16 17 18 19 20 22 23 24 28
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >9 >11 >12 >13 >14 >15 >16 >17 >18 >19 >21 >22 >23 >27

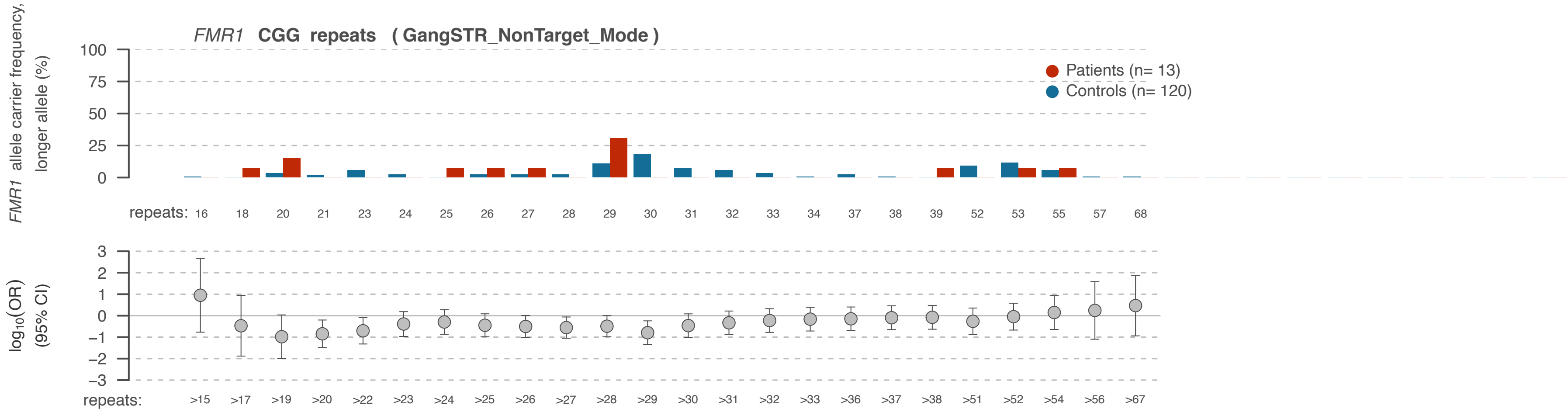

K
C9orf72 GGGGCC repeats (GangSTR_NonTarget_Mode)
C9orf72 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 35)
Controls (n= 136)
repeats: 2 3 4 5 6 7 8 9 10 11 12 13 14 16
$\log_{10}$(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >1 >2 >3 >4 >5 >6 >7 >8 >9 >10 >11 >12 >13 >15
M
CBL CCG repeats (GangSTR_NonTarget_Mode)
CBL allele carrier frequency, longer allele (%)
Patients (n= 41)
Controls (n= 136)
repeats: 11 12 13 14 16 17 18 19 20 21 22 25
repeats: >10 >11 >12 >13 >15 >16 >17 >18 >19 >20 >21 >24
L
CACNA1A CAG repeats (GangSTR_NonTarget_Mode)
CACNA1A allele carrier frequency, longer allele (%)
Patients (n= 69)
Controls (n= 136)
repeats: 7 11 12 13 14
repeats: >6 >10 >11 >12 >13
N
CNBP CCTG repeats (GangSTR_NonTarget_Mode)
CNBP allele carrier frequency, longer allele (%)
Patients (n= 40)
Controls (n= 136)
repeats: 12 13 14 15 16 17 18 19 20
*
*
repeats: >11 >12 >13 >14 >15 >16 >17 >18 >19

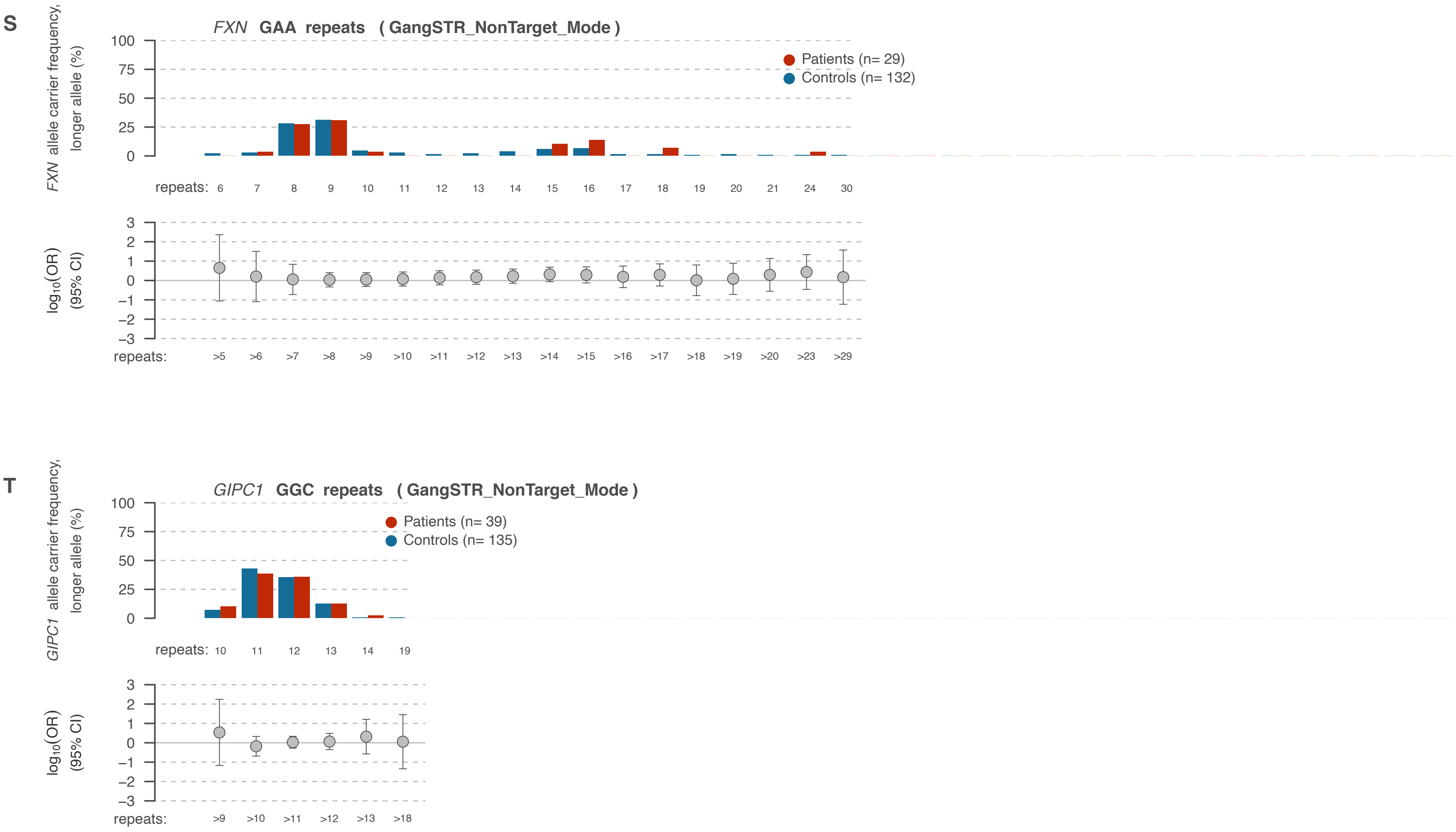

O
CSTB CCCCGCCCCGCG repeats (GangSTR_NonTarget_Mode)
CSTB allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 38)
● Controls (n= 135)
repeats: 2 3
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >1 >2
P
DIP2B CGG repeats (GangSTR_NonTarget_Mode)
DIP2B allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 39)
● Controls (n= 136)
repeats: 7 8 9 11 13 14 15 16 17 18 19 20 21 22 23 25 27
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >7 >8 >10 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >24 >26

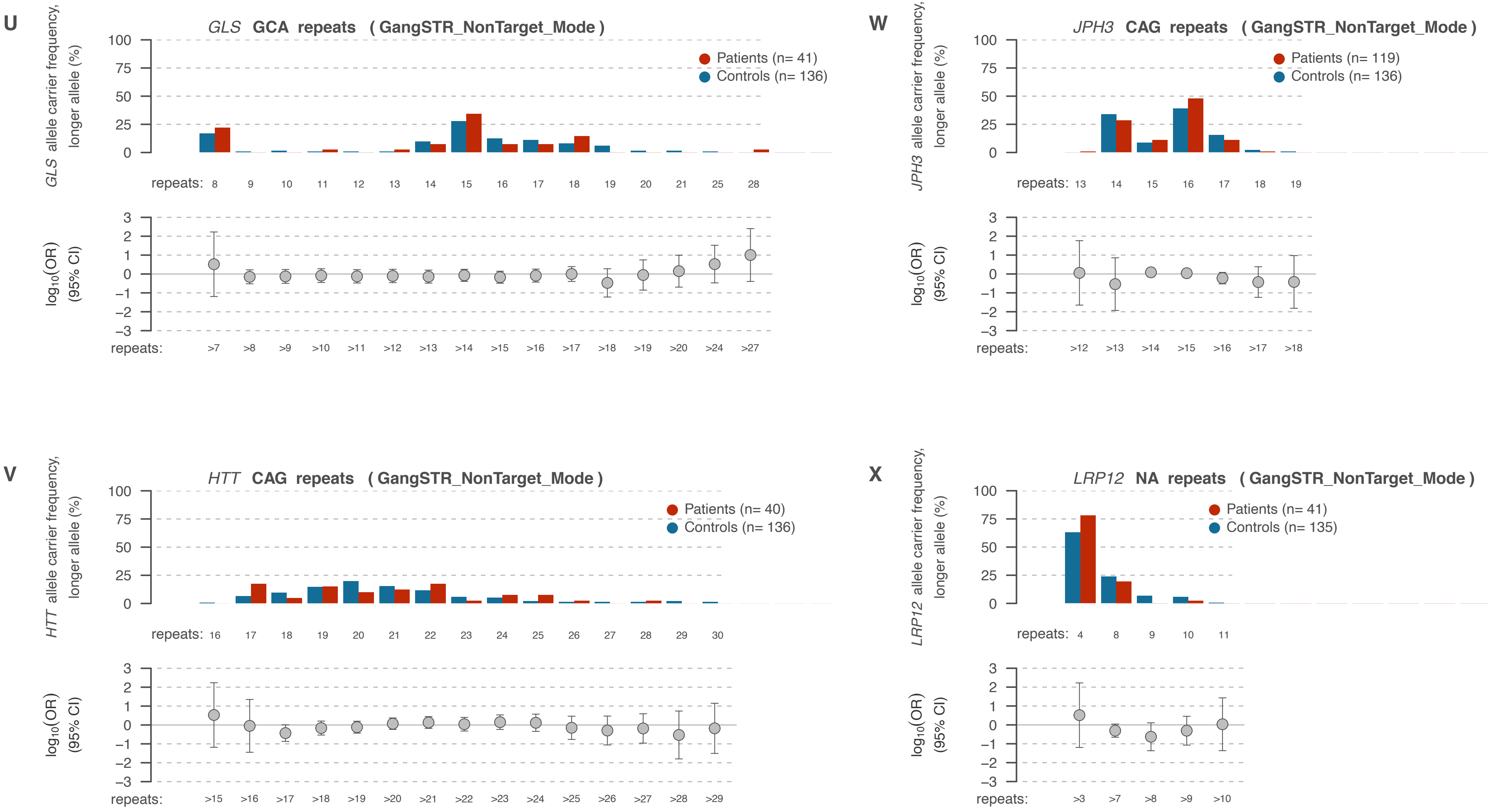
Q


DM1 NA repeats (GangSTR_NonTarget_Mode)
DM1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 40)
Controls (n= 136)
repeats: 5 8 9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 33
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >4 >7 >8 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >27 >32

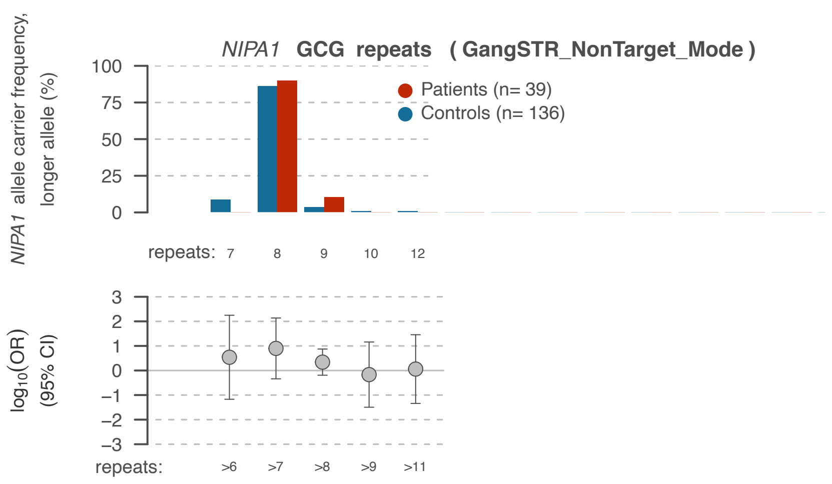
R


FMR1 CGG repeats (GangSTR_NonTarget_Mode)
FMR1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 13)
Controls (n= 120)
repeats: 16 18 20 21 23 24 25 26 27 28 29 30 31 32 33 34 37 38 39 52 53 55 57 68
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >15 >17 >19 >20 >22 >23 >24 >25 >26 >27 >28 >29 >30 >31 >32 >33 >36 >37 >38 >51 >52 >54 >56 >67


S
FXN GAA repeats (GangSTR_NonTarget_Mode)
FXN allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 29)
Controls (n= 132)
repeats: 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 24 30
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >5 >6 >7 >8 >9 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >23 >29
T
GIPC1 GGC repeats (GangSTR_NonTarget_Mode)
GIPC1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 39)
Controls (n= 135)
repeats: 10 11 12 13 14 19
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >9 >10 >11 >12 >13 >18

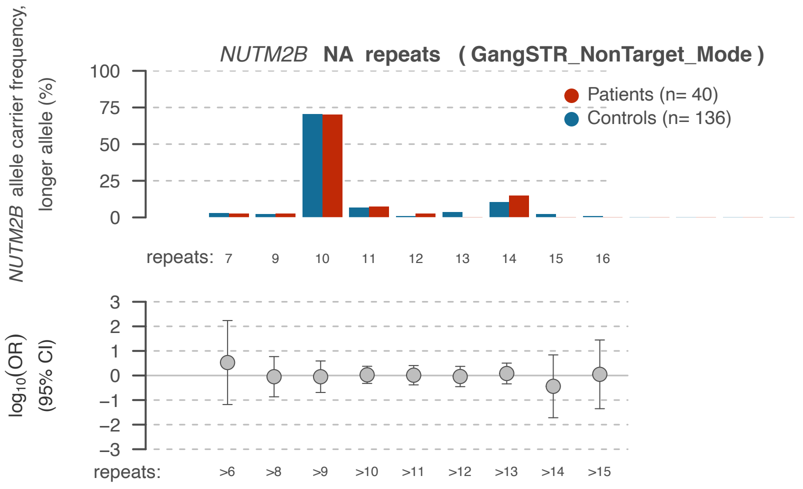

U
GLS GCA repeats (GangSTR_NonTarget_Mode)
GLS allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 8 9 10 11 12 13 14 15 16 17 18 19 20 21 25 28
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >9 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >24 >27
W
JPH3 CAG repeats (GangSTR_NonTarget_Mode)
JPH3 allele carrier frequency, longer allele (%)
Patients (n= 119)
Controls (n= 136)
repeats: 13 14 15 16 17 18 19
log10(OR) (95% CI)
repeats: >12 >13 >14 >15 >16 >17 >18
V
HTT CAG repeats (GangSTR_NonTarget_Mode)
HTT allele carrier frequency, longer allele (%)
Patients (n= 40)
Controls (n= 136)
repeats: 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
log10(OR) (95% CI)
repeats: >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >29
X
LRP12 NA repeats (GangSTR_NonTarget_Mode)
LRP12 allele carrier frequency, longer allele (%)
Patients (n= 41)
Controls (n= 135)
repeats: 4 8 9 10 11
log10(OR) (95% CI)
repeats: >3 >7 >8 >9 >10

Y



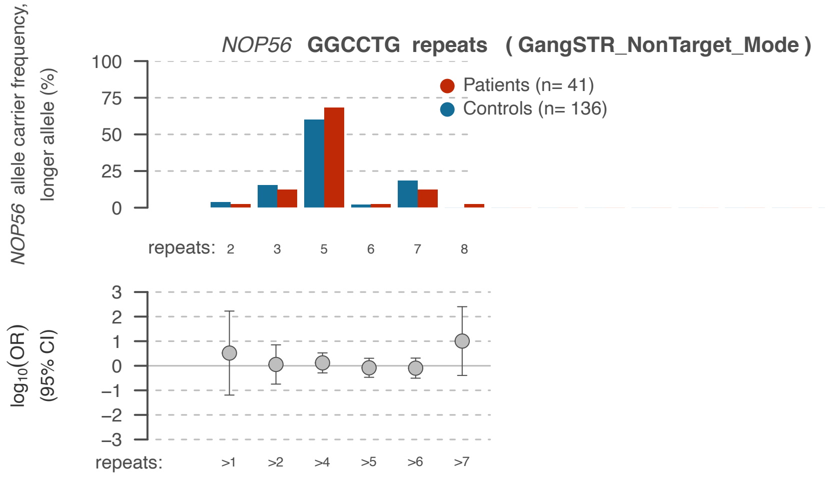
NIPA1 GCG repeats (GangSTR_NonTarget_Mode)
NIPA1 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 39)
● Controls (n= 136)
repeats: 7 8 9 10 12
log$_{10}$(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >7 >8 >9 >11

A.2



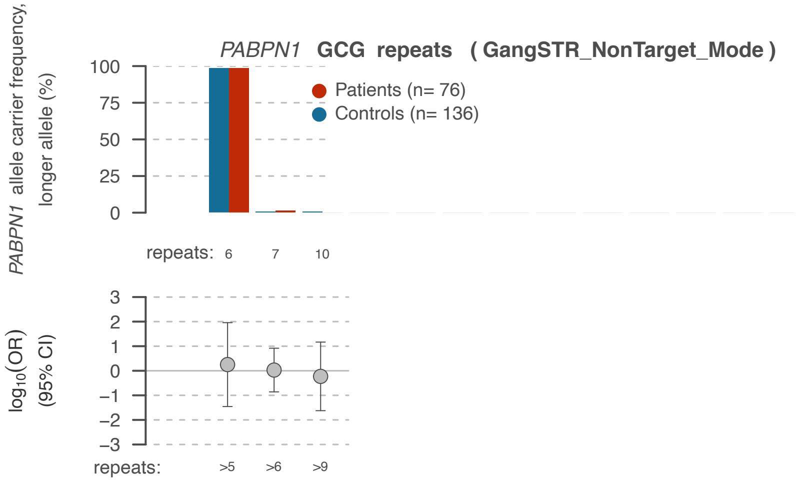
NUTM2B NA repeats (GangSTR_NonTarget_Mode)
NUTM2B allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 40)
● Controls (n= 136)
repeats: 7 9 10 11 12 13 14 15 16
log$_{10}$(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >8 >9 >10 >11 >12 >13 >14 >15

Z



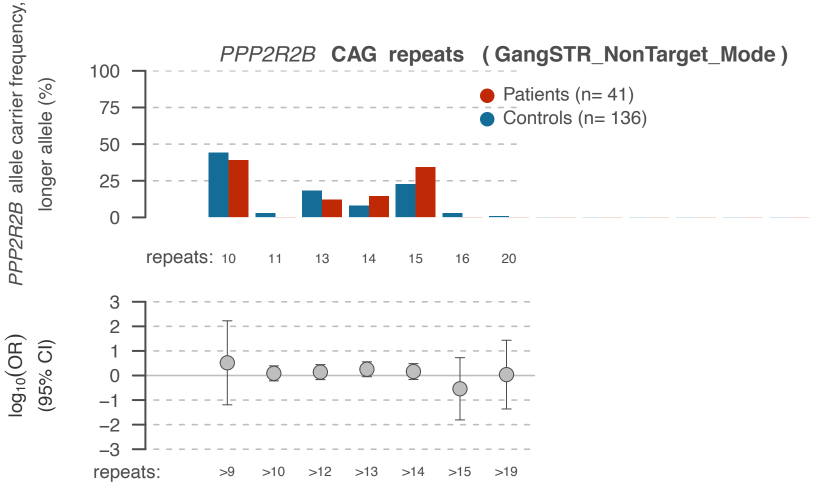
NOP56 GGCCTG repeats (GangSTR_NonTarget_Mode)
NOP56 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 136)
repeats: 2 3 5 6 7 8
log$_{10}$(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >1 >2 >4 >5 >6 >7

B.2

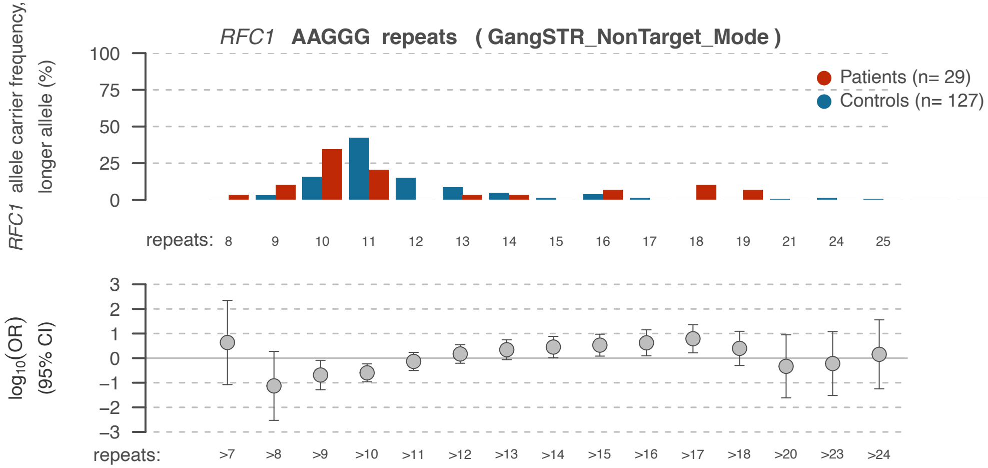


PABPN1 GCG repeats (GangSTR_NonTarget_Mode)
PABPN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 76)
● Controls (n= 136)
repeats: 6 7 10
log$_{10}$(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >5 >6 >9

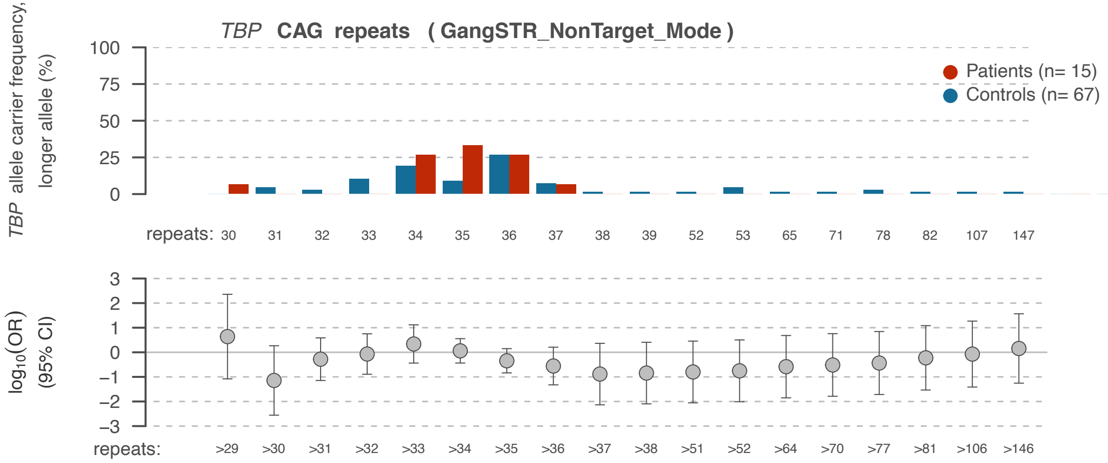
C.2


PPP2R2B CAG repeats (GangSTR_NonTarget_Mode)
PPP2R2B allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 10 11 13 14 15 16 20
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >9 >10 >12 >13 >14 >15 >19

D.2

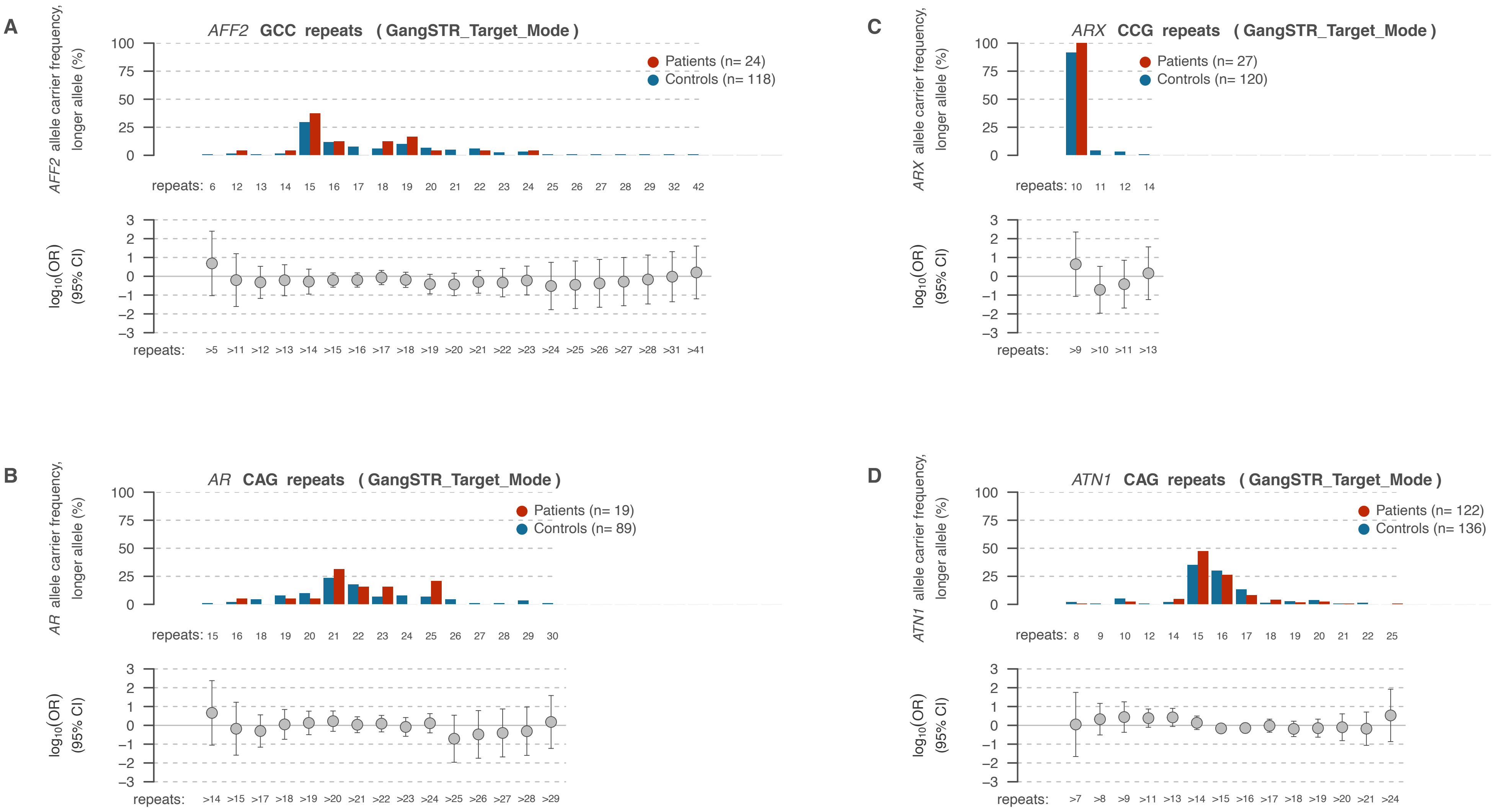

RFC1 AAGGG repeats (GangSTR_NonTarget_Mode)
RFC1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 29)
Controls (n= 127)
repeats: 8 9 10 11 12 13 14 15 16 17 18 19 21 24 25
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >9 >10 >11 >12 >13 >14 >15 >16 >17 >18 >20 >23 >24

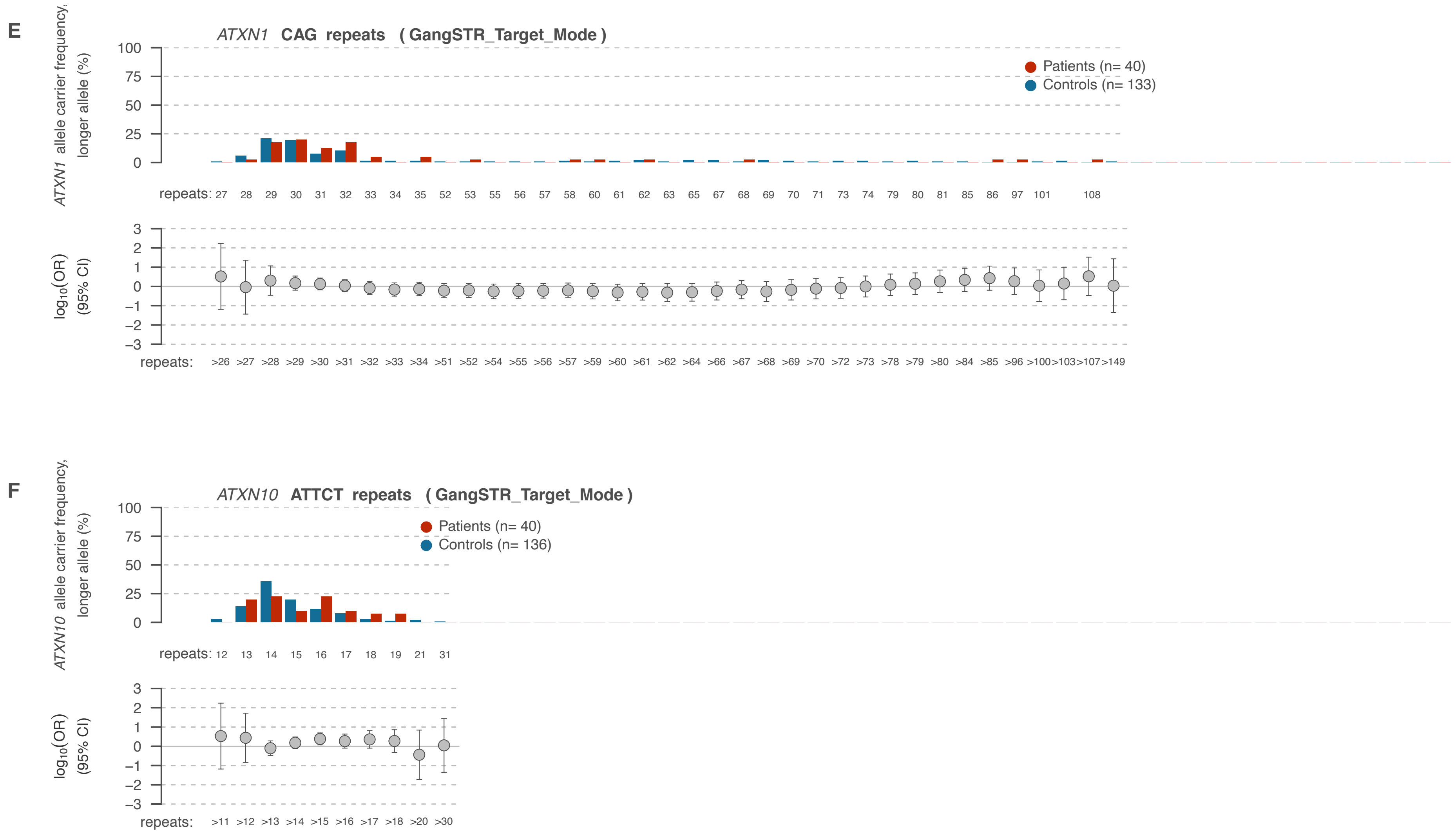
E.2

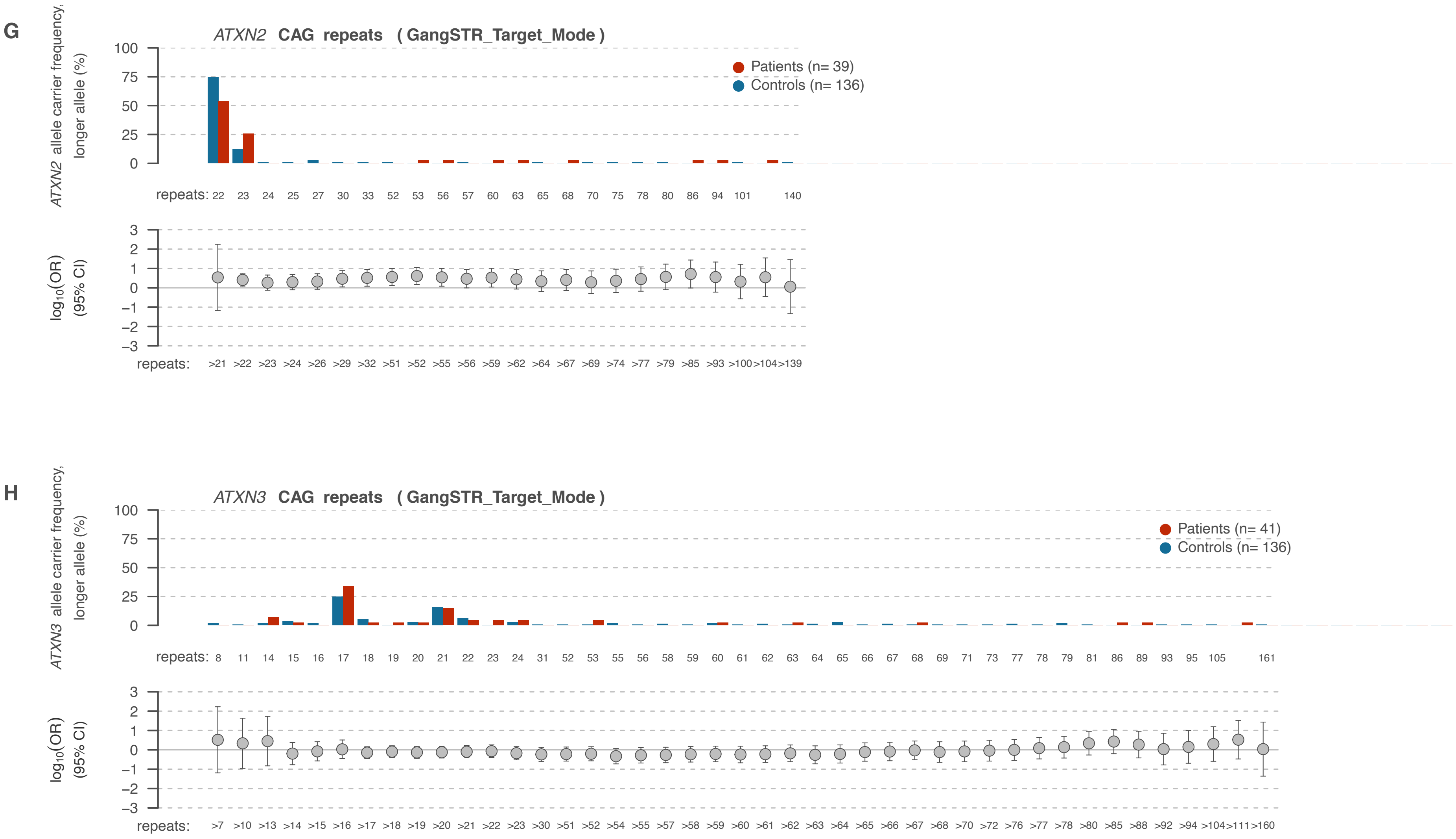


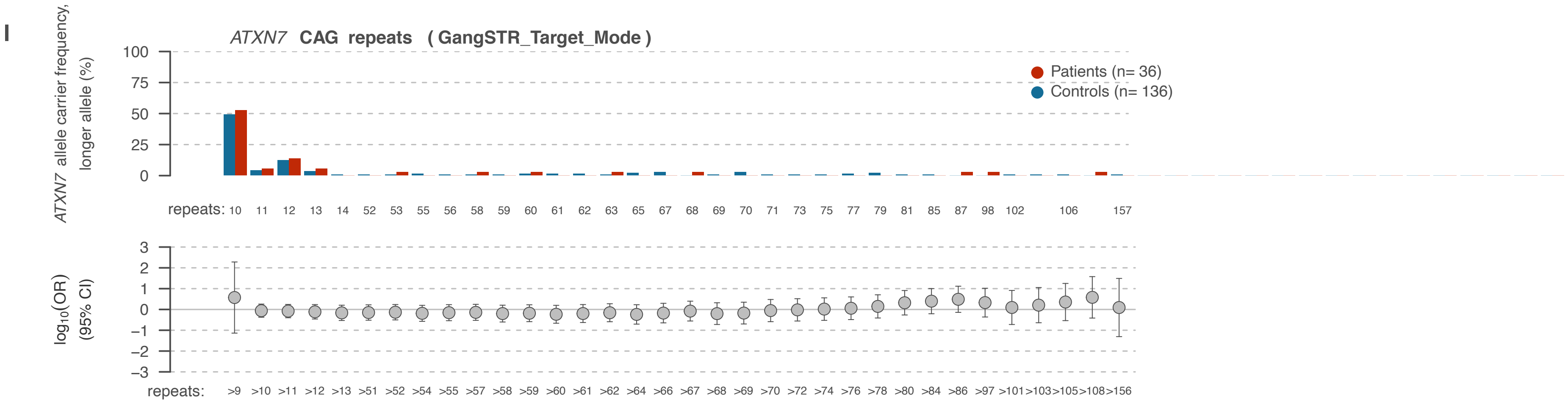
Supplementary Figure 3.15: GangSTR (genome-wide mode) prediction of STR lengths in epilepsy patients

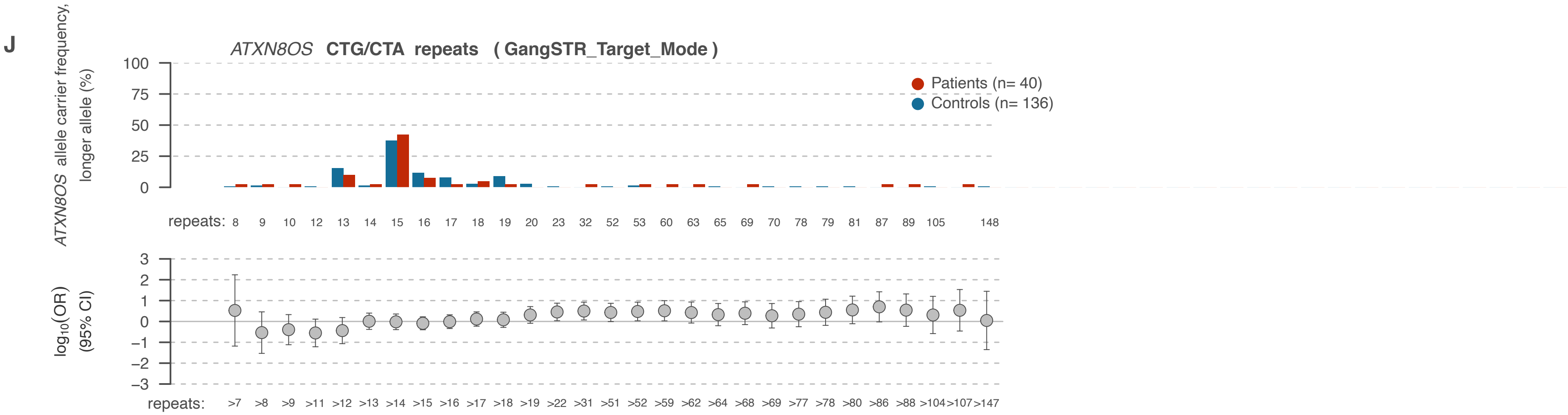
For each gene genotyped with GangSTR (genome-wide mode) the allele lengths in epilepsy patients are compared to 136 Irish controls. The upper plot shows the predicted allele lengths and the lower plot shows the OR. An asterisks indicate a significant OR. The epilepsy results include PCR-free WGS samples, PCR WGS samples and WES sample if an RMSD below one was observed when comparing WES results to WGS results for a given gene.

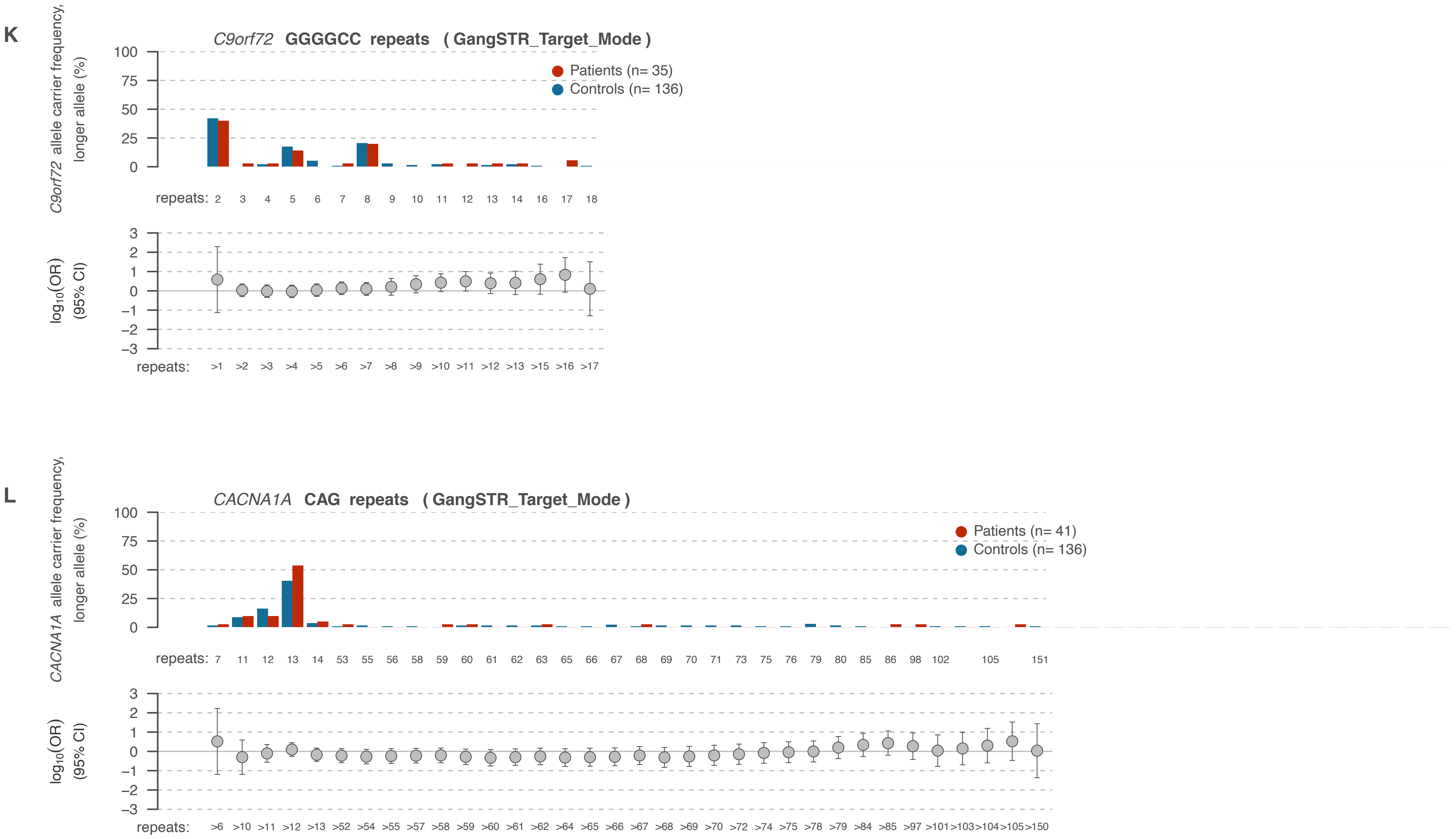

A
AFF2 GCC repeats (GangSTR_Target_Mode)
AFF2 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 24)
Controls (n= 118)
repeats: 6 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 32 42
log$_{10}$(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >5 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >31 >41
B
AR CAG repeats (GangSTR_Target_Mode)
AR allele carrier frequency, longer allele (%)
Patients (n= 19)
Controls (n= 89)
repeats: 15 16 18 19 20 21 22 23 24 25 26 27 28 29 30
repeats: >14 >15 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >29
C
ARX CCG repeats (GangSTR_Target_Mode)
ARX allele carrier frequency, longer allele (%)
Patients (n= 27)
Controls (n= 120)
repeats: 10 11 12 14
repeats: >9 >10 >11 >13
D
ATN1 CAG repeats (GangSTR_Target_Mode)
ATN1 allele carrier frequency, longer allele (%)
Patients (n= 122)
Controls (n= 136)
repeats: 8 9 10 12 14 15 16 17 18 19 20 21 22 25
repeats: >7 >8 >9 >11 >13 >14 >15 >16 >17 >18 >19 >20 >21 >24

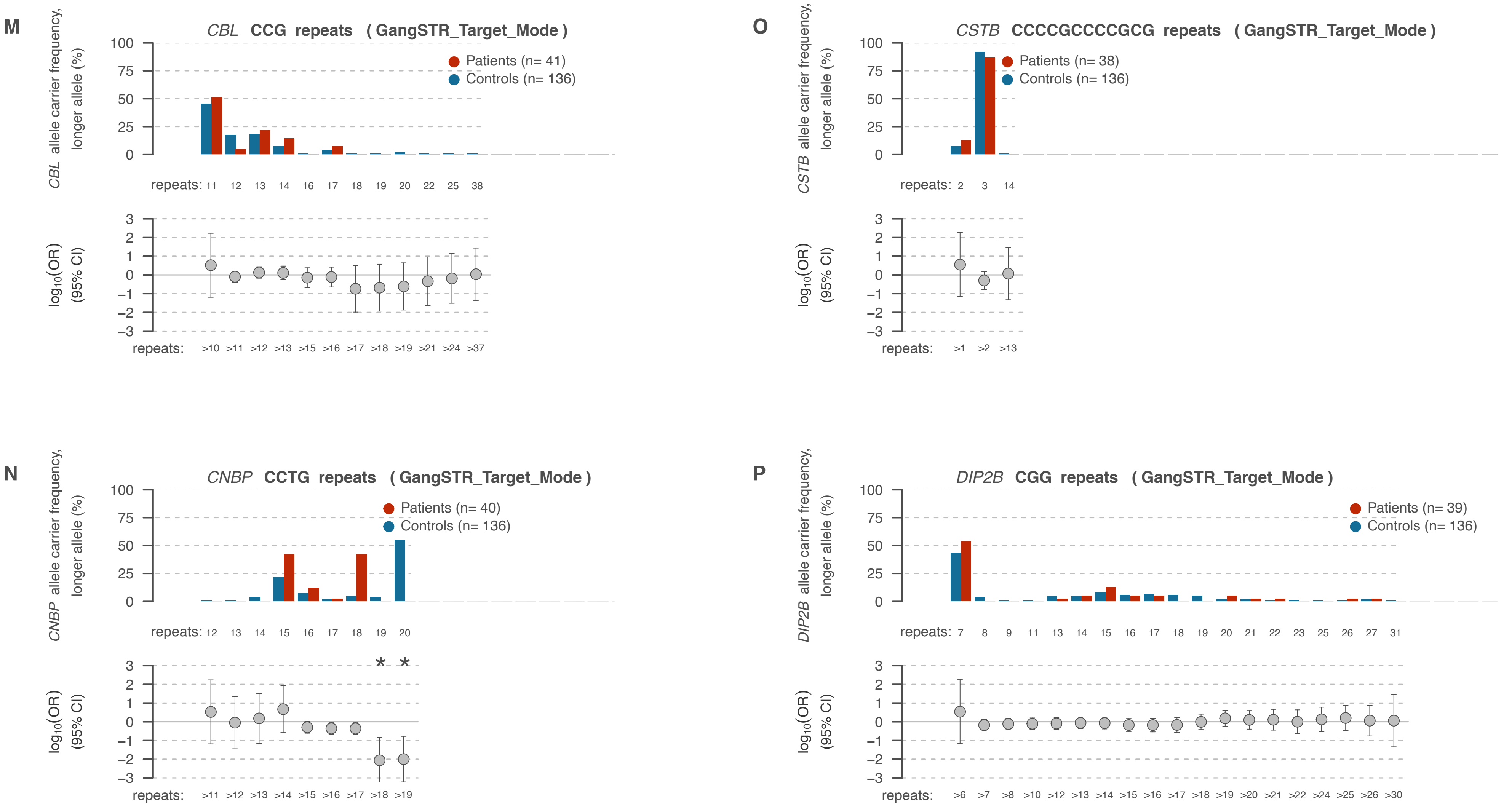

E
ATXN1 CAG repeats (GangSTR_Target_Mode)
ATXN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 40)
● Controls (n= 133)
repeats: 27 28 29 30 31 32 33 34 35 52 53 55 56 57 58 60 61 62 63 65 67 68 69 70 71 73 74 79 80 81 85 86 97 101 108
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >26 >27 >28 >29 >30 >31 >32 >33 >34 >51 >52 >54 >55 >56 >57 >59 >60 >61 >62 >64 >66 >67 >68 >69 >70 >72 >73 >78 >79 >80 >84 >85 >96 >100 >103 >107 >149
F
ATXN10 ATTCT repeats (GangSTR_Target_Mode)
ATXN10 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 40)
● Controls (n= 136)
repeats: 12 13 14 15 16 17 18 19 21 31
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >11 >12 >13 >14 >15 >16 >17 >18 >20 >30


G
ATXN2 CAG repeats (GangSTR_Target_Mode)
ATXN2 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 39)
Controls (n= 136)
repeats: 22 23 24 25 27 30 33 52 53 56 57 60 63 65 68 70 75 78 80 86 94 101 140
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >21 >22 >23 >24 >26 >29 >32 >51 >52 >55 >56 >59 >62 >64 >67 >69 >74 >77 >79 >85 >93 >100 >104 >139
H
ATXN3 CAG repeats (GangSTR_Target_Mode)
ATXN3 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 8 11 14 15 16 17 18 19 20 21 22 23 24 31 52 53 55 56 58 59 60 61 62 63 64 65 66 67 68 69 71 73 77 78 79 81 86 89 93 95 105 161
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >10 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >30 >51 >52 >54 >55 >57 >58 >59 >60 >61 >62 >63 >64 >65 >66 >67 >68 >70 >72 >76 >77 >78 >80 >85 >88 >92 >94 >104 >111 >160

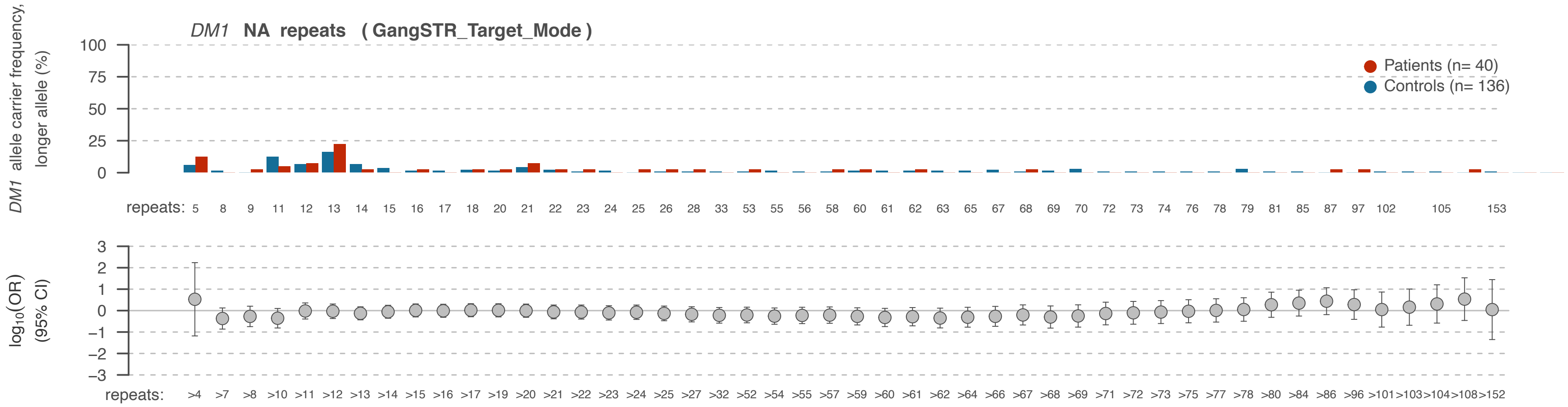

I
ATXN7 CAG repeats (GangSTR_Target_Mode)
ATXN7 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 36)
Controls (n= 136)
repeats: 10 11 12 13 14 52 53 55 56 58 59 60 61 62 63 65 67 68 69 70 71 73 75 77 79 81 85 87 98 102 106 157
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >9 >10 >11 >12 >13 >51 >52 >54 >55 >57 >58 >59 >60 >61 >62 >64 >66 >67 >68 >69 >70 >72 >74 >76 >78 >80 >84 >86 >97 >101 >103 >105 >108 >156


J
ATXN8OS CTG/CTA repeats (GangSTR_Target_Mode)
ATXN8OS allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 40)
Controls (n= 136)
repeats: 8 9 10 12 13 14 15 16 17 18 19 20 23 32 52 53 60 63 65 69 70 78 79 81 87 89 105 148
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >9 >11 >12 >13 >14 >15 >16 >17 >18 >19 >22 >31 >51 >52 >59 >62 >64 >68 >69 >77 >78 >80 >86 >88 >104 >107 >147

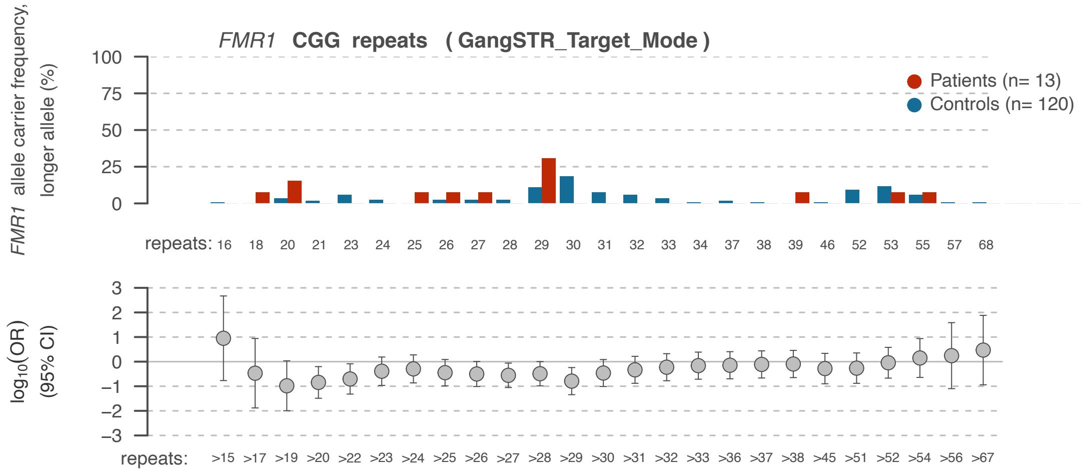

K
C9orf72 GGGGCC repeats (GangSTR_Target_Mode)
C9orf72 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 35)
● Controls (n= 136)
repeats: 2 3 4 5 6 7 8 9 10 11 12 13 14 16 17 18
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >1 >2 >3 >4 >5 >6 >7 >8 >9 >10 >11 >12 >13 >15 >16 >17
L
CACNA1A CAG repeats (GangSTR_Target_Mode)
CACNA1A allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 136)
repeats: 7 11 12 13 14 53 55 56 58 59 60 61 62 63 65 66 67 68 69 70 71 73 75 76 79 80 85 86 98 102 105 151
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >10 >11 >12 >13 >52 >54 >55 >57 >58 >59 >60 >61 >62 >64 >65 >66 >67 >68 >69 >70 >72 >74 >75 >78 >79 >84 >85 >97 >101 >103 >104 >105 >150


M
CBL CCG repeats (GangSTR_Target_Mode)
CBL allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 11 12 13 14 16 17 18 19 20 22 25 38
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >10 >11 >12 >13 >15 >16 >17 >18 >19 >21 >24 >37
O
CSTB CCCCGCCCCGCG repeats (GangSTR_Target_Mode)
CSTB allele carrier frequency, longer allele (%)
Patients (n= 38)
Controls (n= 136)
repeats: 2 3 14
log10(OR) (95% CI)
repeats: >1 >2 >13
N
CNBP CCTG repeats (GangSTR_Target_Mode)
CNBP allele carrier frequency, longer allele (%)
Patients (n= 40)
Controls (n= 136)
repeats: 12 13 14 15 16 17 18 19 20
*
*
log10(OR) (95% CI)
repeats: >11 >12 >13 >14 >15 >16 >17 >18 >19
P
DIP2B CGG repeats (GangSTR_Target_Mode)
DIP2B allele carrier frequency, longer allele (%)
Patients (n= 39)
Controls (n= 136)
repeats: 7 8 9 11 13 14 15 16 17 18 19 20 21 22 23 25 26 27 31
log10(OR) (95% CI)
repeats: >6 >7 >8 >10 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >24 >25 >26 >30

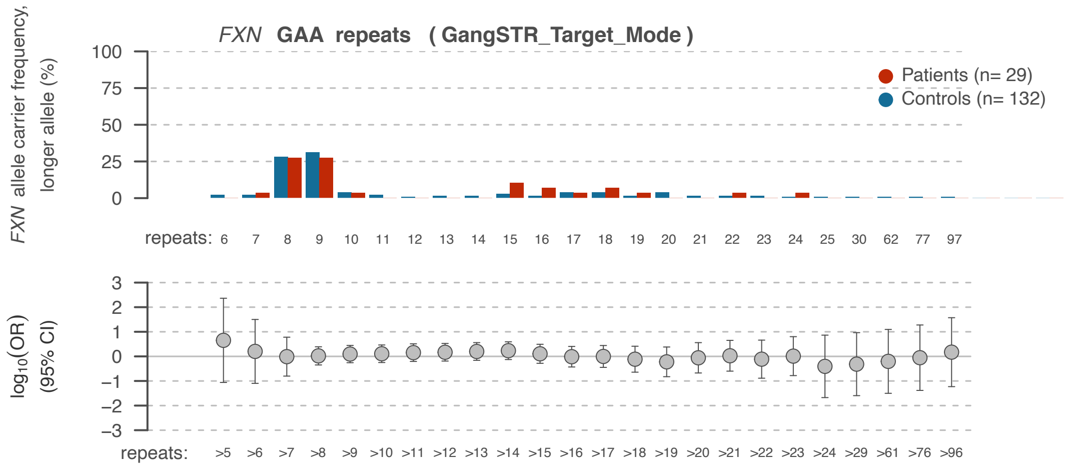
Q


DM1 NA repeats (GangSTR_Target_Mode)
DM1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 40)
Controls (n= 136)
repeats: 5 8 9 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 28 33 53 55 56 58 60 61 62 63 65 67 68 69 70 72 73 74 76 78 79 81 85 87 97 102 105 153
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >4 >7 >8 >10 >11 >12 >13 >14 >15 >16 >17 >19 >20 >21 >22 >23 >24 >25 >27 >32 >52 >54 >55 >57 >59 >60 >61 >62 >64 >66 >67 >68 >69 >71 >72 >73 >75 >77 >78 >80 >84 >86 >96 >101 >103 >104 >108 >152

R

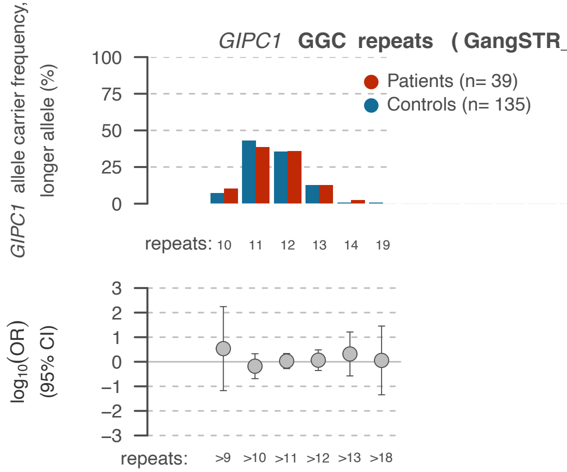

FMR1 CGG repeats (GangSTR_Target_Mode)
FMR1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 13)
Controls (n= 120)
repeats: 16 18 20 21 23 24 25 26 27 28 29 30 31 32 33 34 37 38 39 46 52 53 55 57 68
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >15 >17 >19 >20 >22 >23 >24 >25 >26 >27 >28 >29 >30 >31 >32 >33 >36 >37 >38 >45 >51 >52 >54 >56 >67

S

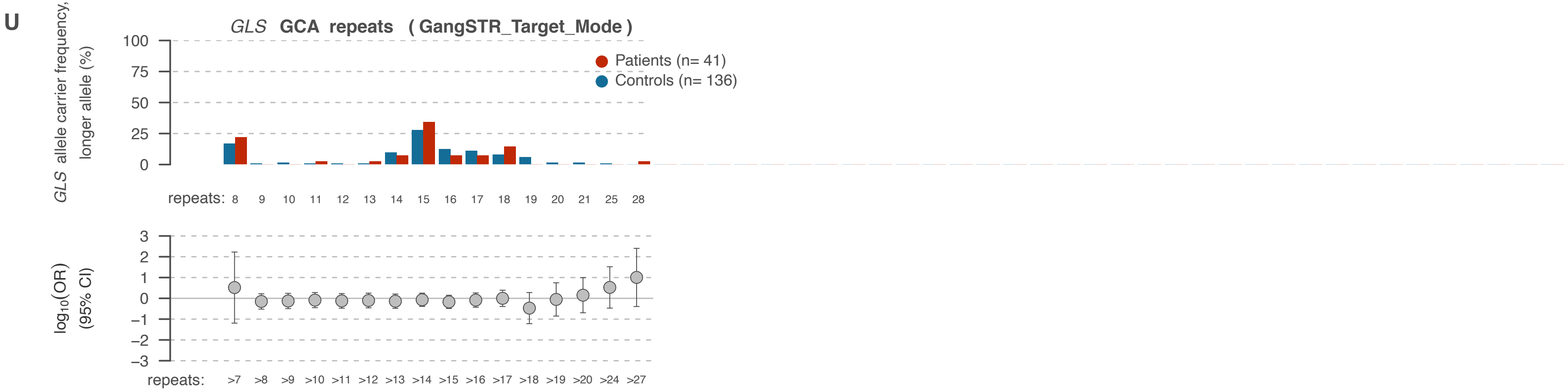


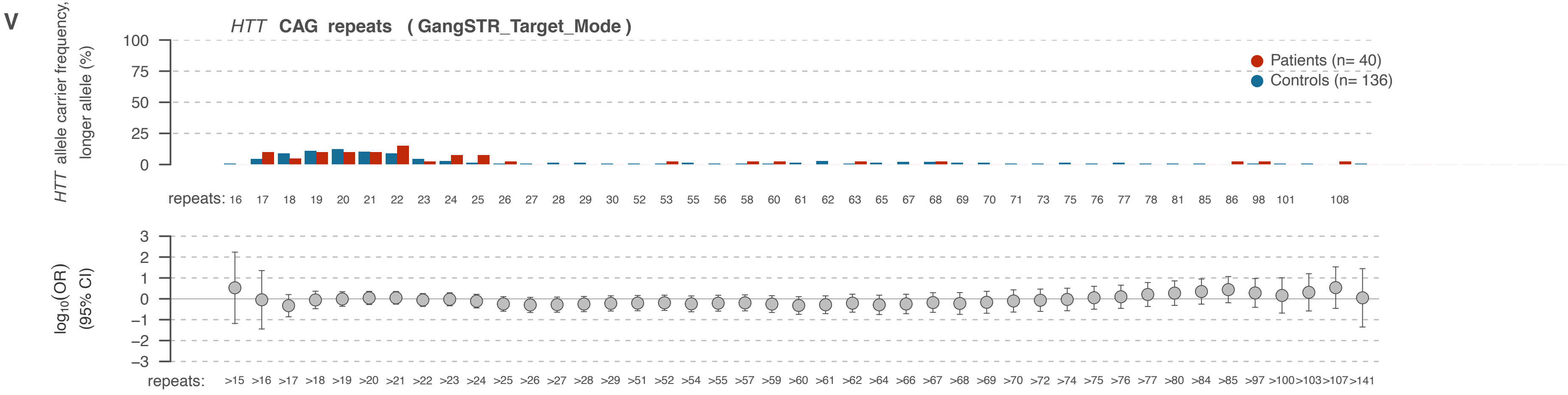
FXN GAA repeats (GangSTR_Target_Mode)
FXN allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 29)
● Controls (n= 132)
repeats: 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 30 62 77 97
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >5 >6 >7 >8 >9 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >29 >61 >76 >96

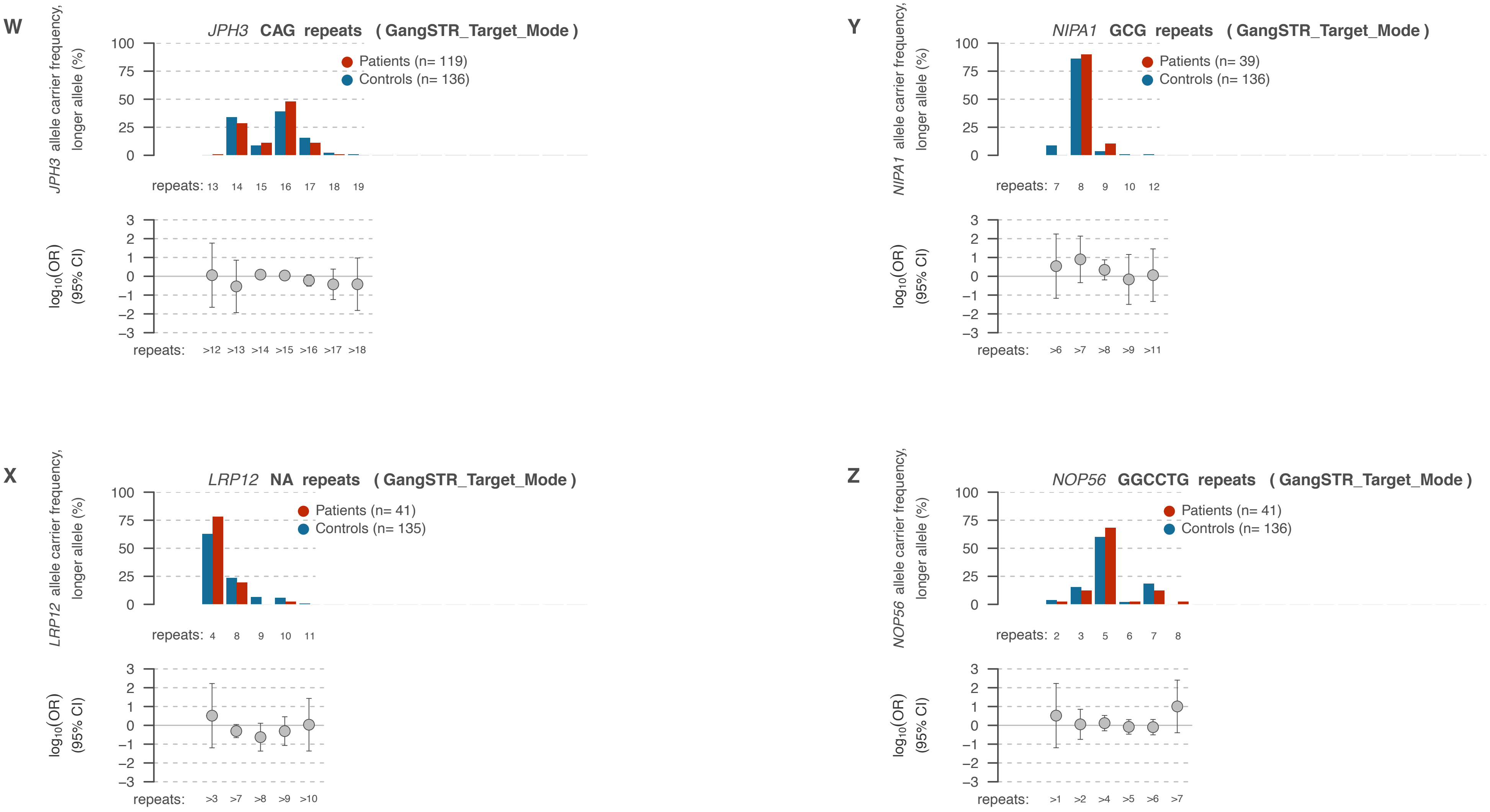
T



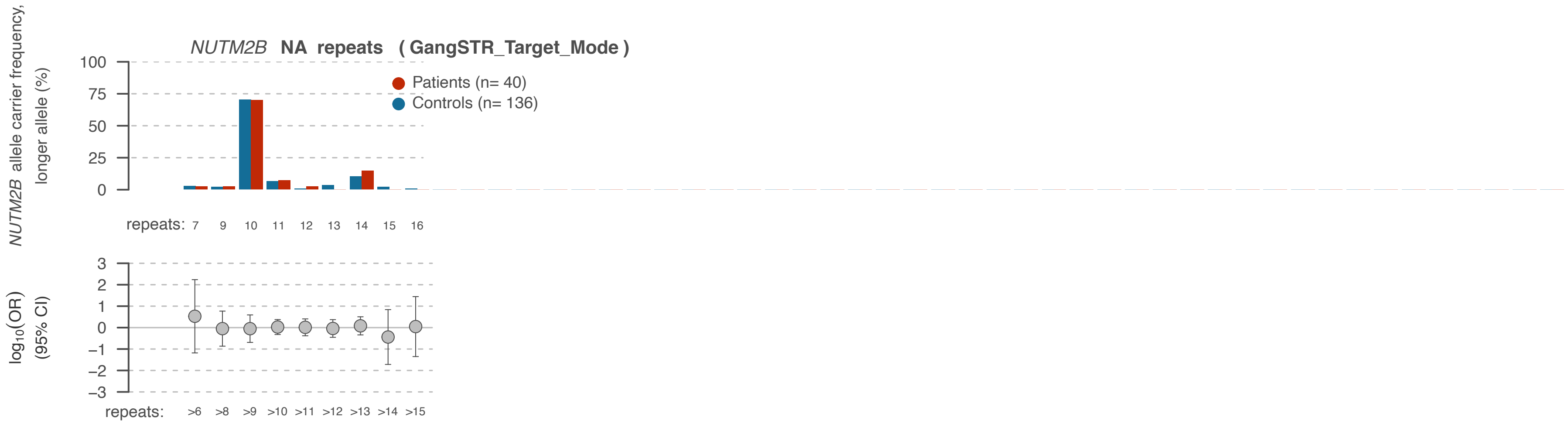
GIPC1 GGC repeats (GangSTR_Target_Mode)
GIPC1 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 39)
● Controls (n= 135)
repeats: 10 11 12 13 14 19
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >9 >10 >11 >12 >13 >18


U
GLS GCA repeats (GangSTR_Target_Mode)
GLS allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 8 9 10 11 12 13 14 15 16 17 18 19 20 21 25 28
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >9 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >24 >27

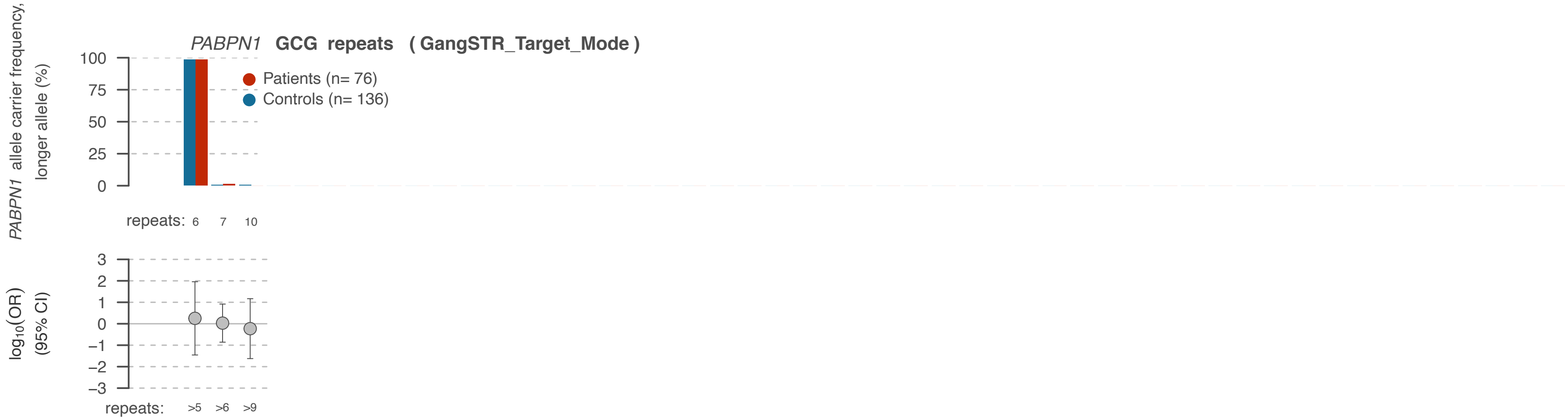

V
HTT CAG repeats (GangSTR_Target_Mode)
HTT allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 40)
Controls (n= 136)
repeats: 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 52 53 55 56 58 60 61 62 63 65 67 68 69 70 71 73 75 76 77 78 81 85 86 98 101 108
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >29 >51 >52 >54 >55 >57 >59 >60 >61 >62 >64 >66 >67 >68 >69 >70 >72 >74 >75 >76 >77 >80 >84 >85 >97 >100 >103 >107 >141

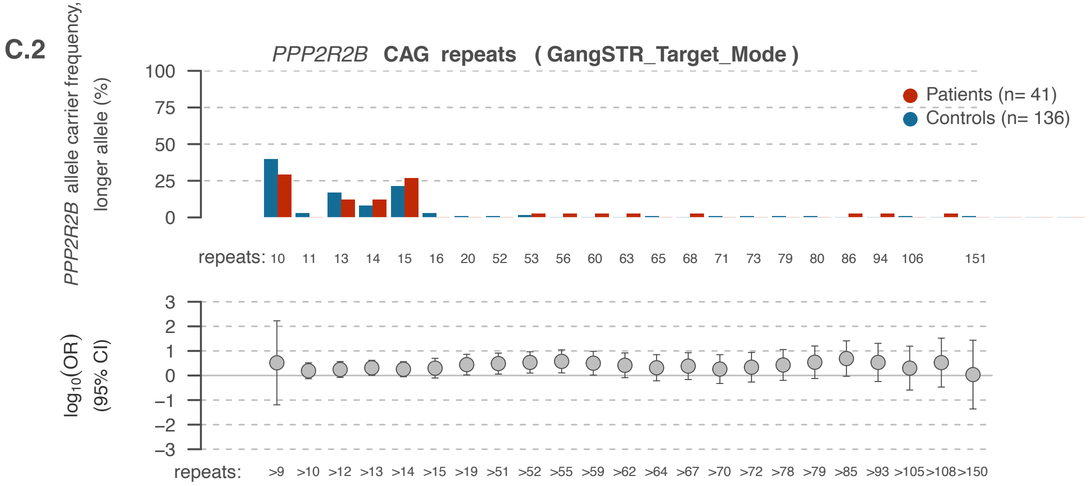

W
JPH3 CAG repeats (GangSTR_Target_Mode)
JPH3 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 119)
Controls (n= 136)
repeats: 13 14 15 16 17 18 19
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >12 >13 >14 >15 >16 >17 >18
Y
NIPA1 GCG repeats (GangSTR_Target_Mode)
NIPA1 allele carrier frequency, longer allele (%)
Patients (n= 39)
Controls (n= 136)
repeats: 7 8 9 10 12
repeats: >6 >7 >8 >9 >11
X
LRP12 NA repeats (GangSTR_Target_Mode)
LRP12 allele carrier frequency, longer allele (%)
Patients (n= 41)
Controls (n= 135)
repeats: 4 8 9 10 11
repeats: >3 >7 >8 >9 >10
Z
NOP56 GGCCTG repeats (GangSTR_Target_Mode)
NOP56 allele carrier frequency, longer allele (%)
Patients (n= 41)
Controls (n= 136)
repeats: 2 3 5 6 7 8
repeats: >1 >2 >4 >5 >6 >7

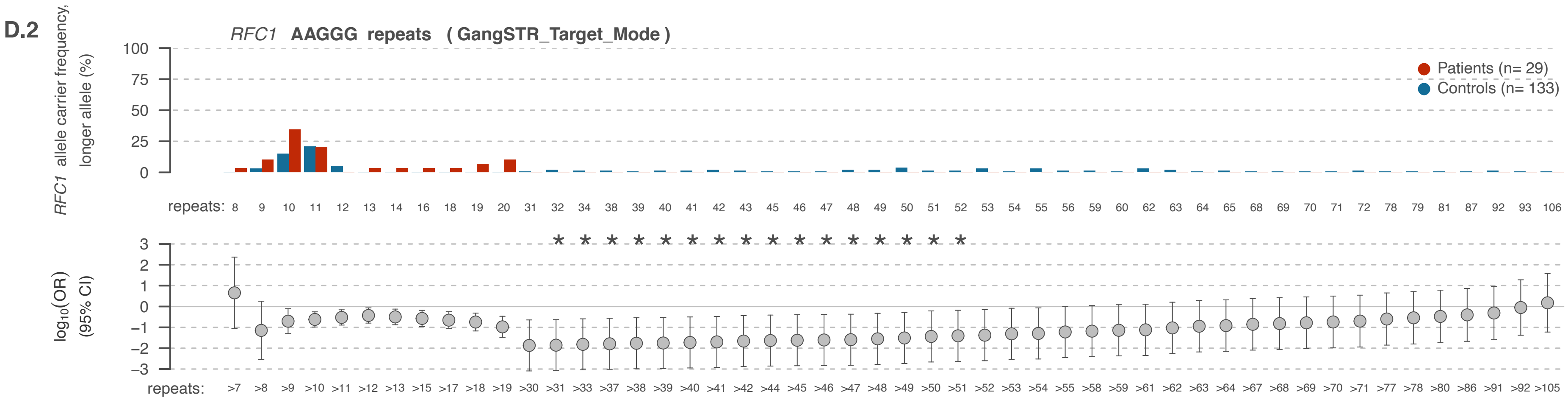
A.2


NUTM2B NA repeats (GangSTR_Target_Mode)
Patients (n= 40)
Controls (n= 136)
NUTM2B allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 7 9 10 11 12 13 14 15 16
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >8 >9 >10 >11 >12 >13 >14 >15

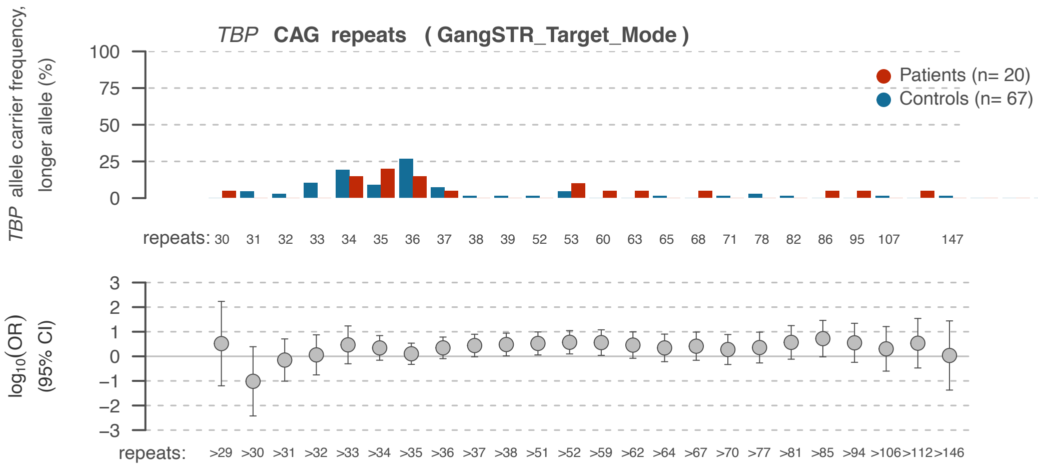
B.2


PABPN1 GCG repeats (GangSTR_Target_Mode)
Patients (n= 76)
Controls (n= 136)
PABPN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 6 7 10
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >5 >6 >9


C.2
PPP2R2B CAG repeats (GangSTR_Target_Mode)
PPP2R2B allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 136)
repeats: 10 11 13 14 15 16 20 52 53 56 60 63 65 68 71 73 79 80 86 94 106 151
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >9 >10 >12 >13 >14 >15 >19 >51 >52 >55 >59 >62 >64 >67 >70 >72 >78 >79 >85 >93 >105 >108 >150

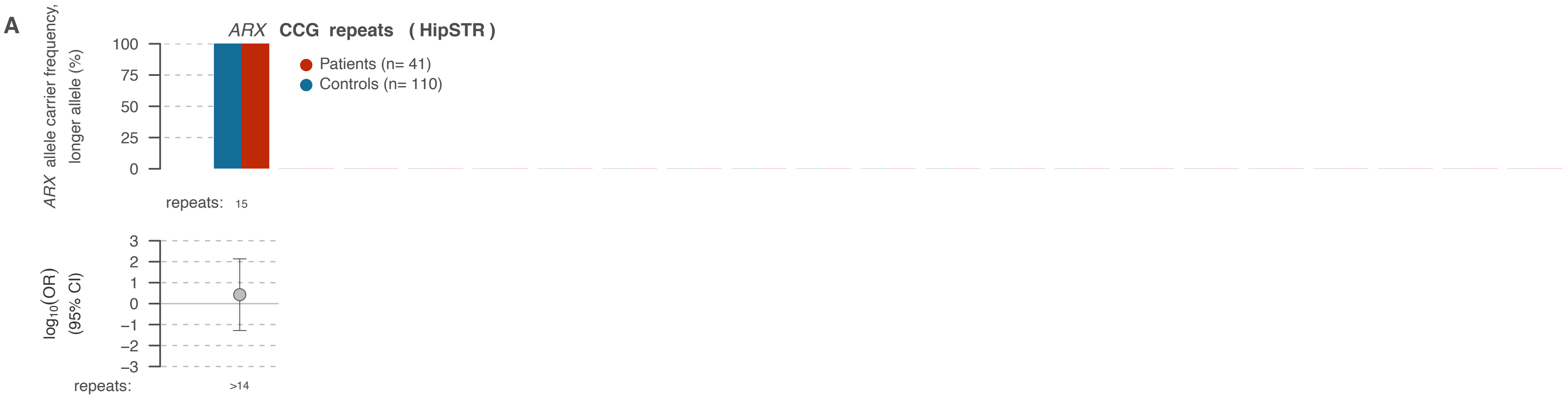

D.2
RFC1 AAGGG repeats (GangSTR_Target_Mode)
RFC1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 29)
Controls (n= 133)
repeats: 8 9 10 11 12 13 14 16 18 19 20 31 32 34 38 39 40 41 42 43 45 46 47 48 49 50 51 52 53 54 55 56 59 60 62 63 64 65 68 69 70 71 72 78 79 81 87 92 93 106
* * * * * * * * * * * * * * * *
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >9 >10 >11 >12 >13 >15 >17 >18 >19 >30 >31 >33 >37 >38 >39 >40 >41 >42 >44 >45 >46 >47 >48 >49 >50 >51 >52 >53 >54 >55 >58 >59 >61 >62 >63 >64 >67 >68 >69 >70 >71 >77 >78 >80 >86 >91 >92 >105

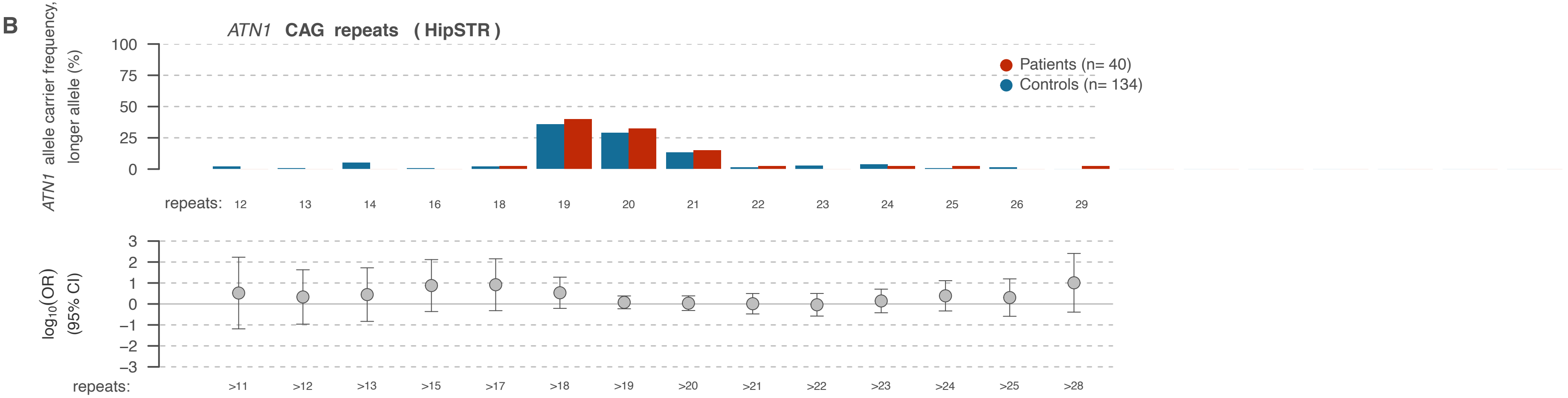
E.2

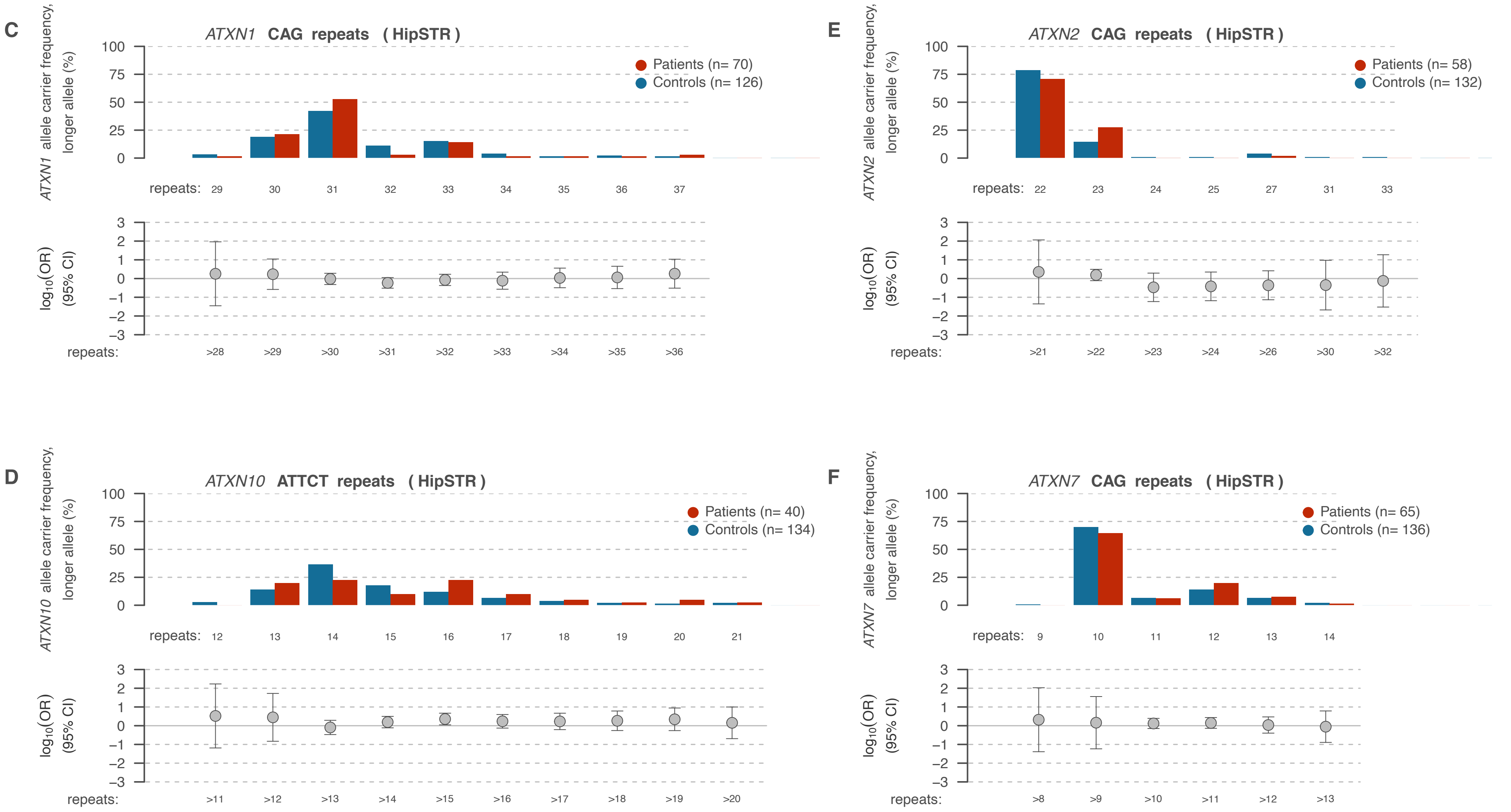


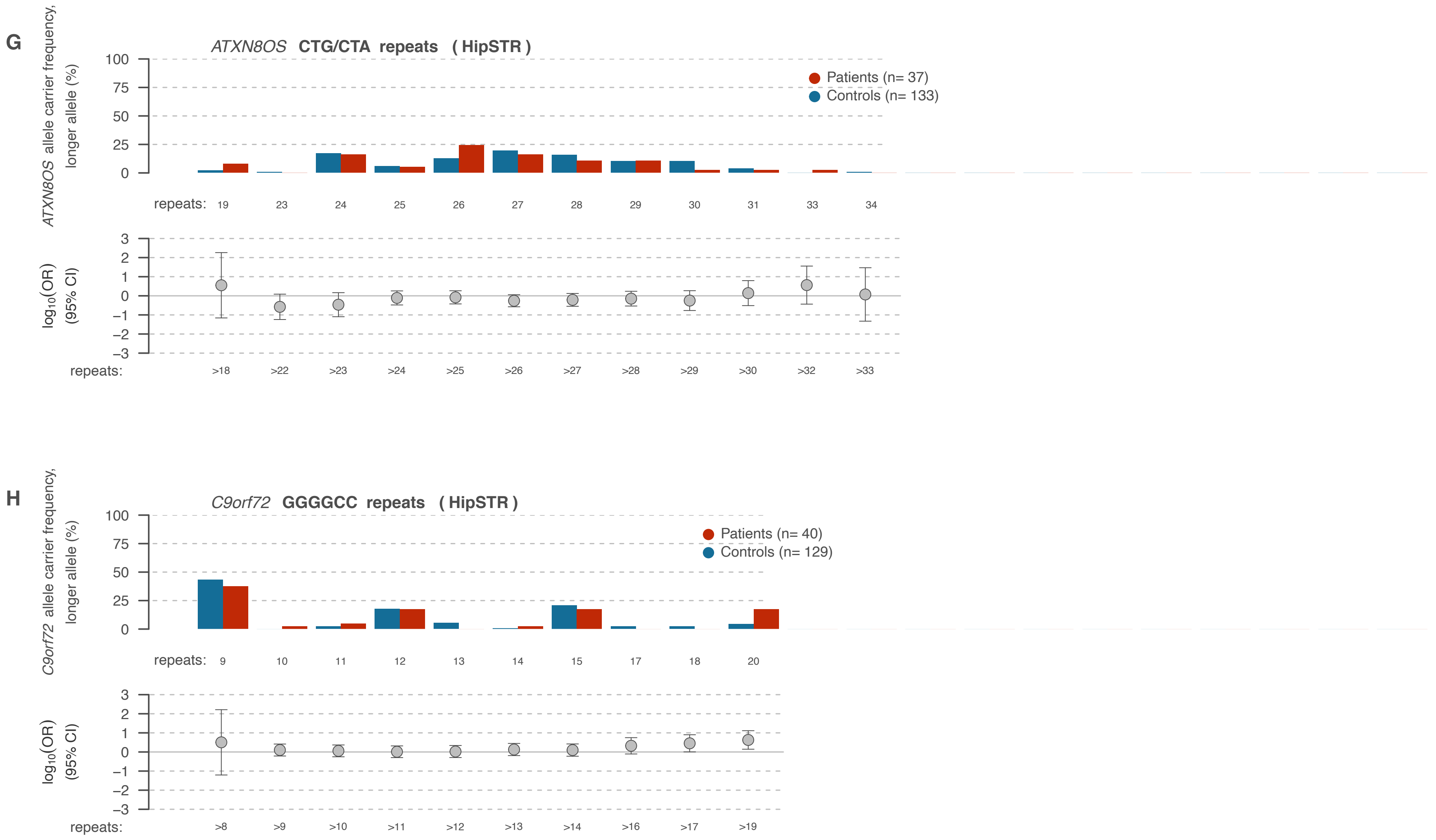
Supplementary Figure 3.16: GangSTR (target mode) prediction of STR lengths in epilepsy patients

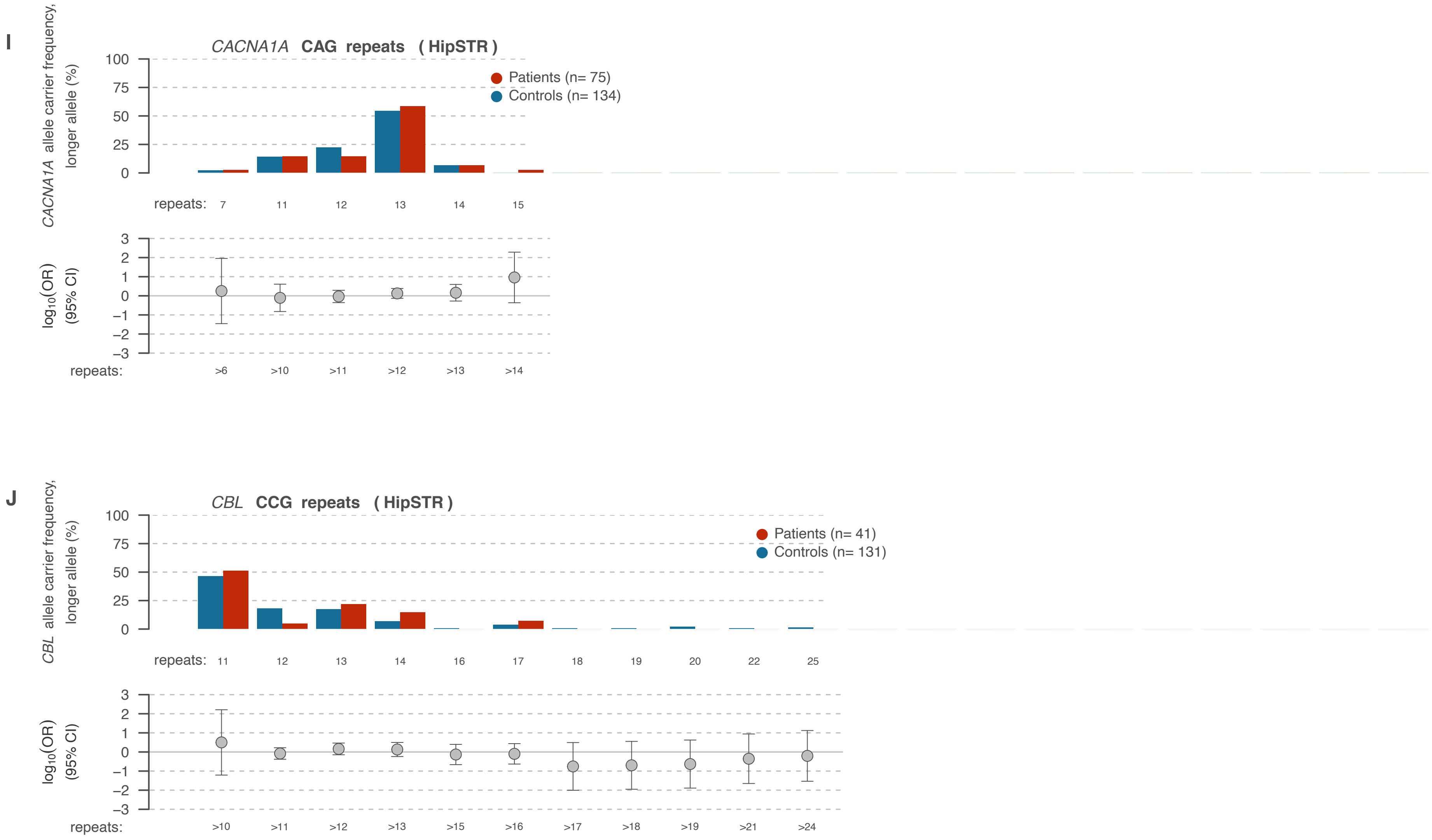
For each gene genotyped with GangSTR (target mode) the allele lengths in epilepsy patients are compared to 136 Irish controls. The upper plot shows the predicted allele lengths and the lower plot shows the OR. An asterisks indicate a significant OR. The epilepsy results include PCR-free WGS samples, PCR WGS samples and WES sample if an RMSD below one was observed when comparing WES results to WGS results for a given gene.

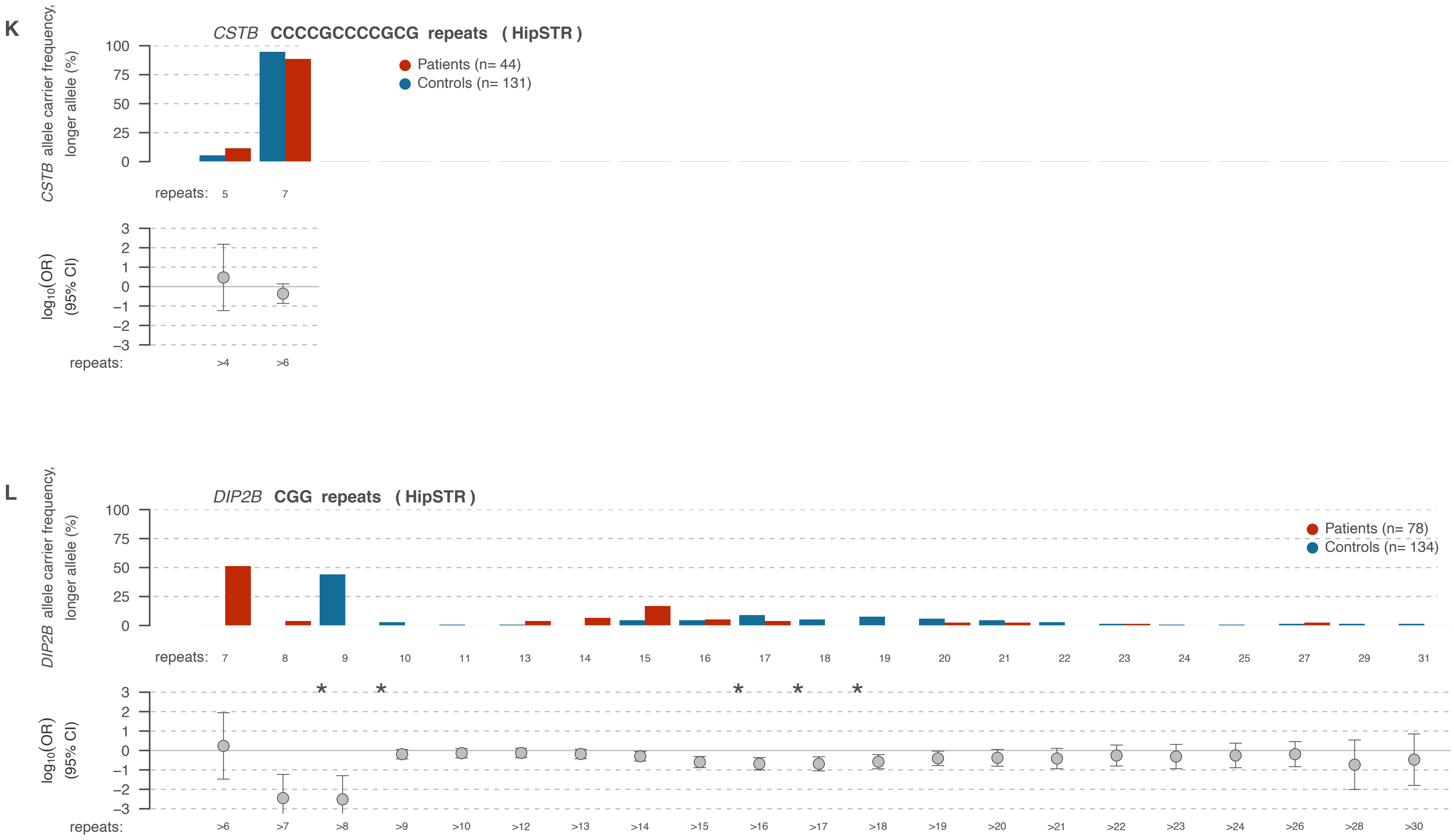

A
ARX CCG repeats (HipSTR)
ARX allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 110)
repeats: 15
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >14

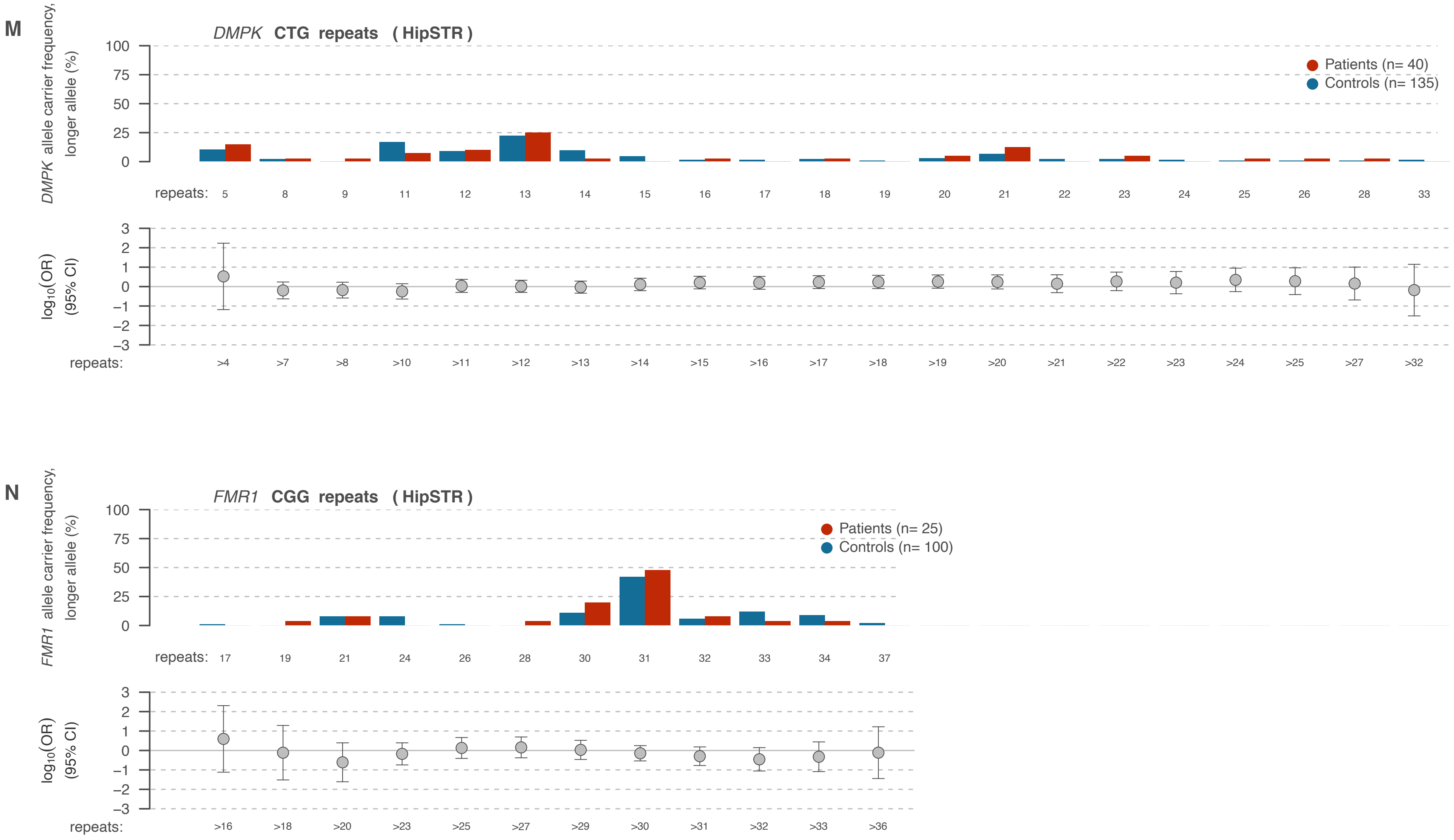

B
ATN1 CAG repeats (HipSTR)
ATN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 40)
Controls (n= 134)
repeats: 12 13 14 16 18 19 20 21 22 23 24 25 26 29
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >11 >12 >13 >15 >17 >18 >19 >20 >21 >22 >23 >24 >25 >28

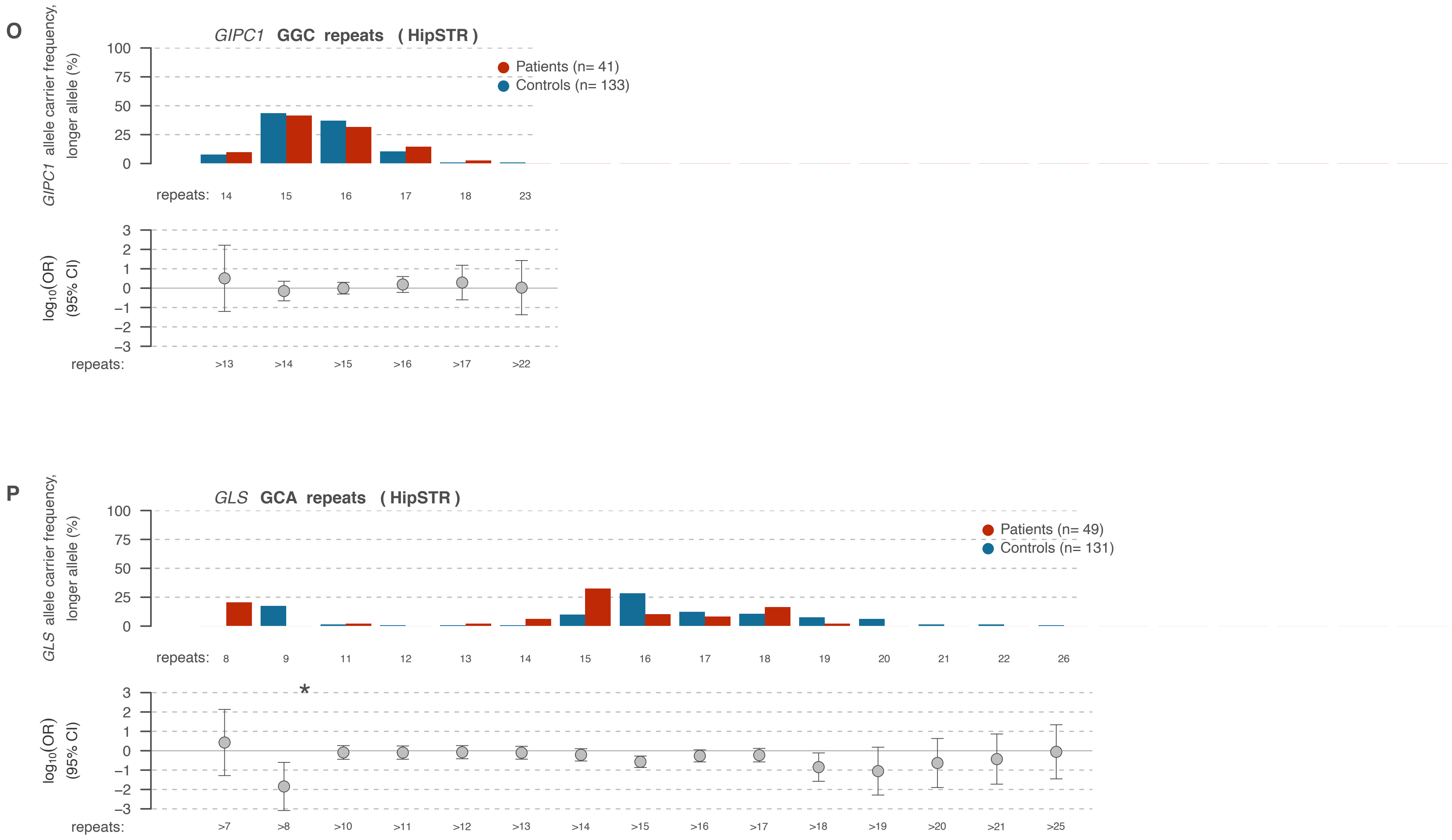

C
ATXN1 CAG repeats (HipSTR)
ATXN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 70)
Controls (n= 126)
repeats: 29 30 31 32 33 34 35 36 37
log10(OR) (95% CI)
3 2 1 0 −1 −2 −3
repeats: >28 >29 >30 >31 >32 >33 >34 >35 >36
E
ATXN2 CAG repeats (HipSTR)
ATXN2 allele carrier frequency, longer allele (%)
Patients (n= 58)
Controls (n= 132)
repeats: 22 23 24 25 27 31 33
repeats: >21 >22 >23 >24 >26 >30 >32
D
ATXN10 ATTCT repeats (HipSTR)
ATXN10 allele carrier frequency, longer allele (%)
Patients (n= 40)
Controls (n= 134)
repeats: 12 13 14 15 16 17 18 19 20 21
repeats: >11 >12 >13 >14 >15 >16 >17 >18 >19 >20
F
ATXN7 CAG repeats (HipSTR)
ATXN7 allele carrier frequency, longer allele (%)
Patients (n= 65)
Controls (n= 136)
repeats: 9 10 11 12 13 14
repeats: >8 >9 >10 >11 >12 >13


G
ATXN8OS CTG/CTA repeats (HipSTR)
ATXN8OS allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 37)
Controls (n= 133)
repeats: 19 23 24 25 26 27 28 29 30 31 33 34
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >18 >22 >23 >24 >25 >26 >27 >28 >29 >30 >32 >33
H
C9orf72 GGGGCC repeats (HipSTR)
C9orf72 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 40)
Controls (n= 129)
repeats: 9 10 11 12 13 14 15 17 18 20
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >9 >10 >11 >12 >13 >14 >16 >17 >19

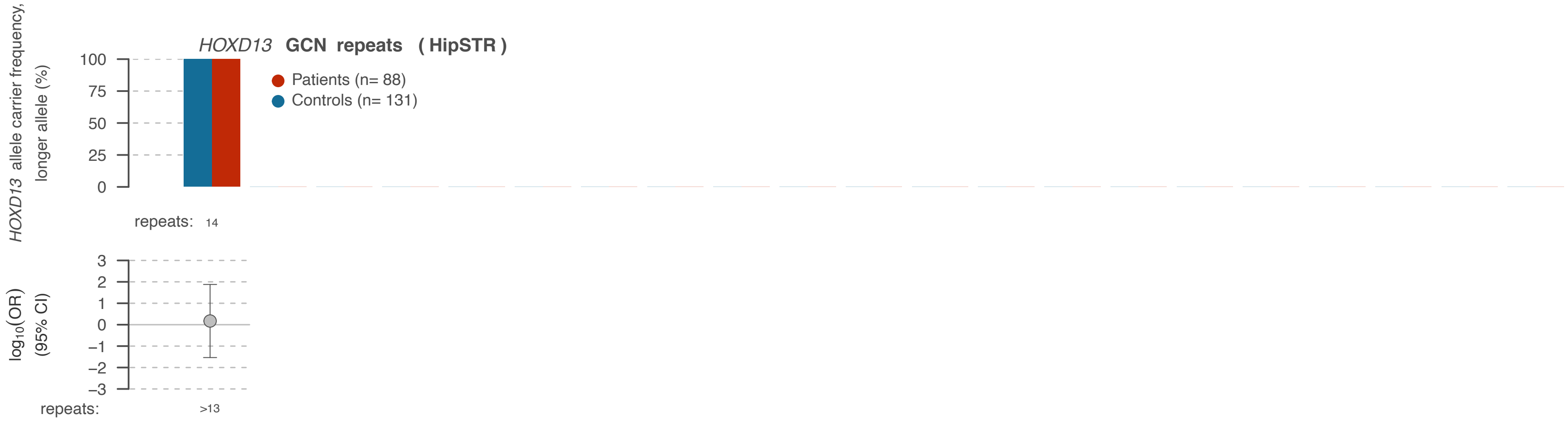

I
CACNA1A CAG repeats (HipSTR)
CACNA1A allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 75)
Controls (n= 134)
repeats: 7 11 12 13 14 15
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >10 >11 >12 >13 >14
J
CBL CCG repeats (HipSTR)
CBL allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 131)
repeats: 11 12 13 14 16 17 18 19 20 22 25
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >10 >11 >12 >13 >15 >16 >17 >18 >19 >21 >24


K
CSTB CCCCGCCCCGCG repeats (HipSTR)
● Patients (n= 44)
● Controls (n= 131)
CSTB allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 5 7
log$_{10}$(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >4 >6
L
DIP2B CGG repeats (HipSTR)
● Patients (n= 78)
● Controls (n= 134)
DIP2B allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 7 8 9 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 27 29 31
log$_{10}$(OR) (95% CI)
3
2
1
0
−1
−2
−3
* * * * *
repeats: >6 >7 >8 >9 >10 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >26 >28 >30

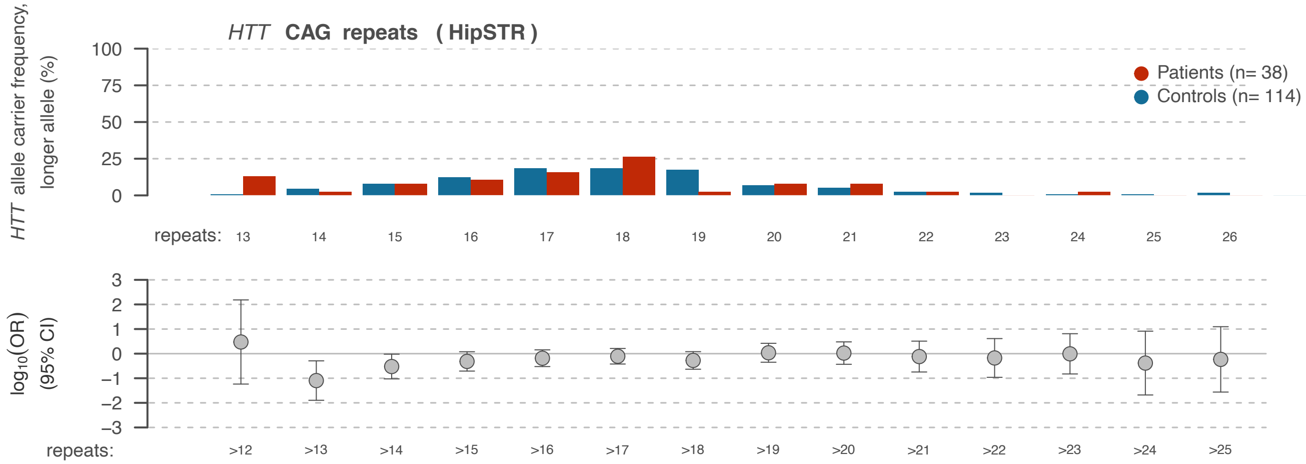

M
DMPK CTG repeats (HipSTR)
DMPK allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 40)
Controls (n= 135)
repeats: 5 8 9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 33
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >4 >7 >8 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >27 >32
N
FMR1 CGG repeats (HipSTR)
FMR1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 25)
Controls (n= 100)
repeats: 17 19 21 24 26 28 30 31 32 33 34 37
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >16 >18 >20 >23 >25 >27 >29 >30 >31 >32 >33 >36


O
GIPC1 GGC repeats (HipSTR)
GIPC1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 133)
repeats: 14 15 16 17 18 23
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >13 >14 >15 >16 >17 >22
P
GLS GCA repeats (HipSTR)
GLS allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 49)
Controls (n= 131)
repeats: 8 9 11 12 13 14 15 16 17 18 19 20 21 22 26
*
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >25

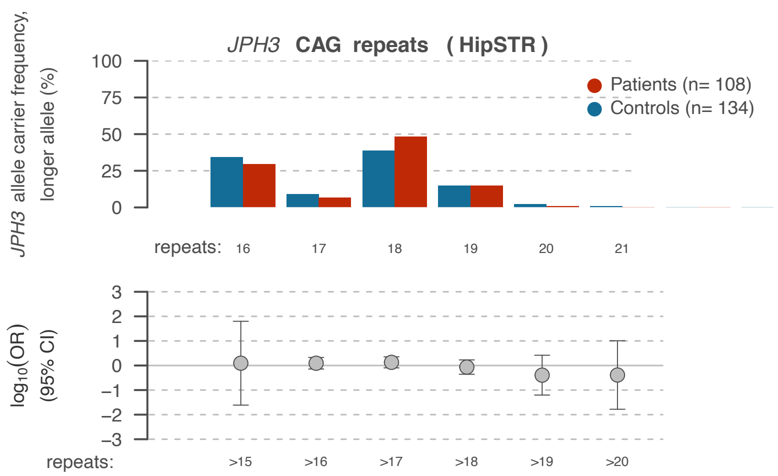
Q


HOXD13 GCN repeats (HipSTR)
HOXD13 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 88)
● Controls (n= 131)
repeats: 14
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >13

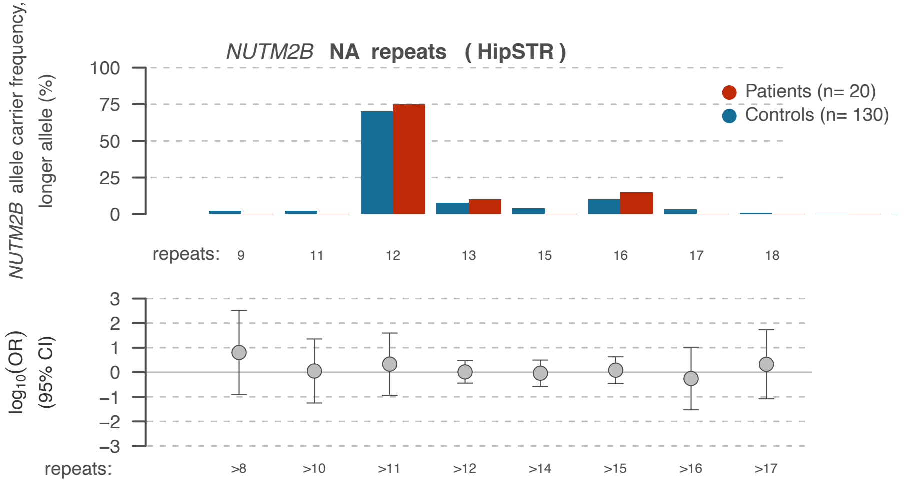
R


HTT CAG repeats (HipSTR)
HTT allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 38)
● Controls (n= 114)
repeats: 13 14 15 16 17 18 19 20 21 22 23 24 25 26
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25

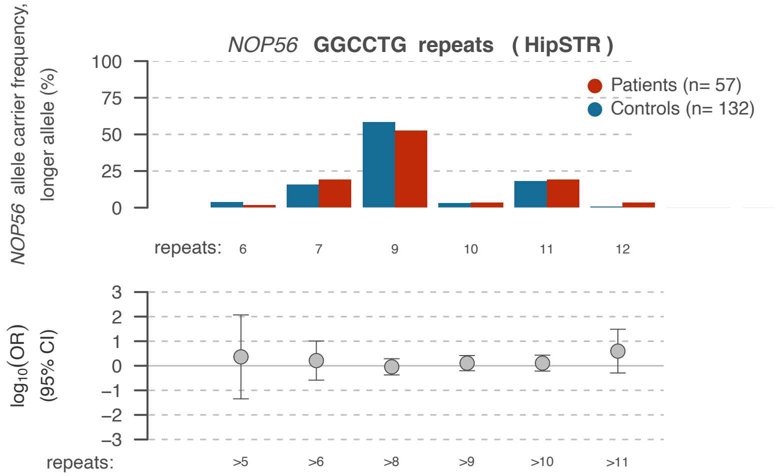
S


JPH3 CAG repeats (HipSTR)
JPH3 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 108)
Controls (n= 134)
repeats: 16 17 18 19 20 21
3
2
1
0
−1
−2
−3
log10(OR) (95% CI)
repeats: >15 >16 >17 >18 >19 >20

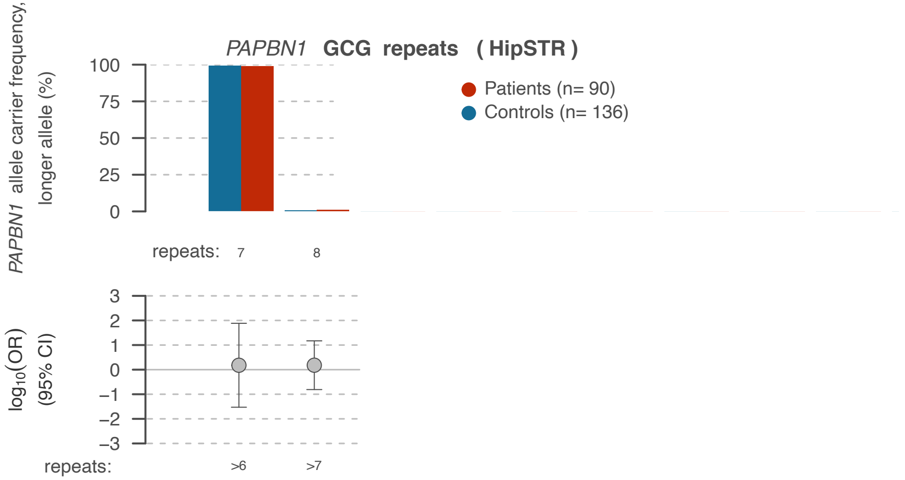
U

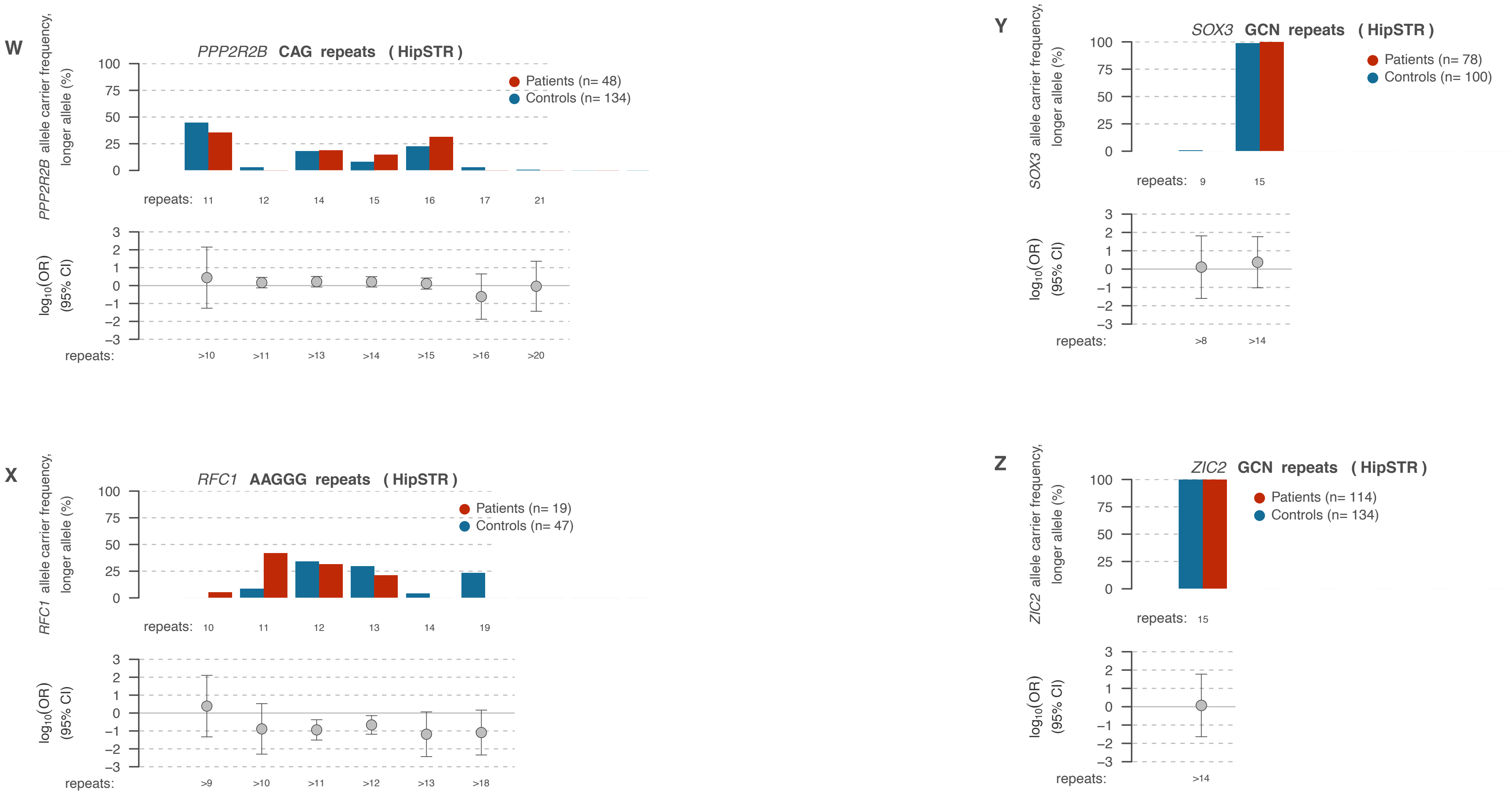

NUTM2B NA repeats (HipSTR)
NUTM2B allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 20)
Controls (n= 130)
repeats: 9 11 12 13 15 16 17 18
3
2
1
0
−1
−2
−3
log10(OR) (95% CI)
repeats: >8 >10 >11 >12 >14 >15 >16 >17

T


NOP56 GGCCTG repeats (HipSTR)
NOP56 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 57)
Controls (n= 132)
repeats: 6 7 9 10 11 12
3
2
1
0
−1
−2
−3
log10(OR) (95% CI)
repeats: >5 >6 >8 >9 >10 >11

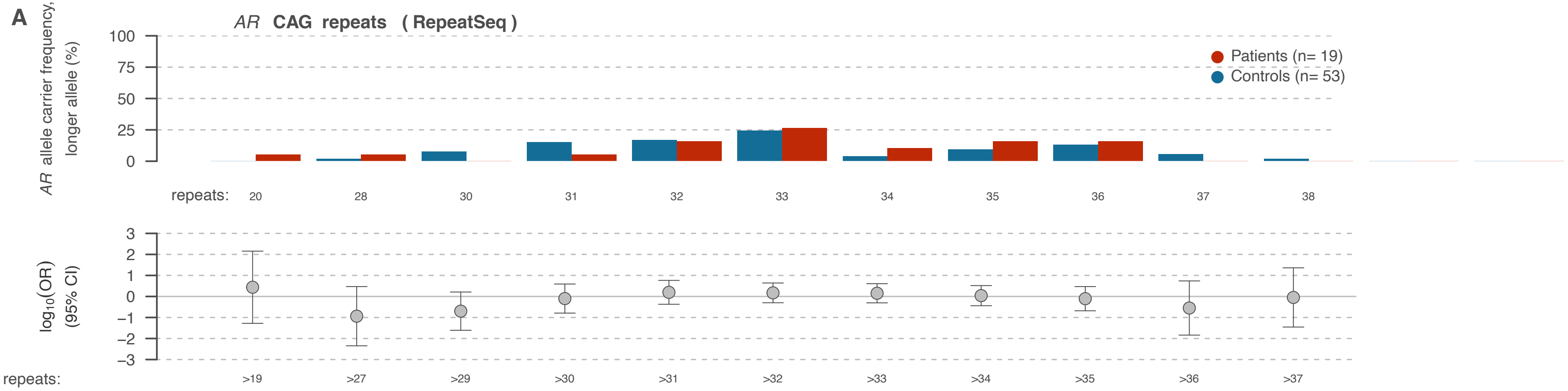
V

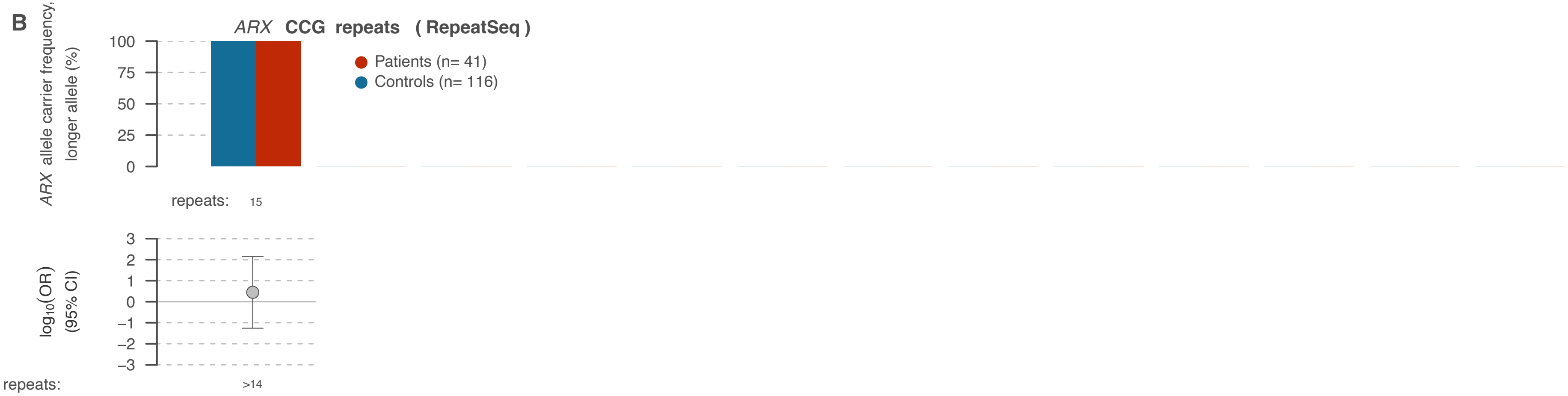

PAPBN1 GCG repeats (HipSTR)
PAPBN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 90)
Controls (n= 136)
repeats: 7 8
3
2
1
0
−1
−2
−3
log10(OR) (95% CI)
repeats: >6 >7

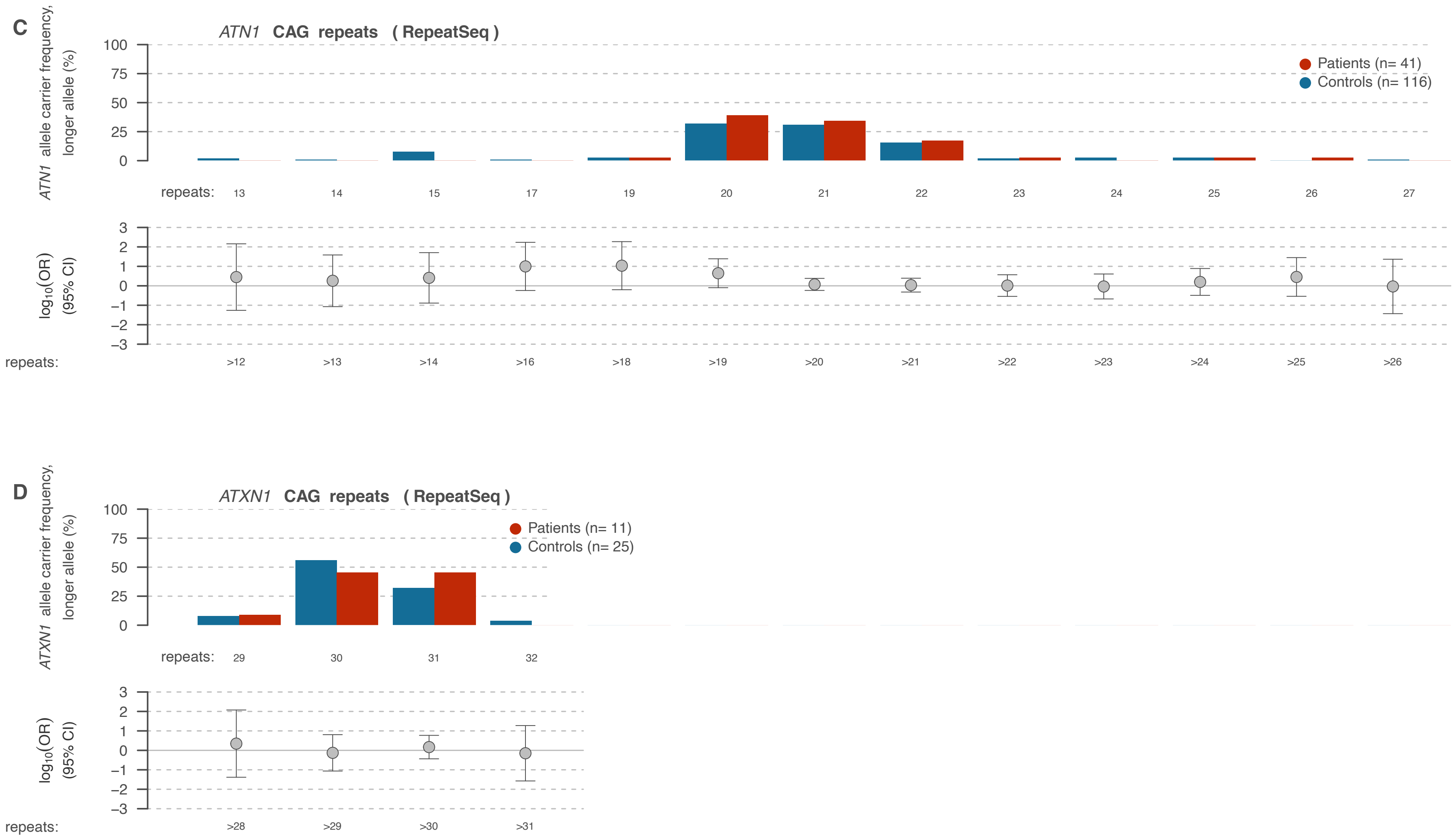


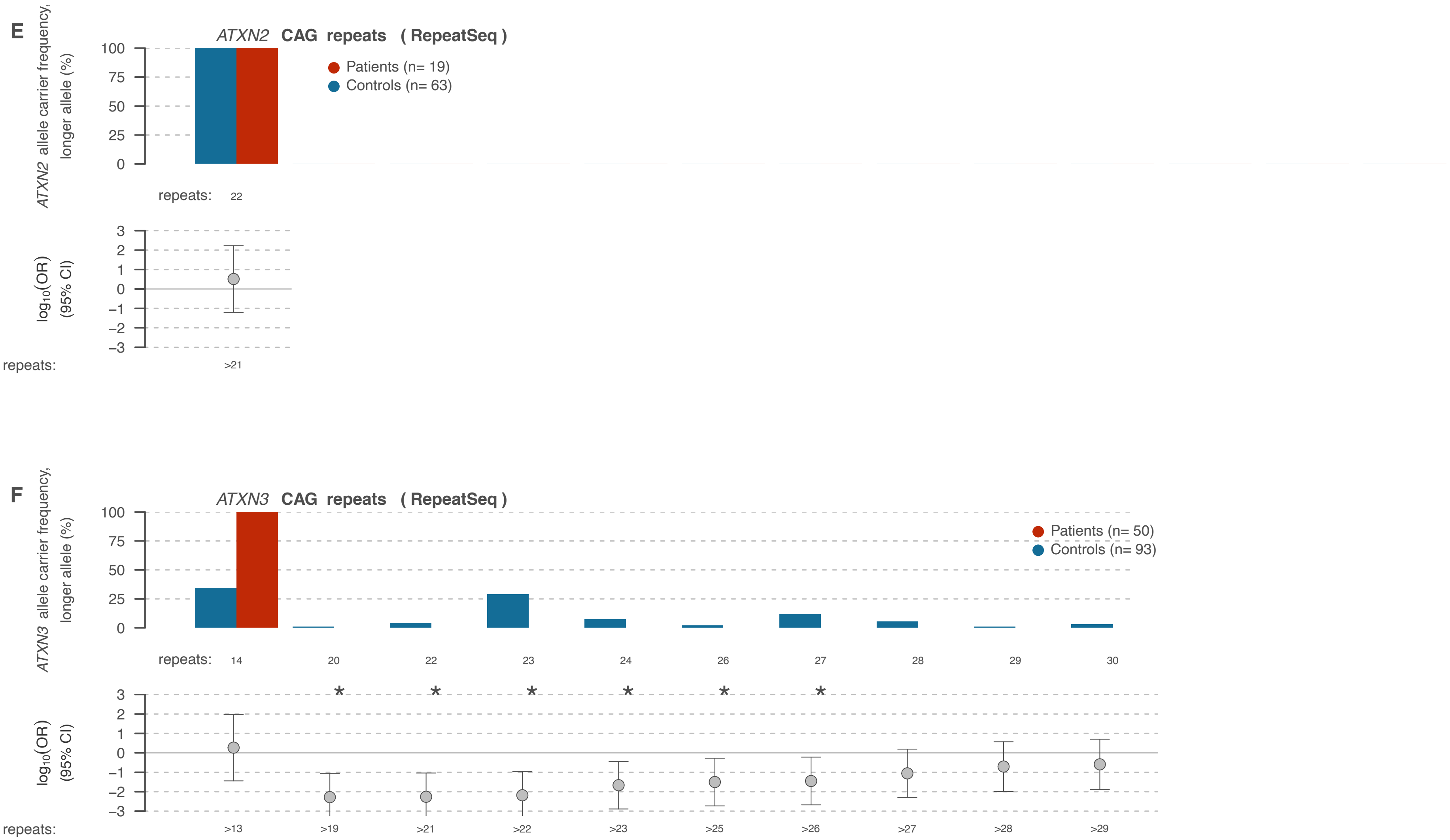
Supplementary Figure 3.17: HipSTR prediction of STR lengths in epilepsy patients

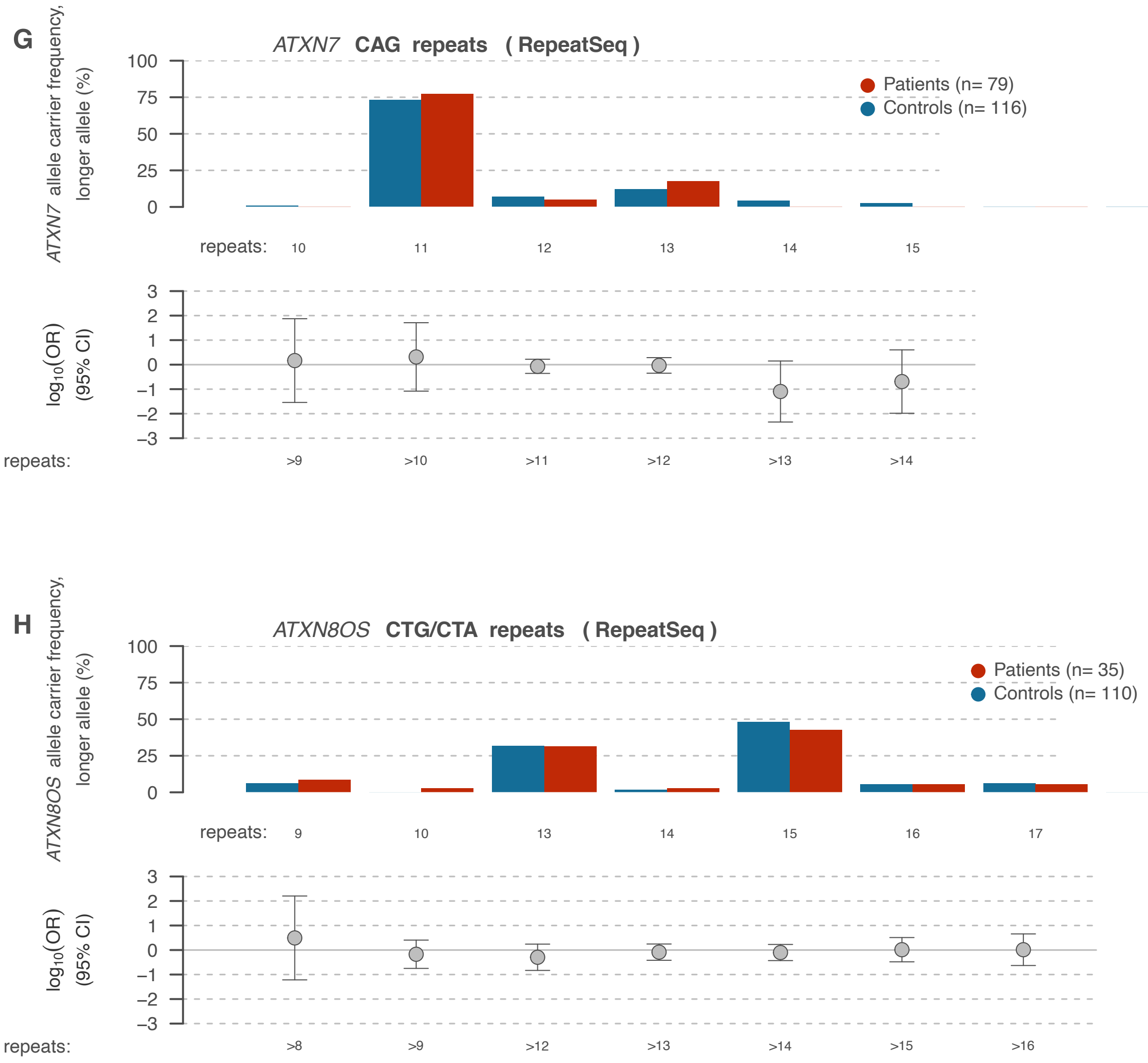
For each gene genotyped with HipSTR the allele lengths in epilepsy patients are compared to 136 Irish controls. The upper plot shows the predicted allele lengths and the lower plot shows the OR. An asterisks indicate a significant OR. The epilepsy results include PCR-free WGS samples, PCR WGS samples and WES sample if an RMSD below one was observed when comparing WES results to WGS results for a given gene.

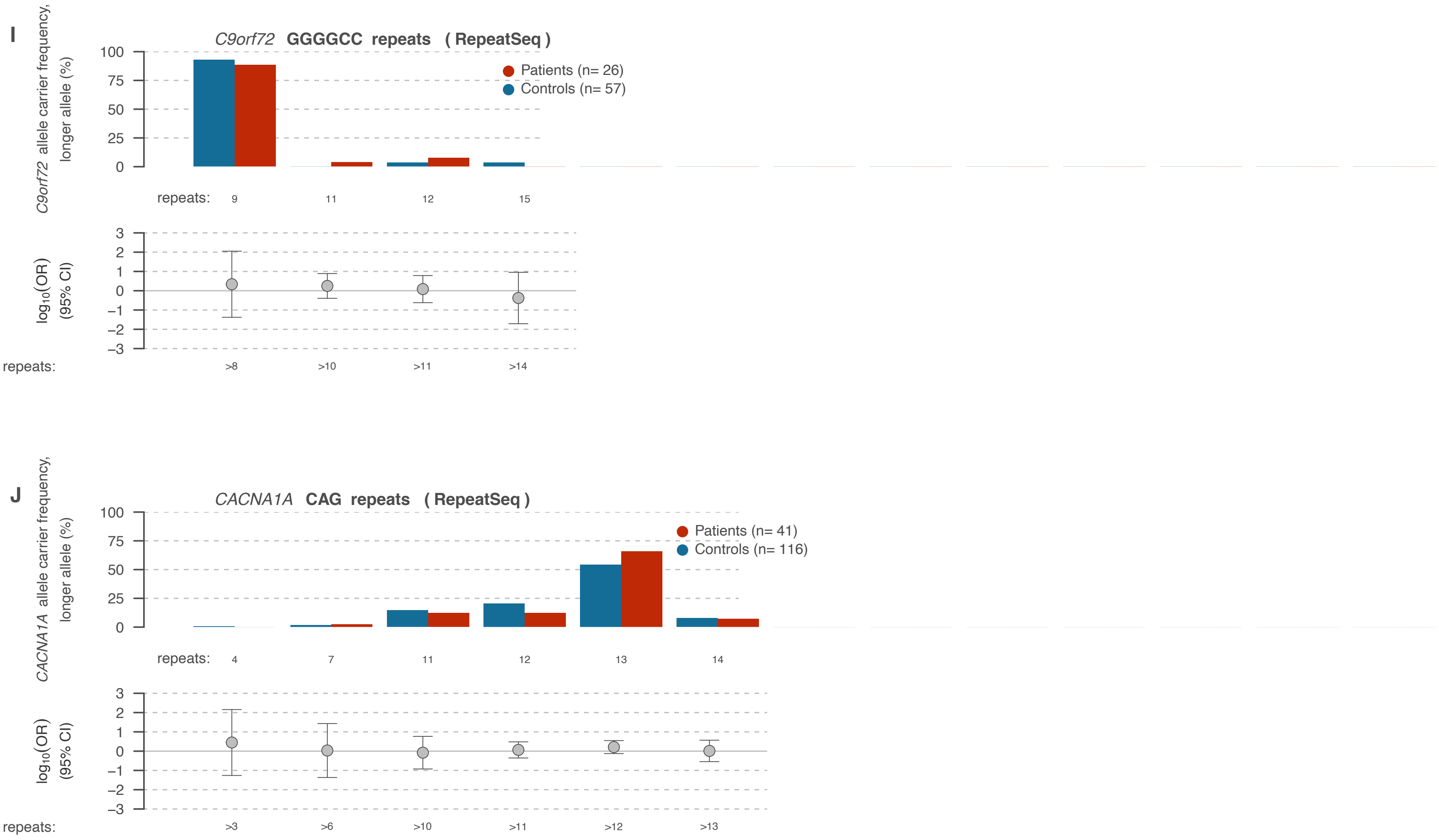

A
AR CAG repeats (RepeatSeq)
AR allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 19)
● Controls (n= 53)
repeats:
20
28
30
31
32
33
34
35
36
37
38
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats:
>19
>27
>29
>30
>31
>32
>33
>34
>35
>36
>37

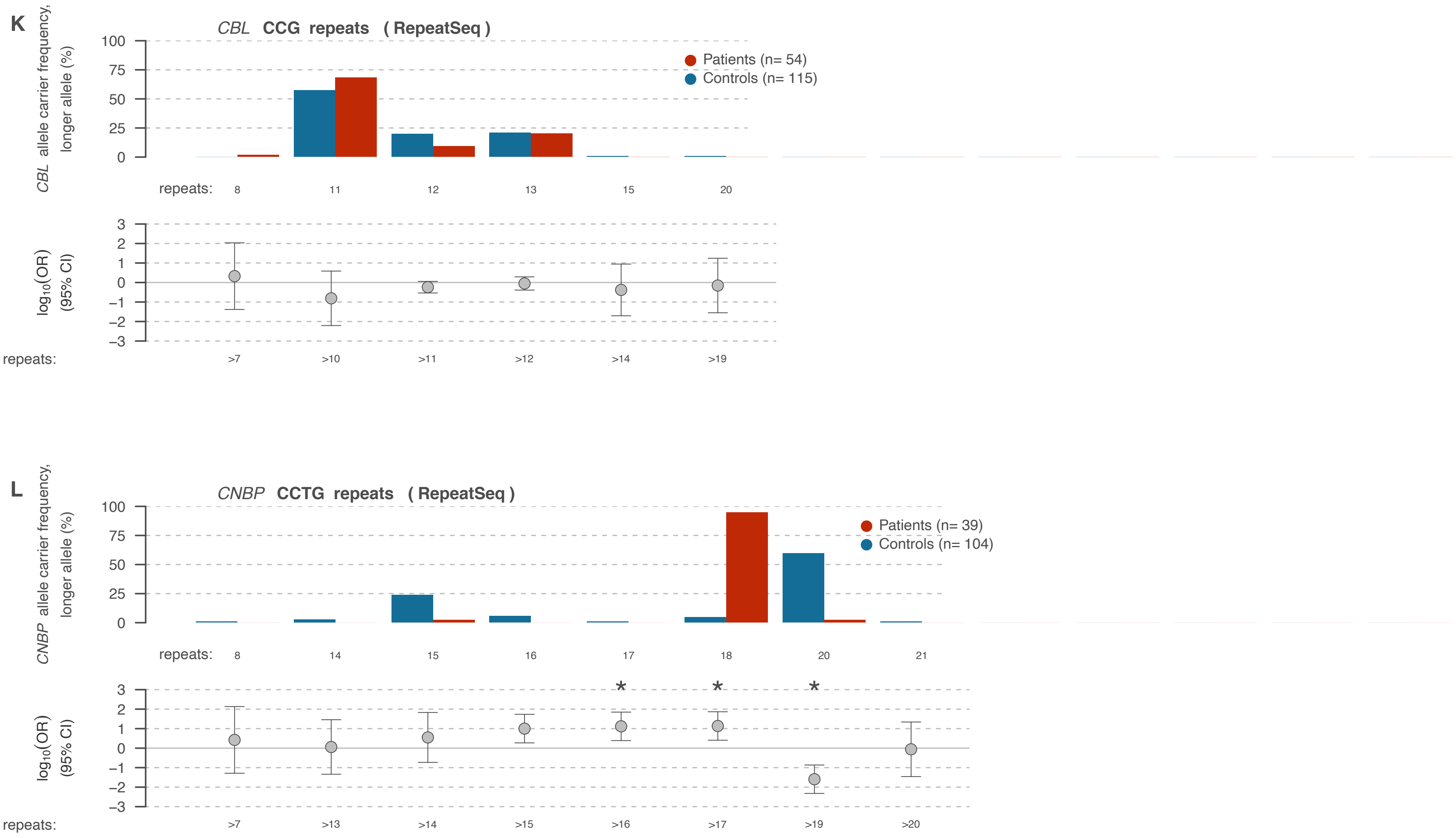

B
ARX CCG repeats (RepeatSeq)
ARX allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 116)
repeats:
15
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats:
>14

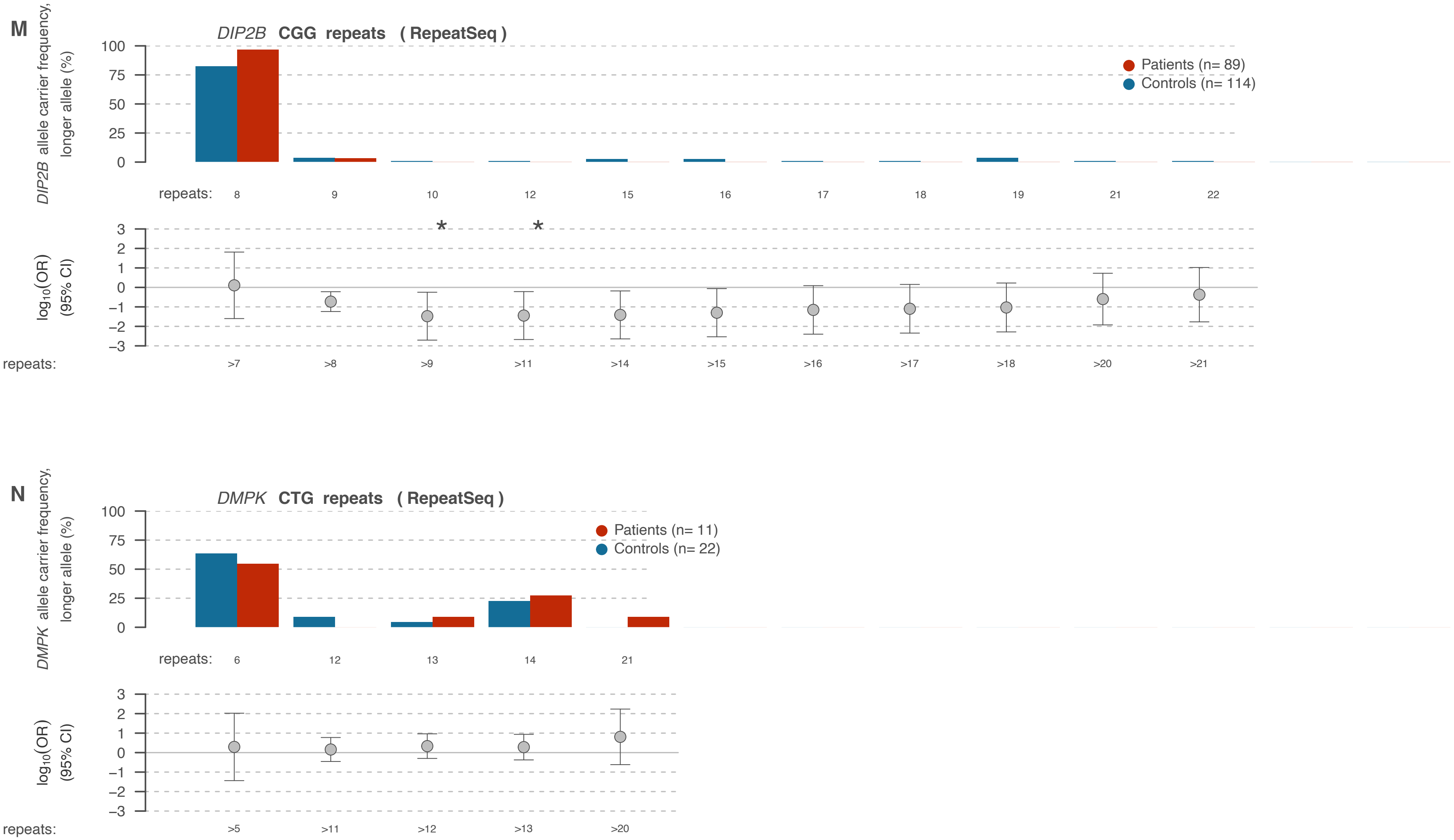

C
ATN1 CAG repeats (RepeatSeq)
ATN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 41)
Controls (n= 116)
repeats: 13 14 15 17 19 20 21 22 23 24 25 26 27
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >12 >13 >14 >16 >18 >19 >20 >21 >22 >23 >24 >25 >26
D
ATXN1 CAG repeats (RepeatSeq)
ATXN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 11)
Controls (n= 25)
repeats: 29 30 31 32
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >28 >29 >30 >31

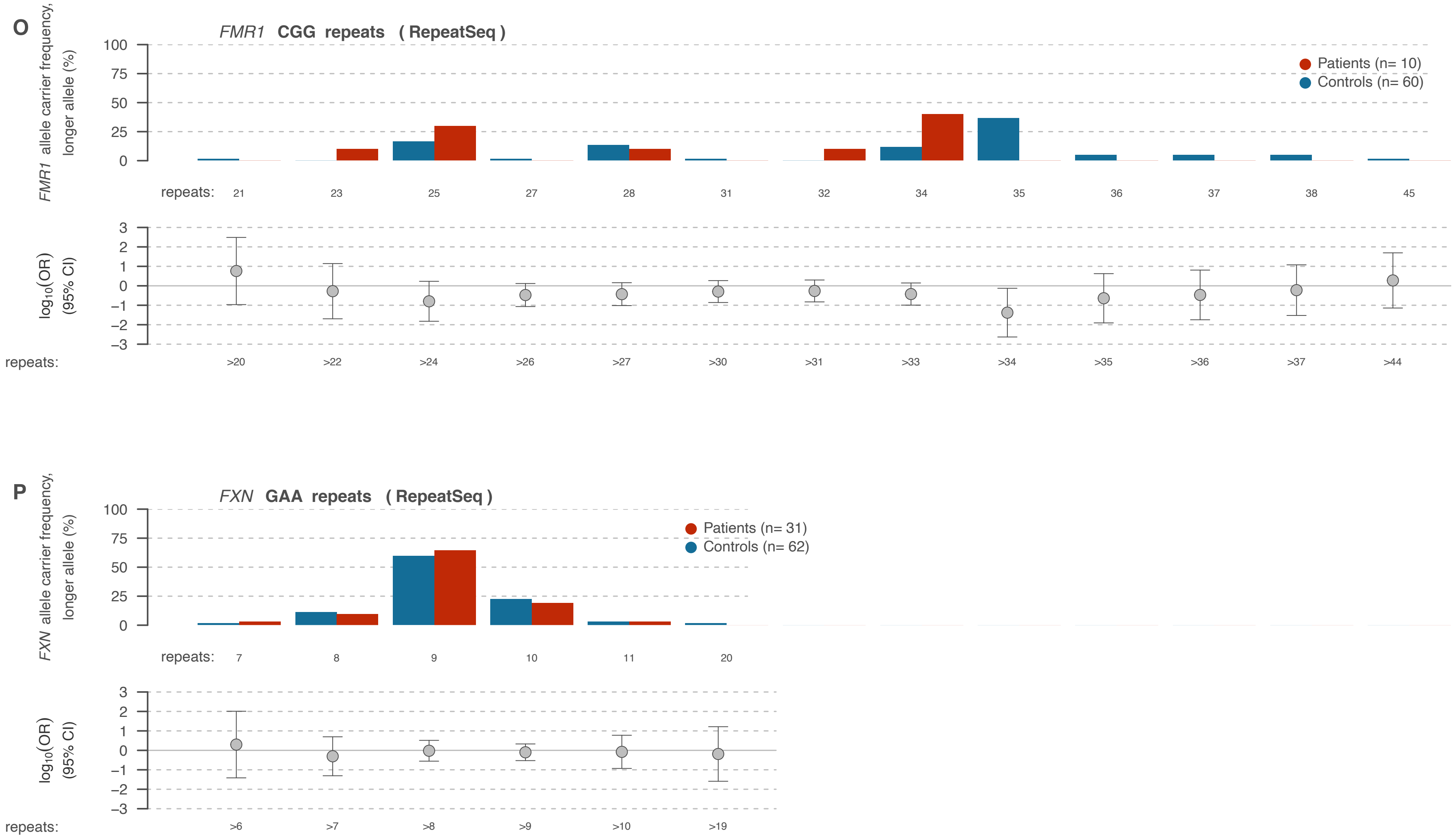

E
ATXN2 CAG repeats (RepeatSeq)
● Patients (n= 19)
● Controls (n= 63)
ATXN2 allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 22
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >21
F
ATXN3 CAG repeats (RepeatSeq)
● Patients (n= 50)
● Controls (n= 93)
ATXN3 allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 14 20 22 23 24 26 27 28 29 30
* * * * * *
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >13 >19 >21 >22 >23 >25 >26 >27 >28 >29

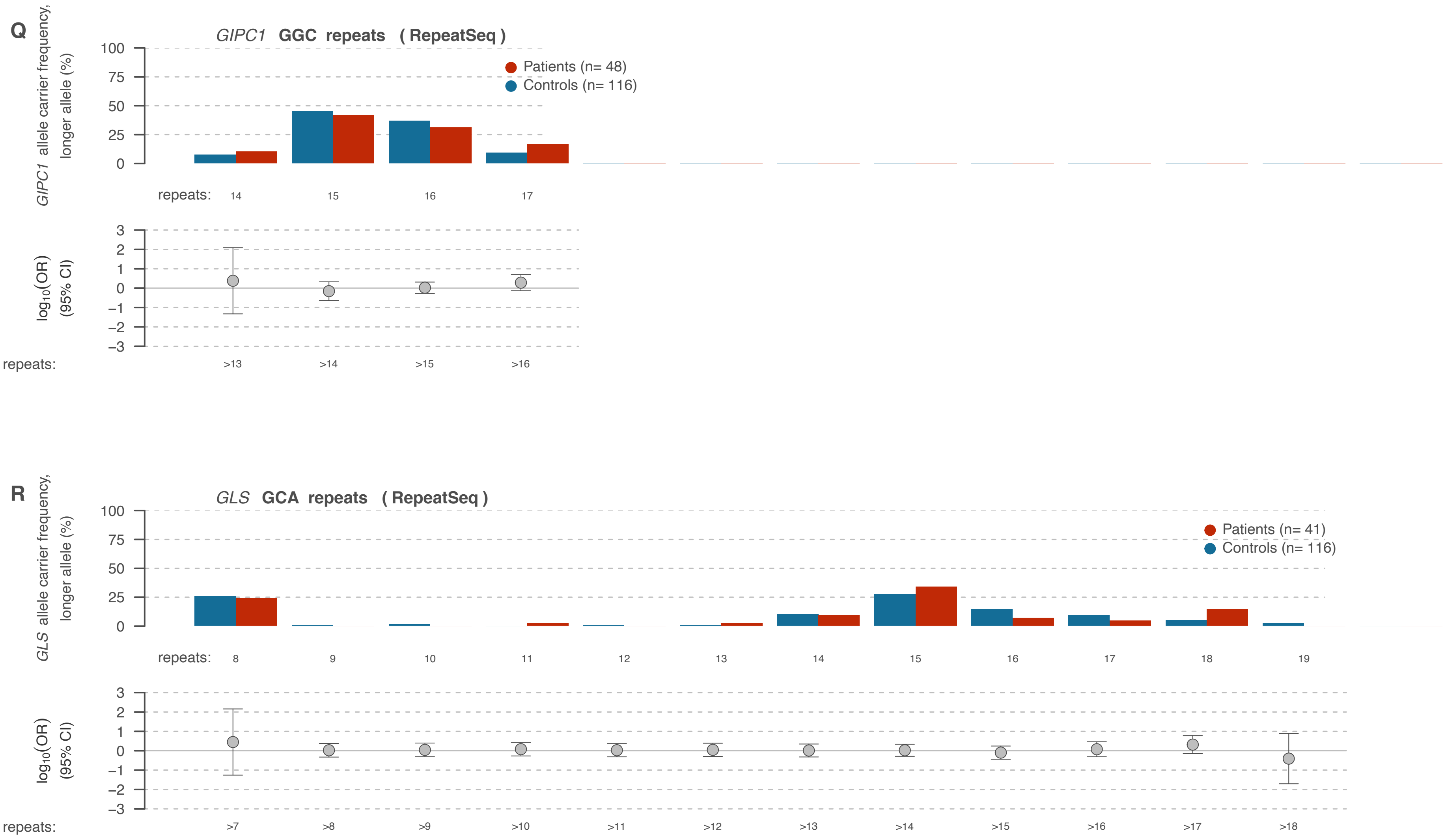

G
ATXN7 CAG repeats (RepeatSeq)
ATXN7 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 79)
● Controls (n= 116)
repeats: 10 11 12 13 14 15
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >9 >10 >11 >12 >13 >14
H
ATXN8OS CTG/CTA repeats (RepeatSeq)
ATXN8OS allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 35)
● Controls (n= 110)
repeats: 9 10 13 14 15 16 17
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >9 >12 >13 >14 >15 >16

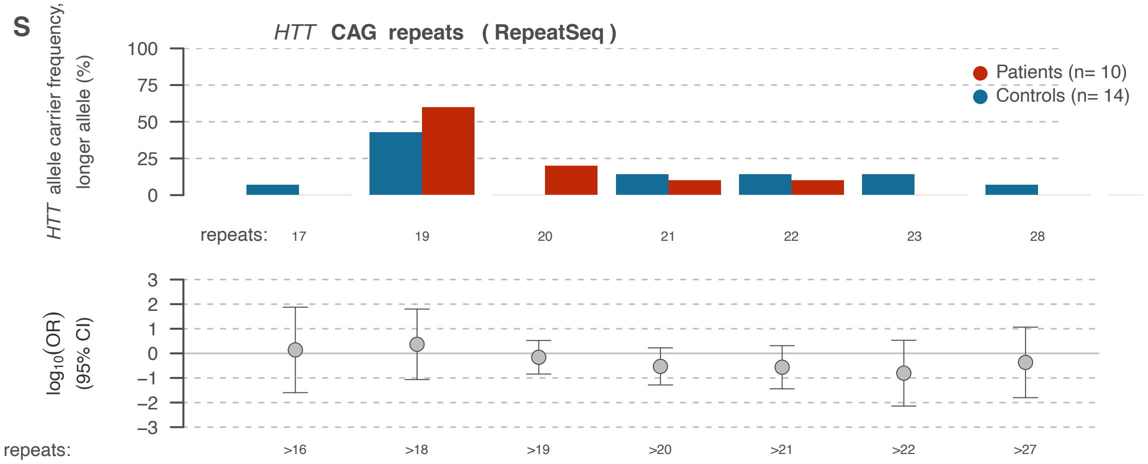

I
C9orf72 GGGGCC repeats (RepeatSeq)
C9orf72 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 26)
● Controls (n= 57)
repeats:
9
11
12
15
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats:
>8
>10
>11
>14
J
CACNA1A CAG repeats (RepeatSeq)
CACNA1A allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 116)
repeats:
4
7
11
12
13
14
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats:
>3
>6
>10
>11
>12
>13

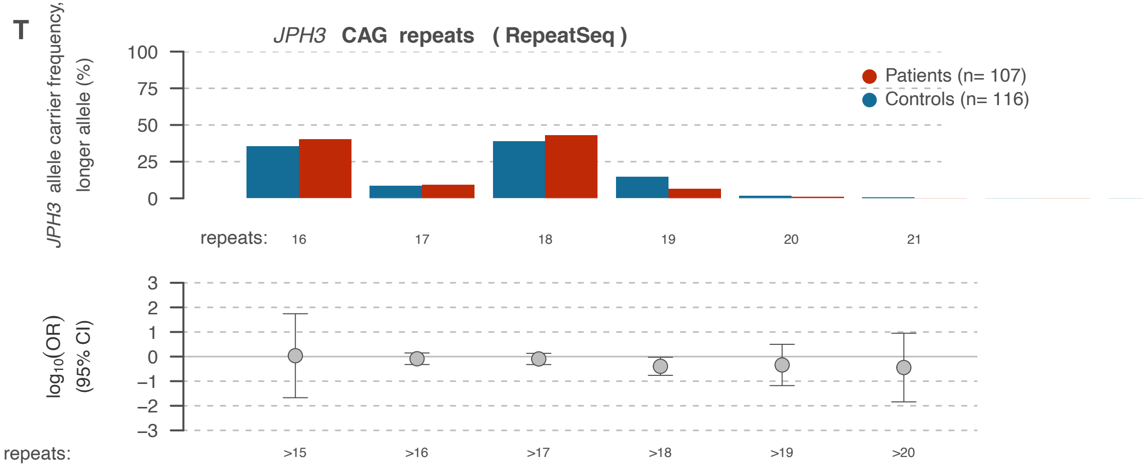

K
CBL CCG repeats (RepeatSeq)
CBL allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 54)
Controls (n= 115)
repeats:
8
11
12
13
15
20
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats:
>7
>10
>11
>12
>14
>19
L
CNBP CCTG repeats (RepeatSeq)
CNBP allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 39)
Controls (n= 104)
repeats:
8
14
15
16
17
18
20
21
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
*
*
*
repeats:
>7
>13
>14
>15
>16
>17
>19
>20

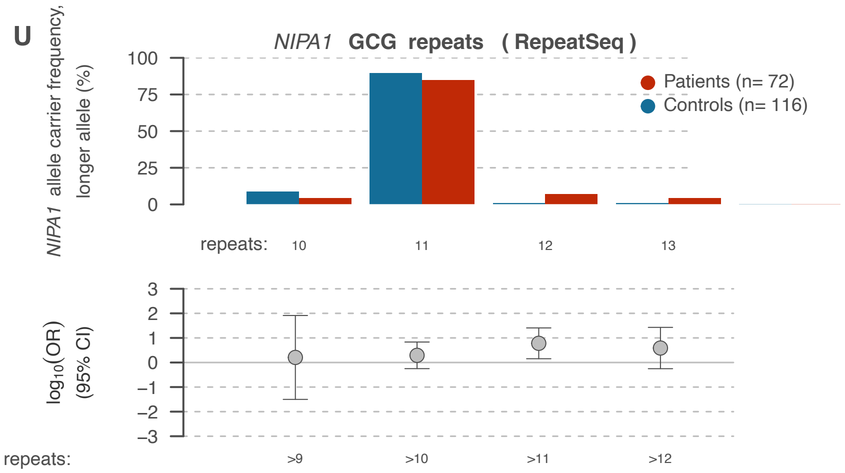

M
DIP2B CGG repeats (RepeatSeq)
DIP2B allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 89)
● Controls (n= 114)
repeats: 8 9 10 12 15 16 17 18 19 21 22
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
*
*
repeats: >7 >8 >9 >11 >14 >15 >16 >17 >18 >20 >21
N
DMPK CTG repeats (RepeatSeq)
DMPK allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 11)
● Controls (n= 22)
repeats: 6 12 13 14 21
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >5 >11 >12 >13 >20

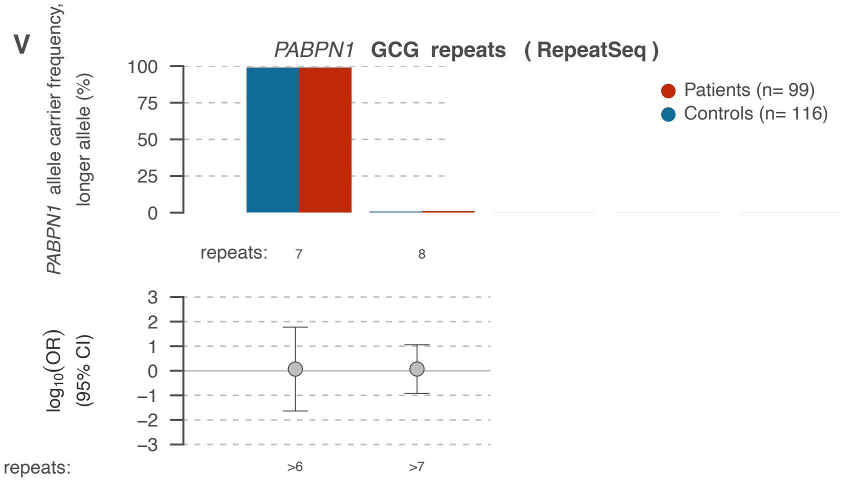

O
FMR1 CGG repeats (RepeatSeq)
FMR1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 10)
Controls (n= 60)
repeats: 21 23 25 27 28 31 32 34 35 36 37 38 45
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >20 >22 >24 >26 >27 >30 >31 >33 >34 >35 >36 >37 >44
P
FXN GAA repeats (RepeatSeq)
FXN allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 31)
Controls (n= 62)
repeats: 7 8 9 10 11 20
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >7 >8 >9 >10 >19


Q
GIPC1 GGC repeats (RepeatSeq)
GIPC1 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 48)
● Controls (n= 116)
repeats:
14
15
16
17
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats:
>13
>14
>15
>16
R
GLS GCA repeats (RepeatSeq)
GLS allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 41)
● Controls (n= 116)
repeats:
8
9
10
11
12
13
14
15
16
17
18
19
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats:
>7
>8
>9
>10
>11
>12
>13
>14
>15
>16
>17
>18

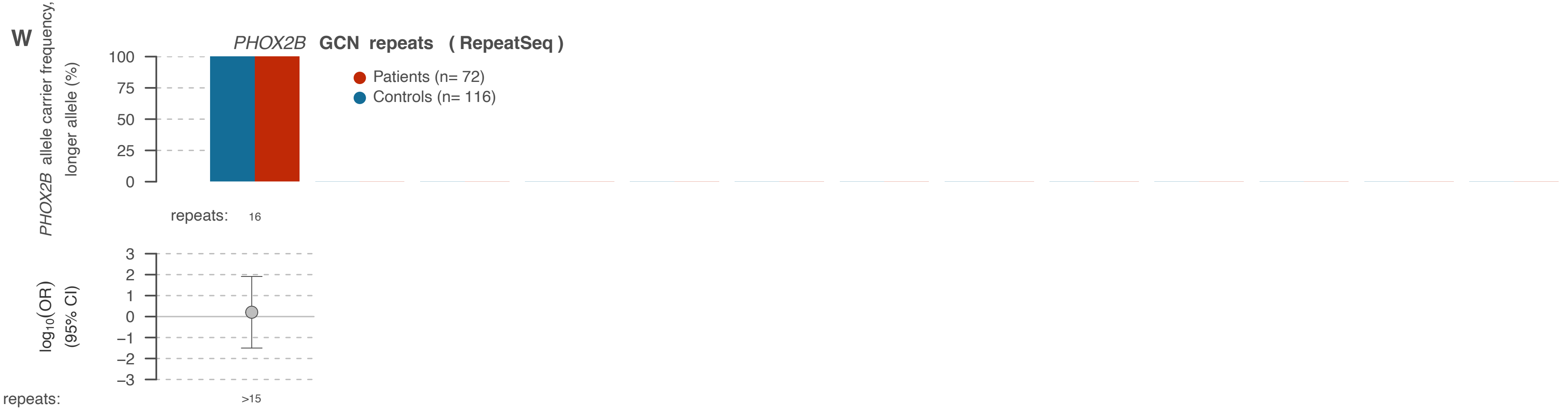

S
HTT CAG repeats (RepeatSeq)
HTT allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 10)
● Controls (n= 14)
repeats: 17 19 20 21 22 23 28
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >16 >18 >19 >20 >21 >22 >27


T
JPH3 CAG repeats (RepeatSeq)
JPH3 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 107)
● Controls (n= 116)
repeats: 16 17 18 19 20 21
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >15 >16 >17 >18 >19 >20


U
NIPA1 GCG repeats (RepeatSeq)
NIPA1 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 72)
● Controls (n= 116)
repeats: 10 11 12 13
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >9 >10 >11 >12

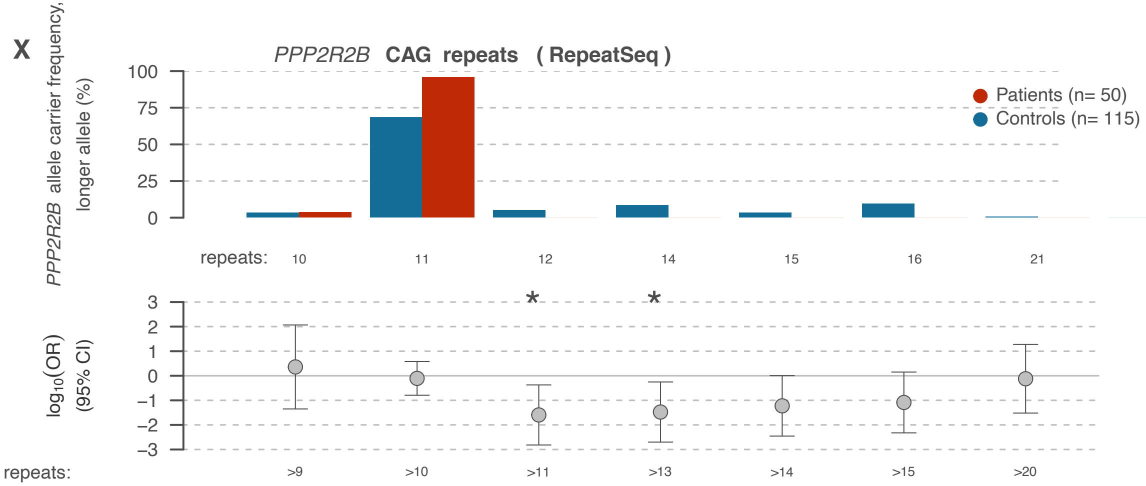

V
PABPN1 GCG repeats (RepeatSeq)
PABPN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 99)
● Controls (n= 116)
repeats: 7 8
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >6 >7

W

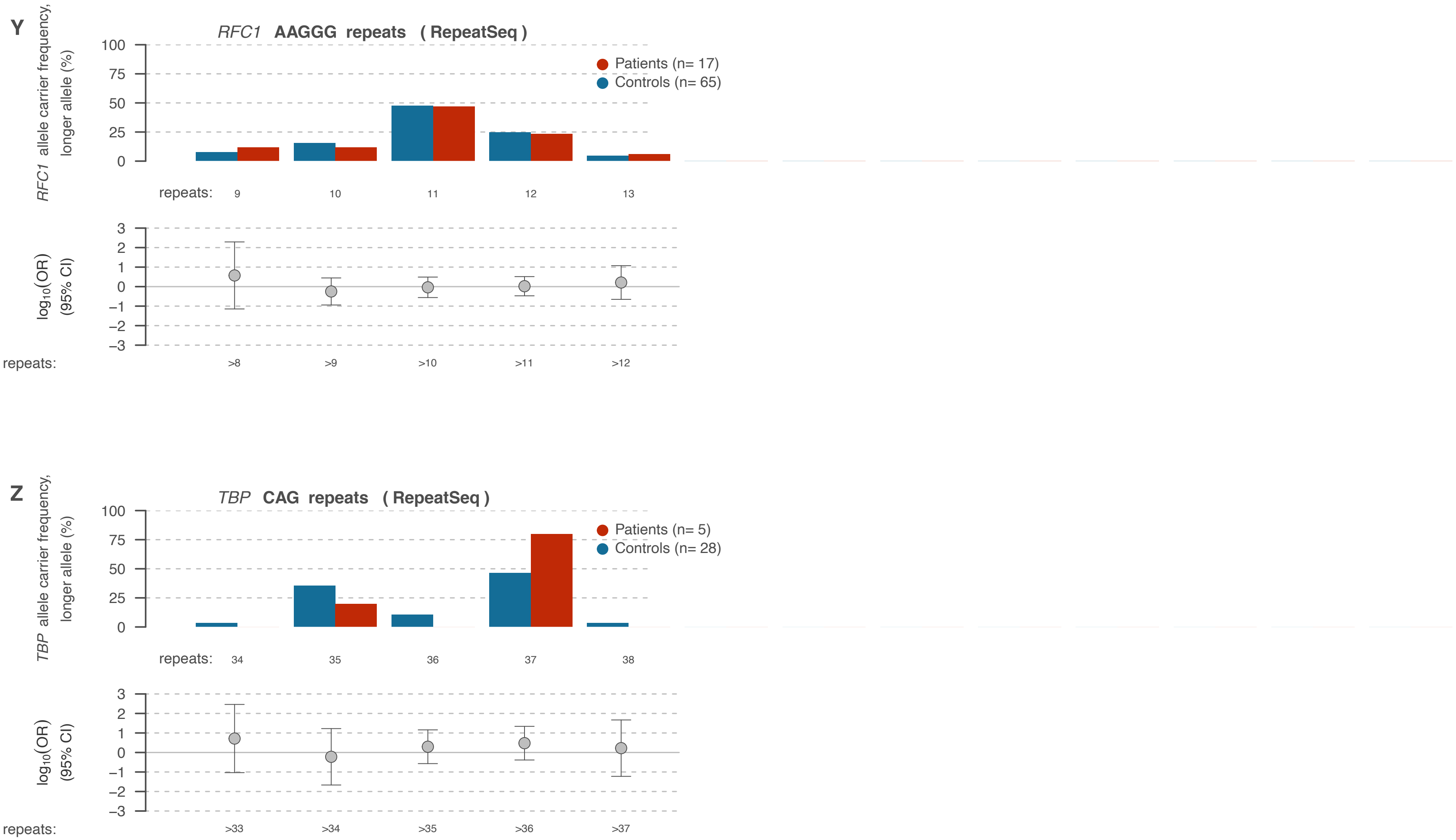


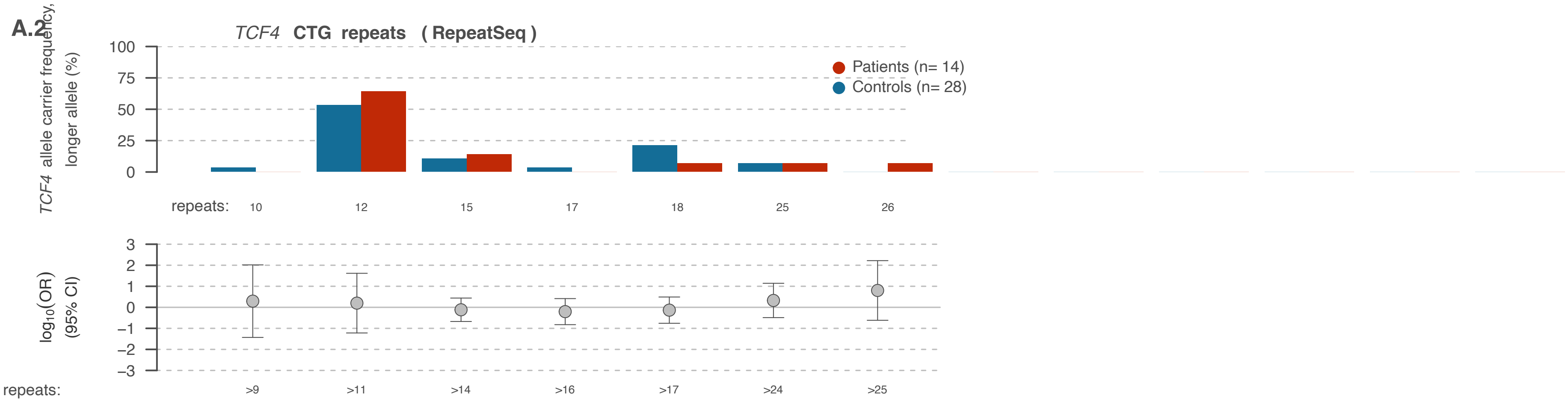
PHOX2B GCN repeats (RepeatSeq)
PHOX2B allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 72)
● Controls (n= 116)
repeats: 16
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >15

X



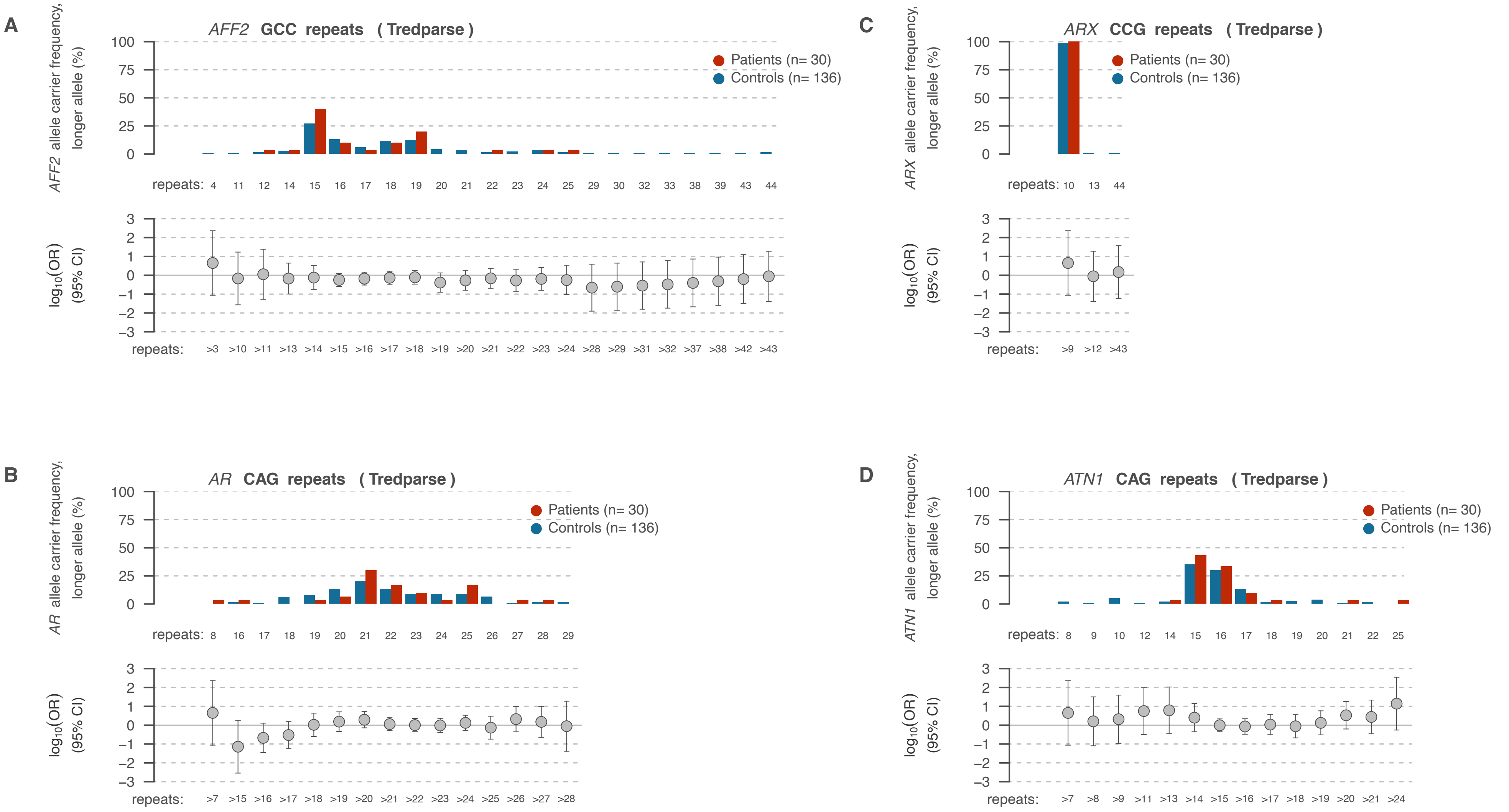
PPP2R2B CAG repeats (RepeatSeq)
PPP2R2B allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 50)
● Controls (n= 115)
repeats: 10 11 12 14 15 16 21
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
*
*
repeats: >9 >10 >11 >13 >14 >15 >20

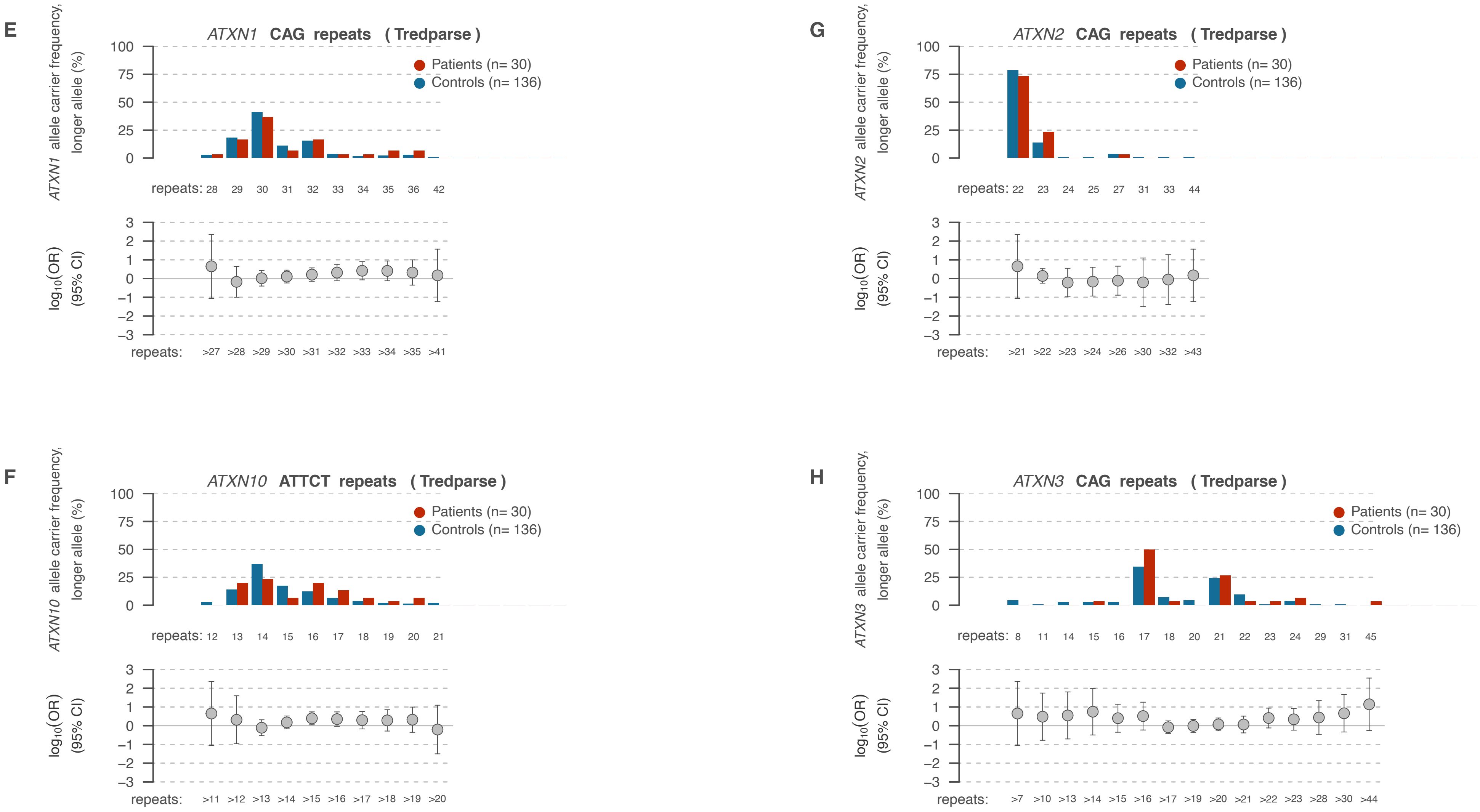

Y
RFC1 AAGGG repeats (RepeatSeq)
RFC1 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 17)
Controls (n= 65)
repeats: 9 10 11 12 13
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >9 >10 >11 >12
Z
TBP CAG repeats (RepeatSeq)
TBP allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 5)
Controls (n= 28)
repeats: 34 35 36 37 38
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >33 >34 >35 >36 >37

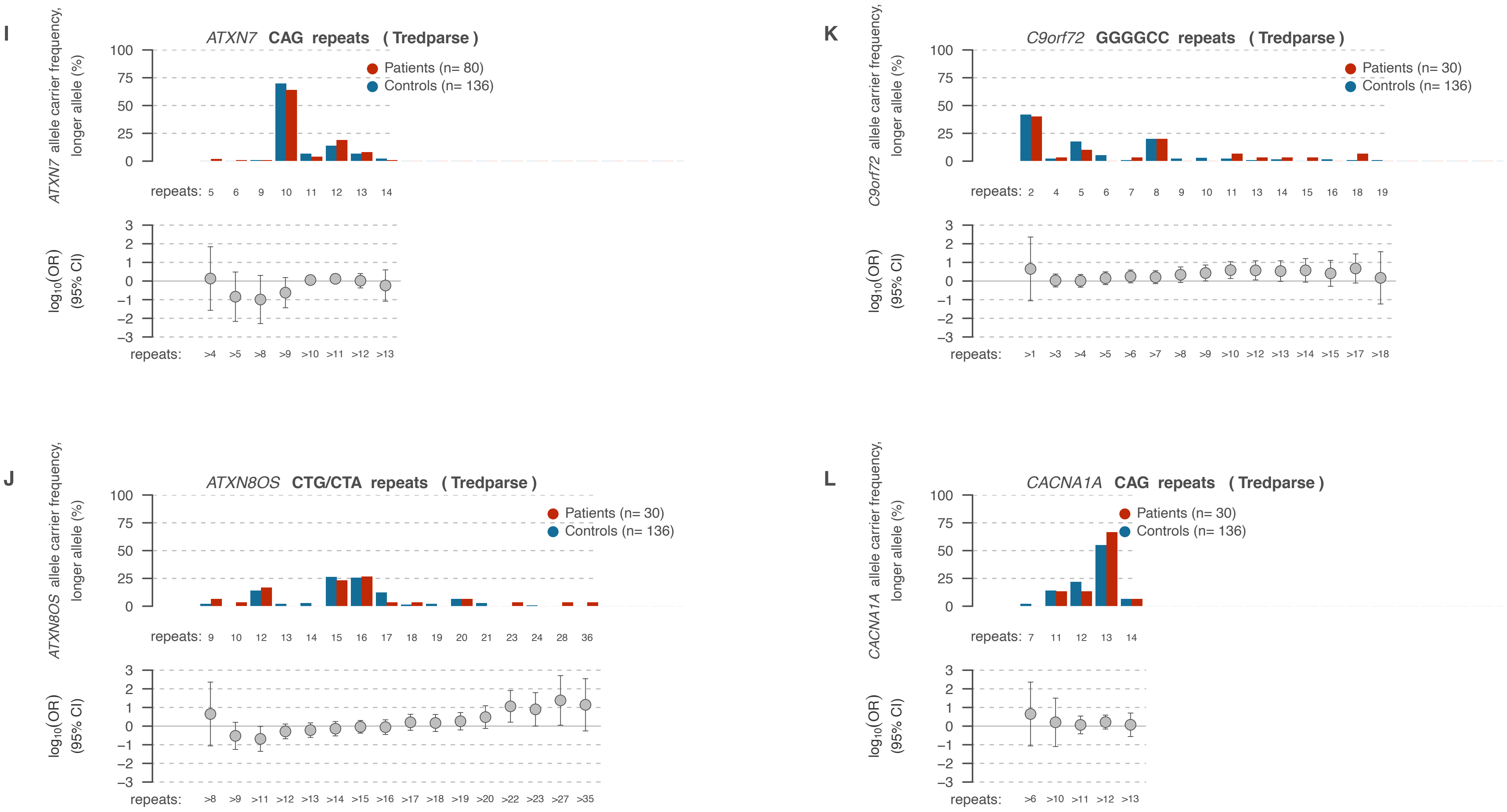


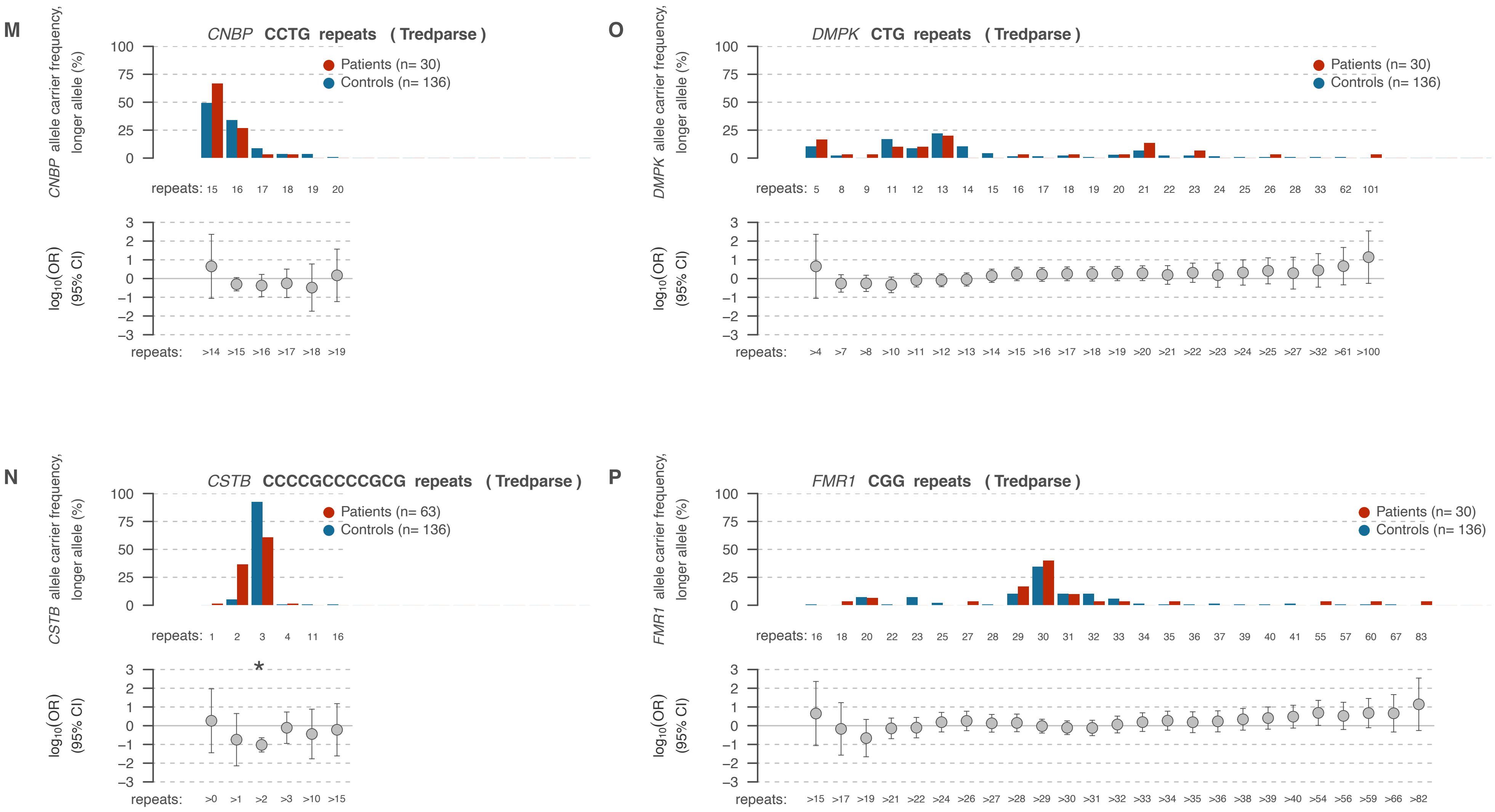
Supplementary Figure 3.18: RepeatSeq prediction of STR lengths in epilepsy patients

For each gene genotyped with RepeatSeq the allele lengths in epilepsy patients are compared to 136 Irish controls. The upper plot shows the predicted allele lengths and the lower plot shows the OR. An asterisks indicate a significant OR. The epilepsy results include PCR-free WGS samples, PCR WGS samples and WES sample if an RMSD below one was observed when comparing WES results to WGS results for a given gene.

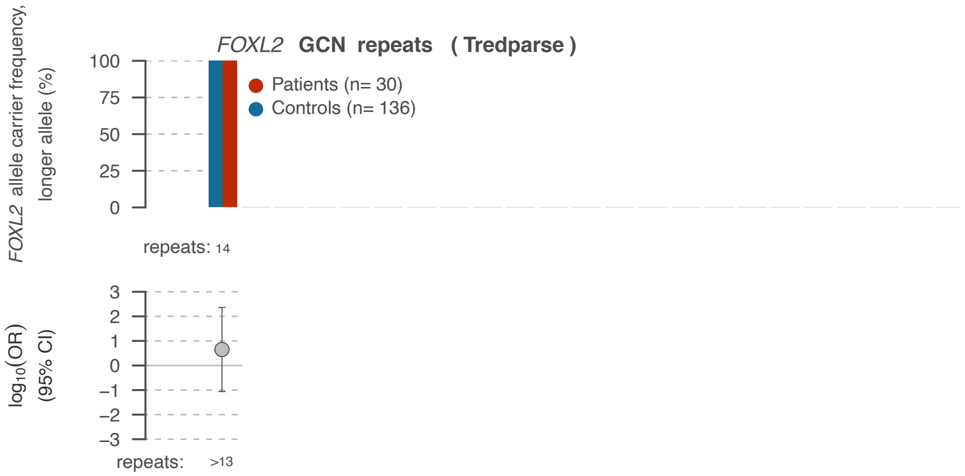

A
AFF2 GCC repeats (Tredparse)
AFF2 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 30)
Controls (n= 136)
repeats: 4 11 12 14 15 16 17 18 19 20 21 22 23 24 25 29 30 32 33 38 39 43 44
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >3 >10 >11 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >28 >29 >31 >32 >37 >38 >42 >43
B
AR CAG repeats (Tredparse)
AR allele carrier frequency, longer allele (%)
Patients (n= 30)
Controls (n= 136)
repeats: 8 16 17 18 19 20 21 22 23 24 25 26 27 28 29
log10(OR) (95% CI)
repeats: >7 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28
C
ARX CCG repeats (Tredparse)
ARX allele carrier frequency, longer allele (%)
Patients (n= 30)
Controls (n= 136)
repeats: 10 13 44
log10(OR) (95% CI)
repeats: >9 >12 >43
D
ATN1 CAG repeats (Tredparse)
ATN1 allele carrier frequency, longer allele (%)
Patients (n= 30)
Controls (n= 136)
repeats: 8 9 10 12 14 15 16 17 18 19 20 21 22 25
log10(OR) (95% CI)
repeats: >7 >8 >9 >11 >13 >14 >15 >16 >17 >18 >19 >20 >21 >24


E
ATXN1 CAG repeats (Tredparse)
ATXN1 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 30)
● Controls (n= 136)
repeats: 28 29 30 31 32 33 34 35 36 42
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >27 >28 >29 >30 >31 >32 >33 >34 >35 >41
F
ATXN10 ATTCT repeats (Tredparse)
ATXN10 allele carrier frequency, longer allele (%)
● Patients (n= 30)
● Controls (n= 136)
repeats: 12 13 14 15 16 17 18 19 20 21
log10(OR) (95% CI)
repeats: >11 >12 >13 >14 >15 >16 >17 >18 >19 >20
G
ATXN2 CAG repeats (Tredparse)
ATXN2 allele carrier frequency, longer allele (%)
● Patients (n= 30)
● Controls (n= 136)
repeats: 22 23 24 25 27 31 33 44
log10(OR) (95% CI)
repeats: >21 >22 >23 >24 >26 >30 >32 >43
H
ATXN3 CAG repeats (Tredparse)
ATXN3 allele carrier frequency, longer allele (%)
● Patients (n= 30)
● Controls (n= 136)
repeats: 8 11 14 15 16 17 18 20 21 22 23 24 29 31 45
log10(OR) (95% CI)
repeats: >7 >10 >13 >14 >15 >16 >17 >19 >20 >21 >22 >23 >28 >30 >44

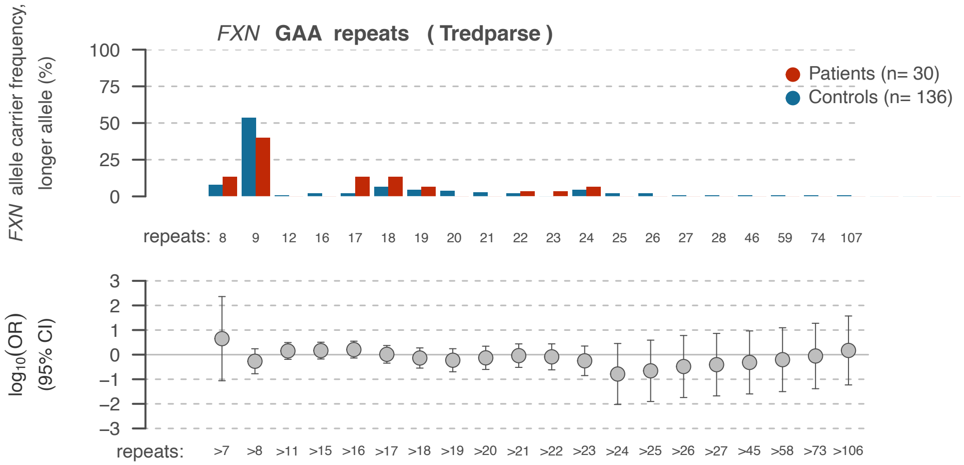

I
ATXN7 CAG repeats (Tredparse)
Patients (n= 80)
Controls (n= 136)
ATXN7 allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 5 6 9 10 11 12 13 14
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >4 >5 >8 >9 >10 >11 >12 >13
K
C9orf72 GGGGCC repeats (Tredparse)
Patients (n= 30)
Controls (n= 136)
C9orf72 allele carrier frequency, longer allele (%)
repeats: 2 4 5 6 7 8 9 10 11 13 14 15 16 18 19
log10(OR) (95% CI)
repeats: >1 >3 >4 >5 >6 >7 >8 >9 >10 >12 >13 >14 >15 >17 >18
J
ATXN8OS CTG/CTA repeats (Tredparse)
Patients (n= 30)
Controls (n= 136)
ATXN8OS allele carrier frequency, longer allele (%)
repeats: 9 10 12 13 14 15 16 17 18 19 20 21 23 24 28 36
log10(OR) (95% CI)
repeats: >8 >9 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >22 >23 >27 >35
L
CACNA1A CAG repeats (Tredparse)
Patients (n= 30)
Controls (n= 136)
CACNA1A allele carrier frequency, longer allele (%)
repeats: 7 11 12 13 14
log10(OR) (95% CI)
repeats: >6 >10 >11 >12 >13

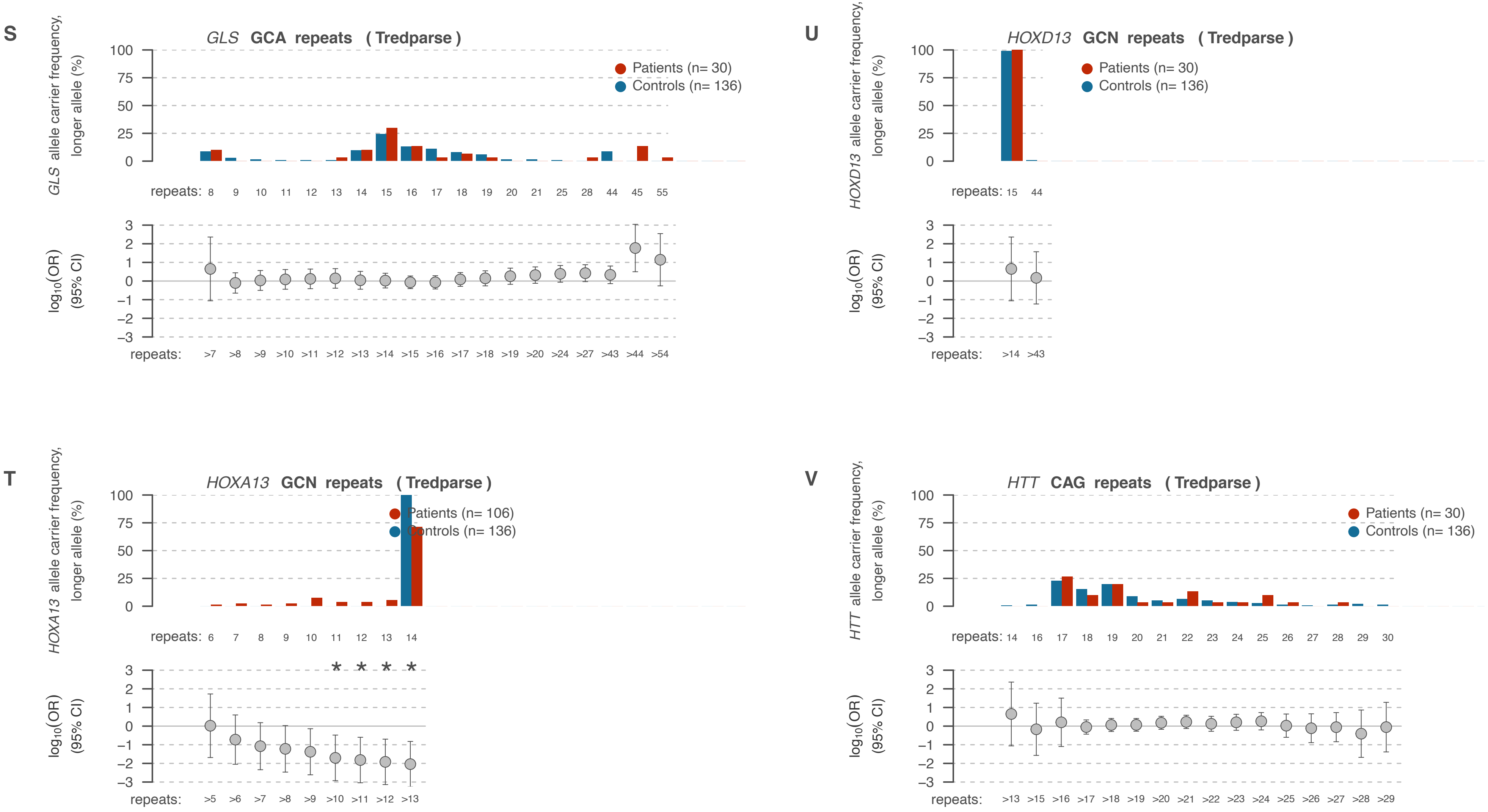

M
CNBP CCTG repeats (Tredparse)
CNBP allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 30)
Controls (n= 136)
repeats: 15 16 17 18 19 20
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >14 >15 >16 >17 >18 >19
O
DMPK CTG repeats (Tredparse)
DMPK allele carrier frequency, longer allele (%)
Patients (n= 30)
Controls (n= 136)
repeats: 5 8 9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 33 62 101
repeats: >4 >7 >8 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >27 >32 >61 >100
N
CSTB CCCCGCCCCGCG repeats (Tredparse)
CSTB allele carrier frequency, longer allele (%)
Patients (n= 63)
Controls (n= 136)
repeats: 1 2 3 4 11 16
*
repeats: >0 >1 >2 >3 >10 >15
P
FMR1 CGG repeats (Tredparse)
FMR1 allele carrier frequency, longer allele (%)
Patients (n= 30)
Controls (n= 136)
repeats: 16 18 20 22 23 25 27 28 29 30 31 32 33 34 35 36 37 39 40 41 55 57 60 67 83
repeats: >15 >17 >19 >21 >22 >24 >26 >27 >28 >29 >30 >31 >32 >33 >34 >35 >36 >38 >39 >40 >54 >56 >59 >66 >82

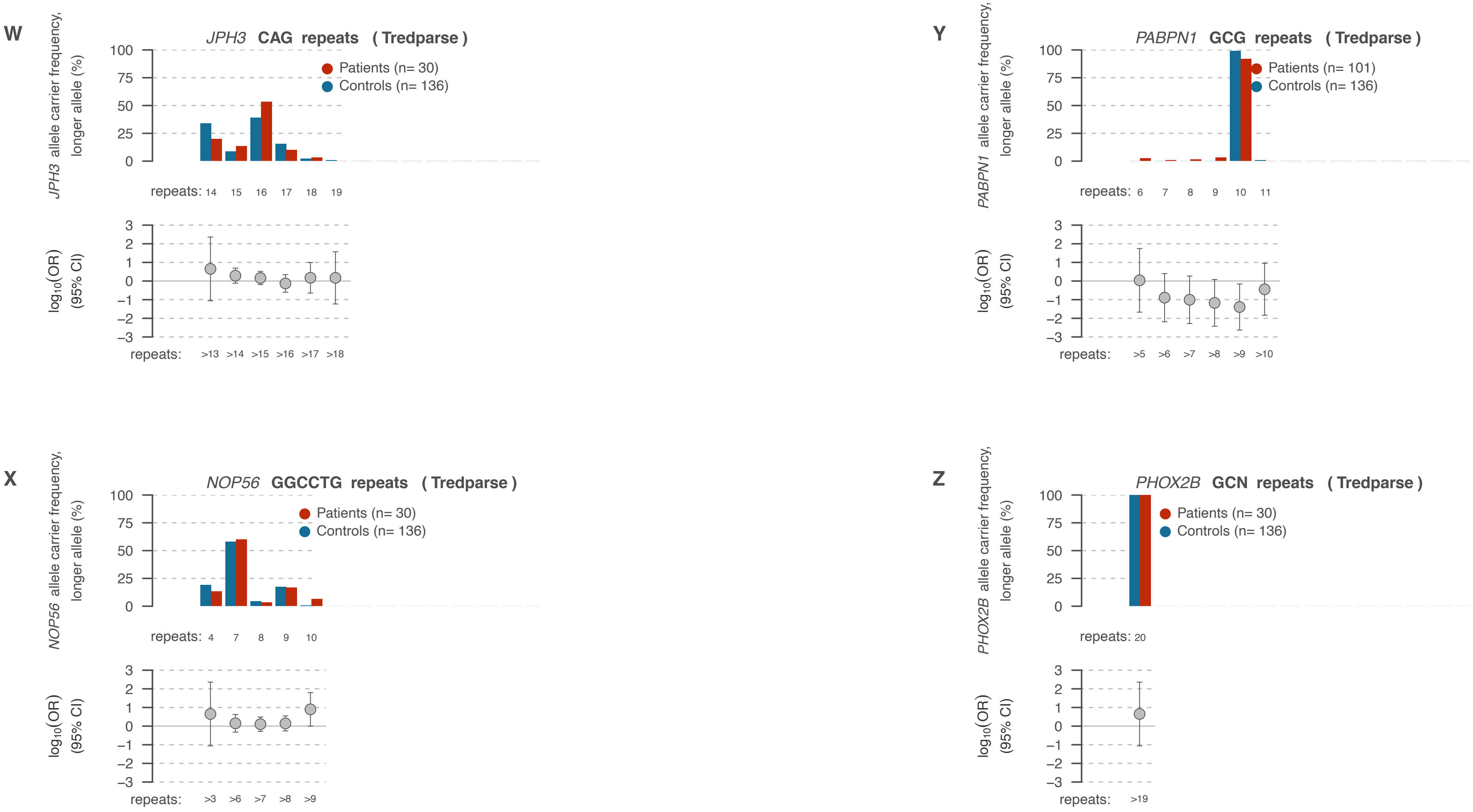
Q


FOXL2 GCN repeats (Tredparse)
Patients (n= 30)
Controls (n= 136)
FOXL2 allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 14
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >13

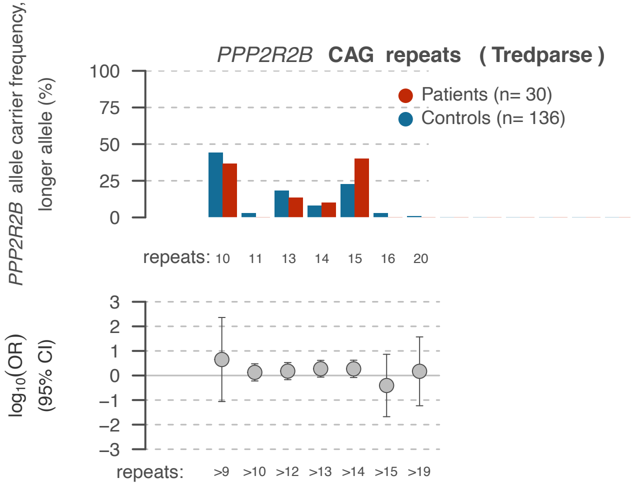
R


FXN GAA repeats (Tredparse)
Patients (n= 30)
Controls (n= 136)
FXN allele carrier frequency, longer allele (%)
100
75
50
25
0
repeats: 8 9 12 16 17 18 19 20 21 22 23 24 25 26 27 28 46 59 74 107
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >7 >8 >11 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >45 >58 >73 >106

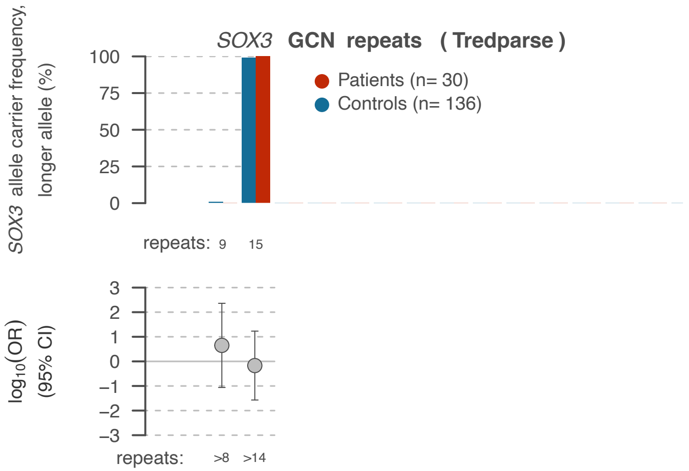

S
GLS GCA repeats (Tredparse)
GLS allele carrier frequency, longer allele (%)
Patients (n= 30)
Controls (n= 136)
repeats: 8 9 10 11 12 13 14 15 16 17 18 19 20 21 25 28 44 45 55
log10(OR) (95% CI)
repeats: >7 >8 >9 >10 >11 >12 >13 >14 >15 >16 >17 >18 >19 >20 >24 >27 >43 >44 >54
U
HOXD13 GCN repeats (Tredparse)
HOXD13 allele carrier frequency, longer allele (%)
Patients (n= 30)
Controls (n= 136)
repeats: 15 44
log10(OR) (95% CI)
repeats: >14 >43
T
HOXA13 GCN repeats (Tredparse)
HOXA13 allele carrier frequency, longer allele (%)
Patients (n= 106)
Controls (n= 136)
repeats: 6 7 8 9 10 11 12 13 14
* * * *
log10(OR) (95% CI)
repeats: >5 >6 >7 >8 >9 >10 >11 >12 >13
V
HTT CAG repeats (Tredparse)
HTT allele carrier frequency, longer allele (%)
Patients (n= 30)
Controls (n= 136)
repeats: 14 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
log10(OR) (95% CI)
repeats: >13 >15 >16 >17 >18 >19 >20 >21 >22 >23 >24 >25 >26 >27 >28 >29
100 75 50 25 0
3 2 1 0 −1 −2 −3


W
JPH3 CAG repeats (Tredparse)
JPH3 allele carrier frequency, longer allele (%)
100
75
50
25
0
Patients (n= 30)
Controls (n= 136)
repeats: 14 15 16 17 18 19
log₁₀(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >13 >14 >15 >16 >17 >18
X
NOP56 GGCCTG repeats (Tredparse)
NOP56 allele carrier frequency, longer allele (%)
Patients (n= 30)
Controls (n= 136)
repeats: 4 7 8 9 10
log₁₀(OR) (95% CI)
repeats: >3 >6 >7 >8 >9
Y
PABPN1 GCG repeats (Tredparse)
PABPN1 allele carrier frequency, longer allele (%)
Patients (n= 101)
Controls (n= 136)
repeats: 6 7 8 9 10 11
log₁₀(OR) (95% CI)
repeats: >5 >6 >7 >8 >9 >10
Z
PHOX2B GCN repeats (Tredparse)
PHOX2B allele carrier frequency, longer allele (%)
Patients (n= 30)
Controls (n= 136)
repeats: 20
log₁₀(OR) (95% CI)
repeats: >19

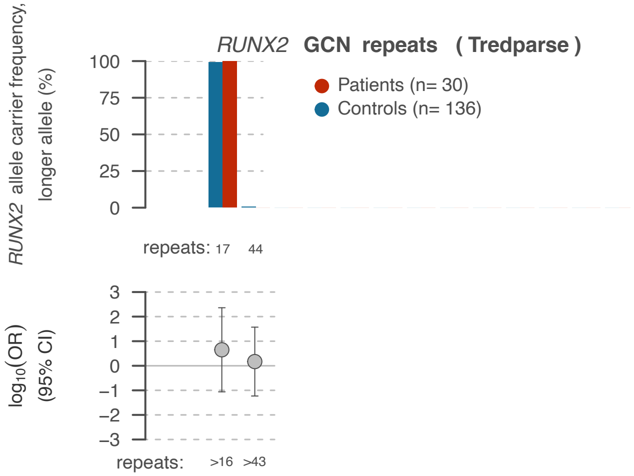
A.2


PPP2R2B CAG repeats (Tredparse)
PPP2R2B allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 30)
● Controls (n= 136)
repeats: 10 11 13 14 15 16 20
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >9 >10 >12 >13 >14 >15 >19

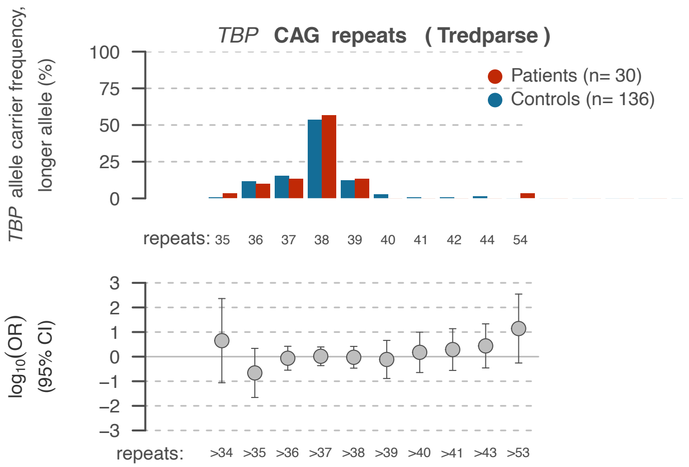
C.2

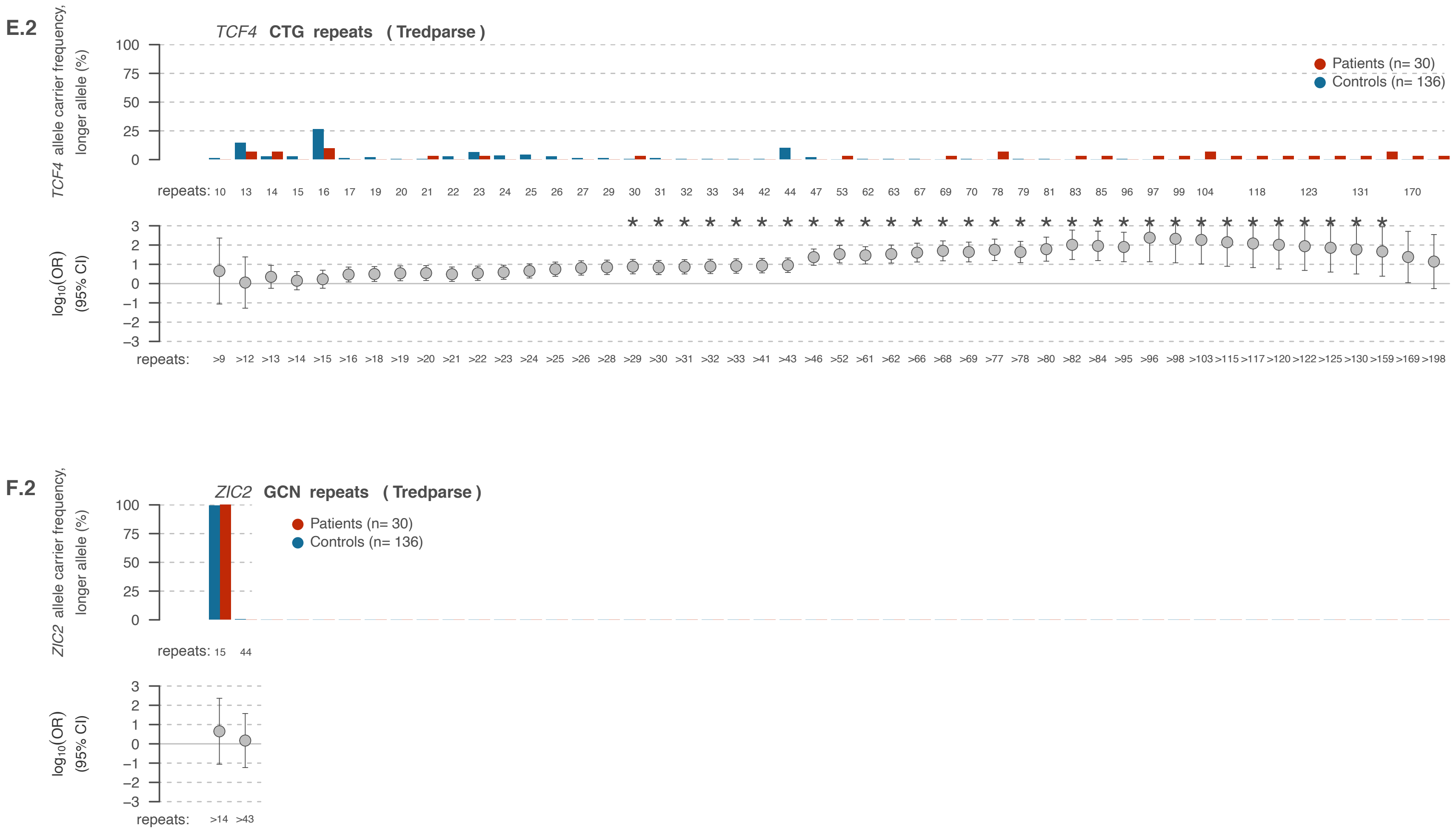

SOX3 GCN repeats (Tredparse)
SOX3 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 30)
● Controls (n= 136)
repeats: 9 15
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >8 >14

B.2


RUNX2 GCN repeats (Tredparse)
RUNX2 allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 30)
● Controls (n= 136)
repeats: 17 44
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >16 >43

D.2


TBP CAG repeats (Tredparse)
TBP allele carrier frequency, longer allele (%)
100
75
50
25
0
● Patients (n= 30)
● Controls (n= 136)
repeats: 35 36 37 38 39 40 41 42 44 54
log10(OR) (95% CI)
3
2
1
0
−1
−2
−3
repeats: >34 >35 >36 >37 >38 >39 >40 >41 >43 >53



Supplementary Figure 3.19: TREDPARSE prediction of STR lengths in epilepsy patients

For each gene genotyped with TREDPARSE the allele lengths in epilepsy patients are compared to 136 Irish controls. The upper plot shows the predicted allele lengths and the lower plot shows the OR. An asterisks indicate a significant OR. The epilepsy results include PCR-free WGS samples, PCR WGS samples and WES sample if an RMSD below one was observed when comparing WES results to WGS results for a given gene.

exSTRa Predicted Significant Repeat Expansions in *LRP12*

Identical Reads	Repeat Count	Sequence
Patient: EP5A		
1	13	GACGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGAG
1	≥12	CCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGAG
13	12	GACGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGAG
1	≥11	GCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGAG
1	≥10	CGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGAG
30	9	GACGCCGCCGCCGCCGCCGCCGCCGCCGCCGAG
1	9	ACGCCGCCGCCGCCGCCGCCGCCGCCGCCGAG
2	≥8	CGCCGCCGCCGCCGCCGCCGCCGCCGAG
1	≥6	CCGCCGCCGCCGCCGCCGCCGAG
1	≥6	CGCCGCCGCCGCCGCCGCCGAG
2	≥5	GCCGCCGCCGCCGCCGAG
1	≥3	CCGCCGCCGCCGAG
1	≥1	CGCCGAG
Patient: EP6A		
13	12	GACGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGAG
1	≥7	CGCCGCCGCCGCCGCCGCCGCCGAG
1	≥7	GCCGCCGCCGCCGCCGCCGCCGAG
25	5	GACGCCGCCGCCGCCGCCGAG
1	≥4	CGCCGCCGCCGCCGAG

Supplementary Figure 3.20 : Exploration of samples with exSTRa predicted LRP12 REs

exSTRa predicts two epilepsy patients to have significant repeats in LRP12. Reads here are directly extracted from the patient bam files as there is insufficient information from other tools to make a conclusion as to the veracity of these repeats. It is seen that while some stutter error is visible, both patients appear to have alleles of 9/12 and 5/12, well within the non-pathogenic range.

exSTRa Predicted Significant Repeat Expansions in *SAMD2*

Identical Reads	Repeat Count	Sequence
Patient: EP13A		
1	≥1	CAAATAAAAT
1	≥10	CAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAA
1	≥10	CAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATA
1	≥14	CAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAAT
3	≥14	CAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAATAAAATA
1	≥15	CAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAA
1	≥16	CAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAA
1	≥16	CAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATA
4	20	CAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAAATGAA
1	≥18	AATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAAATGAA
1	≥17	AATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAAATGAA
1	≥16	AATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAAATGAA
1	≥8	AAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAAATGAA
1	≥1	CAAAAATAAAA
1	≥10	CAAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATA
6	12	CAAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAT
1	≥6	AATAAAATAAAATAAAATAAAATAAAATAAAATAAT
1	≥4	TAAAATAAAATAAAATAAAATAAT
1	≥2	ATAAAATAAAATAAAATAAT
1	≥2	AAATAAAATAAAATAAT
1	≥2	ATAAAATAAAATAAT
1	≥2	TAAAATAAAATAAT
1	N/A	ATAAT

Supplementary Figure 3.21 : Exploration of samples with exSTRa predicted *SAMD12* REs

exSTRa predicts a single epilepsy patient to have a significant repeats in *SAMD12*. Reads here are directly extracted from the patient bam file as there is insufficient information from other tools to make a conclusion as to the veracity of these repeats. It is seen that while some stutter error is visible, the patient appears to have heterozygous 12/20 repeats within the non-pathogenic range.

exSTRa Predicted Significant Repeats in *NUTM2B*

Identical Reads	Repeat Count	Sequence
Patient: EP6A		
1	>7	AGGAAGCGGCGGGGCGGCGGCGGCGG
1	>9	AGGAAGCGGCGGGGCGGCGGCGGCGGCGGC
1	>11	AGGAAGCGGCGGGGCGGCGGCGGCGGCGGCGGCGGC
4	13	AGGAAGCGGCGGGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCCGGGAA
Patient: EP7A		
1	>7	AGGAAGCGGCGGGGCGGCGGCGGCGGC
1	>9	AGGAAGCGGCGGGGCGGCGGCGGCGGCGGCGGCGGCG
1	>11	AGGAAGCGGCGGGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCCGG
8	13	AGGAAGCGGCGGGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCCGGGAA

Supplementary Figure 3.22 : Exploration of samples with exSTRa predicted *NUTM2B* REs

exSTRa predicts two epilepsy patients to have significant repeats in *NUTM2B*. Reads here are directly extracted from the patient bam files as there is insufficient information from other tools to make a conclusion as to the veracity of these repeats. Both patients appear to be homozygous for 13 repeats, well within the non-pathogenic range.

exSTRa Predicted Significant Repeats in *NOTCH2*

Patient: EP5A

Read Count	Repeat Count	Sequence
1	≥1	TGCCCAGGCG
4	≥1	TGCCCAGGCGG
1	≥2	TGCCCAGGCGGC
1	≥2	TGCCCAGGCGGCG
1	≥3	TGCCCAGGCGGCGGCG
2	≥3	TGCCCAGGCGGCGGCGGA
1	≥4	TGCCCAGGCGGCGGCGGCG
30	4	TGCCCAGGCGGCGGCGGCGGA
2	≥5	TGCCCAGGCGGCGGCGGCGGCGG
2	5	TGCCCAGGCGGCGGCGGCGGCGGA
2	≥6	TGCCCAGGCGGCGGCGGCGGCGGCG
2	≥7	TGCCCAGGCGGCGGCGGCGGCGGCGGC
1	≥8	TGCCCAGGCGGCGGCGGCGGCGGCGGCGG
197	7	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGA
7	9	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGCGGCGGA
1	N/A	TGCCCAGGCGGCGTCGGCGGCGGCGGCGGA
1	≥13	TGCCCAGGCGGGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGC
1	≥14	TGCCCAGGCGGGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGC
2	≥15	TGCCCAGGCGGGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGC
1	N/A	TGCCCAGGCTGCGGCCGCGGCCGCTGCGGCGGCGGA

Patient: EP6A

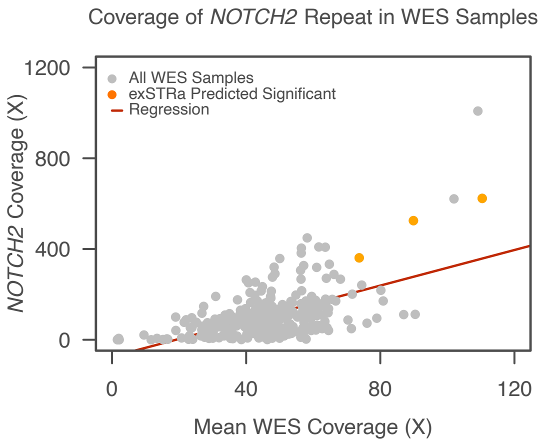
Read Count	Repeat Count	Sequence
2	≥1	TGCCCAGGCG
2	≥2	TGCCCAGGCGGCG
1	≥2	TGCCCAGGCGGCGG
3	≥3	TGCCCAGGCGGCGGC
4	≥3	TGCCCAGGCGGCGGCG
2	N/A	TGCCCAGGCGGCGGCGAGATCGGA
1	≥4	TGCCCAGGCGGCGGCGGC
66	4	TGCCCAGGCGGCGGCGGCGGA
1	≥5	TGCCCAGGCGGCGGCGGCGGC
2	≥6	TGCCCAGGCGGCGGCGGCGGCGGC
1	≥6	TGCCCAGGCGGCGGCGGCGGCGGCG
1	6	TGCCCAGGCGGCGGCGGCGGCGGCGGA
3	≥7	TGCCCAGGCGGCGGCGGCGGCGGCGGC
24	≥8	TGCCCAGGCGGCGGCGGCGGCGGCGGCGG
215	7	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGA
1	≥8	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGC
1	≥8	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGCG
5	9	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGCGGA
8	10	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGCGGCGGA
2	13	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGAG
32	13	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGA
31	14	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGC
1	N/A	TGCCCAGGCGGCTGCGGCGGCGGCGGCGG

Patient: EP7A

Read Count	Repeat Count	Sequence
1	≥1	TGCCCAGGC
2	≥1	TGCCCAGGCG
1	1	TGCCCAGGCGAG
1	≥1	TGCCCAGGCGG
1	≥2	TGCCCAGGCGGC
1	≥2	TGCCCAGGCGGCG
2	≥2	TGCCCAGGCGGCGG
2	≥3	TGCCCAGGCGGCGGC
1	≥3	TGCCCAGGCGGCGGCG
2	≥4	TGCCCAGGCGGCGGCGG
3	≥4	TGCCCAGGCGGCGGCGGC
1	4	TGCCCAGGCGGCGGCGGCAGA
1	≥5	TGCCCAGGCGGCGGCGGCGGC
3	≥6	TGCCCAGGCGGCGGCGGCGGCGG
1	≥6	TGCCCAGGCGGCGGCGGCGGCGGC
3	≥6	TGCCCAGGCGGCGGCGGCGGCGGCG
1	6	TGCCCAGGCGGCGGCGGCGGCGGCGAG
1	6	TGCCCAGGCGGCGGCGGCGGCGGCGGCAG
3	≥7	TGCCCAGGCGGCGGCGGCGGCGGCGGCGG
338	7	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGA
5	≥8	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGCGGA
1	≥9	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGCGGC
15	9	TGCCCAGGCGGCGGCGGCGGCGGCGGCGGCGGCGGA
1	6	TGCCCAGGCGGCGGCGGCGGCGGCGTCGGA
1	6	TGCCCAGGCGGCGTCGGCGGCGGCGGCGGA

Supplementary Figure 3.23 : Exploration of samples with exSTRa predicted NOTCH2 REs

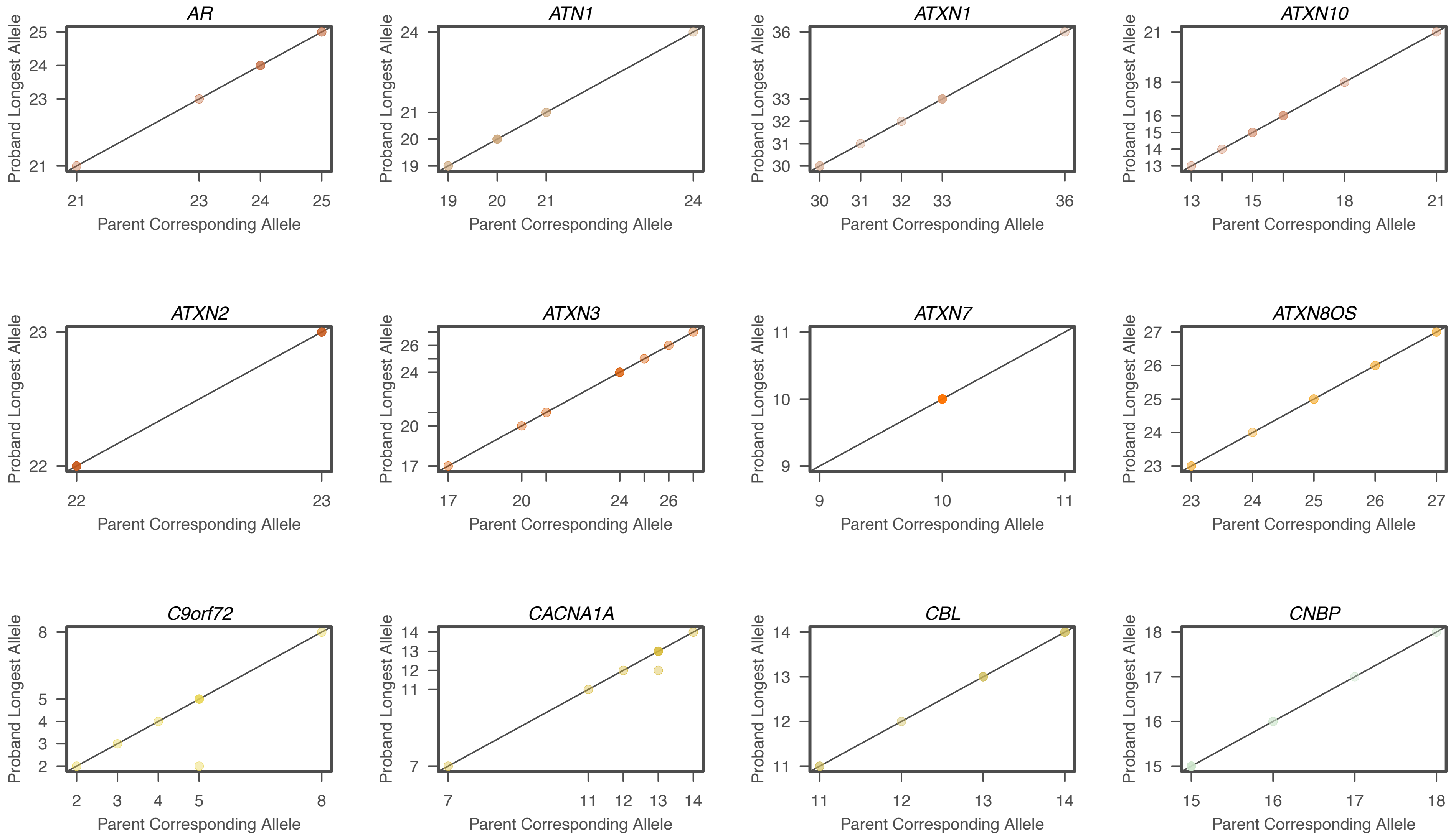
exSTRa predicts three epilepsy patients to have significant repeats in NOTCH2. Reads here are directly extracted from the patient bam files as there is insufficient information from other tools to make a conclusion as to the veracity of these repeats. While there is some variability in reads, likely resulting from stutter error during the sequencing of these WES samples, the reads do not support an expansion at this locus.

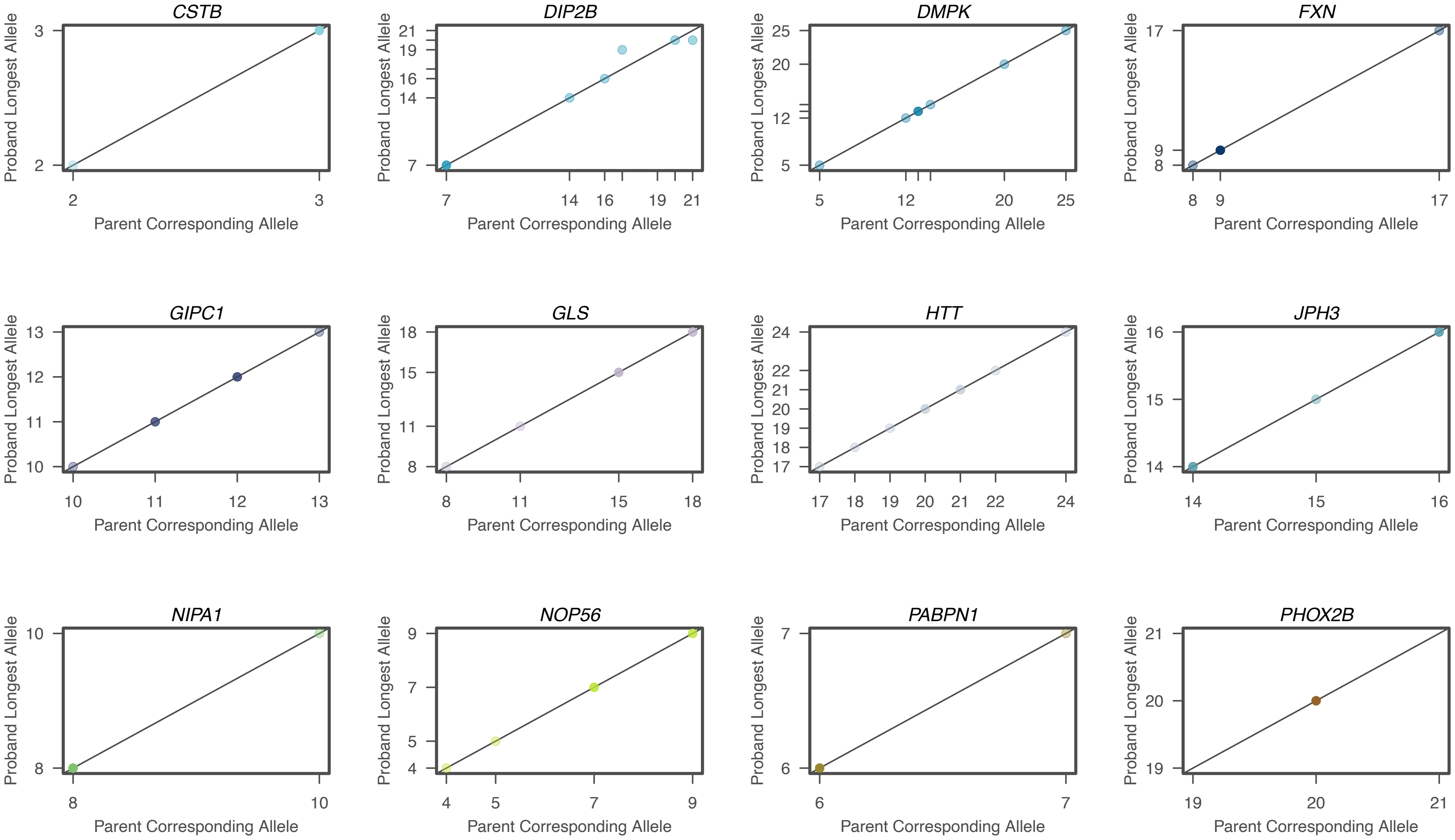


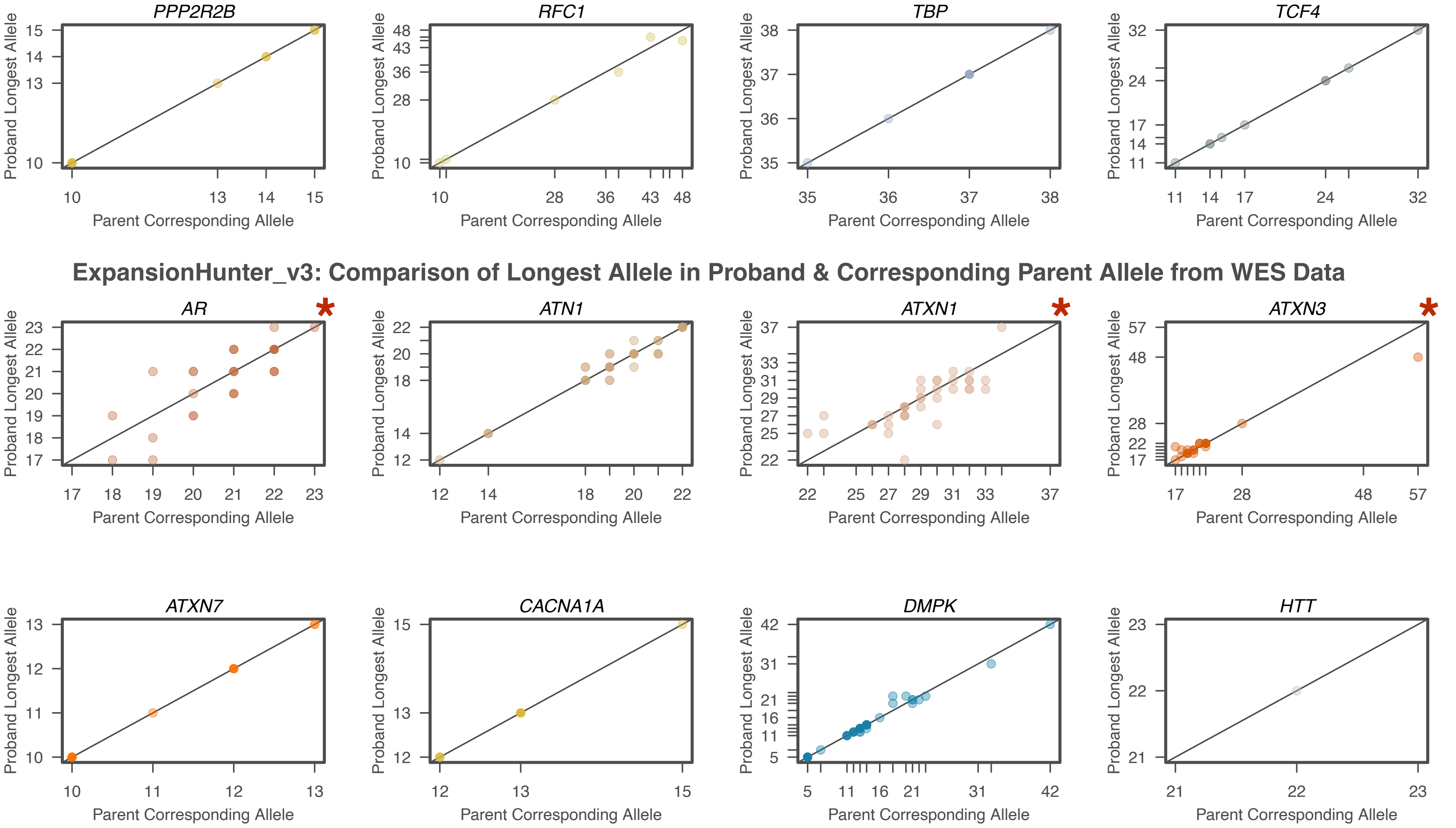
Supplementary Figure 3.24 : Coverage of samples with exSTRa predicted NOTCH2 REs

While samples predicted to have a repeat at this locus (orange) are deeply sequenced at this locus, this is found to be proportional to their overall exome-wide coverage.

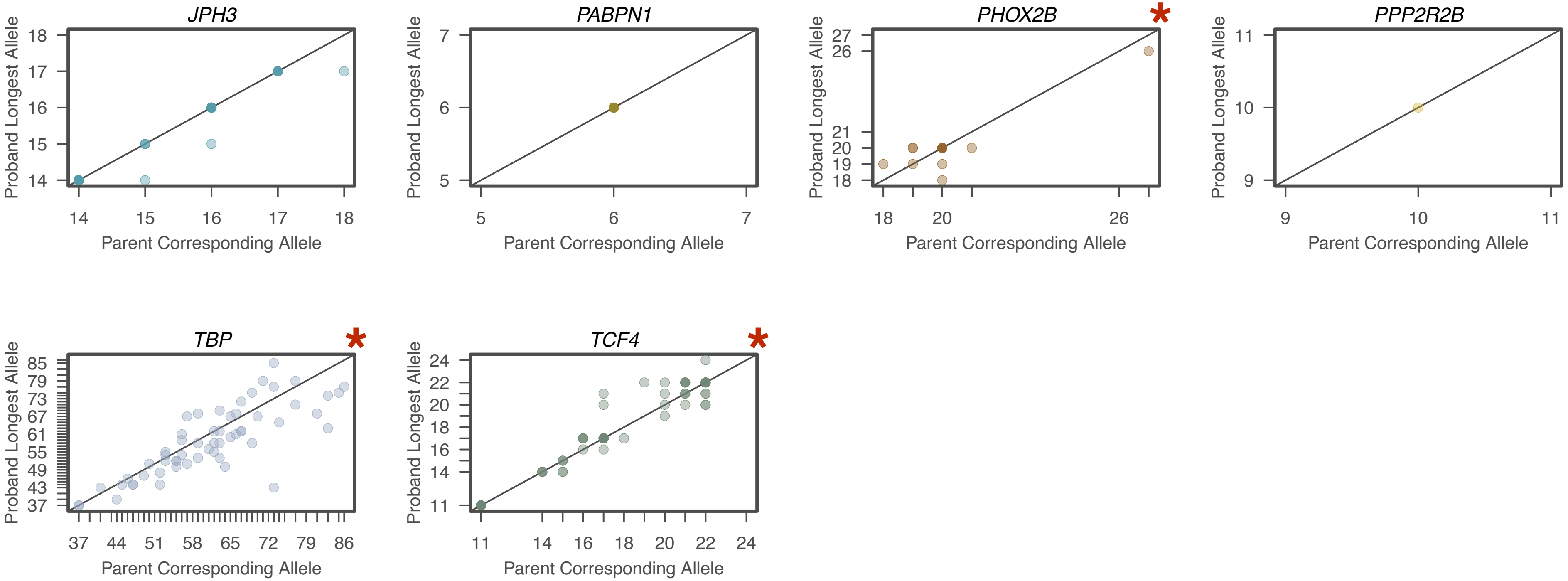
ExpansionHunter_v3: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WGS PCR Data




CSTB
DIP2B
DMPK
FXN
GIPC1
GLS
HTT
JPH3
NIPA1
NOP56
PABPN1
PHOX2B
Proband Longest Allele
Parent Corresponding Allele
2
3
7
14
16
19
21
5
12
20
25
8
9
17
10
11
13
15
18
22
24
4



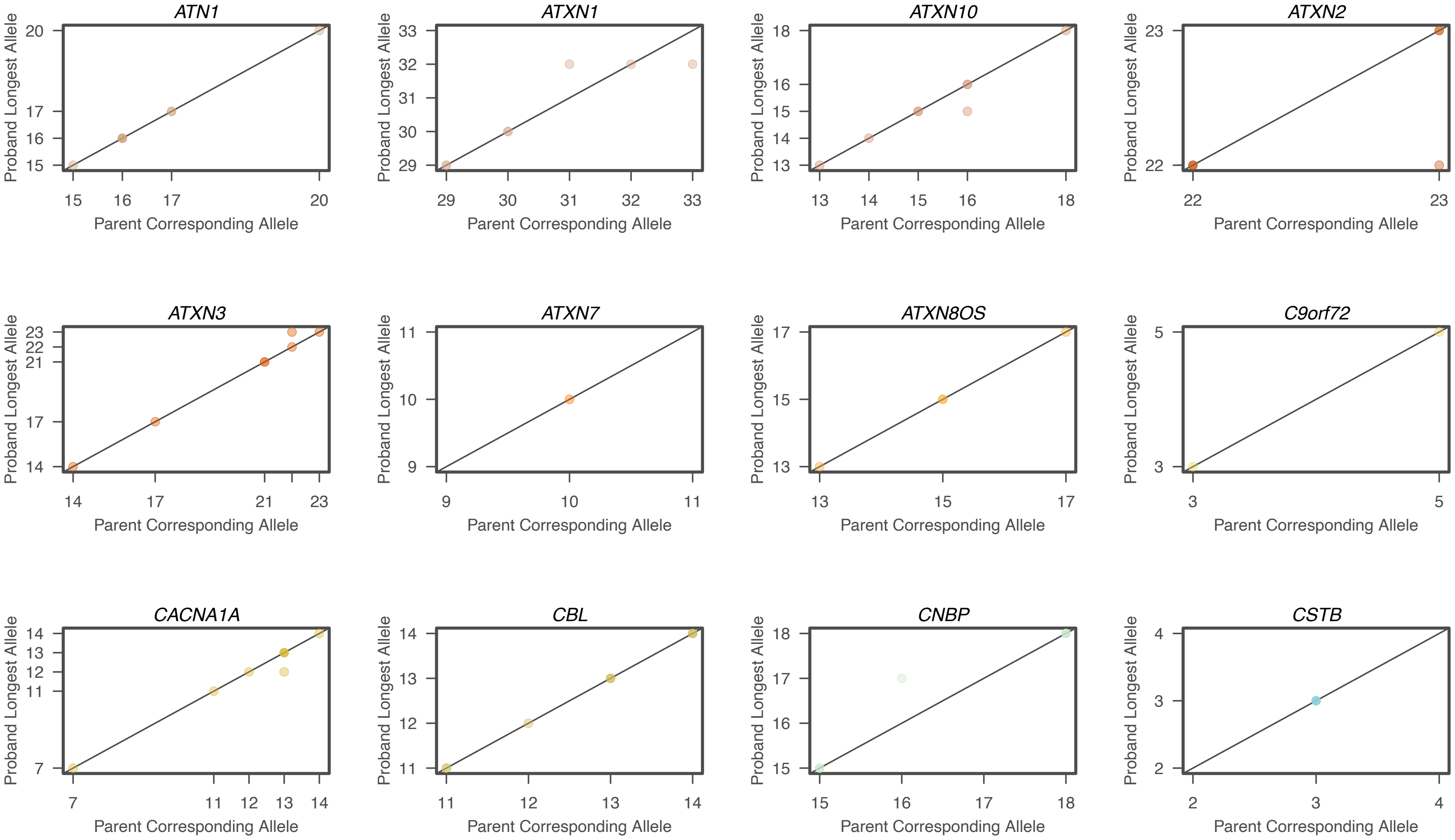
ExpansionHunter_v3: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WES Data

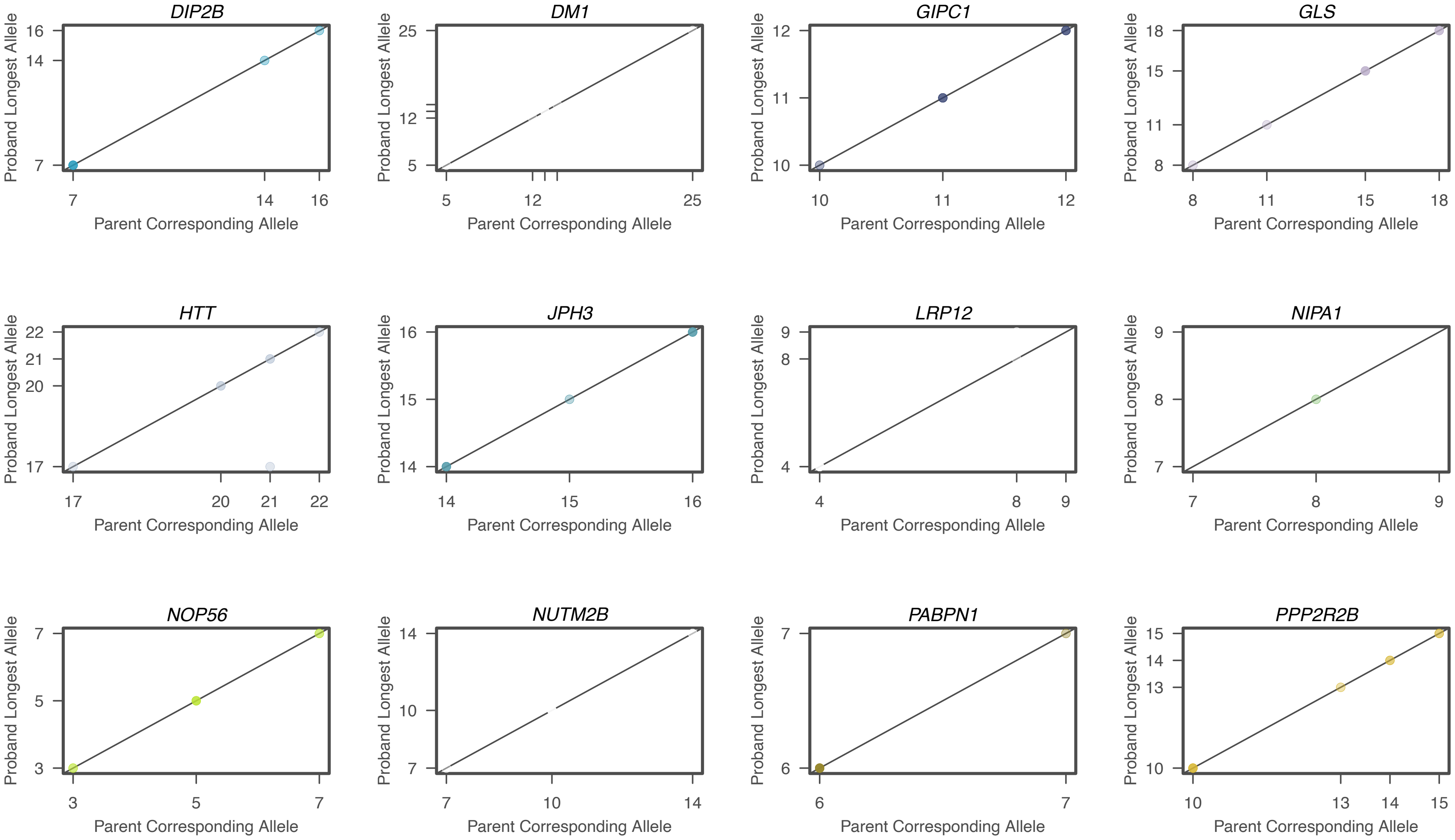


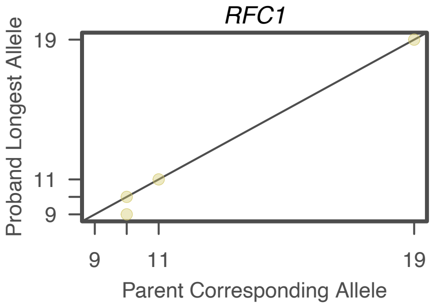
Supplementary Figure 3.25 : ExpansionHunter v3 exploration of potential de novo REs in epilepsy patients

For each patient the longest observed allele at a given locus is compared relative to the longest allele observed in parental samples. A red asterisks indicates that the gene in question had poor concordance when comparing WES and WGS genotypes for the same samples, consequently WES genotypes may not be reliable.

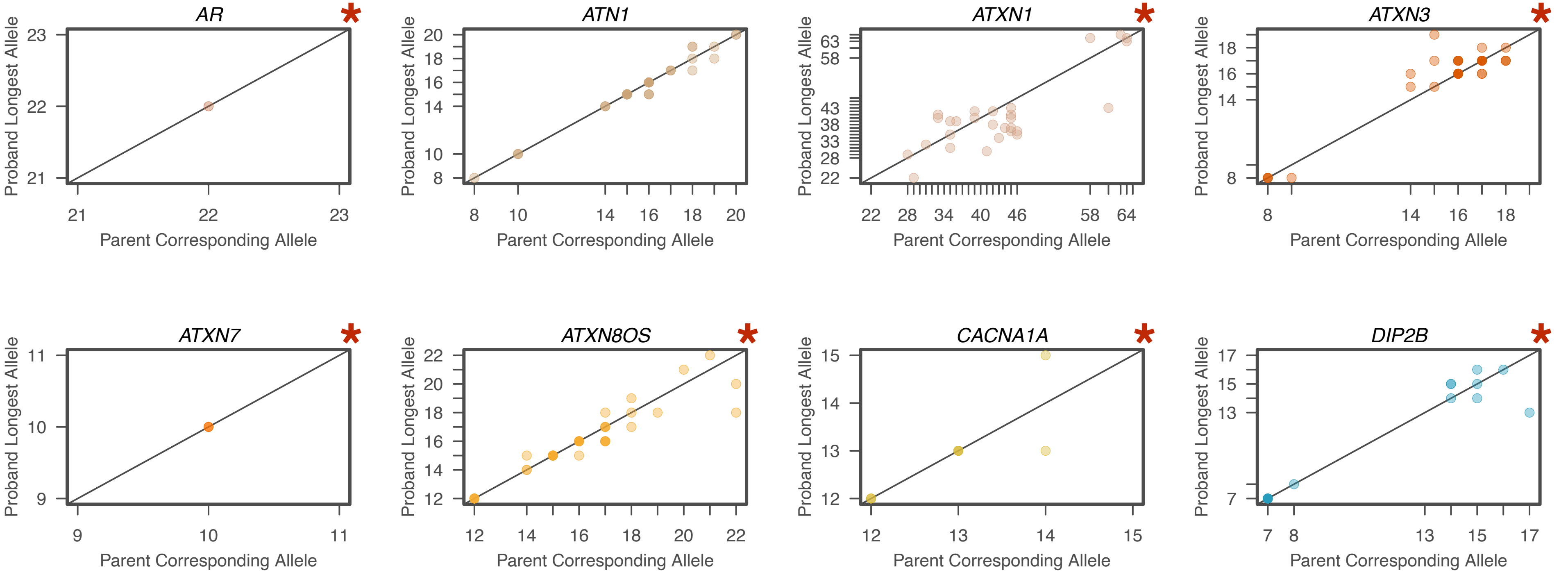
GangSTR_Target_Mode: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WGS PCR Data

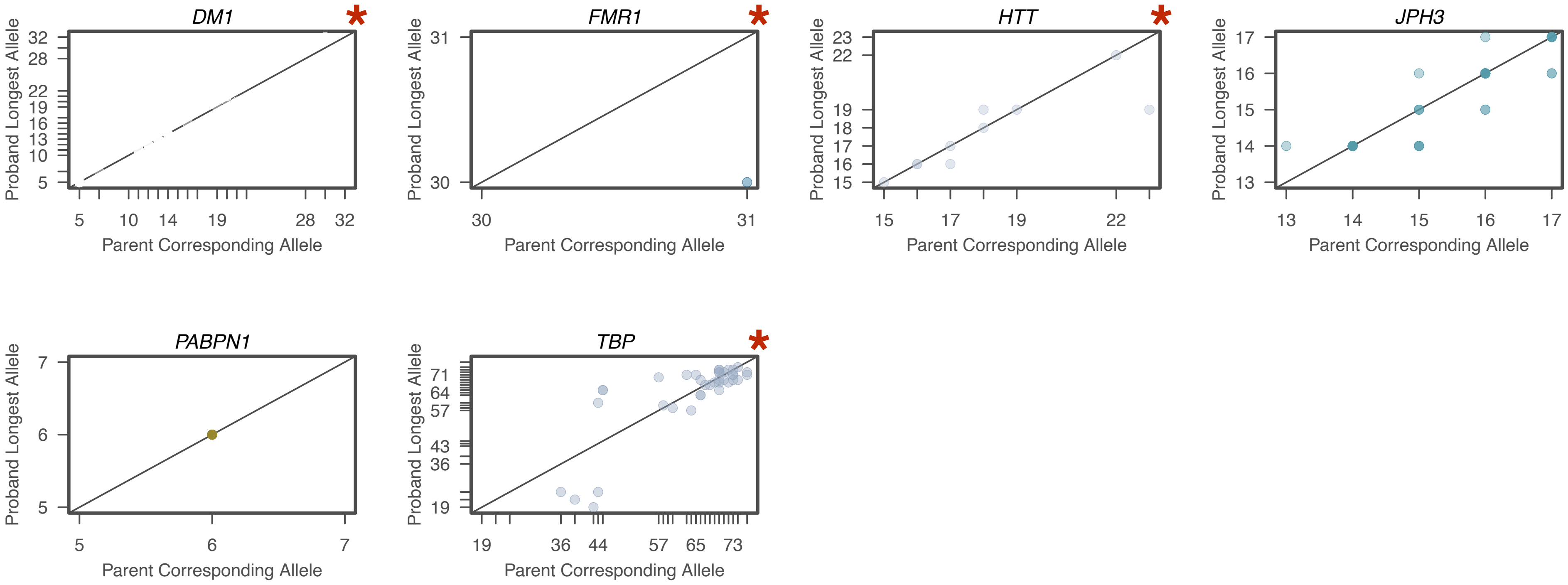



DIP2B
Proband Longest Allele
16
14
7
7
14
16
Parent Corresponding Allele
DM1
Proband Longest Allele
25
12
5
5
12
25
Parent Corresponding Allele
GIPC1
Proband Longest Allele
12
11
10
10
11
12
Parent Corresponding Allele
GLS
Proband Longest Allele
18
15
11
8
8
11
15
18
Parent Corresponding Allele
HTT
Proband Longest Allele
22
21
20
17
17
20
21
22
Parent Corresponding Allele
JPH3
Proband Longest Allele
16
15
14
14
15
16
Parent Corresponding Allele
LRP12
Proband Longest Allele
9
8
4
4
8
9
Parent Corresponding Allele
NIPA1
Proband Longest Allele
9
8
7
7
8
9
Parent Corresponding Allele
NOP56
Proband Longest Allele
7
5
3
3
5
7
Parent Corresponding Allele
NUTM2B
Proband Longest Allele
14
10
7
7
10
14
Parent Corresponding Allele
PABPN1
Proband Longest Allele
7
6
6
7
Parent Corresponding Allele
PPP2R2B
Proband Longest Allele
15
14
13
10
10
13
14
15
Parent Corresponding Allele



GangSTR_Target_Mode: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WES Data

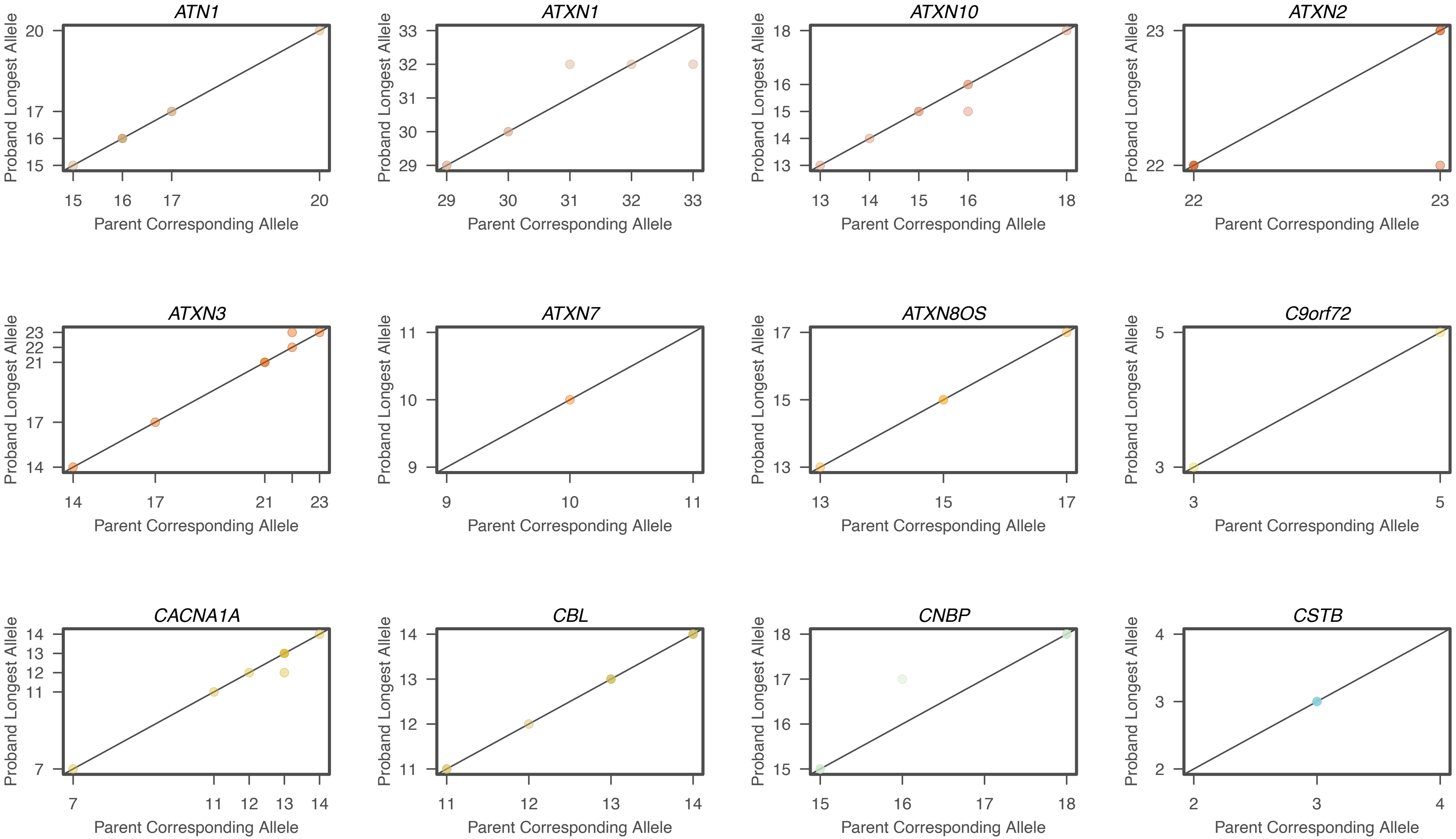


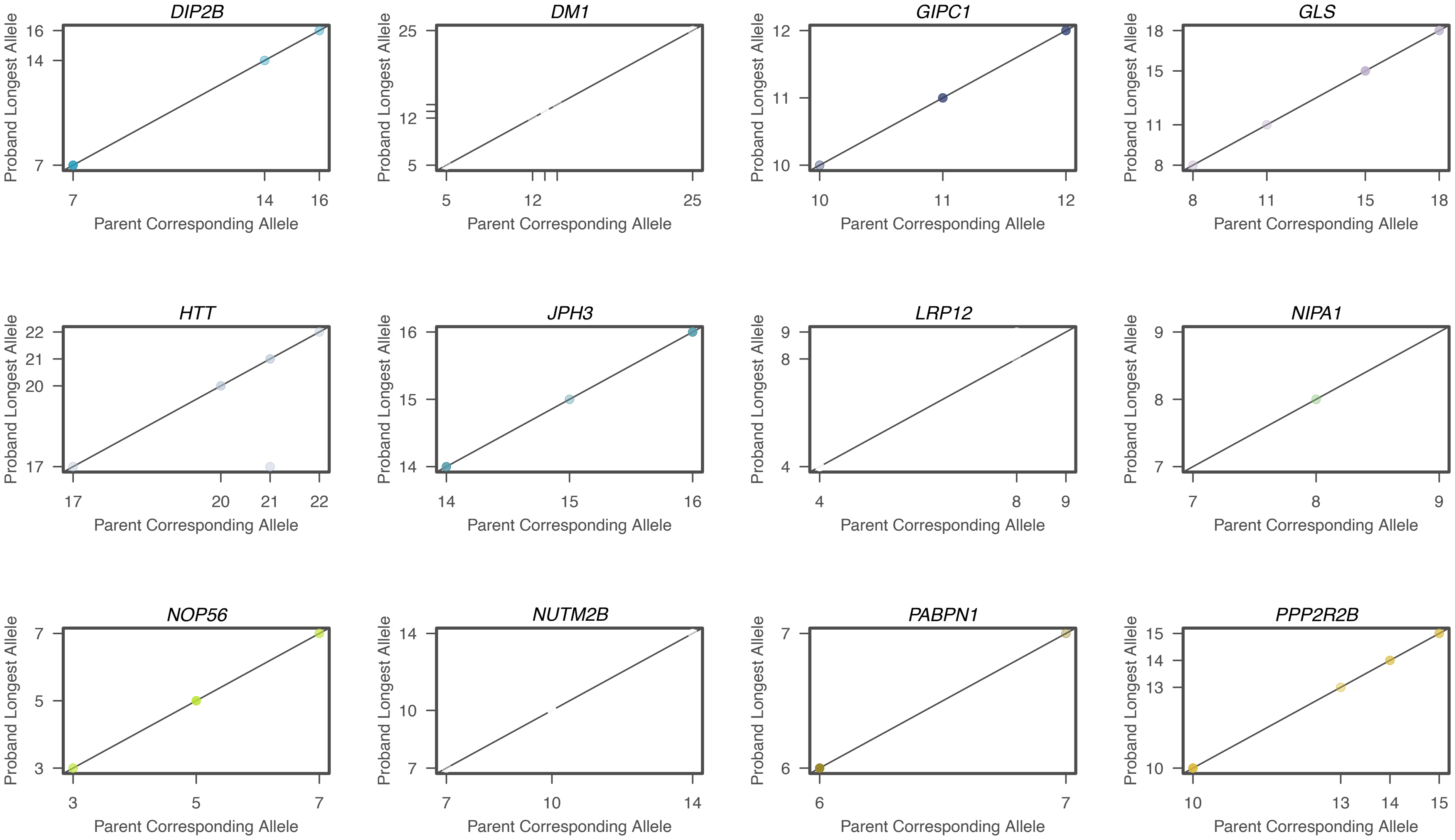


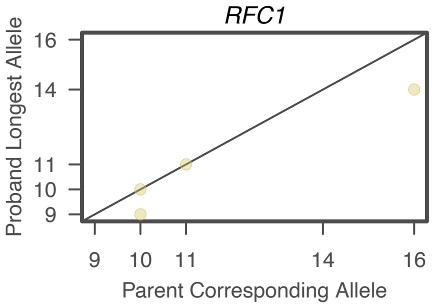
Supplementary Figure 3.26 : GangSTR (Target Mode) exploration of potential de novo REs in epilepsy patients

For each patient the longest observed allele at a given locus is compared relative to the longest allele observed in parental samples. A red asterisks indicates that the gene in question had poor concordance when comparing WES and WGS genotypes for the same samples, consequently WES genotypes may not be reliable. `

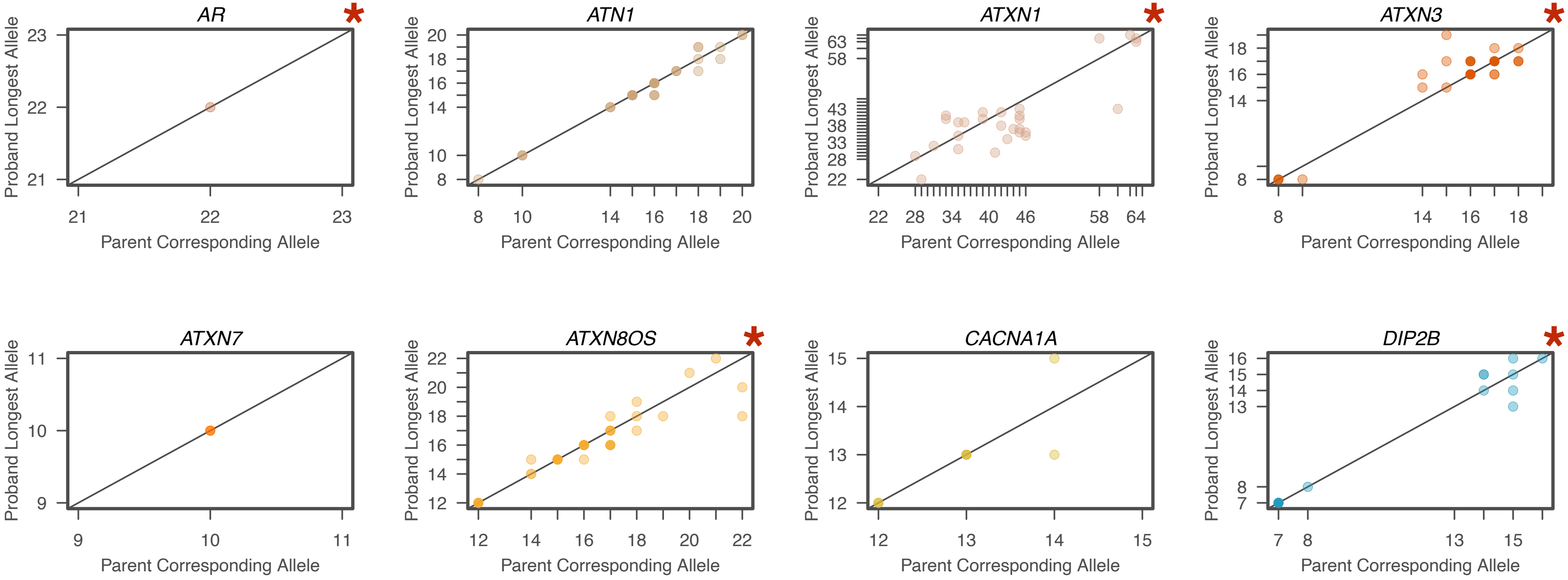
GangSTR_NonTarget_Mode: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WGS PCR Dat

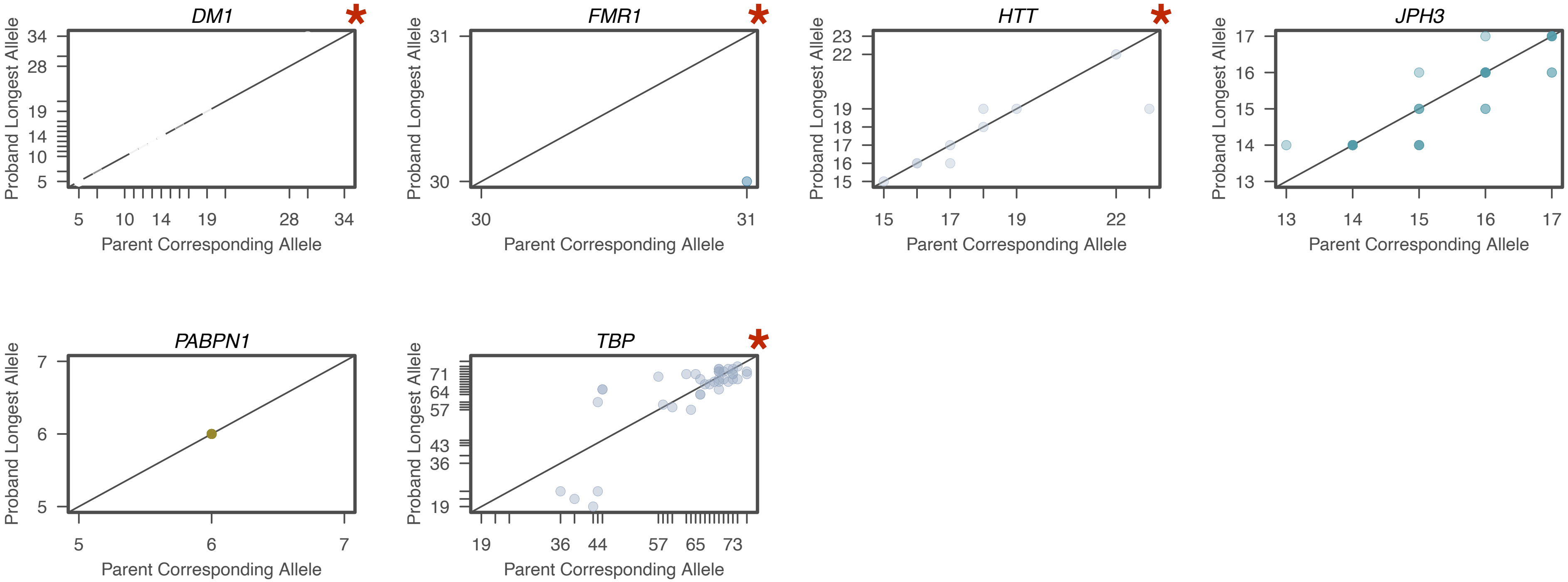



DIP2B
Proband Longest Allele
16
14
7
7
14
16
Parent Corresponding Allele
DM1
Proband Longest Allele
25
12
5
5
12
25
Parent Corresponding Allele
GIPC1
Proband Longest Allele
12
11
10
10
11
12
Parent Corresponding Allele
GLS
Proband Longest Allele
18
15
11
8
8
11
15
18
Parent Corresponding Allele
HTT
Proband Longest Allele
22
21
20
17
17
20
21
22
Parent Corresponding Allele
JPH3
Proband Longest Allele
16
15
14
14
15
16
Parent Corresponding Allele
LRP12
Proband Longest Allele
9
8
4
4
8
9
Parent Corresponding Allele
NIPA1
Proband Longest Allele
9
8
7
7
8
9
Parent Corresponding Allele
NOP56
Proband Longest Allele
7
5
3
3
5
7
Parent Corresponding Allele
NUTM2B
Proband Longest Allele
14
10
7
7
10
14
Parent Corresponding Allele
PABPN1
Proband Longest Allele
7
6
6
7
Parent Corresponding Allele
PPP2R2B
Proband Longest Allele
15
14
13
10
10
13
14
15
Parent Corresponding Allele



GangSTR_NonTarget_Mode: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WES Data

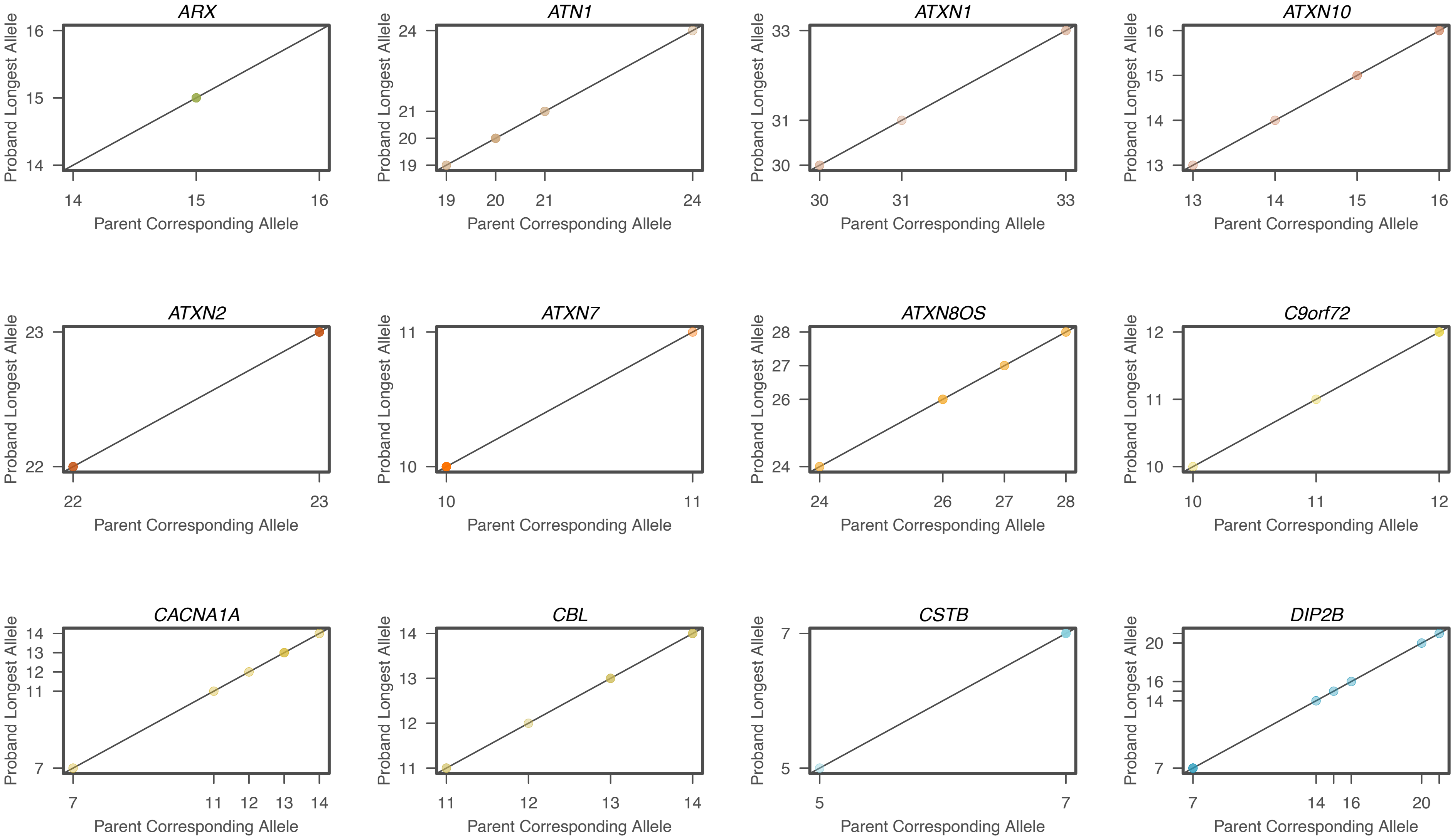


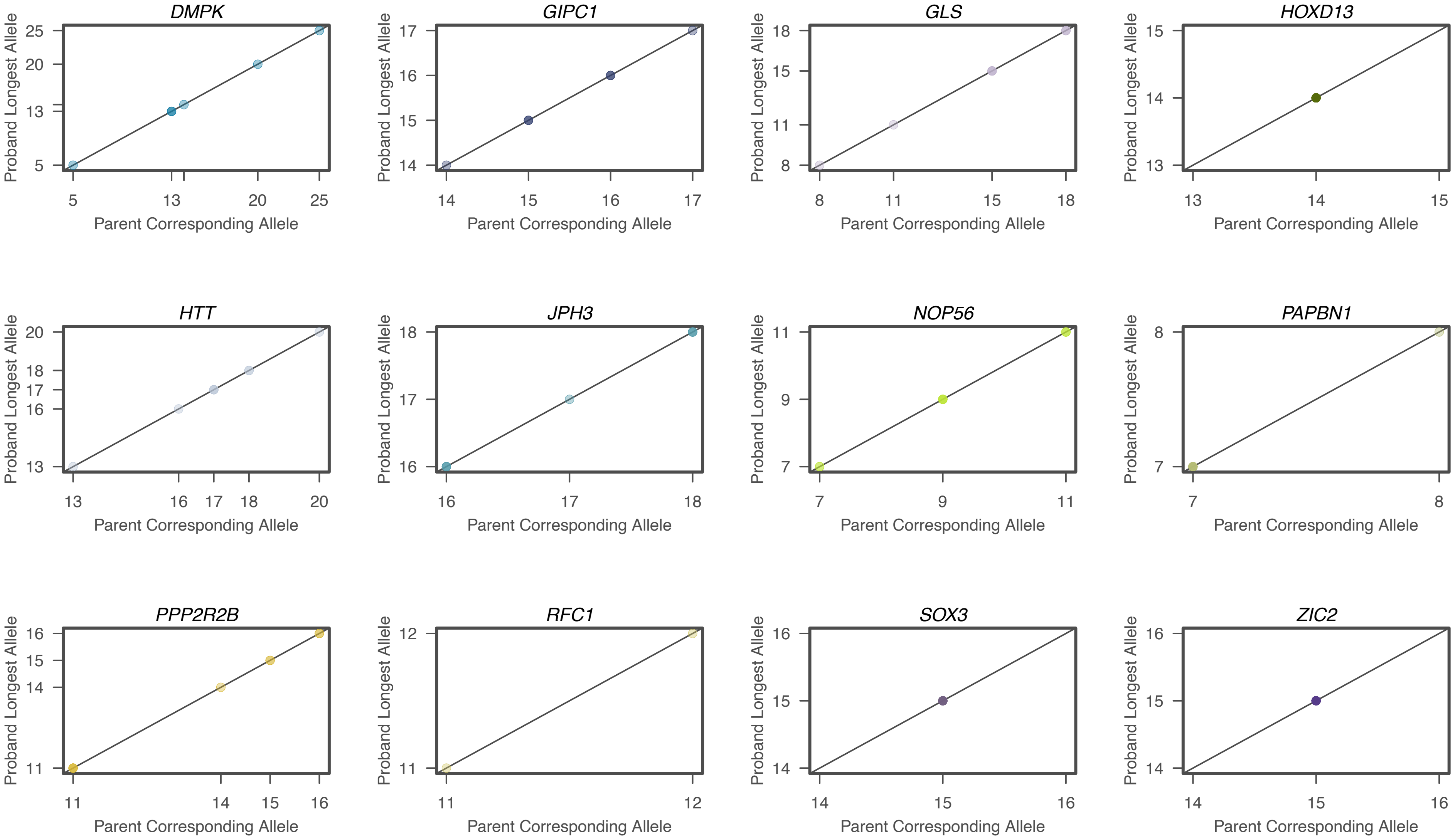


Supplementary Figure 3.27 : GangSTR (Genome-wide Mode) exploration of potential de novo REs in epilepsy patients

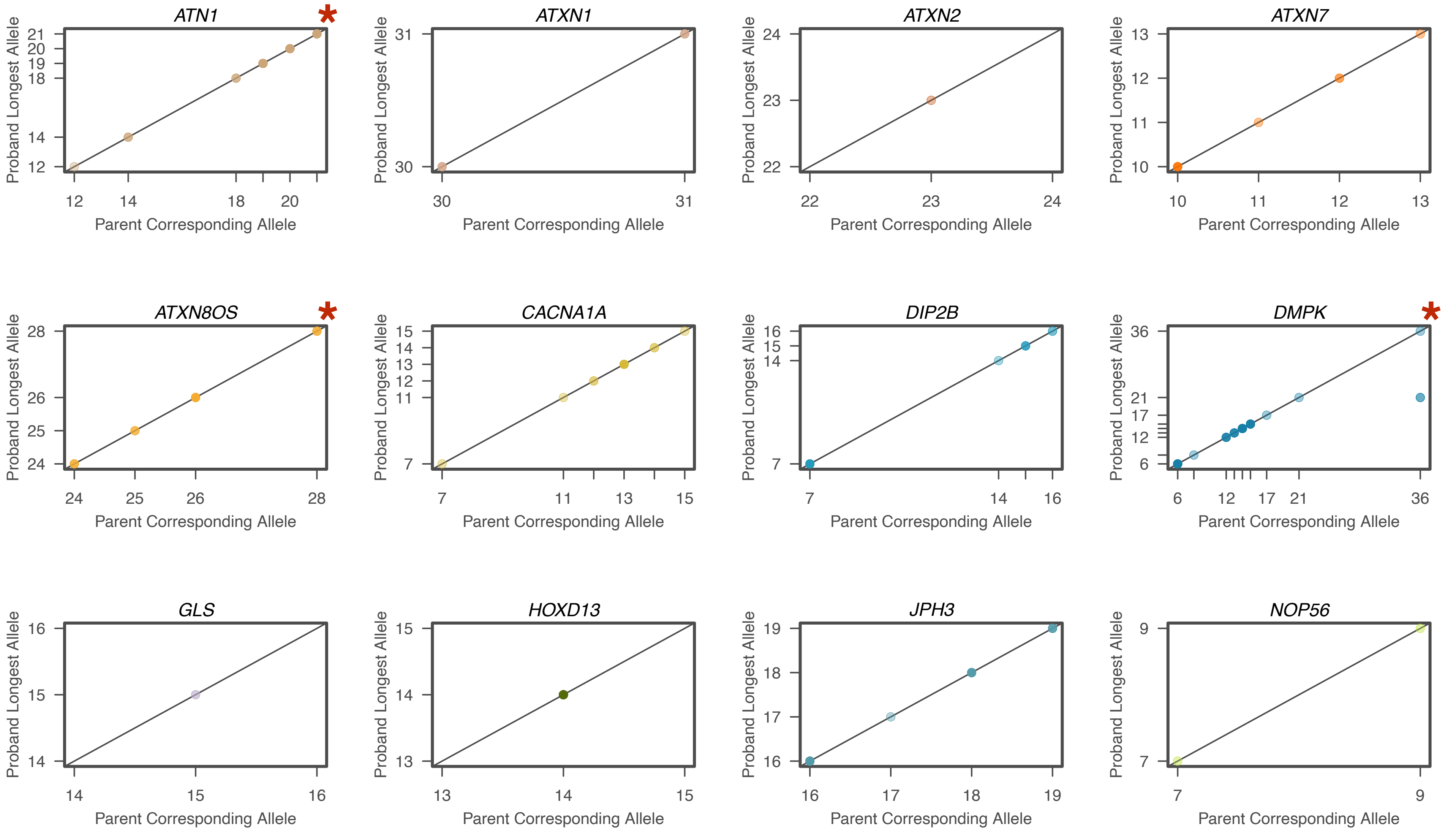
For each patient the longest observed allele at a given locus is compared relative to the longest allele observed in parental samples. A red asterisks indicates that the gene in question had poor concordance when comparing WES and WGS genotypes for the same samples, consequently WES genotypes may not be reliable. `

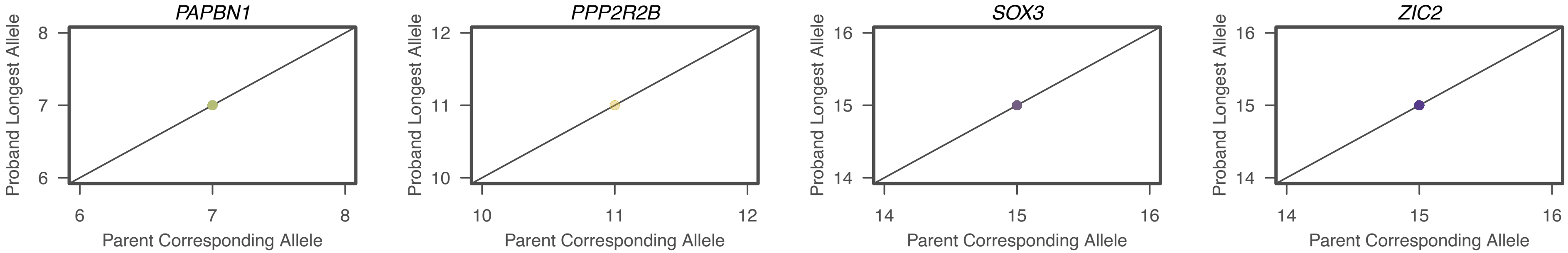
HipSTR: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WGS PCR Data




DMPK
Proband Longest Allele
25
20
13
5
5
13
20
25
Parent Corresponding Allele
GIPC1
Proband Longest Allele
17
16
15
14
14
15
16
17
Parent Corresponding Allele
GLS
Proband Longest Allele
18
15
11
8
8
11
15
18
Parent Corresponding Allele
HOXD13
Proband Longest Allele
15
14
13
13
14
15
Parent Corresponding Allele
HTT
Proband Longest Allele
20
18
17
16
13
13
16
17
18
20
Parent Corresponding Allele
JPH3
Proband Longest Allele
18
17
16
16
17
18
Parent Corresponding Allele
NOP56
Proband Longest Allele
11
9
7
7
9
11
Parent Corresponding Allele
PAPBN1
Proband Longest Allele
8
7
7
8
Parent Corresponding Allele
PPP2R2B
Proband Longest Allele
16
15
14
11
11
14
15
16
Parent Corresponding Allele
RFC1
Proband Longest Allele
12
11
11
12
Parent Corresponding Allele
SOX3
Proband Longest Allele
16
15
14
14
15
16
Parent Corresponding Allele
ZIC2
Proband Longest Allele
16
15
14
14
15
16
Parent Corresponding Allele

HipSTR: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WES Data

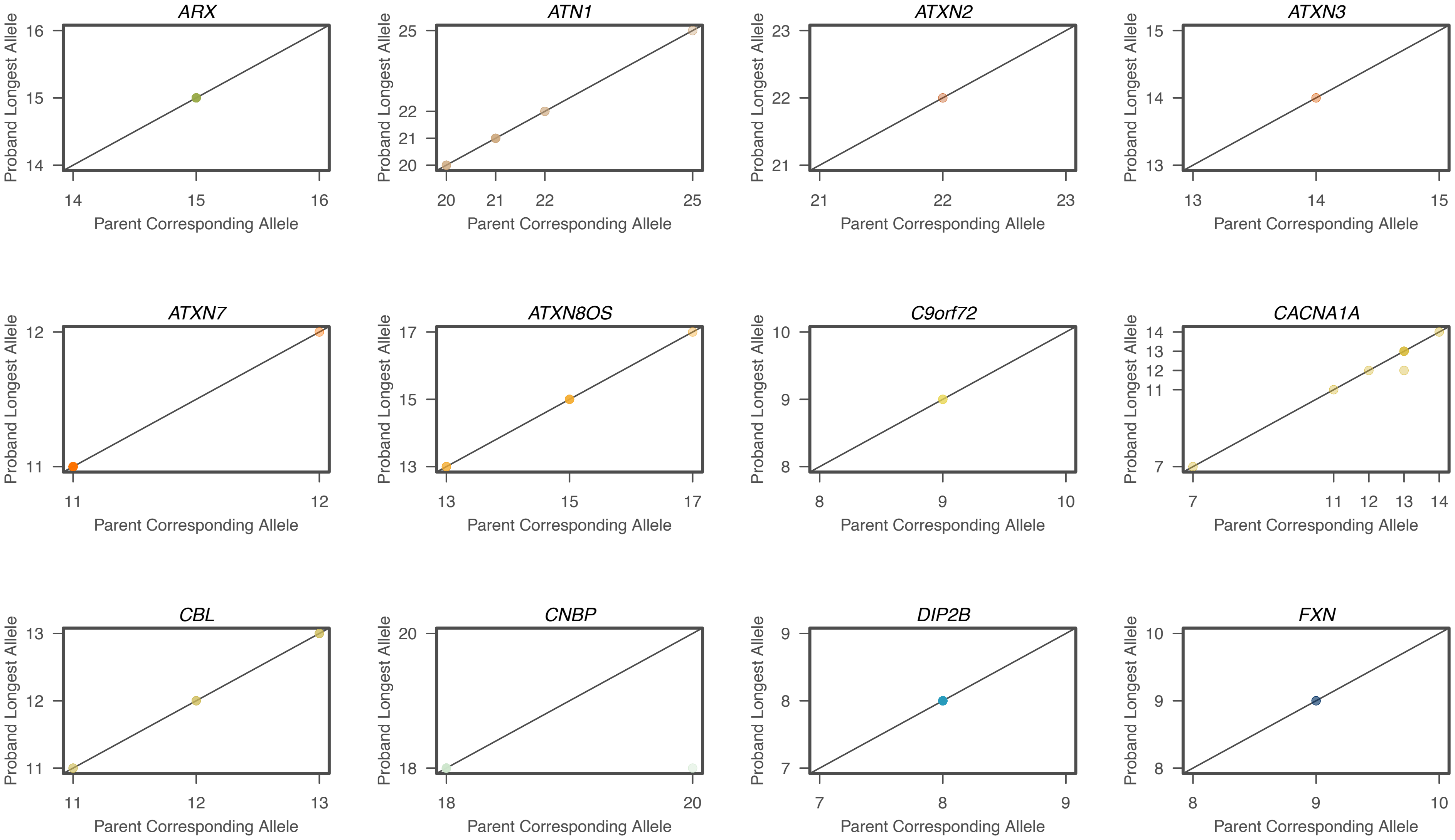


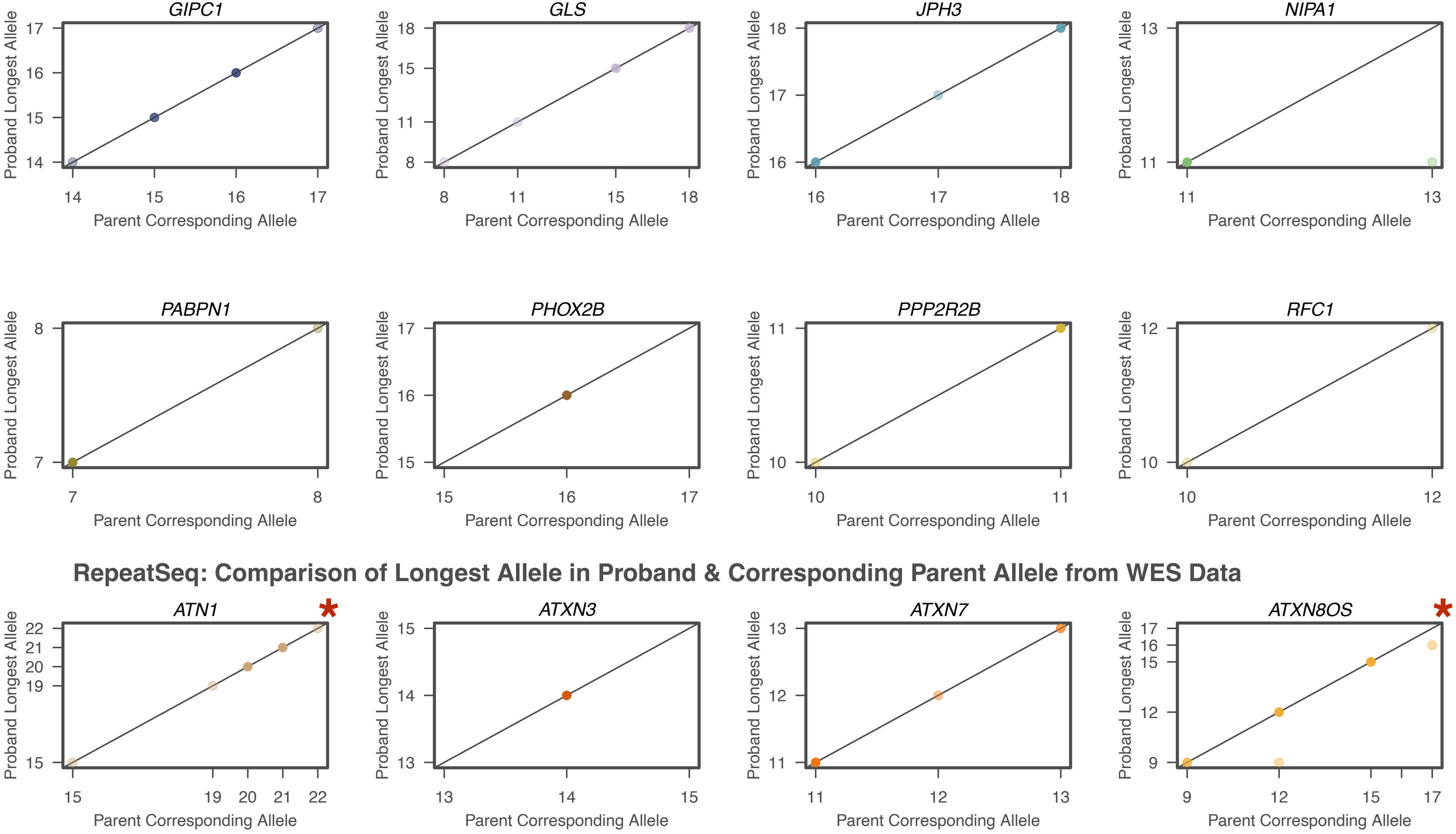


Supplementary Figure 3.28 : HipSTR exploration of potential de novo REs in epilepsy patients

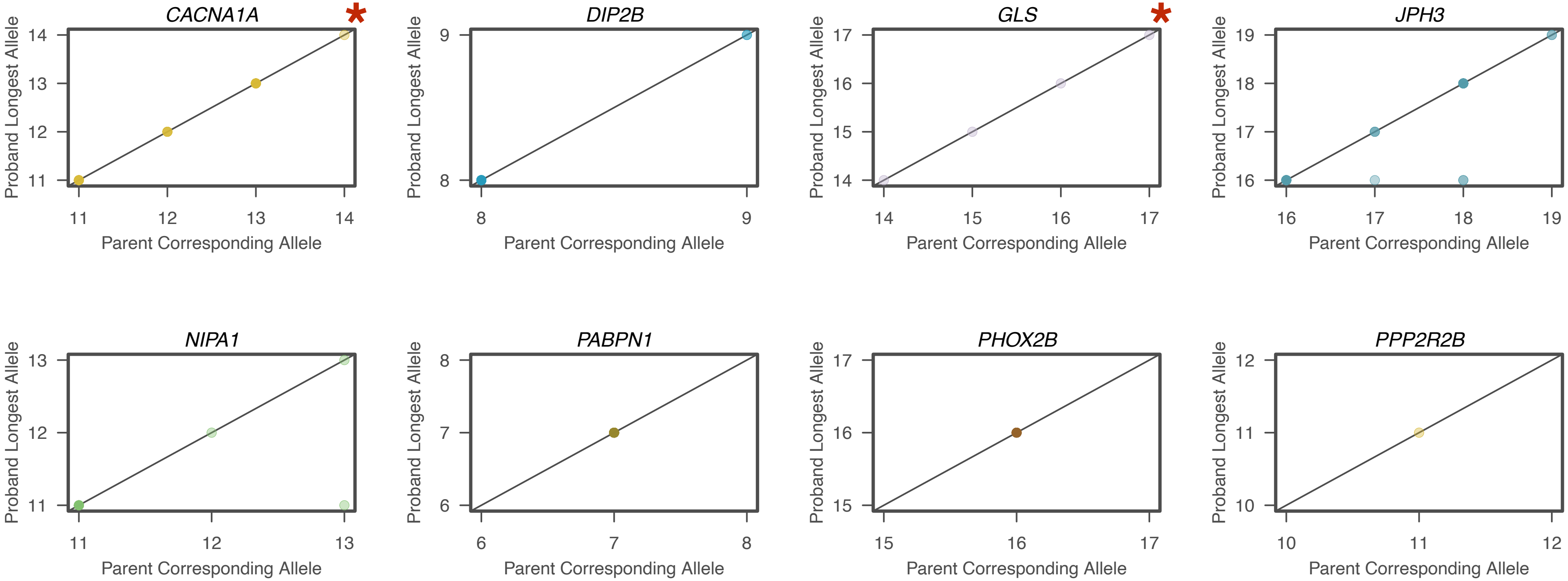
For each patient the longest observed allele at a given locus is compared relative to the longest allele observed in parental samples. A red asterisks indicates that the gene in question had poor concordance when comparing WES and WGS genotypes for the same samples, consequently WES genotypes may not be reliable. `

RepeatSeq: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WGS PCR Data





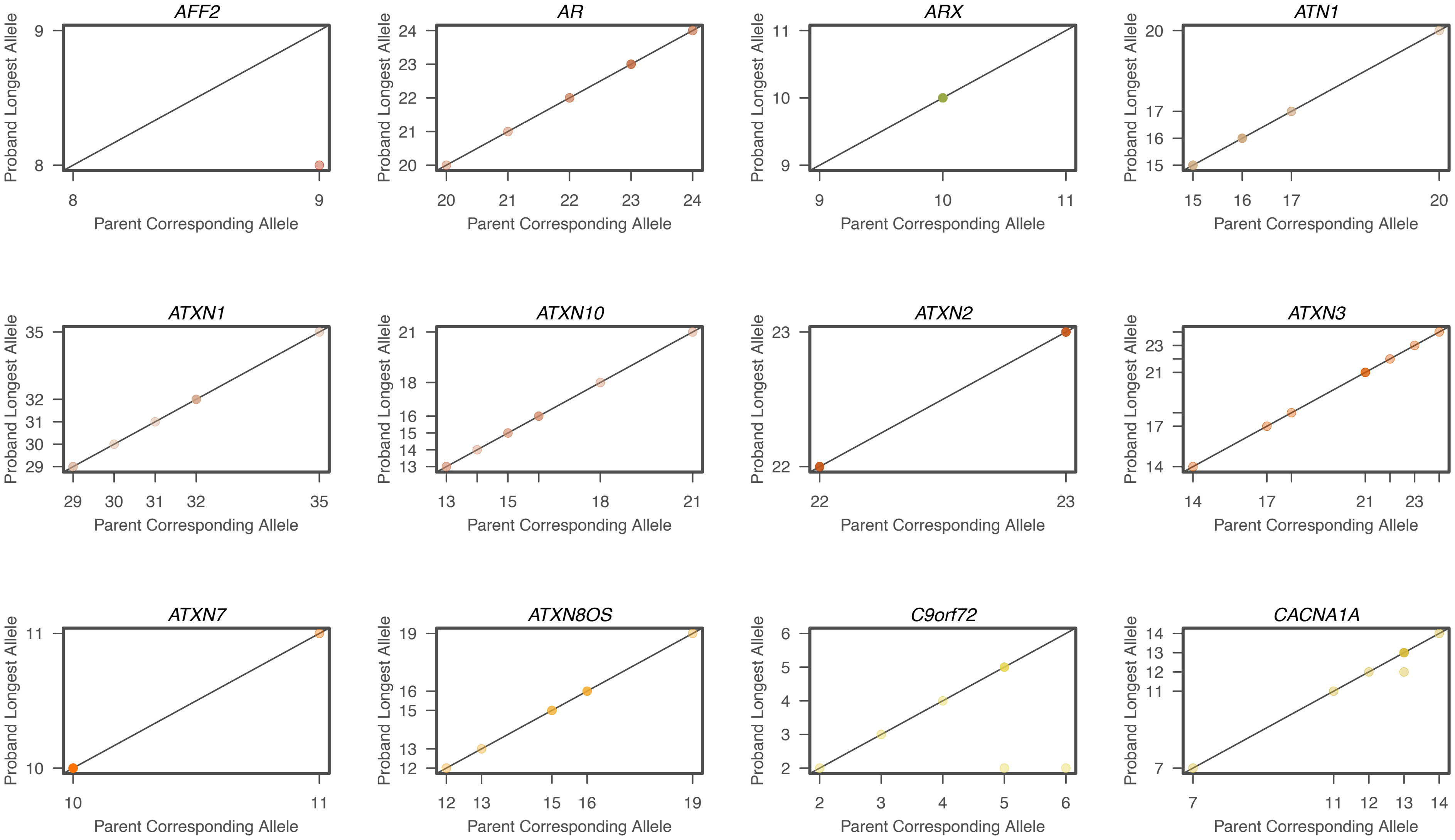
RepeatSeq: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WES Data

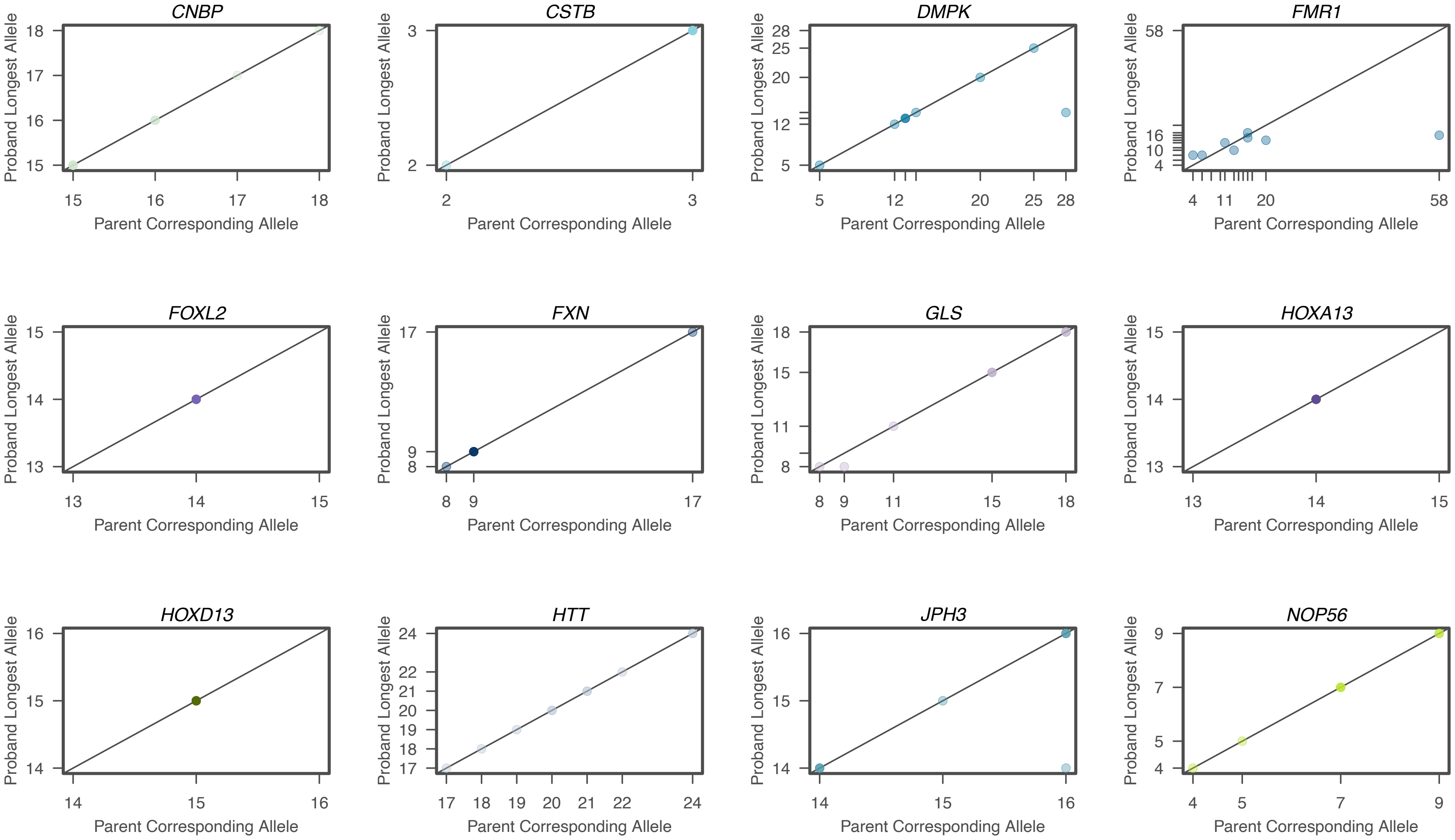


Supplementary Figure 3.29 : RepeatSeq exploration of potential de novo REs in epilepsy patients

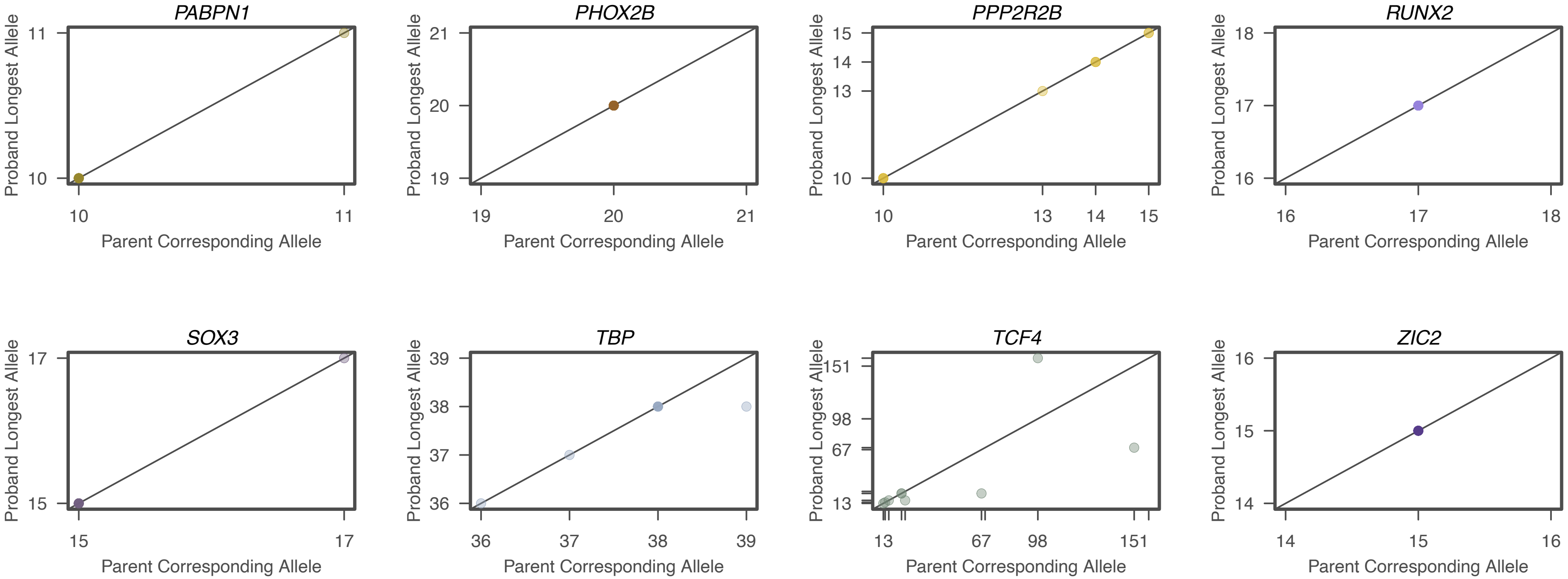
For each patient the longest observed allele at a given locus is compared relative to the longest allele observed in parental samples. A red asterisks indicates that the gene in question had poor concordance when comparing WES and WGS genotypes for the same samples, consequently WES genotypes may not be reliable. `

Tredparse: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WGS PCR Data

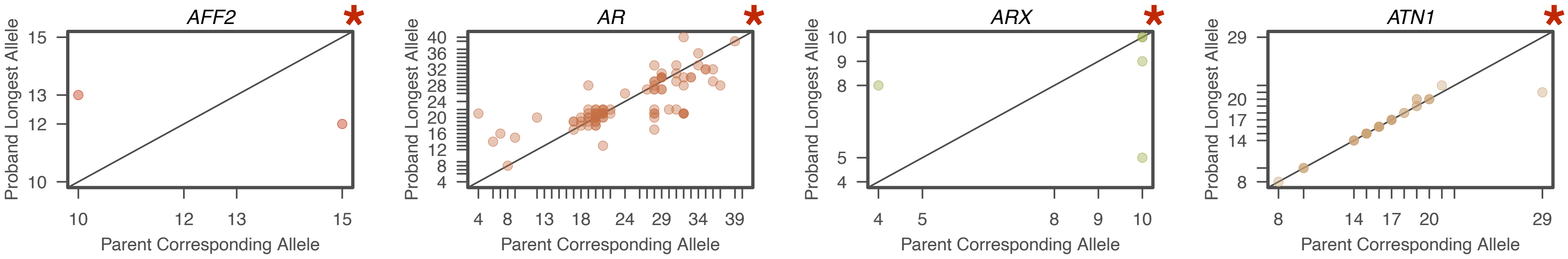


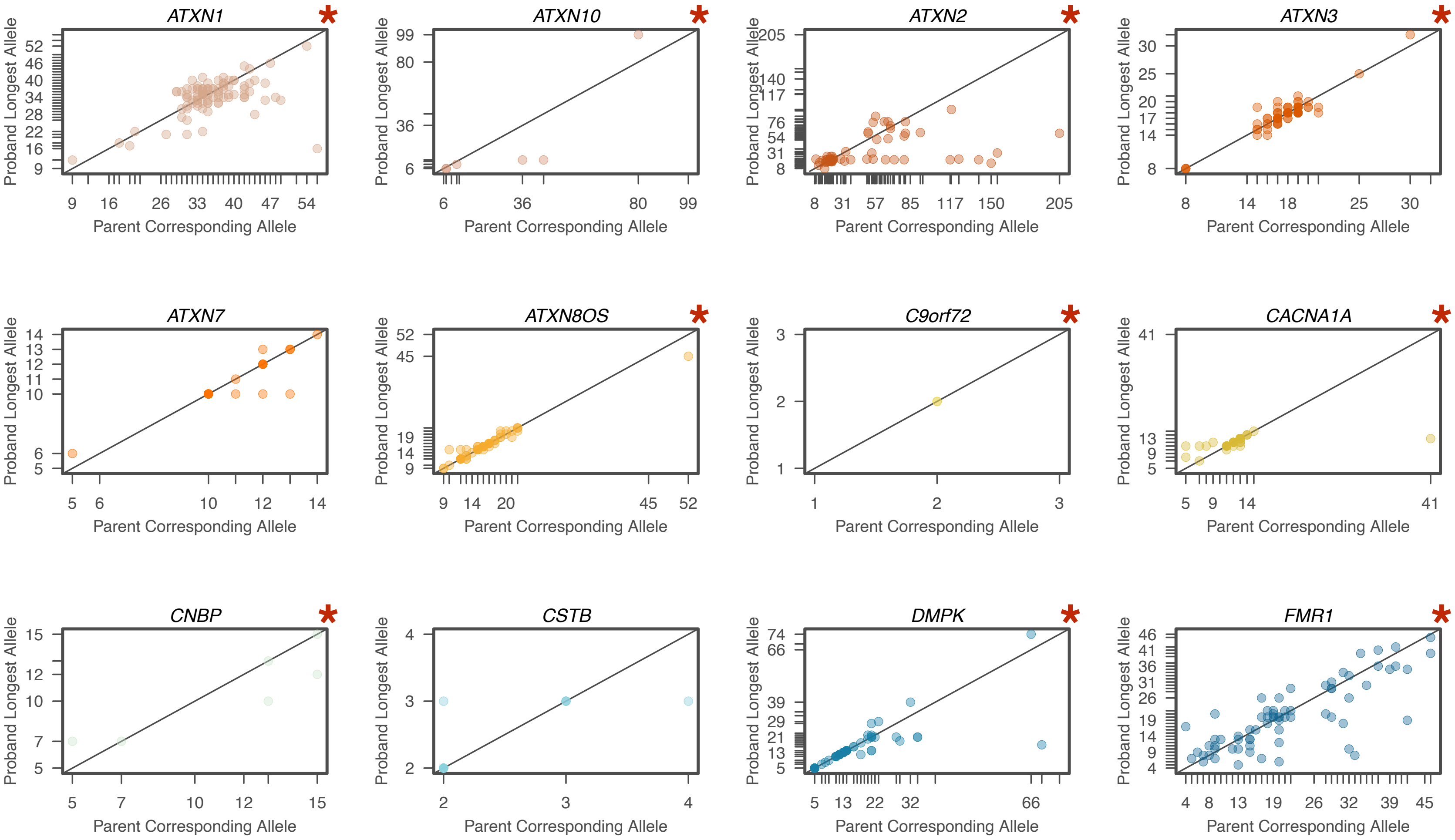


CNBP
Proband Longest Allele
18
17
16
15
15
16
17
18
Parent Corresponding Allele
CSTB
Proband Longest Allele
3
2
2
3
Parent Corresponding Allele
DMPK
Proband Longest Allele
28
25
20
12
5
5
12
20
25
28
Parent Corresponding Allele
FMR1
Proband Longest Allele
58
16
10
4
4
11
20
58
Parent Corresponding Allele
FOXL2
Proband Longest Allele
15
14
13
13
14
15
Parent Corresponding Allele
FXN
Proband Longest Allele
17
9
8
8
9
17
Parent Corresponding Allele
GLS
Proband Longest Allele
18
15
11
8
8
9
11
15
18
Parent Corresponding Allele
HOXA13
Proband Longest Allele
15
14
13
13
14
15
Parent Corresponding Allele
HOXD13
Proband Longest Allele
16
15
14
14
15
16
Parent Corresponding Allele
HTT
Proband Longest Allele
24
22
21
20
19
18
17
17
18
19
20
21
22
24
Parent Corresponding Allele
JPH3
Proband Longest Allele
16
15
14
14
15
16
Parent Corresponding Allele
NOP56
Proband Longest Allele
9
7
5
4
4
5
7
9
Parent Corresponding Allele



Tredparse: Comparison of Longest Allele in Proband & Corresponding Parent Allele from WES Data





Supplementary Figure 3.30 : TREDPARSE exploration of potential de novo REs in epilepsy patients

For each patient the longest observed allele at a given locus is compared relative to the longest allele observed in parental samples. A red asterisks indicates that the gene in question had poor concordance when comparing WES and WGS genotypes for the same samples, consequently WES genotypes may not be reliable.